In-silico tools for the functional annotation of a protein and prediction of cancer biomarkers

By

Sumeet Patiyal

(PhD17204)

Under the Supervision of Prof. Gajendra P.S. Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

August, 2022

In-silico tools for the functional annotation of a protein and prediction of cancer biomarkers

By

Sumeet Patiyal

(PhD17204)

A Thesis

Submitted in Partial Fulfilment of the Requirements for the Degree Of

Doctor of Philosophy

Under the Supervision of Prof. Gajendra P.S. Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

August, 2022

Certificate

This is to certify that the thesis entitled **"In-silico tools for the functional annotation of a protein and prediction of cancer biomarkers"** being submitted by **Mr. Sumeet Patiyal** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy,** is an original research work, carried out by him under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2022

Prof. Gajendra P. S. Raghava
Supervisor Name
Indraprastha Institute of Information Technology Delhi
New Delhi - 110020

Acknowledgements

"No one who achieves success does so without acknowledging the help of others.
The wise and confident acknowledge this help with gratitude."
-Alfred North Whitehead

This is a defining point in my life since I am on the verge of completing my thesis. My heart is racing with a mixture of happiness and grief. I am ecstatic as I am about to complete an important phase of my life but at the same time feeling sad for leaving people and memories behind which transformed my career and life from nothing to something. These are deeply associated with my stay at Indraprastha Institute of Information Technology, Delhi (IIIT-D). I would like to take this opportunity to express my deepest gratitude and regards from the bottom of my heart to people who helped me to move ahead so far. Not only the duration but also the difficult journey of a thesis study itself demands an intellectual fortification in the form of a person whom we refer as the teacher, mentor, thesis guide, supervisor, or doctoral advisor. In my case, such a figure of wisdom, wit, intellect, and acumen is in the form of ***Prof. Gajendra P. S. Raghava,*** *who is always much more concerned about the projects assigned to me, who always supported me in best or worst phase of my life, and who never gave up on me. Providing a research environment of international standards both in terms of infrastructure as well as openness of ideas, he is always keen to make the projects multidimensional. I am indebted for life to him for giving me a chance to be a part of his research team for a long period of five years.*

I would like to express my sincere gratitude to ***Prof. Gajendra P. S. Raghava*** *and* ***Prof. Pankaj Jalote*** *(the then Director, IIIT-D), for granting me the privilege of admission in the PhD programme and subsequent provisions of using the institute's facilities for efficient research work. I am so grateful to* ***Prof. Gajendra P. S. Raghava,*** *who gave me the opportunity to complete my M.Tech. thesis under his supervision along with my PhD thesis. I literally cannot thank him enough. I am grateful to the subsequent Director* ***Dr. Ranjan Bose*** *for advancing the institute infrastructure that facilitated me during my PhD tenure. My sincere thanks to* ***Dr. K. Sriram*** *and* ***Dr. Vibhor Kumar*** *for being in my evaluation committee and guiding me through their continuous support and wisdom throughout this journey. I would also like to thank* ***Dr. Manohar Khushlani*** *and* ***Dr. Pravesh Biyani,*** *for whom I worked as teaching assistant in the course of theatre appreciation and community work. With them I have learnt a lot about the other aspects of life other than science.*

Frankly, it was not easy to change my field from pure experimental to completely computational. The first hurdle that came in my journey was the course work which demanded lots of programming and mathematical understanding. But, the faculties were so supportive and understanding that I passed all the courses with flying colors. I like to extent my gratitude to ***Prof. Gajendra P. S. Raghava, Dr. K. Sriram, Dr. Debarka Sengupta, Dr. Subhadip***

Raychaudhuri, Dr. Bapi Chatterjee, Dr. Angshul Majumdar, *and* ***Dr. Vibhor Kumar,*** *for keeping faith in me during the course work. Finally, I'd want to thank the administrative staff at IIIT-D, particularly* ***Mrs. Priti Patel****,* ***Ms. Shipra Jain****,* ***Ms. Sheetu Ahuja****, and* ***Ms. Anshu Dureja****, for always being available to answer our questions and handle our academic issues promptly. A special thanks to* ***Mr. Imran Khan, Nidhi Mam, Mr. Kapil Dev Garg*** *for releasing my stipend on time. I am also grateful to IIIT-D for providing first-rate infrastructure and facilities. An official note of thanks to the funding agency* ***Department of Biotechnology (DBT)*** *for providing the research fellowship to me to support my doctoral studies.*

My heartfelt thanks to my colleagues in my lab who are always helpful and maintained a healthy and most enjoyable working environment. I am extremely grateful to the people present in the lab, present prior to my joining, and those who came after me. I learned a lot from my seniors ***Dr. Sherry Bhalla, Dr. Piyush Agrawal, Dr. Salman Sadullah Usmani, Dr. Pawan Kumar Raghav, Dr. Akshara Pande, Dr. Lubna Maryam, Dr. Harpreet Kaur, Dr. Vinod Poriya, Dr. Rajesh Kumar, Dr. Anjali Lathwal,*** *and* ***Dr. Chakit Arora.*** *I possibly cannot imagine more cooperative, helpful, and better lab colleagues. I had some greatest enjoyable time with my lab colleagues,* ***Dr. Sherry Bhalla, Dr. Piyush Agrawal, Dr. Salman Sadullah Usmani, Dr. Rajesh Kumar, Dr. Vinod Poriya, Dr. Harpreet Kaur, Dr. Anjali Lathwal, Dr. Chakit Arora, Dr. Dilraj Kaur, Anjali Dhall, Nishant Kumar, Shubham Choudhury, Ritu Tomer, Nisha Bajiya, Shivani Malik, Anand Singh Rathore, Mansi Goel, Dr. Leimarembi Devi Naorem, Neelam Sharma,*** *and* ***Akanksha Arora.*** *My special thanks to* ***Anjali Dhall*** *for her continuous support and help. I cannot imagine a colleague more cooperative, more helpful, or a better illuminant than her, heartiest thanks to her for being an intellectual and academic buttress for me especially in tough and demanding situations. I could not have finished my M. Tech. and Ph.D. thesis writing without her help, she made the compilation and timely submission of this thesis with her immense help during the last-minute rush. I would also like to thank my Ph.D. batchmates along with whom I pursued my pre-Ph.D. course work. Some of the batchmates whom I have always admired and learnt from are* ***Neetesh Pandey, Priyadarshini Rai, Shivam Sharma, Rohit Kumar, Abhishek Aggarwal, Sana Akhtar, Aditi Sharma.*** *I would like to thank* ***Indra Prakash Jha, Neetesh Pandey, Priyadarshini Rai, Dr. Shiju, Dr. Krishan Gupta, Sarita Poonia, Smriti Chawla, Shreya Mishra, Raghava Awasthi, Chitrita Goswami, Omkar R. Chandra,*** *for being supportive and available at all times when I needed them. The most heart-warming feeling for me is to thank my adorable juniors-* ***Nishant, Shubham, Ritu, Nisha, Shivani, Anand, Shruti, Madhu, Samridhi, Vishakha, Sanjay, Aayshi, Sakshi, Shubadeep, Gayatri, Sukriti, Akshya, Pradeep, Alok, Jasmine, and Ramsha,*** *who rejuvenated me and re-infused a carefree attitude of a young learner into me by extending me a mealtime and teatime companionship. I am also thankful to all of these awesome personalities for making me a part of their own important and amazing moments of life, such as, their birthdays and professional achievements.*

I would like to thank the hostel facilities provided by the administrative staff of IIIT-D. A special thanks to ***Rajeev Ji*** *(hostel warden)****, Tinku bhaiya, Naveen Ji, Suneel Ji*** *for their availability and get the job done at the earliest. In no way I can forget to thank the mess and canteen that has fed me day and night. I am indebted for life to these people for providing me homely food*

away from home. I am grateful to the 'Ravi Tea Stall' with its vendor ***Rajesh Ji,*** *who served exhilarating tea in the morning and evening time that lighten up the neurons in my brain after a long day of tiredness.*

Last but not the least, I would like to thank my beloved sister, ***Madhvi Patiyal,*** *with my brother-in-law* ***Loveleen Parmar,*** *and two amazing nieces* ***Nishita and Mishita,*** *who never doubted me for even a single second and had an unshakable faith in me. I found the most unique affection for me in my nieces who is always dissatisfied because of my rare visits of very short time. I would also like to thank my brother* ***Mr. Vikas Patiyal,*** *who always supported me no matter what. Yet, the most difficult part is that of acknowledging the parents for our accomplishments. Nevertheless, the bigger question is whether this is different from acknowledging the almighty.*

I am very grateful to have the best parents in this entire universe, ***Shri Nanak Chand*** *and* ***Smt. Sushma Devi.***

Sumeet Patiyal

Abstract

Proteins play major roles in many biological processes such as enzymes, transporters, replication, transcription, gene expression regulation, repair and building of tissues, energy production, molecule transport, muscle development, wound healing, and muscle and bone restoration. Due to the advancement in sequencing technology, database of protein sequences are growing with exponential rate. Thus, functional annotation of a protein or prediction of function of a protein is one of the major challenges in post genomics era. In this study, a systematic attempt had been made to understand and predict function of a protein. First objective of this thesis is to predict regulatory proteins as they play an essential role in the replication, transcription, regulation of gene expression. In order predict function of a protein using machine learning techniques, a method Pfeature has been developed to predict protein features. Pfeature allows to compute a wide range of features/descriptors from the sequence and structure information of a protein. Further, these features have been used to develop machine learning based models for predicting transcription factors, important regulatory proteins. These proteins coordinate the biological functions by interacting with other molecules. Thus, it is crucial to identify interacting residues in a protein that interact with other biological molecules. Second objective of this study is to determine the protein-molecule interactions. Under this objective three prediction methods (NAGbinder, DBPred & Pprint2) have been developed for predicting interacting residues in a protein. NAGbinder developed for predicting N-acetyl glucosamine (NAG) binding residues. PPRINT2 for predicting RNA interacting residues in a protein. DBPred for predicting DNA binding sites in a proteins. In addition to functional annotation of proteins, an attempt has been made to identify cancer associated mutations in genome to understand the cancer pathogenesis. In order to achieve this objective, first we examining and compare mutation calling techniques. In this study four major mutation calling techniques (Mutect2, MuSE, Varscan2 & SomaticSniper) have been benchmarked and identify the best mutation calling technique. Finally, we developed a method for the identification of prognostic and diagnostic biomarkers using the mutation profile of liver cancer patients. All methods developed during this study are freely available to scientific community in form of web services and standalone software packages.

List of Publications

Thesis Related Publications

- **Patiyal S**, Agrawal P, Kumar V, Dhall A, Kumar R, Mishra G, et al. NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. ***Protein Sci.*** 2020 Jan;29(1):201–10.

- **Patiyal S**[#], Dhall A[#], Raghava GPS. Prediction of risk-associated genes and high-risk liver cancer patients from their mutation profile: Benchmarking of mutation calling techniques, ***Biology Methods and Protocols***, 2022;bpac012.

- **Patiyal S**, Dhall A, Raghava GPS**.** DBPred: A deep learning method for the prediction of DNA interacting residues in protein sequences. ***Brief Bioinform***. 2022.

- Pande A[#], **Patiyal S**[#]**,** Lathwal A, Arora C, Kaur D, Dhall A, et al. Computing wide range of protein/peptide features from their sequence and structure. ***Journal of Computational Biology.*** 2022.

- **Patiyal S,** Dhall A, Bajaj K, Sahu H, Raghava GPS**.** Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. ***Brief Bioinform***. 2022. (in press)

- **Patiyal S,** Tiwari P, Ghai M, Dhapola A, Dhall A, Raghava GPS**.** A hybrid approach for predicting transcription factors. (Submitted)

Other Publications

- Singh SK, **Patiyal S**, Kaur R, Kumar A. Prediction of metal ion binding site in truncated globin of Myxococcus xanthus DK1622 in homologous model. ***MOJ Proteomics Bioinforma.*** 2017;5(2017):7–12.

- Agrawal P, **Patiyal S**, Kumar R, Kumar V, Singh H, Raghav PK, Raghava GPS**.** ccPDB 2.0: an updated version of datasets created and compiled from Protein Data Bank. ***Database.*** 2019;2019.

- Kumar R, **Patiyal S**, Kumar V, Nagpal G, Raghava GPS. In silico analysis of gene expression change associated with copy number of enhancers in pancreatic adenocarcinoma. **Int J Mol Sci.** 2019;20(14):3582**.**

- Kaur D, **Patiyal S**, Sharma N, Usmani SS, Raghava GPS. PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands. ***Database*.** 2019;2019.

- **Patiyal S**[#], Kaur D[#], Kaur H[#], Sharma N[#], Dhall A, Sahai S, et al. A Web-Based Platform on Coronavirus Disease-19 to Maintain Predicted Diagnostic, Drug, and Vaccine Candidates. ***Monoclon Antib Immunodiagn Immunother.*** 2020;39(6):204–16.

- Dhall A, **Patiyal S**, Kaur H, Bhalla S, Arora C, Raghava GPS. Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics. ***Front Genet.*** 2020; 11:22.

- Kumar R, Lathwal A, Kumar V, **Patiyal S**, Raghav PK, Raghava GPS. CancerEnD: A database of cancer associated enhancers. ***Genomics***. 2020;112(5):3696–702.

- Kumar V, Kumar R, Agrawal P, **Patiyal S,** Raghava GPS. A Method for Predicting Hemolytic Potency of Chemically Modified Peptides from Its Structure**.** ***Front Pharmacol***. *2*020; 11:54.

- Dhall A[#], **Patiyal S**[#], Sharma N, Usmani SS, Raghava GPS. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. ***Brief Bioinform***. 2021 Mar;22(2):936–45.

- Sharma N, **Patiyal S**, Dhall A, Pande A, Arora C, Raghava GPS**.** AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. ***Brief Bioinform***. 2021;22(4):bbaa294.

- Chaudhary A[#], Bhalla S[#], **Patiyal S**, Raghava GPS, Sahni G. FermFooDb: A database

of bioactive peptides derived from fermented foods. ***Heliyon*****.** 2021;7(4):e06668.

- Kumar V, **Patiyal S**, Kumar R, et al. B3Pdb: an archive of blood–brain barrier-penetrating peptides. ***Brain Struct. Funct.*** 2021; 226:2489–2495.

- Kumar V, **Patiyal S**, Dhall A, Sharma N, Raghava GP**.** B3Pred: A Random-Forest-Based Method for Predicting and Designing Blood–Brain Barrier Penetrating Peptides. Vol. 13, ***Pharmaceutics***. 2021

- Dhall A, **Patiyal S**, Sharma N, Devi NL, Raghava GPS**.** Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. ***Comput Biol Med.*** 2021;137(October 2021):104780

- Sharma N, **Patiyal S**, Dhall A, Naorem DL, Raghava GPS**.** ChAlPred: A web server for prediction of allergenicity of chemical compounds. ***Comput Biol Med.*** 2021;136(September 2021):104746.

- Attila Gabor, Alice Driessen, Jovan Tanevski, Baosen Guo, Wencai Cao, He Shen, et al. Cell-to-cell and type-to-type heterogeneity of signaling networks: insights from the crowd. ***Mol Syst Biol.*** 2021;17(10).

- Tarca AL, Pataki BA, Romero R, Sirota M, Guan Y, Kutum R, et al. Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. ***Cell Reports Med***. 2021;2(6)

- Kaur D**, Patiyal S**, Arora C, Singh R, Lodhi G, Raghava GPS. In silico tool for predicting, scanning and designing defensins. ***Front Immunol.*** 2021;4817.

- Jain S, Dhall A, **Patiyal S**, Raghava GPS**.** IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. ***Comput Biol Med.*** 2021;137(October 2021):104780.

- Dhall A, **Patiyal S**, Raghava GPS. HLA_{nc}Pred: A method for predicting promiscuous non-classical HLA binding sites. ***Brief Bioinform***. 2022; bbac192.

- **Patiyal S**, Singh N, Ali MZ, Pundir DS, Raghava GPS. Sigma70Pred: A highly accurate method for predicting sigma70 promoter in Escherichia coli K-12 strains. ***Front Microbiol.*** 2022.

- Kumar N, **Patiyal S**, Choudhury S, Tomer R, Dhall A, Raghava GPS. DMPPred: A tool for Identification of Antigenic Regions Responsible for Inducing Type 1 Diabetes Mellitus. ***Brief Bioinform***. 2022;.

URL of Computational Resources

Following resource have been developed during this study.

Name	Web-server	Dataset	Standalone
Pfeature	*pfeature	-	#Pfeature
TransFacPred	*transfacpred	-	#transfacpred
NAGbinder	*nagbinder	*nagbinder/download.php	#nagbinder
DBPred	*dbpred	*dbpred/download.php	*dbpred/stand.html
PPRInt2	*pprint2	*pprint2/dataset.php	#pprint2
Mutation Bench	-	-	#mutation_bench

***: https://webs.iiitd.edu.in/raghava/; #: https://github.com/raghavagps/**

Table of Content

List of Abbreviations

Acronym	Full Form
1D-CNN	One-Dimensional Convolutional Neural Network
3D	3 dimensional
AAB	Amino Acid Binary Profile
AAC	Amino Acid Composition
Acc	Accuracy
ACR	Autocorrelation
AMPs	Antimicrobial Peptides
APAAC	Amphiphilic Pseudo Amino Acid Composition
ATC	Atomic Composition
AUROC	Area Under Receiver Operating Characteristic
BLAST	Basic Local Alignment Search Tool
BRCA	Breast invasive carcinoma
BTC	Bond Composition
C-index	Concordance index
CD-HIT	Cluster Database at High Identity with Tolerance
CeTD	Composition enhanced-Transition Distribution
CHOL	Cholangiocarcinoma
CI	Confidence interval
Cox-PH	Cox proportional hazard
CSS	Cascading Style Sheets
CTD	Conjoint Triad Distribution
CV	Cross-Validation
DDR	Distance Distribution of Residues
DNA	Deoxyribose Nucleic Acid
DPC	Di-Peptide Composition
DSSP	Define Secondary Structure of Protein
DT	Decision Tree

DTR	Decision Tree Regressor
E-value	Expect value
ENT	Elastic Net Regressor
ET	Extra Tree
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
GBM	Glioblastoma multiforme
GNB	Gaussian Naïve Bayes
HR	Hazard ratio
HTTP	Hyper Text Transfer Protocol
KNN	K Nearest Neighbors
LAS	Lasso Regressor
LIHC	Liver hepatocellular carcinoma
LPC	Ligand Protein Contacts
LR	Logistic Regression
LR	Linear Regression
MAE	Mean Absolute Error
MCC	Matthew's Correlation Coefficient
MLP	Multi-Layer Perceptron
NAG	N-acetylglucosamine
NB	Naive Bayes
NS	Negative Samples
OS	Overall Survival
PAAC	Pseudo Amio Acid Composition
PCB	Physico-chemical Properties Binary Profile
PCP	Physico-chemical Properties
PDB	Protein Data Bank
PHP	Personal Home Page
PS	Positive Samples
PSI-BLAST	Position-Specific Iterated BLAST
PSSM	Position Specific Score Matrices

QSO	Quasi Sequence Order
RF	Random Forest
RFR	Random Forest Regressor
RID	Ridge Regressor
RMSE	Root Mean Square Error
RNA	Ribose Nucleic Acid
RRI	Residue Repeats Information
SASA	Solvent Accessibility Surface Area
Sens	Sensitivity
SEP	Shannon Entropy for Proteins
SER	Shannon Entropy for Residues
SMILES	Simplified Molecular Input Line Entry System
SOCN	Sequence Order Coupling Number
SPC	Shannon Entropy for Physico-chemical Properties
Spec	Specificity
SQL	Structured Query Language
SVC	Support Vector Classifiers
SVM	Support Vector Machine
SVR	Support Vector Regressor
TCGA	The Cancer Genome Atlas
TN	True Negative
TP	True Positive
TS	Total Samples
TSL	Two Sample Logo
XGB	eXtreme Gradient Boosting

List of genes and their description

Gene	Description
ADGRF4	Adhesion G Protein-Coupled Receptor F4
ALPP	Alkaline Phosphatase, Placental
ARHGEF11	Rho Guanine Nucleotide Exchange Factor 11
ATG9B	Autophagy Related 9B
BIRC6	Baculoviral IAP Repeat Containing 6
BRINP2	BMP/Retinoic Acid Inducible Neural Specific 2
BRSK2	BR Serine/Threonine Kinase 2
CACNG7	Calcium Voltage-Gated Channel Auxiliary Subunit Gamma 7
CAD	Carbamoyl-Phosphate Synthetase 2, Aspartate Transcarbamylase, And Dihydroorotase
CLDN20	Claudin 20
CLK2	CDC Like Kinase 2
CLMP	CXADR Like Membrane Protein
CNTN5	Contactin 5
CSMD3	CUB And Sushi Multiple Domains 3
DHX8	DEAH-Box Helicase 8
DNAJC9-AS1	DnaJ Heat Shock Protein Family (Hsp40) Member C9 And MRPS16 Antisense RNA 1
EPHA3	EPH Receptor A3
EVC2	EvC Ciliary Complex Subunit 2
FAM160A2	Family With Sequence Similarity 160 Member A2
FAM187B	Family With Sequence Similarity 187 Member B
HAUS5	HAUS Augmin Like Complex Subunit 5
ITGB8	Integrin Subunit Beta 8
KIAA2026	Uncharacterized Protein KIAA2026
KIF26B	Kinesin Family Member 26B
KTN1	Kinectin 1
LAMC3	Laminin Subunit Gamma 3
LINC00304	Long Intergenic Non-Protein Coding RNA 304

LINC00972	Long Intergenic Non-Protein Coding RNA 972
LINC02210-CRHR1	LINC02210-CRHR1 Readthrough
LOC100287329	Uncharacterized LOC100287329
LOC100420587	SHC Binding And Spindle Associated 1 Pseudogene
LOC101929073	Uncharacterized LOC101929073
LRP1B	LDL Receptor Related Protein 1B
MGAT4EP	MGAT4 Family Member E, Pseudogene
NOMO3	NODAL Modulator 3
NR2C2AP	Nuclear Receptor 2C2 Associated Protein
NYNRIN	NYN Domain And Retroviral Integrase Containing
OR52B6	Olfactory Receptor Family 52 Subfamily B Member 6
OR5AS1	Olfactory Receptor Family 5 Subfamily AS Member 1
OR6C76	Olfactory Receptor Family 6 Subfamily C Member 76
P4HTM	Prolyl 4-Hydroxylase, Transmembrane
PAX7	Paired Box 7
PCDH15	Protocadherin Related 15
PDE11A	Phosphodiesterase 11A
PIGO	Phosphatidylinositol Glycan Anchor Biosynthesis Class O
PLCB1	Phospholipase C Beta 1
S100A12	S100 Calcium Binding Protein A12
SIPA1L3	Signal Induced Proliferation Associated 1 Like 3
SNHG10	Small Nucleolar RNA Host Gene 10
SNTG1	Syntrophin Gamma 1
SPDYA	Speedy/RINGO Cell Cycle Regulator Family Member A
SUPT20H	SPT20 Homolog, SAGA Complex Component
SYDE1	Synapse Defective Rho GTPase Homolog 1
TAS1R2	Taste 1 Receptor Member 2
TBX3	T-Box Transcription Factor 3
TG	Thyroglobulin
TM4SF18	Transmembrane 4 L Six Family Member 18
TOP2A	DNA Topoisomerase II Alpha

TP53	Tumor Protein P53
TYK2	Tyrosine Kinase 2
WIZ	WIZ Zinc Finger
XIRP2	Xin Actin Binding Repeat Containing 2
ZNF521	Zinc Finger Protein 521

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Biological molecules are the essential entity of living organisms. The flawless coordination between these molecules is responsible for the well-being of living organisms, and a minute disturbance in this system may lead to various disorders (Pizzino et al., 2017). The major types of biomolecules are carbohydrates, lipids, nucleic acids, and proteins, which are responsible for diverse functions (Abarca-Cabrera, Fraga-Garcia, & Berensmeier, 2021). Of note, nucleic acids (DNA and RNA) are the fundamental units of life which encode unique genetic information to provide distinctive phenotypes to each organism. These nucleotides code for various amino acids, the basic units of proteins that perform different essential and complex functions. Proteins are made-up of twenty amino-acids and involved in various biological functions. For instance, antibodies are large Y-shaped proteins offers protection against foreign invaders, contractile proteins are responsible for muscle contractions and movements. In addition, proteins also acts like enzymes and carry out various catalytic reactions in the body. Whereas, storage proteins store the fundamental units such as nucleotides and amino acids for future use, and transport proteins are responsible for the movement of crucial molecules such as oxygen, throughout the body (Zaretsky & Wreschner, 2008).

One of the most important class of proteins are regulatory proteins, that are crucial accessory proteins which can regulate the various biological processes in the cell, for example transcription in which transcription factors bind to the specific DNA segments and control gene expression. Transcription factors are important regulatory proteins which can control cell differentiation, gene expression, gene-regulatory pathways, and several immunological responses (Jenner et al., 2009; P. Li, Spolski, Liao, & Leonard, 2014; Mariani, Lohning, Radbruch, & Hofer, 2004). The improper regulation of transcription factors and their associated binding areas may lead to the development of various diseases like neurodegenerative disorders, autoimmunity diseases, cardiovascular diseases, and cancer (X.-F. Chen, Zhang, Xu, & Bu, 2013; Lee & Young, 2013; Y. Yang et al., 2017). The mutations associated with the transcription factor binding sites can lead to aberrant gene expression, which may result in cancer progression and proliferation (Bushweller, 2019; M. Lambert, Jambon, Depauw, & David-Cordonnier, 2018). For instance, the mis-regulation by the NF-KB transcription factor may cause several inflammatory autoimmune disorders (Aksentijevich & Zhou, 2017; Miraghazadeh & Cook, 2018; Y. Zhou et al., 2020). In the past, researchers and clinicians attempted to inhibit or target the transcription factors and their binding sites during

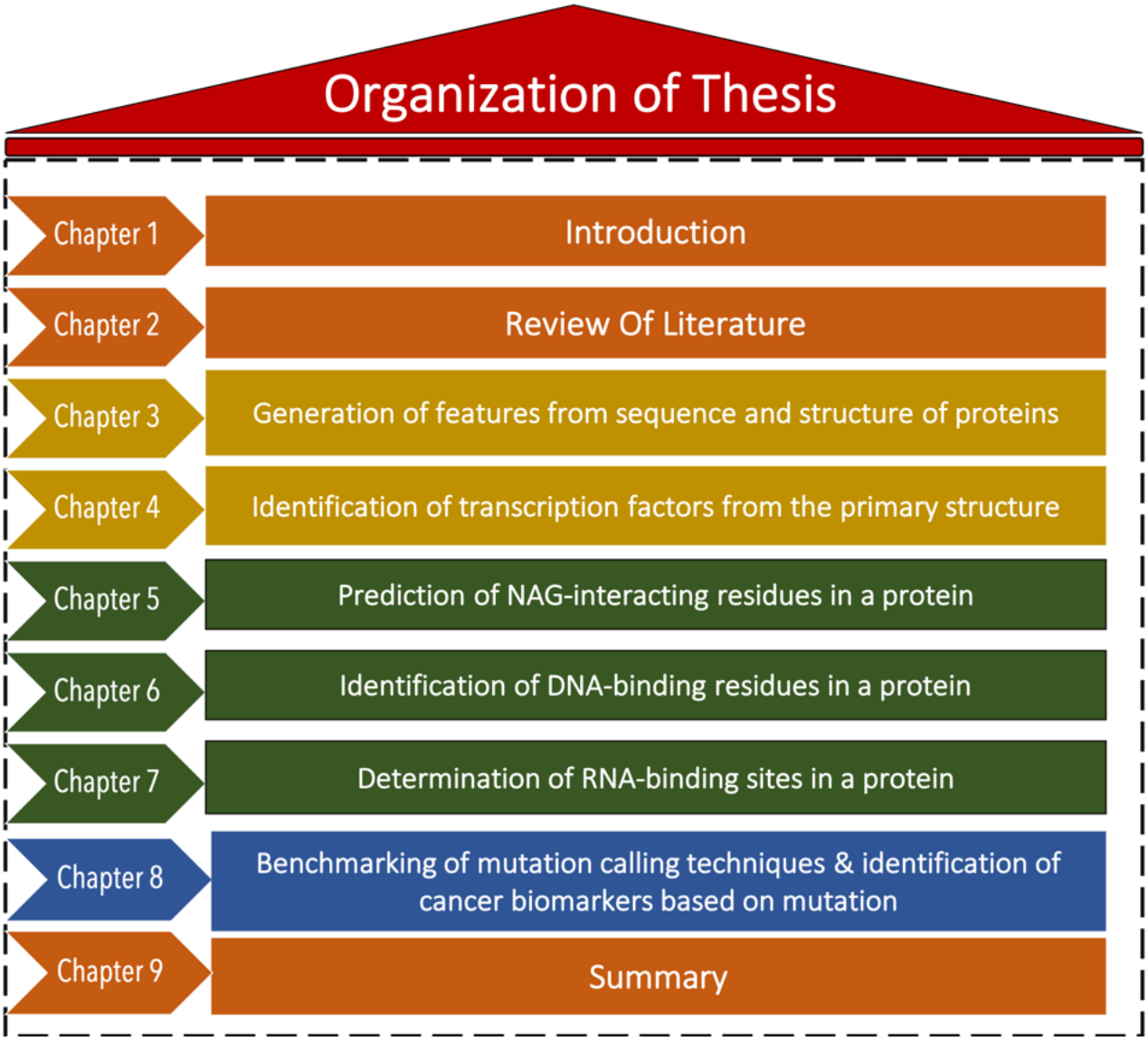


Figure 1.1: Organization of thesis in different chapters

the disease development for a better cure (Cox, Platt, & Zhang, 2015; Didiasova, Schaefer, & Wygrecka, 2018; Huh, Kim, Jeong, & Park, 2019; M. Lambert et al., 2018; H. Li et al., 2020).

Important biological, chemical, and metabolic processes (including enzyme catalysis and signal transmission) are carried out via protein-ligand interactions (Du et al., 2016). Additionally, knowledge of the processes behind protein-ligand interaction and binding, such as the lock-and-key method, the induced-fit method, and conformational selection method, might aid in the identification and development of more effective therapeutic compounds (Du et al., 2016; Gallina, Bork, & Bordo, 2014). Our life entirely depends on molecular interactions such as DNA-protein, RNA-protein, and protein-protein (Cozzolino, Iacobucci, Monaco, & Monti, 2021). DNA-protein interactions are very essential and play significant biological functions, such as transcription, regulation of gene expression, and splicing. Three-dimensional (3D) structure of protein-DNA complex allows the researchers to capture the essential information on protein-DNA interaction, which can be exploited to develop better diagnostics and treatments (Teng et al., 2021; C. Wang et al., 2014). These types of interactions are identified by experimental methods like x-ray crystallography and nuclear magnetic resonance (NMR) technologies (Y.-X. Wang, Zuo, Wang, Yu, & Butcher, 2010). Identification of 3D information of interacting residues is very crucial in structure-based drug discovery (Batool, Ahmad, & Choi, 2019). By understanding the interactions between protein and its ligands, one can design novel drugs for the number of disorders (Syriopoulou, Markopoulos, Tzakos, & Mavromoustakos, 2021). Several studies revealed that RNA-protein interactions are majorly involved in the development of human cancers, neurological disorders like Alzheimer's sclerosis, and genetic disorders (Rybak-Wolf & Plass, 2021; Schuschel et al., 2020; J. P. Taylor, Brown, & Cleveland, 2016). These interactions are also necessary for protein synthesis, post-translation modifications, viral assembly and replication.

In addition, there are several ligands like ATP (J. Hu et al., 2021), GTP (Chauhan, Mishra, & Raghava, 2010a), NAD (Ansari & Raghava, 2010), SAM (Agrawal, Mishra, & Raghava, 2020), and NAG (Patiyal et al., 2020a), which interact with the protein and perform several biological activities. N-acetyl glucosamine is an essential biomolecule that is ubiquitous in nature and presents in organisms ranging from unicellular like bacteria to multicellular organisms like humans (Naseem, Parrino, Buenten, & Konopka, 2012). A recent study suggests that NAG can be used to treat autoimmune disorders (J.-K. Chen, Shen, & Liu, 2010). In humans, NAG-signalling allows the coexistence of microorganisms ranging from bacteria to

fungi in the human gut (Nicholson et al., 2012). Many oncogenes and tumour suppressor gene products, including c-Myc, SV40 large T antigen, and p53, have been demonstrated to be changed by O-GlcNAc (Shimizu, Shibuya, & Tanaka, 2022). O-GlcNAc is a type of protein glycosylation that is nearly exclusively present in eukaryotic cells' cytoplasm and nucleus (Y. Gao, Wells, Comer, Parker, & Hart, 2001). Addition or removal of O-GlcNAc in oncoproteins, tumour suppressor proteins, and other tumor-related proteins' may play a significant role in the aetiology of cancers (Hart, Slawson, Ramirez-Correa, & Lagerlof, 2011). Cancer is intensely heterogeneous and causes millions of deaths around the entire globe. The treatments available nowadays are ineffective and may vary from patient to patient (Zaorsky et al., 2017). Cancer is associated with genetic mutations; hence, several target therapies are available that mainly target the mutated genes in order to develop new therapeutics (Jin et al., 2019). There are several mutation calling and detection techniques available to identify the genetic variants such as SNP, indel, SV, etc (Xu, 2018). In addition, TCGA (Zhining Wang, Jensen, & Zenklusen, 2016), GEO (Clough & Barrett, 2016), and ICGC (Junjun Zhang et al., 2011) provide cancer-specific mutation and transcriptomic profiles to understand the severity of the disease on the survival of the patients. In the past, several studies performed survival analyses to identify the genes associated with the mutations and aberrant gene expression, which play a crucial role in the survival of cancer patients.

1.2 Proposal's origin

In the last few decades, several attempts have been made to annotate proteins at the structural and functional levels. Annotation, building in-silico tools, and utilizing biological understanding to grasp the mechanistic point have all been extensively investigated to study the behaviour of biological macromolecules. The literature has witnessed a tremendous gain in the number of protein sequences due to the advancements in the current sequence technologies. A massive amount of sequence data has been generated and submitted to the major databases. However, the fundamental challenge is to annotate and capture meaningful information from these sequences. For instance, UniProt comprises around 231,921,734 protein sequences; however, only 567,483 sequences are manually curated in UniProtKB/Swiss-Prot database. This clearly indicates a wide gap between the number of sequences and their annotations. In order to fill the gap, it is essential to understand the function of proteins using the primary sequence information. In the past, a number of computational methods have been generated to assign the functions to protein sequences using prediction models. These models require

numerical features or descriptors representation of sequences to predict their roles. A number of features/descriptors exist in the literature that encodes the sequence information but in a scattered form. In order to aid the researchers working in the field of proteomics, it is important to compile and provide a single platform for the computation of a wide range of features using protein sequences and structures. One of the most important classes of proteins is transcription factors which control the rate of transcription of genes and play a significant role in cell differentiation, intracellular signaling, and maintaining cell-cyle. In brief, it is crucial to identify these important DNA-binding proteins. The transcription factors execute their role by interacting with DNA, and knowledge of these interactions is important to understand the overall mechanism while developing new therapeutics. Similarly, proteins accomplish their functions by interacting with other macromolecules such as DNA, RNA, and proteins or by interacting with other molecules such as ATP, ADP, GTP, NAD, NAG, SAM, etc. Accurate identification of these interactions requires structure information, but it is a tedious and time-expensive process. Of note, it is the need of the hour to develop better computational approaches to identify the biological interaction using sequence information. The biological interactions can be hampered in disease conditions, which could be due to the mutation at the genome level. In the literature, it has been reported that cancer is significantly associated with genetic mutations. The mutations in the oncogenes can reduce the survival of cancer patients. So, the number of mutation callers has been developed to capture the mutation at the genome level. It is important to choose the accurate mutation calling technique to identify prognostic biomarkers using the mutation information. Hence, to understand the advantages and disadvantages of the mutation calling techniques, it is essential to benchmark them to identify the diagnostic and prognostic biomarkers in the cancer cohorts. Innovative prognostic techniques can be employed to give more exact risk prediction and hence more effective therapeutic planning.

1.3 Objective of thesis

In order to overcome the drawbacks mentioned above, we have made an effort to understand and explore the protein annotation and usage of mutation profiles to develop the prognostic model. The present work is majorly divided into three parts: (i) Functional annotation of Protein, (ii) Identification of Protein-molecules interaction, and (iii) Prediction of cancer-associated mutations.

Due to the continuous improvement in the sequencing technologies, the respective repositories are flooded with the sequence information. But at the same time, a large proportion of the sequences in the databases are remaining unannotated. To address this issue computationally, number of methods have been generated in the last few decades which are based on various numerical representation of the structures and sequences such as composition, binary profile, alignment information, and many more. Throughout the literature, myriads of sequence and structure based features are present, but in the scattered form. Hence, it is the need of the hour to bring all the possible features of a protein at one platform. Moreover, optimal implementation of these features to functionally annotate crucial proteins, such as transcription factors, is an important task. Several experimental procedures are available to functionally annotate a protein as a transcription factor, but they are laborious and time-consuming in nature. On the other hand, the computational approaches based on structure and sequences features are easy to use and reliable. However, structure-based method comes with the drawback of the availability of the 3D structure, i.e., structure-based methods need the tertiary structure of proteins to make the predictions. On the other side, sequence-based method are fast and perform equivalent to the structure-based method. Therefore, it is crucial to develop sequence-based method to functionally annotate a protein into transcription factors with the reliable accuracy and performance. To fulfil the first objective, the aforementioned issues needed to be addressed.

In the biological systems, proteins are the functional unit which execute their function by interacting with either other macromolecules such as nucleic acids like DNA or RNA, or with small molecules like NAG. Understanding of these interactions is a must to explore the underlying mechanism cascade, and knowledge of these interactions is an important aspect of drug-designing. These interactions can be explored via the 3D protein-molecule complex structures. X-ray crystallography and nucleic magnetic resonance are the few techniques to determine the tertiary structures, but they depend upon the stability of the complexes in the in-vitro conditions. Due to the limited availability of 3D structure of the proteins, the application of the existing structure-based approaches is limited. On the other hand, due to the advancement of the sequencing technologies, ample of sequences are available to train and develop the accurate prediction models. Therefore, it is the absolute necessity to develop the sequence-based prediction models with the ability to annotate the proteins at the residue level. In order to overcome the above mentioned drawbacks, second objective of our study was formulated.

As mentioned above, most of the biological functions are the outcome of the interactions between the macromolecules and various molecules. A slight alteration in these interactions may lead to fatal diseases like cancer. One of the major reasons for the disruption or abnormal behaviour of these interactions is mutation at the genome level. Hence, it is important to identify these mutations with high precision. For the same, number of mutation calling techniques were developed based on different statistical algorithms and concepts. But, it is difficult to choose one technique among the others for a particular task. Hence, benchmarking of the existing methods is an essential need of the moment. Further, there are number of methods developed to identify the diagnostic and prognostic biomarkers in various cancers using different types of genomic profiles. Similarly, it is important to explore the role of mutations to determine the prognostic and diagnostic biomarkers in different cancer types. The third part of the study is designed to overcome the issues mentioned above.

1.4 Organization of Chapters

In this study, we have tried to address the aforementioned issues in three major objectives. At first, we tried to annotate the functions of proteins at the sequence and residue levels. For this, under first objective, we have compiled all the possible features of protein sequences and structures reported in the literature and developed an in-silico tool named “Pfeature”. In order to annotate the regulatory protein, such as the transcription factor, we developed a computational method, “TransFacPred” to predict the transcription factors using sequence information. To fulfil the second objective, we have compiled three chapters to explore the interactions between biomolecules and proteins, we have developed user-friendly web servers and standalone packages for DNA (DBPred), RNA (PPRInt2), and NAG (NAGbinder). In addition, to annotate the risk-associated biomarkers using mutation profiles, we benchmarked different mutation calling techniques and provided a python-based pipeline on GitHub, which can be further used on other cancers to analyse and predict high-risk associated genes. We have organized the entire thesis into nine chapters. The information it contains is as follows:

Chapter 1: This chapter introduces the importance of biological molecules in the living systems. It highlights the gap between the rate of increase of protein sequences in the respective database and their annotation. Hence, this chapter provides the importance of annotation of proteins. Protein molecules execute their functions by interacting either interacting with itself or with other molecules, therefore, this chapter also provides the overview regarding the

necessity of studying interaction between the proteins and other molecules. Further, it introduces the importance of mutations in the identification of diagnostic and prognostic biomarkers in the cancer patients, as cancer is the second leading cause of death worldwide.

Chapter 2: This chapter provides the review of the literature on tools for annotation of proteins using sequence and structure information. Importance of protein molecule interaction and tools to predict the same, tools developed for predicting interaction between proteins and nucleic acids such as DNA and RNA, and finally end the chapter with the role of mutations in cancer and techniques to detect the mutations at the genome level. In a nut shell, this chapter explains why this study was conducted.

Chapter 3: This chapter focuses on the first objective of the thesis, which is the development of a platform known as "Pfeature", to calculate a wide range of features from protein sequences and structures. In this chapter, first we tried to explore the literature to gather the information about the features that can be generated using protein structure and sequences. Moreover, tried to come up with the novel features based on the patterns sequences follows pertaining to a particular function. This chapter provides the detailed description of features computed by each module of Pfeature such as composition, binary profiles, evolutionary information, structure, and patterns. Moreover, it also provides the information about the model building module which do not calculate features but develop classification and regression models based on the features calculated using other modules and submodules of Pfeature. Pfeature provides different types of features for sequences and structures, for instance, it can calculate around 200000 features from a single sequence. It provides the largest number of features known to date. This chapter also discusses the utility of the developed resource for developing prediction models to functionally annotate the proteins at sequence and residue level. It is provided to the scientific community in various forms such as web-server, Python-based standalone, Python-based library, and Python-based scripts.

Chapter 4: This chapter is about the annotation of transcription factors using primary structure information. We created a computational approach to predict the transcription factors from a provided protein sequences. Here, we have provided a freely-accessible and user-friendly web-interface named "TransFacPred", where users are allowed to submit multiple sequences in FASTA format. This tool will predict if the submitted protein is a transcription factor or not. It has two models implemented at the back-end, one uses amino acid composition as input feature to make prediction, and other implemented hybrid approach which is a combination of

alignment-free approach, i.e., machine-learning based model and alignment-based approach by implementing BLAST. Other than the web-interface, this method is also available as standalone on TransFacPred web-server and GitHub.

Chapter 5: This chapter explores the interaction between proteins and N-acetylglucosamine (NAG), as NAG is a ubiquitous in nature and performs several important functions in organisms ranging from bacteria to humans. We created a bioinformatic-ware to predict the NAG-interacting residues in a protein sequence. It also has a web-server named as "NAGbinder", where users can submit multiple sequences in the FASTA format and NAG-interacting residues gets highlighted in the result page, which is downloadable in various formats. It has two major modules for the prediction, where one uses binary profile to identify the NAG-interacting residues in a protein sequence. The other module calculates PSSM profile to determine which residue can interact with NAG and which cannot. This method is also available as the Perl- and Python-based standalone accessible at web-server and GitHub.

Chapter 6: This chapter is also about the annotation of a protein but at a residue level to determine the DNA binding residues in a protein sequence. This method deployed as web-interface for the easy accessibility, other than that Python-based standalone is provided for making prediction for a huge dataset or when the internet is not available. This method is provided by the name "DBPred". We have implemented 1D-CNN based models develop using features like binary profile, physico-chemical properties profile, PSSM profile, and their combination termed as hybrid profile. It also has a user-friendly web-server, where users are allowed to submit multiple sequences in the FASTA format. DNA-interacting residues gets highlighted in the result page, which can be downloadable as text document, image, and PDF file. It has three major modules, first one uses binary profile and physico-chemical properties profile for determining DNA-interacting residues, where second module computes PSSM profile, and third module uses hybrid profile to make prediction.

Chapter 7: Similar to chapter 6, this chapter also traverse annotation at the residue level for the identification of RNA-binding sites in a protein sequence. It is an updated version of Pprint tool. We have used the largest benchmark dataset in this study. To serve the scientific community, we have provided this method as web-server named "Pprint2", and standalone. In this study, we have also implemented 1D-CNN to predict the RNA-interacting residues in a protein sequence, by generating three different features as binary profile, physico-chemical properties profile, and PSSM profile. It also has a user-friendly web-server, where users are

allowed to submit multiple sequences in the FASTA format and RNA-interacting residues gets highlighted in the result page, which is downloadable in .txt, .png, and .pdf formats. It has two module, first one uses binary profile and physico-chemical properties profile for determining RNA-interacting residues, where second module computes PSSM profile to make prediction.

Chapter 8: As it has been demonstrated in the literature that cancer is associated with the genetic mutations. Hence, it is of utmost importance to explore the prognostic role of mutations to understand its influence on the overall survival of the cancer patients. Several methods has been designed in the past to accurately detect the mutations at the genome level but which one should be used is not stated anywhere, hence it depends upon the users' intuition to decide which mutation calling technique to be used. Therefore, we tried to benchmark four widely used mutation calling techniques such as Mutect2, MuSE, Varscan2, and SomaticSniper, using VCF and MAF files of liver cancer patients from TCGA cohort, to explore the prognostic potential of the mutation profiles. After that we tried to identify the diagnostic and prognostic biomarkers for liver cancer patients' mutation profiles by developing classification and regression models. We have provided the Python-based end-to-end pipeline which needed VCF/MAF files to determine diagnostic and prognostic biomarkers available on GitHub.

Chapter 9: In this chapter, we have summarized the thesis work by providing a quick overview of the study and its contribution to the area of research. Figure 1.1 provide the overall organization of the thesis.

Chapter 2
Review of literature

2.1 Overview

The cell is the smallest and fundamental unit of life that makes up the living organisms ranging from bacteria to humans. Several biological processes inside the cell are responsible for the well-being of an organism which is regulated by various biomolecules. Proteins are the essential biomolecule that play crucial roles in the body. Most of the work inside the cell is done by the proteins and is responsible for the structure, function, and regulation of various processes. There are number of interactions occurs in each organism at any given point of time. Proteins are the most important molecules that engage in these interactions. Proteins are big and complex molecules composed of many smaller components known as amino acids that are linked together in a lengthy chain. Proteins make up about 20% of the human body's substance and are a vital component of cells, muscles, tissues, ligaments, and so on. Proteins control all necessary biological tasks such as cell replication and regulation, enzymatic reactions, energy production, molecule transport, muscle development, wound healing, and muscle and bone restoration. To fulfil these activities, proteins create complexes with other molecules such as proteins, peptides, small molecules, ions, and so on. Proteins can act as antibodies and protects against foreign particles such as bacteria and viruses. They also behave as enzymes and regulates the chemical reactions inside the cell. In addition, they can act as the messenger molecules like hormones to transmit signals during various biological processes. Moreover, proteins provide structural support, allow them to move, and work as transporter and storage units. In order to develop new therapeutics to treat various diseases, one has to study the interactions between the proteins and molecules. By using a competitive inhibitory mechanism, these compounds compete with naturally binding molecules for binding at the active site. Various kinase inhibitors, including as imatinib, sorafenib, and nilotinib, have been engineered to impede ATP binding (Jianming Zhang, Yang, & Gray, 2009). Similarly, Ras inhibitors such as SML-8-73-1 and SML-10-70-1, which mimic GDP and bind to Ras, exist (Lim et al., 2014; Shima et al., 2015). Mutations at the genomic level can hamper these interactions, which may lead to various diseases including cancer. Understanding of these mutations at the genomic level can be an important aspect to understand the progression of a disease. Figure 2.1 represents the different types of protein annotation based on functional annotation and interactions.

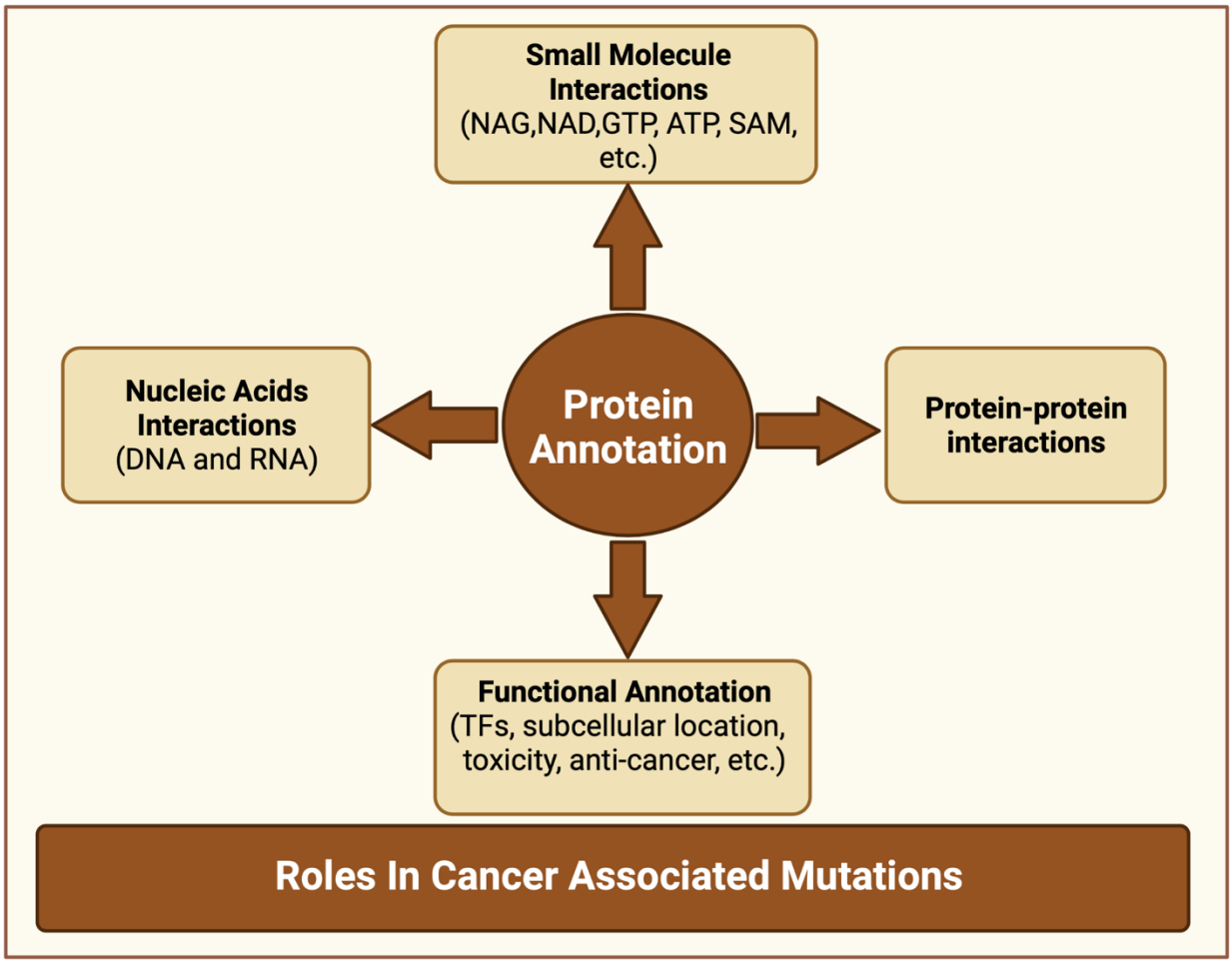


Figure 2.1: Different ways to annotate a protein based on its function and interactions with small molecules, nucleic acids, and proteins

2.2 Feature generation for annotation

A protein sequence can be annotated by using similarity search against well-annotated databases but the rate of increase in the number of sequences is far ahead than its rate of annotation. Hence, only similarity search based methods are not sufficient to meet the pace of sequence generation, therefore, there is a need of prediction models which can learn the patterns or features in the sequences belonging to a particular class of proteins and can use the same to make the prediction for uncharacterized proteins. However, to develop the models there is a pre-requisite to represent the sequence information via some numerical vectors. Therefore, number of efforts have been made in the last two decades for the development of platforms which can calculate and provide the numerical representation for various biological macromolecules such as DNA, RNA, and Proteins, and for molecules such as chemicals. Table

2.1 provides the brief description of the tools developed for calculate features for different molecules. PROFEAT (Z. R. Li et al., 2006) was one of the first webserver developed in 2006, for calculating structural and physico-chemical properties using amino acid sequences of proteins and peptides, this tool computed 6 types of features that include 51 descriptors and 1447 descriptor values. The same tool was updated in 2011 (Rao, Zhu, Yang, Li, & Chen, 2011) and able to calculate 11 types of features from amino acid sequences, 400 features for small molecules, and additional features for protein-protein and protein-small molecule interactions; which is again updated in 2017 (P. Zhang et al., 2017) to add the facility of calculating network descriptors which added another 329 network descriptors and protein-protein interaction descriptors. Then, PyDPI (Cao, Liang, et al., 2013)was developed which is a Python package developed to calculate features to understand the drug-protein interaction and hence devoted for chemoinformatics, bioinformatics, and chemogenomics studies. PyDPI provided 9890 features belonged to 52 types of features for proteins, 615 features belonged to 13 types of descriptors for drug, and seven type of molecular fingerprints. Further, Propy (Cao, Xu, & Liang, 2013)abbreviation for protein in Python was developed in 2013, that generated different modes of pseudo amino acid composition discovered by Chou Fasman, along with that it provided 5 feature groups composed of 13 features. Till now, methods were developed particularly for proteins, then Pse-in-One (B. Liu, Liu, et al., 2015) came in 2015 with the ability to handle DNA, RNA, and proteins and generated large number of features using 28 different modes. Its update was developed in 2017 by the name Pse-in-One 2.0 (B. Liu et al., 2017) which was improved by adding 23 new pseudo component modes along with new features for analysis. Other than Python, R was also used to generate features for proteins in 2015 in tool Protr/ProtrWeb (Xiao, Cao, Zhu, & Xu, 2015) which provided 22 different types of features that constitutes 22700 feature values. Further, PDBparam (Nagarajan et al., 2016) was developed to extract more than 50 different structure based features for any submitted structure. BioTriangle (Dong et al., 2016) was another tool developed by Dong et. al. to represent the molecules like nucleic acids, proteins, and chemicals, and their interaction using different feature vectors. POSSUM (J. Wang et al., 2017) was another very interesting tool which represents the protein sequences via 21 different types of PSSM based descriptors. Then, very powerful method named BioSeq-Analysis (B. Liu, 2019) was developed which was able to complete three main steps of machine learning at a single platform, i.e., it extracts features from sequences of DNA, RNA, and protein, construct the optimal predictor, and evaluate the performance of the predictor. Its update was developed by the name BioSeq-Analysis 2.0 (B. Liu, Gao, & Zhang, 2019) in 2019 to analyse the sequences of DNA, RNA, and proteins at

residue and sequence level by providing 26 features at residue level and 90 features at sequence level. Other than that, another tool iFeature (Z. Chen et al., 2018) was made available, which can calculate the wide range of features for protein and peptide sequences, other than that this programme was able to incorporate 12 various types of regularly used feature clustering, selection, and dimensionality reduction techniques, considerably easing machine-learning model training, analysis, and benchmarking. Its update was also developed recently by the name iFeatureOmega (Z. Chen et al., 2022) in which Chen et al. widened the spectrum by providing the facility of analysing and visualizing 189 representations for biological sequences, structures and ligands. Other than these tools, Chen et. al. built another powerful tool iLearn (Z. Chen et al., 2020) and its update iLearnPlus (Z. Chen et al., 2021) to provide wide spectrum of features for sequence for proteins and nucleic acids, along with the facility of model building and visualization. PyFeat (Muhammod et al., 2019) is another very crucial Python-based toolkit which can be used to calculate features for DNA, RNA, and protein sequences, which was built to provide the local information by capturing information about the interaction of neighbouring residues, which further select the best features using AdaBoost classifier, hence provide the effective features.

Table 2.1: List of computational resources developed for computing features of protein using direct and indirect methods

Name	Description	Weblink	Year	Working Status
PROFEAT	Computes features like Structural and physico-chemical using amino acid sequence	http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi	2011	No
PyDPI	Python Package for computing features for chemoinformatics, bioinformatics, and chemogenomics Studies	https://sourceforge.net/projects/pydpicao/	2013	Yes
Propy	Protein in python (propy) for calculating the widely used structural and physicochemical features	http://code.google.com/p/protpy/downloads/list	2013	Yes
Pse-in-One	Features based on pseudo components of DNA, RNA, and protein sequences	http://bioinformatics.hitsz.edu.cn/Pse-in-One/	2015	No
Protr/ProtrWeb	R based package and web server for generating various numerical representation schemes of protein sequences	http://protrweb.scbdd.com/	2015	Yes
PDBparam	Calculates structural parameters for proteins	http://www.iitm.ac.in/bioinfo/pdbparam/	2016	Yes

BioTriangle	Various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions	http://biotriangle.scbdd.com	2016	Yes
POSSUM	PSSM profiles based feature generation	https://possum.erc.monash.edu/	2017	Yes
BioSeq-Analysis	Machine learning approaches based analysis of DNA, RNA and protein sequence	http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/	2017	Yes
Pse-in-One 2.0	Improved package for calculating features based on pseudo components of DNA, RNA, and protein sequences	http://bliulab.net/Pse-in-One2.0/	2017	Yes
iFeature	Python Toolkit and Web Server for Calculating Structural and Physicochemical Feature from Protein and Peptide Sequences	https://ifeature.erc.monash.edu/	2018	Yes
PyBioMed	Python library for feature calculation for chemicals, proteins and DNAs and their interactions	http://projects.scbdd.com/pybiomed.html	2018	Yes
iLearn	Feature generation, machine learning analysis, and model development for DNA, RNA, and Protein sequences	https://ilearn.erc.monash.edu/	2019	Yes
PyFeat	Feature generation for DNA, RNA, and Protein sequence using Python	https://github.com/mrzResearchArena/PyFeat/	2019	Yes
BioSeq-Analysis2.0	Improved package for machine learning approaches based analysis of DNA, RNA and protein sequence	http://bliulab.net/BioSeq-Analysis2.0/home/	2019	Yes
Seq2Feature	A comprehensive tool for the feature-extraction	https://www.iitm.ac.in/bioinfo/SBFE	2019	Yes
iLearnPlus	Sequence analysis, prediction and visualization of DNA, RNA, and Proteins	https://ilearnplus.erc.monash.edu/	2021	Yes
iFeatureOmega	Python Toolkit and Web Server for engineering, visualization and analysis of features from molecular sequences, structural and ligand datasets	http://ifeatureomega.erc.monash.edu	2022	Yes

2.3 Annotation of transcription factors

Transcription factors are the one of the most essential class of proteins which regulates transcription inside the cell and therefore controls many processes such as cell proliferation, expression of genes, and so on. Moreover, a minute alteration in the activity of transcription factors may lead to bad impact on the activity of genes that play a role in the cell division cycle, and hence, can be a significant contributor in oncogenesis. Therefore, it is important to annotate the transcription factors to understand the underlying mechanism which can be exploited to

develop new therapeutics for several diseases. There are several databases which are developed to store and provide access to the information related to the transcription factors in various organisms, such as Jasper (Castro-Mondragon et al., 2022) which was recently updated and is a freely-accessible database for the non-redundant transcription factor binding profiles for transcription factors for six major taxonomic groups. Other than that, there are databases which contained information specific to the organisms, such as, TRANSFAC (Wingender, Dietze, Karas, & Knuppel, 1996) which is specific to the eukaryotes and comprise information about the transcription factors and their DNA binding sites; Riken Transcription Factor Database (Kanamori et al., 2004) which provides the information about the transcription factor genes and related genes in mouse; DRTF (G. Gao et al., 2006) presents putative transcription factors in *Oryza sativa L. ssp. indica* and *ssp. Japonica*; CollecTF (Kilic, White, Sagitova, Cornish, & Erill, 2014) provides information for transcription factor binding sites across the bacteria domain; RedFly (Rivera, Keranen, Gallo, & Halfon, 2019) is the database of known insect transcriptional cis-regulatory modules, cis-regulatory module segments, predicted cis-regulatory modules , and transcription factor binding sites; ReMap (Cheneby et al., 2020) is the database of Homo sapiens and Arabidopsis thaliana transcriptional regulators; PlantPAN (Chow et al., 2019) is the database of transcription factor binding sites, corresponding transcription factor, and other important regulatory elements in plants; Yeastract (Monteiro et al., 2020) provides information on regulatory associations between transcription factors and target genes in Saccharomyces cerevisiae; and FlyMine (Lyne et al., 2007) is the database for genomic and protein data for Drosophila site-specific transcription factors.

Other than the knowledgebases, several prediction tools are present in the literature with the ability to classify protein sequences into transcription factors. Zheng et al. developed TFMiner (Zheng et al., 2008) in 2008 for the classification of transcription factors by combining two different algorithms such as SVM and error-correcting output coding. They were able to achieve the highest accuracy of 97.83% for their method. SABINE method was then developed by implementing support vector regression to predict the binding specificities of transcription factors. Another widely used tool TFPredict and SABINE (Eichner et al., 2013) was developed in 2013 which performs four major operations to perform classification, such as, first is to identify the transcription factors among the other proteins, second is to determine the structural superclass of transcription factors; in third step DNA binding domains are identified; and then it predicts cis-acting DNA motifs. TFpredict was developed as a companion to SABINE, and it predicts the DNA-motif bound by a transcription factor based on its amino acid sequence,

superclass, DNA-binding domains, and organism. BART (Zhenjia Wang et al., 2018) uses human and mouse ChIP-seq data to predict the functional factors that can bind to the cis-regulatory regions. It implemented an optimized feature selection algorithm with Adaptive Lasso at the back end. The diffTF (Berest et al., 2019) tool classifies the transcription factors into activators and repressors. Further, deep learning based method developed recently named as DeepTFactor (Kim, Gao, Palsson, & Lee, 2021) with the ability to classify the protein sequences into transcription factors, with a very high accuracy. They have also provided the list of 4674808 transcription factors predicted from 73873012 protein sequences in 48346 genomes. There are source specific tools available in literature such as PredicTF (Oliveira Monteiro et al., 2022) which can prediction and categorize the potential bacterial TF in complex microbial communities as well as single species. This method have used BacTFDB which is curated from UniProt and CollecTF, to develop the deep learning model to predict transcription factors and their families in genomes and metagenomes. The details and working status of each tool is compiled in Table 2.2.

Table 2.2: Computational resources developed for predicting transcription factor

Name	Description	Weblink	Year	Working Status
TFMiner	Identification and classification of transcription factor using SVM and ECOC	http://itfp.biosino.org/itfp/TFMiner	2008	No
SABINE	Prediction of the binding specificity of transcription factors using support vector regression	http://www.cogsys.cs.uni-tuebingen.de/software/SABINE/downloads/index.htm	2008	Yes
TFPredict	Machine learning based classification and structural characterization of transcription factors	https://github.com/draeger-lab/TFpredict	2013	Yes
BART	Predicts functional factors such as transcription factors and chromatin regulators that bind at cis-regulatory regions	http://bartweb.org/	2018	Yes
diffTF	Helps in analysing differential Transcription factors activity and in classification of Transcription Factors as activator or repressor	https://git.embl.de/grp-zaugg/diffTF	2019	Yes
DeepTFactor	Predicts whether a protein is a transcription factor	https://bitbucket.org/kaistsystemsbiology/deeptfactor/src/master/	2021	Yes
PredicTF	Predicts bacterial transcription factors in complex microbial communities	https://github.com/mdsufz/PredicTF	2021	Yes

2.4 Protein-molecule interaction

The interactions between proteins and other molecules perform critical biological, molecular, and metabolic processes such enzyme catalysis and signal transmission. Furthermore, knowing the mechanisms involved in protein-ligand interaction and binding, such as the lock and key method, induced-fit approach, or conformational selection approach, will aid in drug discovery and design (Du et al., 2016). The methods available for determining the interactions can be broadly divided into three categories such as, sequence-based, structure-based, and hybrid methods. Due to the dependency of structure-based approaches and hybrid methods on the availability of structures, their applications are limited as determination of the structure for complexes using experimental approaches like x-ray crystallography and NMR is a very tedious and time-expensive processes. On the other hand, due to the improvements in the sequencing technologies the sequences are increasing at an exponential rate. Therefore, the sequence-based methods are becoming popular as they can perform equivalently as structure-based and hybrid approaches. Thus, number of methods have been developed in the past using various machine-learning and deep-learning techniques. Sequence based methods can be classified into two classes such as similarity search methods and pattern based methods.

In case of similarity based approaches, the amount of homology between the homolog and the query sequence is significantly weighted in similarity search algorithms. If two proteins have a significant sequence similarity, they are likely to have the same structure and binding pocket. In order to detect such homologs in various databases, tools such as BLAST, PSI-BLAST, and HMM are commonly utilised. If the sequence similarity is low, domains or protein regions with the highest similarity are examined for annotation. iPfam is one such example, in which domains are searched in a given protein sequence and the domain with the highest similarity to a bound ligand is presented (Finn, Miller, Clements, & Bateman, 2014). The algorithm based on this approach is quick and simple to understand.

Methods entail using structural information to locate binding pockets and interaction residues are highly accurate. However, many protein structures are not accessible in the structural databank due to experimental method restrictions, and the majority of them are membrane proteins, which are the first option as therapeutic targets for most pharma firms and academics. Rapid advances in sequencing technology have resulted in millions of sequences that remain unannotated. In that situation, sequence-based algorithms that can predict the interaction

residue in a given protein sequence were required. Prediction models are built using sequence characteristics such as evolutionary profile, binary profile, sequence-derived structural features such as secondary structure, surface accessibility area, and so on. These models are then used to forecast which residues will interact in a new protein sequence. ATPint (Chauhan, Mishra, & Raghava, 2009a), TargetATPsite (Yu, Hu, Huang, et al., 2013), NAGbinder (Patiyal et al., 2020b). SAMbinder (Agrawal et al., 2020), ATPsite (K. Chen, Mizianty, & Kurgan, 2011), GTPBinder (Chauhan, Mishra, & Raghava, 2010b), FADPred (N. K. Mishra & Raghava, 2010a), NsitePred (K. Chen, Mizianty, & Kurgan, 2012), VitaPred (Panwar, Gupta, & Raghava, 2013), TargetVita (Yu et al., 2014), and others have been developed in the literature for predicting small molecule interactions in proteins.

2.4.1 Methods for predicting DNA interacting residues

Being one of the most important interactions, the binding site prediction of DNA on protein is always a topic of interest in the scientific community. Ample of methods have been developed in the past based on various algorithms, datasets, programming languages, etc. to identify the interacting residues on protein as DNA-protein interactions are important in a variety of biological functions such as transcription, gene expression control, and splicing. The DNA-protein interactions are the fundamental type of interactions for almost all biological activities and processes, such as transcription, gene expression regulation, repair, packaging of chromosomal DNA, and splicing (Aeling et al., 2007; Schonbach et al., 2011; Si, Zhao, & Wu, 2015; Wong, Li, Peng, & Wong, 2016). The first DNA-protein complex interaction was observed in 1984 using X-ray crystallography (Y.-H. Cai & Huang, 2012). To date, with the advancements of high-throughput experimental technologies such as protein binding microarray (PBM) (Berger et al., 2006), ChIP-seq, ChIP-chip (Collas, 2010), mChIP (Furlan-Magaril, Rincon-Arano, & Recillas-Targa, 2009), nuclear magnetic resonance (NMR) spectroscopy (Ponting, Schultz, Milpetz, & Bork, 1999), electrophoretic mobility shift assays (EMSAs) (S Jones, van Heyningen, Berman, & Thornton, 1999) and protein microarray assays (Ho, Jona, Chen, Johnston, & Snyder, 2006), etc. Several experimental methods are used to confirms interactions between protein and DNA-binding residues. The availability of empirical data on 3D structures of protein-DNA complexes and binding residues; supports biologists and researchers to reveal the essential knowledge on protein-DNA interactions identification. This analysis reveals information regarding amino-acid properties, conformational changes of DNA molecules, the importance of hydrogen bonds, electrostatic,

van der Waals interactions, etc. (Jayaram, McConnell, Dixit, Das, & Beveridge, 2002; Lejeune, Delsaux, Charloteaux, Thomas, & Brasseur, 2005; Nadassy, Wodak, & Janin, 1999; Nagarajan, Ahmad, & Gromiha, 2013). Currently, a huge number of experimentally curated DNA-proteins interactions have registered in the protein data bank (PDB) (Rose et al., 2015). It is a time-consuming, costly and labour-intensive process of capturing the protein-DNA interaction information using experimental techniques. Today thousands of proteins have been sequenced and submitted in public repositories, and it becomes very challenging to identify such types of interactions using empirical methods. Due to the biological significance of protein-DNA interactions, it is essential to capture the correct information from the available sequencing data. Therefore, from the last few decades, many in silico methods have been generated to predict and identify protein-DNA binding interactions. (Miao & Westhof, 2015; Schmidtke & Barril, 2010; Si et al., 2015; Liangjiang Wang & Brown, 2006). Based on input features, these tools were classified into four major categories, i.e., sequence-based methods (Hwang, Gou, & Kuznetsov, 2007), structure-based methods (Susan Jones, Barker, Nobeli, & Thornton, 2003; Tjong & Zhou, 2007), evolutionary methods (Chowdhury, Shatabda, & Dehzangi, 2017), and hybrid methods (B.-Q. Li, Feng, Ding, & Cai, 2014; R. Liu & Hu, 2013).

In the recent years, a number of machine learning techniques, such as the Hidden Markov Model (HMM) (Shanahan, Garcia, Jones, & Thornton, 2004), random forest (K. K. Kumar, Pugalenthi, & Suganthan, 2009; W.-Z. Lin, Fang, Xiao, & Chou, 2011; Nimrod, Schushan, Szilagyi, Leslie, & Ben-Tal, 2010; Szilagyi & Skolnick, 2006), support vector machines (B. Liu, Wang, & Wang, 2015), Naive Bayes classifier (Lou et al., 2014), etc., have been used for the better prediction of DNA-binding residues (29097781). Several protein-DNA binding residues prediction methods are available in the form of webservers, standalone packages, and online services, including DBS-Pred (Ahmad, Gromiha, & Sarai, 2004), DisoRDPbind (Peng, Wang, Uversky, & Kurgan, 2017), and ProteDNA (Chu et al., 2009), BindN (Liangjiang Wang & Brown, 2006), BindN+ (Liangjiang Wang, Huang, Yang, & Yang, 2010), BindN-RF (Liangjiang Wang, Yang, & Yang, 2009), hybridNAP (Jian Zhang, Ma, & Kurgan, 2019), funDNApred (Amirkhani, Kolahdoozi, Wang, & Kurgan, 2020), ProNA2020 (Qiu et al., 2020) etc. The prediction methods like ProteDNA, DP-Bind, BindN. are entirely dependent on the Support vector machine. In contrast, DISIS, DBS-Pred, and DBS-PSSM algorithms developed on the k-nearest neighbor technique. Whereas, few most popular methods, such as hybridNAP, DRNApred (J. Yan & Kurgan, 2017), and DisoRDPbind (Peng et al., 2017), use regression models to predict DNA-binding residues. EL_PSSMRT (J. Zhou, Lu, Xu, He, & Wang, 2017),

DNAbind (M. Kumar, Gromiha, & Raghava, 2007) utilizes evolutionary information (Position Specific Score Matrix (PSSM)) for the better prediction of DNA-binding residues in the amino-acid sequences. Few studies have shown that the structure-based methods and hybrid methods (which use both structure and sequence information) perform better than sequence-based methods.

But, structure-based and hybrid methods fail when corresponding 3D information is not available for a given query protein sequence. Although some homology-based structure prediction methods are currently available; but, there is a considerable difference between the actual structure (i.e., experimental) and the predicted structure. Due to advancements in new technologies, there is a substantial increment of the protein sequences, so the gap between structure and sequencing data becomes more voluminous. Therefore, it is very challenging to capture the 3D information for all the protein sequences available today. Besides this, computational models are more efficient, economic, and convenient for the prediction of DNA-binding residues using sequence-based features. Table 2.3 provides the list of the prediction methods developed for the identification of the DNA interacting residues in a protein sequences.

Table 2.3: Methods developed for predicting DNA-interacting residues in a protein from its sequence and structure (Partially adopted from Patiyal, Dhall, & Raghava, 2021)

Name	Description	Weblink	Year	Working Status
DBS-Pred	Neural network based method	http://www.netasa.org/dbs-pred/	2004	No
DBS-PSSM	PSSM-based prediction method	http://www.netasa.org/dbs-pssm/	2005	No
Pro-DNA	Structure-based prediction method	http://proteomics.bioengr.uic.edu/pro-dna	2005	No
BindN	SVM based DNA/RNA-binding site prediction	http://bioinformatics.ksu.edu/bindn/	2006	No
DNABindR	Naïve Bayes classifier based method	http://turing.cs.iastate.edu/PredDNA/predict.html	2006	No
DP-Bind	PSSM-based prediction method	http://lcg.rit.albany.edu/dp-bind/	2006, 2007	Yes
BindN-RF	RF-based prediction method DNA-interacting residues	http://bioinfo.ggc.org/bindn-rf/	2009	No
DBindR	Evolutionary information based prediction method	http://www.cbi.seu.edu.cn/DBindR/DBindR.htm	2009	No

BindN+	PSSM-based prediction method	http://bioinfo.ggc.org/bindn+/	2010	No
MetaDBSite	Integrative tool for the prediction	http://projects.biotec.tu-dresden.de/metadbsite/ http://sysbio.zju.edu.cn/metadbsite	2011	No
DNABR	RF-based prediction method	http://www.cbi.seu.edu.cn/DNABR/	2012	No
DNABind	Structure-based prediction method	http://mleg.cse.sc.edu/DNABind/	2013	Yes
SPOT-Seq (DNA)	Structure-based prediction method	http://sparks-lab.org/	2014	Yes
PDNAsite	SVM and ensemble learning based prediction method	http://hlt.hitsz.edu.cn:8080/PDNAsite/	2016	No
CNNsite	Convolutional Neural Network based method	http://hlt.hitsz.edu.cn:8080/CNNsite/	2016	No
TargetDNA	Evolutionary information based prediction method	http://csbio.njust.edu.cn/bioinf/TargetDNA/	2017	Yes
HybridNAP	DNA-, RNA-, protein-binding residue prediction method	http://biomine.cs.vcu.edu/servers/hybridNAP/	2017	Yes
funDNApred	Fuzzy cognitive map prediction model	http://biomine.cs.vcu.edu/servers/funDNApred/	2018	Yes
iProDNA-CapsNet	Neural network based prediction method	https://github.com/ngphubinh/iProDNA-CapsNet	2019	Yes
DNAPred	Ensembled Hyperplane-Distance-Based SVM	http://202.119.84.36:3079/dnapred/	2019	Yes
SVMnuc & NucBind	Support vector machine-based ab-initio method	https://yanglab.nankai.edu.cn/NucBind/	2019	Yes
ProNA2020	Neural network based prediction method	www.predictprotein.org	2020	Yes
NCBRPred	Multi-label learning framework method	http://bliulab.net/NCBRPred/	2021	Yes
GraphBind	Structure-based hierarchical graph neural network method	http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/	2021	Yes
GraphSite	AlphaFold2 based prediction using graph transformer method	https://biomed.nscc-gz.cn/apps/GraphSite	2022	Yes

2.4.2 Identification of RNA binding sites

The interaction between RNA and proteins plays a major role in the regulation of many biological processes such as gene expression regulation, viral assembly & replication, posttranscriptional modification, and protein synthesis (Gangloff, Soustelle, & Fabre, 2000; B. Lin & Pang, 2019; Pattnaik et al., 2018; Payne, Khalid, & Wagner, 2018; Standart & Jackson, 1994; Turner & Diaz-Munoz, 2018). The involvement of RNA-protein interaction in the development of cancers and neurological disorders has been reported in the literature (Carey

& Wickramasinghe, 2018; Idda, Munk, Abdelmohsen, & Gorospe, 2018; Kwiatkowski et al., 2009; Tsai, Spitale, & Chang, 2011; M. Zhou, Zhao, Wang, Sun, & Su, 2019). Other than that, these interactions play a significant role in genetic and infectious disorders. It is vital to get knowledge on the RNA-protein interaction residues in order to comprehend the roles and processes of any biological activity. Several structures of protein-RNA interacting residues have been identified and described in Protein Data Bank as a result of advances in experimental methods like as X-ray crystallography and nuclear magnetic resonance (NMR). However, these experimental procedures are both costly and time demanding. Computational techniques based on sequence information, on the other hand, are particularly practical and cost-effective in the identification of RNA-binding residues. Therefore, number of methods have been developed in the past as shown in Table 2.4, which are able to predict the RNA-interacting residues in a protein with reliable accuracy. These methods were developed by implementing various algorithms such as machine-learning and deep-learning using distinct representation of proteins. BindN (Liangjiang Wang & Brown, 2006) was one of the first methods which employed SVM based algorithm to identify the DNA and RNA interacting residues on a protein using amino acid sequences. The model was developed using three important features such as side chain pK(a) value, hydrophobicity index and molecular mass of an amino acid. The method was further updated to BindN+ (Liangjiang Wang et al., 2010) with better performance by using PSSM and biochemical properties of amino acids. RNABindR (Terribilini et al., 2007) was developed using the distance cut-off to identify which amino acid can contact RNA and implemented Naive Bayes classifier to train the model on the non-redundant dataset. Pprint (M. Kumar, Gromiha, & Raghava, 2008) was another very interesting method which was developed on 86 proteins using PSSM profile based SVM model. Likewise, PRINTR (Y Wang, Xue, Shen, & Xu, 2008) was also developed using the combination of PSSM and SVM classifier. PRNA (Z.-P. Liu, Wu, Wang, Zhang, & Chen, 2010) was developed using random forest algorithm to train the model by using interaction propensity and other sequence- and structure based features of proteins. RPISeq (Muppirala, Honavar, & Dobbs, 2011) was another method which predicts the interaction between RNA and protein sequence, where normalized 3-mer composition vector was created to represent proteins and 4-mer composition vector was created to represent RNA. RNABindRPlus (Walia et al., 2014) combined the power of machine learning and homology based method to make the reliable prediction for RNA interacting residues. There are other methods like DR_bind (Y. C. Chen, Sargsyan, Wright, Huang, & Lim, 2014) which uses the evolutionary and energetic features to make prediction, where RBScore&NBench (Miao & Westhof, 2016) is an combination of web-server and databases,

which uses scoring scheme to make the predictions. SNBRFinder (X. Yang, Wang, Sun, & Liu, 2015) is based on a the hybrid algorithm which combined SVM based model based on sequence profile with the homology based model developed using hidden Markov model profile. Further, very powerful methods such as hybridNAP and ProNA2020 (Qiu et al., 2020; Jian Zhang et al., 2019), with the ability to find the interacting residues for DNA, RNA, and proteins on a protein at a single platform were developed using neural networks with big dataset.

Table 2.4: In-silico tools for the prediction of RNA binding residues in a protein

Name	Description	Weblink	Year	Working Status
BindN	SVM based method for predicting DNA and RNA binding sites	http://bioinformatics.ksu.edu/bindn/	2006	No
RNABindR	Distance cut-off based prediction	http://bindr.gdcb.iastate.edu/RNABindR	2007	Yes
Pprint	SVM and PSSM profile based prediction method	https://webs.iiitd.edu.in/raghava/pprint/	2008	Yes
PRINTR	SVM and PSSM profile based prediction method	http://210.42.106.80/printr/	2008	No
BindN+	SVM based DNA or RNA-binding site prediction using PSSM profile	http://bioinfo.ggc.org/bindn+/	2010	No
PRNA	Random forest based approach using combined features	http://www.sysbio.ac.cn/datatools.asp	2010	No
RPISeq	Sequence information prediction method for RNA-protein interaction	http://pridb.gdcb.iastate.edu/RPISeq/	2011	Yes
RNABindRPlus	Machine Learning and Sequence Homology-Based Methods	http://ailab-projects2.ist.psu.edu/RNABindRPlus/	2014	Yes
DR_bind	Evolutionary conserved structural and energetic features based prediction	https://drbind.limlab.ibms.sinica.edu.tw/	2014	Yes
RBScore &NBench	Scoring scheme based linking of feature values with nucleic acid binding probabilities	http://ahsoka.u-strasbg.fr/rbscorenbench/	2015	No
SNBRFinder	Prediction of nucleic acids binding residues using hybrid algorithm	http://ibi.hzau.edu.cn/SNBRFinder	2015	No
DRNAPred	Fast sequence-based method that accurately predicts RNA interacting residues	http://biomine.cs.vcu.edu/servers/DRNApred/	2017	Yes

HybridNAP	Relative solvent accessibility (RSA), evolutionary conservation and propensity of amino acids (AAs) of binding based prediction of RNA interacting residues	http://biomine.cs.vcu.edu/servers/hybridNAP/	2017	Yes
RPI-Bind	A structure-based method for accurate identification of RNA-protein binding sites	http://ctsb.is.wfubmc.edu/publications/RPI-Bind-Pred.php	2017	No
iDeepS	Deep convolutional and recurrent neural networks based prediction method	https://github.com/xypan1232/iDeepS	2018	Yes
RPiRLS	Quantitative matrix based predictions of RNA interaction	http://bmc.med.stu.edu.cn/RPiRLS	2018	No
SVMnuc & NucBind	Support vector machine-based ab-initio method	https://yanglab.nankai.edu.cn/NucBind/	2019	Yes
ProNA2020	Standard neural networks based method for per-residue binding prediction	www.predictprotein.org	2020	Yes
NCBRPred	Multi-label learning framework method	http://bliulab.net/NCBRPred/	2021	Yes

2.5 Role of mutations in cancer

The occurrence and spread of cancer are linked to mutation accumulation. However, only a small percentage of a patient's mutations are shown to be the cause of cellular changes that result in cancer. Molecular profiles of malignancies may be useful in predicting the patients' clinical outcomes. Mutations play a major role in cancer research and one of the most accurate indicators of a mutation's driver status in patients is its recurrence. The genome is not affected uniformly by DNA damage and repair processes, and some mutations are more likely to develop than others. The type of malignancy also affects mutational probability (Brown, Li, Goncearenco, & Panchenko, 2019; Chang et al., 2016; Loeb, Loeb, & Anderson, 2003; I. P. Tomlinson, Novelli, & Bodmer, 1996; I. Tomlinson, Sasieni, & Bodmer, 2002; Wyrick & Roberts, 2015).

2.5.1 Benchmarking of mutation calling techniques

Due to advancements in computational technologies a number of mutations calling tools are developed in the past. A number of somatic mutations calling algorithms (e.g. Mutect2, Varscan2, SomaticSniper, MuSE, Strelka2, etc.) have been developed. For instance, MuTect is a technique that uses a Bayesian algorithm to identify somatic point mutations using genomic data with high specificity. Mutect2 uses the local assembly of haplotypes to identify somatic mutations such single-nucleotide mutations and indels. It uses a Bayesian method to evaluate the genotype likelihoods in the tumor and normal samples. The SomaticSniper pipeline used for the discovery of somatic SNVs using exomes and whole-genome sequencing data. Varscan2 mutation calling algorithm to identify germline variations, somatic mutations, and copy number variants in tumor-normal data. In addition, MuSE is a Markov Substitution model for evolution that finds novel mutations in data from extensive tumor sequencing. In Table 2.5 contains the details of the existing tools which are widely used to detect the somatic mutations at genome level. The germline mutations are mutations associated with the germline cells (sperm and eggs). Such type of mutations transferred from parents to offspring. Studies report that 8.5 percent of the children and adolescent's cancer carry germline mutations in genes (Jinghui Zhang et al., 2015). On the other side, somatic mutations do not affect offspring. These mutations occurred throughout an organism's life cycle either spontaneously as a result of mistakes in the DNA repair systems or as a direct reaction to stress or as a normal aspect of ageing. Somatic mutations are mostly identified during human cancers where a variety of mutations in oncogenes, tumor suppressor genes, and DNA repair processes have influenced the survival of patient (Miles & Tadi, 2022). In metastatic colon cancer patient's high rate of mutations was observed in APC, MUTYH, and TP53 genes (Ozdemir at al., 2021). A study also reports that mutations in TP53, BRCA1, BRCA2, BRAF, KRAS, BIRC5 genes associated with survival of cancer patients (Kaubryte & Lai, 2022).

Table 2.5: Mutation calling techniques for identifying mutations in genomes from next-generation sequencing data

Name	Description	Weblink	Year	Working Status
deepSNV	Detection and quantification of sub clonal SNVs in mixed populations	https://github.com/gerstung-lab/deepSNV	2012	Yes

Strelka	Somatic SNV and small indel detection method	ftp://strelka@ftp.illumina.com/	2012	Yes
VarScan2	Identify of SNV and CNV using sequencing data	http://varscan.sourceforge.net	2012	Yes
SomaticSniper	Somatic mutation identification using whole genome data	http://gmt.genome.wustl.edu/somatic-sniper/current/	2012	Yes
MuTect2	Mutation identification using exome and genome sequences	https://software.broadinstitute.org/cancer/cga/mutect	2013	Yes
MuSE	Markov Substitution model for Evolution algorithm for mutation calling	https://bioinformatics.mdanderson.org/public-software/muse/	2016	Yes
VarDict	Identify SNV, MNV, InDels, complex and structural variants	https://github.com/AstraZeneca-NGS/VarDict	2016	Yes
MuTect2	Local assembly and realignment tool for detection of SNVs and indels	https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-mutect2	2019	Yes

2.5.2 Mutation based cancer biomarkers

A genetic disorder such as, cancer, occurs due to genetic mutations. Mutations in genes are involved in the development and progression of several types of cancers (Lu et al., 2020). There are two major types of mutations acquired/somatic mutation and germline mutations. Somatic mutations are most common in cancer and occur due to damage/changes in genes. Some of the mutations are germline mutations which cause hereditary disorders for instance mutations in BRCA1 and BRCA2 increases the risk of breast cancers in women. In 1988, first time researchers showed a link between the KRAS mutations and the emergence of tumor (Vogelstein et al., 1988). Cancer biomarkers are majorly classified into prognostic, predictive and diagnostic. Prognostic biomarkers are mainly linked with the disease outcome and survival of cancer patients. Somatic mutation of the KRAS gene is associated with the development of colorectal cancer. In addition, literature reports that proliferation and survival of CRC patients are stimulated by EGFR (via signalling involving the MAPK, PIK3 and JAK/STAT pathways) (Coppede, Lopomo, Spisni, & Migliore, 2014; Voon & Kong, 2011). Patients are more likely to develop breast cancer if they have mutations in the genes BRCA1, BRCA2, ATM, and P53, whose products are involved in DNA repair (Nalejska, Maczynska, & Lewandowska, 2014). Zhang et al., shows lung carcinoma patients are at higher risk with mutation in KRAS, RB1, SMAD4, TP53, EGFR, and BRAF genes (S. Zhang, Zeng, Lin, Liang, & Huang, 2022). In

additions, mutations in genes such as DRD3, SETX, ZNF560, DNAJC2, GMPPA, and MMRN2, significantly linked with the overall survival of lung adenocarcinoma patients (Cho, Lee, Ji, & Lee, 2018). Wang et al., report in liver cancer patients TP53 and LRP1B mutated genes significantly reduces the survival of cancer patients and increase tumor mutation burden (Longrong Wang et al., 2019).

2.6 Conclusion

The continuous advancement in the sequencing technology has flooded the respective databases with new sequences belongs to various organisms but their annotation is still a challenge. Besides the sequence technology, there is an enormous improvement in the computational technology too, which made the researchers capable of developing highly accurate prediction models to functionally annotate the proteins. The efficiency of the models lies in the features that is used as the input to train the models, and these features are scattered in the literature, hence there is a need to develop a comprehensive platform to bring all the possible features at one place and to generate novel features. To annotate an important class of proteins such as transcription factors , experimental approaches are available but they are cumbersome and time-expensive. On the other hand, computational approaches are easy, fast, and reliable. Among, computational approaches two types of methods are available such as structure-based and sequence-based method, however, structure-based method comes with the requirement of structure availability which is again a time-consuming process. On the other hand, sequence-based methods are fast and perform equivalently. Therefore, it is important to develop a highly accurate method to classify the sequences into transcription factors using sequence information only. Proteins accomplish their respective functions by interacting with other macromolecules such as DNA/RNA, and molecules like NAG, SAM, ATP. So, to understand and exploit the underlying mechanism, it is important to have a precise knowledge of the interacting residues on a protein. The protein sequences are increasing at an exponential rate in the databases, which provide more data to train the models better. Though number of approaches are available to predict the interactions, but they are either trained on the limited dataset or trained using old technology. Therefore, more accurate computational methods are required to predict the interacting residues on a protein which are trained on larger datasets. The flawless coordination between these molecules or macromolecules is responsible for the well-being of an organism, whereas a small flaw in these coordination may lead to number of diseases like cancers. Mutation at the genome level is one of the major cause for the disruption

of this coordination. In order to handle these situation, it is important to accurately identify the mutations. Albeit, number of mutation detection techniques are available but the choice of optimum technique for different purposes is depends on the researcher's conscience. Therefore, it is important to benchmark the existing mutation calling techniques to understand the merits and demerits of each technique and to get clarity on their usage for different investigations. Moreover, usage of the mutation profiles to identify the diagnostic and prognostic biomarkers is still an unexplored field. Therefore, it is necessary to explore the role of mutations in the diagnosis and prognosis of a disease like cancer. Thus, in-silico tools to explore the influence of mutations on the survival of cancer patients is the need of the hour.

Chapter 3

Generation of features from sequence and structure of proteins

3.1 Introduction

Due to the advancement in the next-generation sequence technologies, a vast amount of high-throughput data in the field of proteomics and genomics is generated, making data analysis a strenuous task (Slatko, Gardner, & Ausubel, 2018). Annotation remains a mighty challenge to the scientific community to interrelate unknown proteins. Proteins have diversified roles from structural component to defense as well as messenger to storage, thus acting as a basis of cellular life. The experimental functional annotation of a newly discovered protein demands vigorous effort and outrageous cost and can be achieved in minimal throughput, but computational algorithms can infer their function with extreme throughput and low cost. The last two decades have witnessed the paradigm shift from small molecule to peptide-based therapeutics, and peptide annotation is as important and exciting as protein annotation (Craik, Fairlie, Liras, & Price, 2013).

In the last few decades, several protein sequence annotation tools have been developed (Cao, Xiao, Xu, & Chen, 2015; Cao, Xu, et al., 2013; Z. Chen et al., 2022, 2018; Dong et al., 2018; Z. R. Li et al., 2006; Xiao et al., 2015), ranging from subcellular localization to the structure and therapeutic properties of a protein. Identifying and classifying a protein is the first step to understanding its significance. CoPId (M. Kumar, Thakur, & Raghava, 2008) identifies the proteins based on their composition, GPCRpred (Bhasin & Raghava, 2004b) predicts family and superfamilies of G-protein coupled receptor (GPCR), whereas GPCRsclass predict amine type of GPCR by using dipeptide compositions. Several other tools, such as OXYpred (Muthukrishnan, Garg, & Raghava, 2007), CytoPred (S & GP, 2008), CyclinPred (Kalita et al., 2008), etc., aid in protein annotations. Recognizing the subcellular localization of a protein is vital in understanding its function. TargetP (Emanuelsson, Nielsen, Brunak, & von Heijne, 2000) is one of the initial subcellular localization prediction tools based on the N-terminal amino acid sequence. ESLpred (Bhasin & Raghava, 2004a) is an SVM-based method that uses dipeptide composition and PSI-BLAST as an input feature to predict the subcellular localization of eukaryotic proteins, whereas ESLpred2 (Garg & Raghava, 2008) is an improved version trained on a larger dataset. Besides this, several other tools such as, APSLAP (Saravanan & Lakshmi, 2013), SCLPred (Mooney, Wang, & Pollastri, 2011), HSLpred (Garg, Bhasin, & Raghava, 2005), RSLpred (Kaundal & Raghava, 2009), MultiLoc2 (Blum, Briesemeister, & Kohlbacher, 2009), WoLF PSORT (Horton et al., 2007) also aid in annotating

the subcellular location of proteins. VICMpred (Saha & Raghava, 2006) uses a hybrid method using tetrapeptide information with amino acid composition (AAC) to predict the functional proteins of gram-negative bacteria.

Similarly, numerous peptide classification and annotation tools have been developed in the past, which utilize compositional features of peptide sequences. AHTpin (R. Kumar et al., 2015), TumorHPD (A. Sharma et al., 2013) uses AAC while ToxinPred (Gupta et al., 2013), IL10pred (Nagpal et al., 2017) uses dipeptide composition as input features. AntiTBpred (Usmani, Bhalla, & Raghava, 2018) uses a hybrid feature based on AAC and N5C5 binary pattern, AntiCP (Agrawal, Bhagat, Mahalwal, Sharma, & Raghava, 2021; Tyagi et al., 2013) uses AAC and binary patterns, AntiFP (Agrawal, Bhalla, et al., 2018) uses the binary pattern of terminal residues, while CellPPD (Gautam et al., 2013) utilizes binary profile as well as motif information as input feature. Besides this, several other in silico tools use peptide features to predict a variety of therapeutic peptides such as IL4pred (Dhanda, Gupta, Vir, & Raghava, 2013), IL6Pred (Dhall, Patiyal, Sharma, Usmani, & Raghava, 2021), IFNepitope (Dhanda, Vir, & Raghava, 2013), iBCE-EL (Manavalan, Govindaraj, Shin, Kim, & Lee, 2018), PIP-EL (Manavalan, Shin, Kim, & Lee, 2018b), MLCPP (Manavalan & Patra, 2022), AIPpred (Manavalan, Shin, Kim, & Lee, 2018a), mAHTPred (Manavalan, Basith, Shin, Wei, & Lee, 2018), and many more.

As discussed above, sequence level annotation mainly focused on feature extraction from whole or segment of protein, which led to the development of many prediction tools and novel discoveries. Nevertheless, residue level annotation has its own importance as it provides the information of the functional site in the protein (Nagel, Jimeno-Yepes, & Rebholz-Schuhmann, 2009). Residue level annotation was initially applied in predicting the secondary structure of the residues where property and profile of individual amino acids were considered while predicting the same. For example, Chou-Fasman utilizes the combination of statistical and heuristic approaches to predict the residue's secondary structure state. Its successful implementation encourages researchers to use residue level features in predicting irregular secondary structures like alpha turns (Yan Wang, Xue, & Xu, 2006), beta turns (Fuchs & Alix, 2005; H. Kaur & Raghava, 2003; Kountouris & Hirst, 2010; H. Singh, Singh, & Raghava, 2015), gamma turns (Guruprasad & Rajkumar, 2000; Jahandideh, Sarvestani, Abdolmaleki, Jahandideh, & Barfeie, 2007), beta hairpins (de la Cruz, Hutchinson, Shepherd, & Thornton, 2002) and beta-barrel (Freeman & Wimley, 2012). These methods are based on several features

such as evolutionary information in the form of PSSM matrix, binary profiles, predicted secondary structure state, solvent accessibility surface area, and many more. Unlike sequence level annotation, residue level annotation requires a feature of individual residue in the form of patterns, which can be of varying length. However, the existing feature generation methods do not provide the facility of providing features for individual residue. Also, they do not consider the distribution of low complexity regions in a protein which could be a potential feature in protein annotation function. In a nut shell, features can be broadly classified into four categories such as composition-based features, binary-profile based features, evolutionary information-based features, and features based on the structures.

An ample of methods have been developed in the last decade in the form of web servers, standalone, and libraries like iFeature (Z. Chen et al., 2018), Rcpi (Cao et al., 2015), protr (Xiao et al., 2015), propy (Cao, Xu, et al., 2013), PyBioMed (Dong et al., 2018), PROFEAT (Z. R. Li et al., 2006), to compute descriptors using sequence information. However, a number of essential features are still missing from these platforms. Annotation is the next step once we get the numerical representation of proteins. In the absence of numerical features, the annotation is done using traditional approaches such as similarity search. BLAST (Johnson et al., 2008) is one of the widely used methods for annotating proteins using similarity search, but it fails when there is no or low similarity between the query sequence and sequences in the databases. To overcome such limitations, machine learning techniques have emerged as a powerful means to annotate the structure as well as the function of both proteins and peptides by utilizing effective mathematical expressions, which actually represent the feature of the corresponding protein or peptide. Recently, numerous methods with the option of calculating features from protein sequences also allow building the models using the machine learning models, such as iFeature (Z. Chen et al., 2018), iFeatureOmega (Z. Chen et al., 2022), BioSeq-Analysis (B. Liu, 2019), BioSeq-Analysis 2.0 (B. Liu et al., 2019), iLearn (Z. Chen et al., 2020), and iLearnPlus (Z. Chen et al., 2021). These tools help novice users or users with no computer background to annotate their proteins of interest or build the prediction models for further use.

To complement all the existing methods, we have made an attempt to put all the feature extraction methods used in the past, as well as some new features, on a single platform, Pfeature. Besides feature calculation, it also allows users to develop and save the prediction models using various classifiers or regressors. Pfeature comprises six effective modules:

Composition, Binary Profiles, Evolutionary Information, Structure, Pattern, and Model Building. In composition modules, we have integrated all the existing features along with some new features like Shannon-entropy based features, composition of atoms & bonds, residue repeat information, and distance distribution based features. The binary Profiles module provides profiles at the level of amino acids, dipeptides, atom & bond, and amino acid index. Additionally, evolutionary information provides the PSSM profile and allows different normalization operations on the generated matrix. The structure module facilitates feature extraction using the structures in the form of fingerprints, smiles, secondary structures, and solvent accessibility. The pattern module enables the generation of patterns of the extracted features. Moreover, the model building module is for building the classification or regression model after applying various operations on the feature matrix. We have provided this tool in the form of web server, python-based standalone, and library. Figure 3.1 shows the complete architecture of the Pfeature.

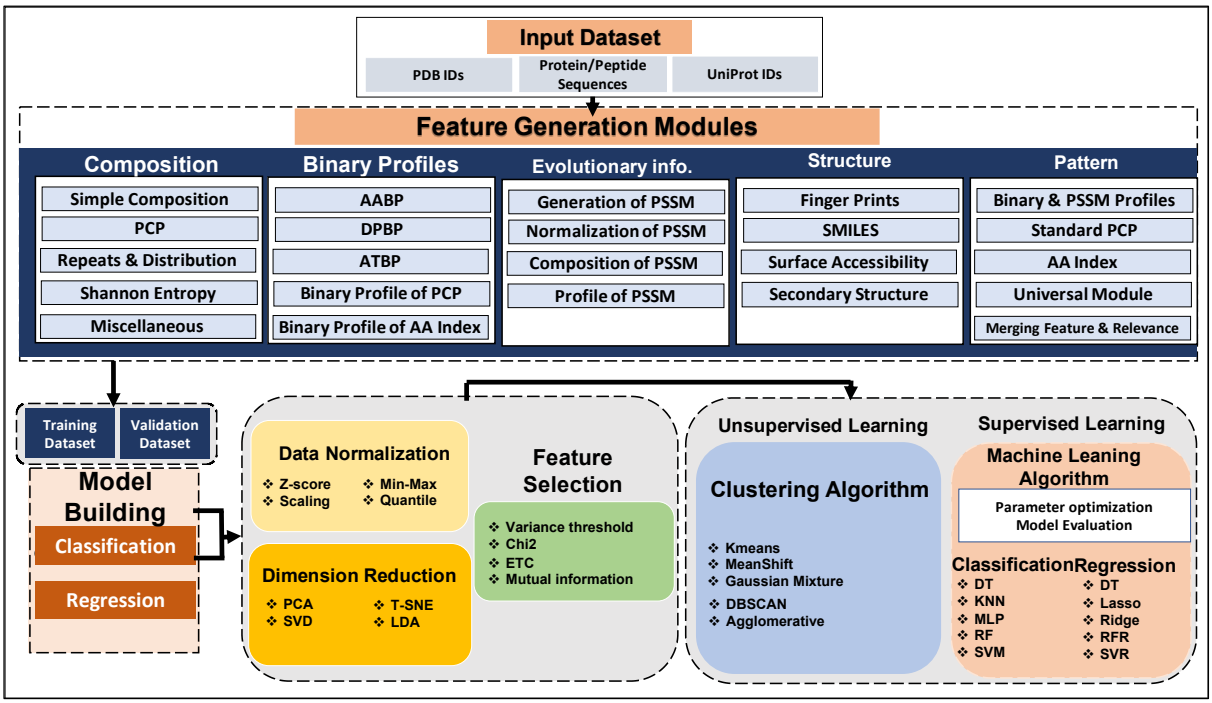


Figure 3.1: Illustration of overall architecture of Pfeature including menus and sub-menus for computing wide-range of protein features

3.2 Composition-based module

Several composition-based features have been created in the last two decades to annotate a protein sequence. Fixed-length vector output of composition-based features enables

researchers to apply various machine learning algorithms to build a model as one of the demerits of machine learning techniques is the requirement of a fixed-length vector as the input feature. One of the widely used features is the amino acid composition, in which the percent proportion of each residue is calculated, and it provides a vector of size 20 for each input sequence. Similarly, the composition of all possible dipeptides and tripeptides in a sequence is computed and referred to as dipeptide- and tripeptide-composition, respectively. Other than that, composition based on physicochemical properties is based on the nature of the amino acids. Autocorrelation-based descriptors are based on the distribution of the amino acid properties, which is determined using amino acid indices (http://www.genome.ad.jp/dbget/aaindex.html) and provide three different correlations. A number of other features are also reported in the Pfeature, which are available on other platforms, such as conjoint triad calculation, pseudo amino acid, amphiphilic pseudo amino acid, composition enhanced transition distribution, quasi sequence order, sequence order coupling number, and many more. Table 3.1 exhibits the number of features calculated by each sub-module of composition-based module. Moreover, it also provides the comprehensive comparison of Pfeature with existing methods at the level of type of descriptors it calculates and type of platform it provides such as analysis and prediction.

Table 3.1: Comprehensive comparison of features integrated in Pfeature with existing platform/software at the level of type of features availability and task. These descriptors are computed at protein level, can be used to compute overall function/structure of a protein (adopted from Pande et al., 2019)

Features Integrated in Pfeature		Feature Calculation Software	Analysis and Prediction Platform
Type of Descriptors	**Number**		
Amino Acid	20	All*	All
Dipeptide	400	All	All
Tripeptide	8000	All	All
Atom and Bond	9	None	None
Physicochemical (Standard + Advanced)	30	Most	Most
AAIndex	566	Most	Most
Autocorrelation (3 × AAIndex)	1698	Most	Most
Shannon Entropy (Overall+ Each residue)	21	None	None
Shannon Entropy of Property	24	None	None

Distance Distribution	20	None	None
Repeats of Residues	20	None	None
Repeats of Physicochemical Properties	19	None	None
Pseudo Amino acid	20 + λ	Most	BSA2, iLP
Amphiphilic Pseudo Amino acid	20 + λ×3	Most	BSA2, iLP
Conjoint Triad Calculation	343	Most	BSA2, iLP
Composition enhanced Transition Distribution (CETD)	189	Most	Most
Sequence Order Coupling Number (SOCN)	λ×2	Most	BSA2, ILP
Quasi-Sequence Order (QSO)	40+(λ×2)	Most	BSA2, iLP
PSSM Composition	400	POSSUM	BSA2
Number of descriptors for whole protein (only single value of λ = 5)			**11879**
Total descriptors (Whole protein + N-Term + C-Term + RN-Term + Split etc.)			**95137**

***All: All software; Most: Most of the software; None: Only Pfeature; BSA2: BioSeq-Analysis 2.0; iLP: iLearnPlus**

3.3 Profile-based module

Composition based feature consider only the sequence and provide number of amino acid in a protein or peptide. But the order of amino acid is also essential in deciphering the role of a protein or peptide. To achieve this, binary profiling of sequences has been done in the past, which encapsulate information of both composition as well as order of sequence. Pfeature incorporated the function which is used to compute amino acid binary profiles, dipeptide binary profiles, a binary profile corresponding to atomic and bond of each amino acid residue of the sequence, physicochemical properties based, and AA Index binary profile. Number of studies has used binary profile for the generated patterns to make prediction at the residue level, such as prediction of secondary structure, DNA/RNA interacting residues prediction.

3.4 Evolutionary information-based module

Evolutionary information is an important and widely used feature of a protein or peptide sequence, therefore we have provided modules to generate PSSM profiles from evolutionarily conserved patterns of protein sequences. Pfeature calculate PSSM profile by implementing the PSI-BLAST which hits query sequence against the SwissProt database. Pfeature provide function to generate the PSSM profile as well as to normalize PSSM profile. We have provided

four different types of normalization method. In addition, function to calculate the composition of PSSM profile has been incorporated.

3.5 Structure-based module

The modules mentioned above can handle sequences with natural amino acids but fail to calculate features for peptides with non-natural amino acids or peptides with chemical modifications. As peptides with modifications cannot be represented by sequences, hence structures are required to calculate their features. In order to develop models to explore the therapeutic potential of a chemically-modified peptide/protein, structural descriptors are used as input features for various classifiers. To facilitate the users, we have incorporated four different sub-modules such as fingerprints, SMILES , secondary structure, and solvent accessibility. In the fingerprints sub-modules, we have implemented PaDEL software (Yap, 2011) in the back end, which takes structure as input and calculates a vector of size 14532 values for each structure. Pfeature does not have the provision to predict the structure of chemically modified peptides or proteins, for a state-of-the-art method like PEPstrMOD (S. Singh et al., 2015) can be used. SMILES sub-module converts the structure into a simple linear notation which reduces the space requirement by 50-70 percent, hence taking less time to process by the software. SMILES notation encodes the complete information into atoms, bonds, branches, ring closure, and disclosure. The secondary structure sub-module calculates the average helix, beta-sheet, and coil composition using the input structure. The computation is achieved by DSSP software. And solvent accessibility sub-module calculates the measures of solvent accessibility by implementing NACCESS software in the backend. Table 2 comprises the comparison between the existing methods and Pfeature for features involved in binary profile, evolutionary information, and structure module.

Table 3.2: Comprehensive comparison of features belongs to binary profile, evolutionary information, and structure module of Pfeature with different platform/software. These descriptors are suitable for predicting function of residues in a protein and function of chemically modified proteins. (adopted from Pande et al., 2019)

Features Integrated in Pfeature		Feature Calculation Software	Analysis and Prediction Platform
Type of Descriptors	**Number**		
Binary Profiles			
Amino Acid	L × 20	None	BSA2, iLP

Dipeptides	L × 400	None	None
Physicochemical Properties	L × 25	None	None
AAIndex	L × 566	None	BSA2
Atom + Bond	L × 20 × 9	None	None
PSSM Profiles			
PSSM Raw Profile	L × 21	POSSUM	BSA2
Normalized PSSM Profile	L × 21	POSSUM	BSA2
Structural Descriptors			
Fingerprints	14532	None	None
Similes Format	L × 15	None	None
Surface Accessibility	L	None	BSA2
Average Secondary Structure	3	None	BSA2
Total descriptors, if we take protein length (L = 100)			**139435**

***None: Only Pfeature; BSA2: BioSeqAnalysis 2.0; iLP: iLearnPlus**

3.6 Pattern module

The pattern of specified window size plays a significant role in assigning a function to an uncharacterized protein/peptide. Various reports suggest that N- the terminal signal sequence influence the protein thermal stability (Booth et al., 2018), protein-protein interactions (Hayes, Alarcon-Hernandez, & Setlow, 2001), and intracellular localization of enzyme activity (Aggarwal & Mondal, 2006). C- terminus protein residues contain targeting signals and are involved in inhibiting luminal ER proteins' excretion (Munro & Pelham, 1987). Therefore, we have provided this module in which patterns of a specified length are generated. It contains five sub-modules. In binary profiles, which generate the patterns of user-specified window size using the amino acid binary profile, the entire vector of the binary profile is converted to overlapping patterns of defined size. A similar process is done for profiles of PSSM, physicochemical properties, and AA index generated using Pfeature. The universal sub-module of the pattern module takes input in either string file or the CSV format and generates overlapping patterns.

3.7 Model building module

After getting the features, several operations can be applied to the feature matrix to make it more suitable for developing a prediction model with the capability to annotate or classify a protein. There are numerous user-friendly tools have recently been created that enable users to

create models by simply uploading their data. The model building module is further subdivided into four sub-modules like merging features which concatenate the features column-wise to combine two or more types of features; the feature relevance sub-module is to provides feature importance to each feature based on performance measures; the classification sub-module applies different operations on the input matrix and generate classification model for further use; similarly, regression sub-module generates regression model. The classification sub-module allows the implementation of various operations like normalization, dimension reduction using feature extraction, feature selection using six different methods, clustering, and model generation using k-fold cross-validation, parameter-optimization using grid search, and development of a model using five different classifiers. The regression sub-module also allows all the operations mentioned above.

3.8 Features specific to Pfeature

We have compiled all the possible features for proteins that exist in the literature. Other than that, we have also introduced some novel features in some modules. Such as, in the composition module, we have introduced higher-order dipeptide composition, which differs in generating dipeptides from the sequences. Existing methods calculate traditional dipeptide composition in which overlapping patterns of length two are generated, and then the composition is calculated, which results in a vector of size 400; in higher-order dipeptide patterns are generated by skipping residue in between based on the order of dipeptide. Also, features based the repeats of amino acids are calculated using equations 1, 2 and 3.

$$RRI_i = \frac{\sum_{j=1}^{N} (R_j)^2}{\sum_{j=1}^{N} R_j} \quad (1)$$

$$DDOR_i = \frac{(R_{NT})^2 + \sum_{j=1}^{N} (R_j)^2 + (R_{CT})^2}{(L - F_i) + 1} \quad (2)$$

$$PRI_i = \frac{\sum_{j=1}^{N} (P_j)^2}{\sum_{j=1}^{N} P_j} \quad (3)$$

where $\boldsymbol{RRI_i}$, $\boldsymbol{DDOR_i}$, $\boldsymbol{PRI_i}$ are residue repeat information, distance distribution of residue and property repeat information of type $\boldsymbol{i}$, respectively. N and R_j are maximum number of occurrences and number of runs/repeats of property type j, respectively. R_{NT}, R_j, R_{CT}, L and F_i are residue distance from N-terminal, inter-distance between residue type i, residue distance from C-terminal, total length of protein sequence and frequency of residue type i, respectively.

Other than that, Shannon-entropy based features was calculated at sequence and residue level by using equation 4 and 5.

$$HS = -\sum_{i=1}^{20} p_i log_2 p_i \quad (4)$$

$$HR_i = -p_i log_2 p_i \quad (5)$$

where ***HS*** is Shannon entropy of a protein sequence and ***HR$_i$*** is entropy of a residue of type *i*. p_i is the probability of a given amino acid in the sequence. This equation was extended to compute entropy of a particular type of property like charge, polarity, hydrophobicity, etc. in a protein sequence.

We have also introduced a feature type which calculates the composition and binary profile based on the atom and bonds present in the amino acids. Other than that, we have provided the facility to calculate the binary profiles as the level of dipeptide which results in the vector size of L*400; at the level of AA index which provides output of size L*566, and at the level of physicochemical property which results in the vector size of L*25 for each sequence of length L.

3.9 Subset of sequences

The majority of protein classification techniques calculate characteristics from the whole protein. Split amino acid composition (SAAP)-based methods outperform whole sequence composition-based methods, according to published research. Proteins are divided into a number of pieces for SAAP, and the composition of each portion is then calculated. It has been noted in the past that a protein's N-terminal region also contributes to how it functions, as is the case with traditional secretory proteins that include signal peptides. Pfeature enables users to calculate a variety of characteristics in specific protein regions, such as the terminal (C- or N- terminal), rest, and split. One benefit of choosing an area is that the user may create a binary and composition profile for this fixed-length region.

3.10 Service to the scientific community

The aim of the study is to provide all the feature generation techniques at a single platform. We have tried to incorporate all the methods used in past. To serve the scientific community, all these methods will be freely available at following platforms.

- **Web Server.** We have implemented all the feature generation methods, in a freely available webserver, named as "**Pfeature**" (https://webs.iiitd.edu.in/raghava/pfeature/).

All the methods have been covered under five main heads such as compositions, binary profiles, evolutionary information, structure and pattern based features. Utilizing the Apache software, a web server has been created on the Linux/Ubuntu operating system. Using HTML, PHP5, and CSS3, this server's web pages have been created. Wide-ranging device compatibility has been achieved through responsive web design (e.g., iPad, Smart Phone, Laptop, Desktop). Users can submit protein sequences in either FASTA format or as a single line using the submission page. PDB IDs or UniProt IDs may be supplied by the users. Results can be downloaded in CSV format in addition to being displayed as HTML pages by the server. Figure 3.2 provides the screenshot of the homepage of Pfeature web-server.

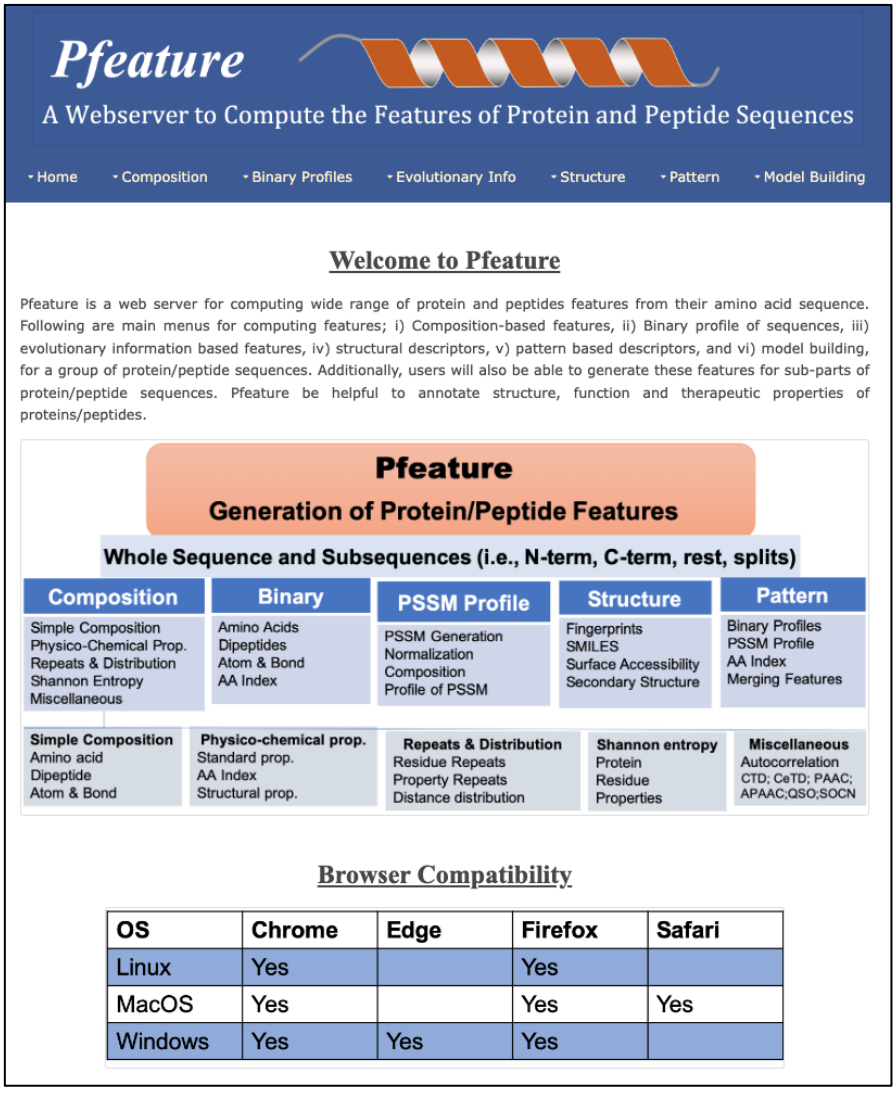


Figure 3.2: Screenshot of homepage of Pfeature server

(URL https://webs.iiitd.edu.in/raghava/pfeature/)

- **Standalone:** We have developed python-based standalone for various platforms such as windows, mac, fedora, ubuntu and centos. The output of standalone enables one to

visualize all the features together belong to a particular module. Three different standalones are developed based on composition binary profiles, and PSSM module. Figure 3.3, 3.4 and 3.5 comprises of screenshots of complete usage of Pfeature standalones for composition-, binary profile-, and PSSM-module, respectively. (https://github.com/raghavagps/Pfeature)

```
# generate the composition based features
python pfeature_comp.py [-h] -i INPUT [-o OUTPUT]
                        [-j {AAC, DPC, TPC, ATC, BTC, PCP, AAI, RRI, PRI, DDR, SEP, SER, SPC, ACR, CTC, CeTD, PAAC, APAAC, QSO,
                            SOC, ALLCOMP}]
                        [-n N_TERMINAL] [-c C_TERMINAL] [-nct NC_TERMINAL]
                        [-rn REST_N] [-rc REST_C] [-rnc REST_NC] [-s SPLIT]
                        [-d LAG] [-w WEIGHT] [-t PWEIGHT]
Optional arguments:
 -h, --help
            show this help message and exit
 -i INPUT
            Input: protein or peptide sequence in FASTA format or single sequence per line in single letter code
 -o OUTPUT
            Output: File for saving results by default pfeature_result.csv
 -j {AAC,DPC,TPC,ATC,BTC,PCP,AAI,RRI,PRI,DDR,SEP,SER,SPC,ACR,CTC,CeTD,PAAC,APAAC,QSO,SOC,ALLCOMP}, by default AAC
 -n N_TERMINAL
            Window Length from N-terminal: by default 0
 -c C_TERMINAL
            Window Length from C-terminal: by default 0
 -nct NC_TERMINAL
            Residues from N- and C-terminal: by default 0
 -rn REST_N
            Number of residues removed from N-terminal, by default 0
 -rc REST_C
            Number of residues removed from C-terminal, by default 0
 -rnc REST_NC
            Number of residues removed from N- and C-terminal, by default 0
 -s SPLIT
            Number of splits a sequence divided into, by default 0
 -d LAG
            This represents the order of gap, lag or dipeptide, by default 1
 -w WEIGHT
            Weighting Factor for QSO: Value between 0 to 1, by default 0.1
 -t PWEIGHT
            Weighting factor for pseudo and amphiphilic pseudo amino acid composition: Value between 0 to 1, by default 0.05
```

Figure 3.3: Complete command line usage for generating composition-based features using Pfeature standalone pfeature_comp.py

```
# generate the binary profile based features
python pfeature_bin.py [-h] -i INPUT [-o OUTPUT]
                       [-j {AAB,DPB,ATB,BTB,PCB,AIB,ALLBIN}] [-n N_TERMINAL]
                       [-c C_TERMINAL] [-nct NC_TERMINAL] [-rn REST_N]
                       [-rc REST_C] [-s SPLIT] [-d LAG]

Optional arguments:
 -h, --help
            show this help message and exit
 -i INPUT
            Input: protein or peptide sequence in FASTA format or single sequence per line in single letter code
 -o OUTPUT
            Output: File for saving results by default pfeature_result.csv
 -j {AAB,DPB,ATB,BTB,PCB,AIB,ALLBIN}, by default AAB
 -n N_TERMINAL
            Window Length from N-terminal: by default 0
 -c C_TERMINAL
            Window Length from C-terminal: by default 0
 -nct NC_TERMINAL
            Residues from N- and C-terminal: by default 0
 -rn REST_N
            Number of residues removed from N-terminal, by default 0
 -rc REST_C
            Number of residues removed from C-terminal, by default 0
 -s SPLIT
            Number of splits a sequence divided into, by default 0
 -d LAG
            This represents the order of gap, lag or dipeptide, by default 1
```

Figure 3.4: Complete command line usage for generating binary profile-based features using Pfeature standalone pfeature_bin.py

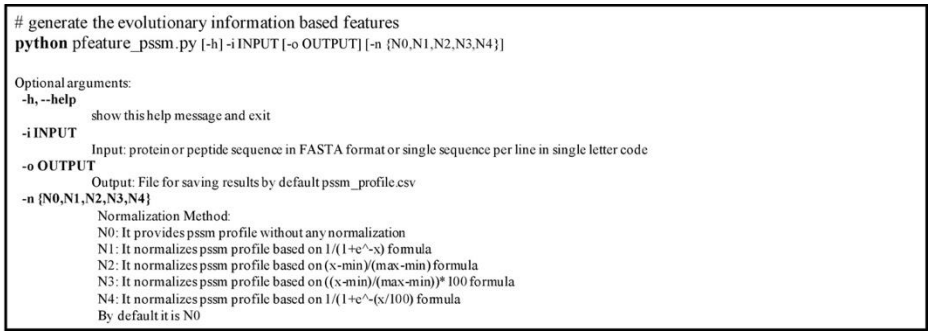


Figure 3.5: Complete command line usage for generating PSSM-based features using Pfeature standalone pfeature_pssm.py

- **Pfeature library:** We have also provided our modules as the functions of python library. These functions can easily be imported and computes the desired features by passing the desired variables. (https://github.com/raghavagps/Pfeature)

3.11 Utility of Pfeature

Pfeature provides all the feature selection methods used in the past. Feature selection is an important step in developing prediction models. The utility of Pfeature can be easily understood by following case studies. Figure 3.6 represents the putative utility of the features in Pfeature.

3.11.1 Peptide classification and protein annotation methods

In past, various peptide classification methods have been developed such as AntiCP (Agrawal et al., 2021; Tyagi et al., 2013) and MLACP (Manavalan et al., 2017) which differentiate anti-cancerous peptide from others, CellPPD (Gautam et al., 2013) and MLCPP (Manavalan, Subramaniyam, Shin, Kim, & Lee, 2018) for classifying or predicting cell penetrating peptide, AntiTbPred (Usmani et al., 2018) to differentiate antitubercular peptide from antibacterial as well as non-antibacterial peptides, Antifp (Agrawal, Bhalla, et al., 2018) to classify antifungal peptides etc. The universal methodology of all these prediction models is to generate a relevant dataset and then generate and select the best feature which can easily differentiate the desired peptide classes with others. A user can generate the varied range of features, being AAC, DPC, TPC, binary profiling, PSSM matrix etc., used in previous studies to develop the prediction model. Similarly, feature selection techniques have played vital role in deciphering the subcellular localization of proteins as well annotating their functional role.

3.11.2 Residue level annotation

Residue level annotation is significant in assigning secondary structure to a protein. Various methods like alphapred (Harpreet Kaur & Raghava, 2004), betapred3 (H. Singh et al., 2015) etc. uses residue level feature extracting techniques. Other newly developed methods like AntiMPmod (Agrawal & Raghava, 2018), CellPPDmod (V. Kumar et al., 2018) etc. implemented SMILES, fingerprints etc. to calculate the differentiating features. PSSM profiling has been extensively used in several ligand binding prediction methods like ATPint (Chauhan, Mishra, & Raghava, 2009b), NADbinder (Ansari & Raghava, 2010), GTPbinder (Chauhan et al., 2010a), DNAbinder (M. Kumar et al., 2007), etc.

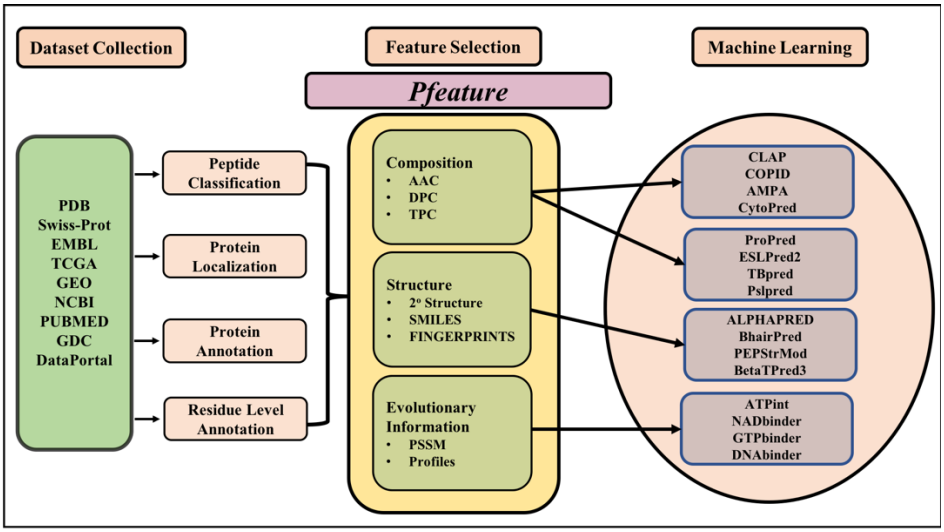


Figure 3.6: Putative usage of features and model building module of Pfeature

3.12 Comparison with existing methods

To understand the application and utility of a newly developed method, it is of great importance to compare it with the already existing methods at different levels. In the last two decades, several methods (Z. Chen et al., 2018; Cao et al., 2015; Xiao et al., 2015; Cao et al., 2013; Dong et al., 2018; Z. R. Li et al. 2006) have been designed to calculate the features for various biological macromolecules, and Pfeature is developed to complement these methods. Comparison at each feature type is not possible as the existing methods calculate features of various biomolecules such as DNA, RNA, Proteins, and chemicals, whereas Pfeature only provides features for the proteins. We have made a comprehensive comparison of features provided in composition-based module with existing tools in Table 3.1, where we tried to segregate the existing methods in two categories, such as, methods for calculating features and

methods to perform analysis/prediction. One can definitely fetch features by using analysis/prediction methods, but the task is quite tedious (e.g., iFeature, iFeatureOmega, iLearn, iLearnPlus, BioSeq-Analysis, BioSeq-Analysis2.0, etc.). There are number of features which are specific to Pfeature that we generated by observing the different trends in the biology, for example, repeats of residues in a sequence of a particular protein is important for its function. Features like repeats residue information, distance distribution of residues, Shannon entropy at the level of whole protein, residues, and physicochemical properties, composition based on atoms & bonds in a protein sequence, and dipeptide composition based on higher order, are unique to Pfeature and could be of significant importance.

Other than that, except binary profile at the level of amino acids which is also available as one-hot encoding feature in existing methods, all other sub-modules are unique to Pfeature which includes binary profile based on dipeptide, atom & bond, physico-chemical properties, and amino acid index. To capture the evolutionary information in the sequences we calculate PSSM profile, although POSSUM provide the facility to compute wide range of PSSM based features but it does not provide the facility to calculate other types of features. Table 3.2 provides the overall comparison of Pfeature with existing methods for binary profile, evolutionary information, and structure-based features. In addition, Pfeature facilitates to compute the features for the segments of a proteins such as N-terminal, C-terminal, rest, split, NC-terminal, etc., which is not available in any of the available methods. All these salient features signify that Pfeature complements the existing methods by providing the features which are not incorporated in the existing tools. Table 3.3 provides the comparison of Pfeature and existing methods in terms of their availability as web-server, standalone, or library.

Table 3.3: Comparison of different software/platform with Pfeature in terms of their availability (adopted from Pande et al., 2019)

Software/platform	Year	Web Server	Standalone	Library	Features	Prediction
Pfeature	2022	Yes	Yes	Python	Direct	Yes
iFeatureOmega	2022	Yes	Yes	Python	Indirect	Yes
iLearnPlus	2021	Yes	Yes	No	Indirect	Yes
iLearn	2019	Yes	Yes	No	Indirect	Yes
PyFeat	2019	No	Yes	No	Direct	Yes
BioSeq-Analysis2.0	2019	Yes	Yes	No	Indirect	Yes
iFeature	2018	Yes	Yes	No	Indirect	Yes

PyBioMed	2018	No	Yes	Python	Direct	No
POSSUM	2017	Yes	Yes	No	Direct	No
BioSeq-Analysis	2017	Yes	Yes	No	Indirect	Yes
Pse-in-One 2.0	2017	Yes	Yes	No	Direct	No
PDBparam	2016	Yes	No	No	Direct	No
BioTriangle	2016	Yes	No	No	Direct	No
Pse-in-One	2015	Yes	Yes	No	Direct	No
Protr/ProtrWeb	2015	yes	Yes	R	Direct	No
PyDPI	2013	No	Yes	Python	Direct	No
Propy	2013	No	Yes	No	Direct	No
PROFEAT	2011	Not active	No	No	Direct	No

*Direct: provide only features; Indirect: develop prediction models and features has to extract from the results

3.13 Discussion and Conclusion

One of the oldest platform developed for computing the structural features and physico-chemical properties of a protein from its amino acid sequence information is PROFEAT, which was developed in year 2006 (Li et al., 2006). Then, it was updated in the year 2011, and it included network, segment descriptors and topological descriptors. Pfeature lacks the network, segment, and topological based features, which are available in PROFEAT. Further, Python-based standalone package named PyDPI (Cao, Liang, et al., 2013) was developed in 2013, which is able to compute 52 types of protein features from six different feature groups. Similarly, Python based library by the name PyBioMed (Dong et al., 2018) was made available that can calculate wide range of features for important molecules such as DNA, chemicals, and Proteins. In the recent years, one of the significant packages called iFeature was developed which can calculate 53 different types of features from protein sequences, moreover, it also provides twelve various types of commonly used feature clustering, selection, and dimensionality reduction algorithms, along with that facilitates feature generation, analysis, training and benchmarking of machine-learning models and predictions. Its updated version iFeatureOmega is developed in 2022, which provided features at three different levels such as sequences which extract descriptors for DNA, RNA and Protein sequences; structure which extract features for protein structures; and ligand which calculates descriptors for small organic molecules such as ligands. Similarly, other platforms such as, iLearn and its update iLearnPlus, BioSeq-analysis and BioSeq-analysis 2.0, were also developed for the predict or analyse the protein data. In a nutshell, several methods have been developed to explore the features from

various biological and chemical molecules, and each method possesses unique features and all methods complement each other.

In the same path, we have developed Pfeature to complement the existing platforms/software/library/standalone, so that users may get more facilities working in the field of bioinformatics. We developed Pfeature with the aim of providing a platform for the functional annotation of a protein at the level of sequence and residue level. Best of our knowledge, none of the existing software provides the facility of calculating features for the chemically modified proteins. Moreover, Pfeature uniquely provides the option of calculating features for various segments of proteins. In this tool, we have integrated not only the features available in the literature but also provided some novel features based on the concepts of biology. Pfeature is a comprehensive, easy-to-use and open-source python package which computes a large number of features, and allows users to calculate various features of protein/peptide based on the sequence, structure and physiochemical properties. Pfeature calculates the several descriptors and gives the information of not only at functional level but also at the residue level. It is a toolkit for combined feature calculation at the functional level, residue level and also from the various profiles such as binary profiles and PSSM profiles. We believe that freely available webservice "https://webs.iiitd.edu.in/raghava/pfeature/" will be very useful to the biologist, with limitation of programming knowledge. In addition, free-availability of python library, standalone, as well as source code will help the researcher to compute wide range of protein and peptide feature from their sequence and structure on a larger scale. Ideally, feature generation tools should provide or calculate features that has the ability to classify different types of protein classes. Unfortunately, there is no single method for providing the features which is best for the different classification tasks. In the present scenario, different types of features are being used for the prediction of different protein classes. However, it is an important task to figure out the universal features with the ability to classify different classes of proteins. Recently, features computed using Pfeature has been used for developing models for predicting pattern recognition receptors, interleukin-6 inducing peptides, and allergenic peptides (Dhall, Patiyal, Sharma, Usmani, & Raghava, 2020; D. Kaur, Arora, & Raghava, 2020; N. Sharma et al., 2020). We have also provided the model building module that can be used independently to analyse the descriptors and build different machine learning based models for classification and regression.

Chapter 4
Identification of transcription factors from the primary structure

4.1 Introduction

Transcription factors (TF) control gene expression via binding to specific DNA segments (S. A. Lambert et al., 2018; Miyazaki & Miyazaki, 2021; Ortet, De Luca, Whitworth, & Barakat, 2012) and act as regulatory molecules in controlling cell differentiation, immune responses and gene regulatory pathways (Fong & Tapscott, 2013; Lee & Young, 2013; H. Singh, Khan, & Dinner, 2014). TFs play major role in the understanding of transcription regulatory mechanism 33372147. Several disorders, including Rubinstein-Taybi, Coffin-Siris, CHOPS syndromes etc., are occurred by improper regulation and mutations in TFs (Kircher et al., 2019; Lee & Young, 2013; Sim, White, & Lockhart, 2015; Weinstein, Blanchard, Moake, Vosburgh, & Moise, 1989). Additionally, a number of biological processes, including aberrant gene expression, chromosomal translocation, mutations linked to non-coding DNA, point mutations, significantly alter TFs binding sites in different cancers (Bushweller, 2019; Dasberg, 1991; Jiramongkol & Lam, 2020; Kishtagari, Levine, & Viny, 2020; Kleinjan & van Heyningen, 2005). In addition, the incorrect regulation of the NF-kB transcription factor is linked to a number of inflammatory and autoimmune illnesses as well as poor immunological development (Hayden & Ghosh, 2012). Literature also prove that one can regulate the expression of genes by genetic manipulations in the transcriptional regulators (Kemmeren et al., 2014; Lee & Young, 2013; Munsky, Neuert, & van Oudenaarden, 2012). Several clinical attempts have been made to target, block, or regulate transcription factor DNA-binding activity in disease states (Cheng et al., 2019; Colak & Ten Dijke, 2017; H. Li et al., 2020).

Additionally, a variety of techniques have been created to discover TFs using the large genome sequencing data (Pereira, Oliveira, & Sousa, 2020). Numerous in silico methods have been developed to annotate TFs at the genome scale in attempt to get over these limitations (Odom, 2011). A method to predict various classes of TFs, such as Helix-turn-helix, Beta-scaffold, and zinc-coordinating DNA binding domains, was developed by Zheng et al. (Zheng et al., 2008). In order to determine a protein's DNA-binding domains and deduce its DNA motif, Eichner et al. created a four-step workflow (Eichner et al., 2013). BRAT uses ChIP-seq dataset for the prediction of transcription factors (S. F. Cai & Levine, 2019). DeepTFactor is the most recently developed deep neural network method allow the prediction of TFs. The greatest dataset feasible was created and used in this study in order to develop an accurate and dependable methodology. The methods that are now available are computationally intensive and demand domain expertise.

We created an improved strategy for accurately predicting transcription factor in order to get beyond the shortcomings of previous techniques. We initially created prediction techniques based on homology or alignment. If the target TF in the database and the query TF are highly comparable, these alignment-based approaches perform well. These strategies are ineffective if a query TF has a high resemblance to a non-TF or a low similarity to known TFs in the database. We created an alignment-free methodology in order to get around these restrictions. Different machine learning methods have been utilised to create prediction models for alignment-based procedures. In the current method, we calculated the compositions of TFs associated and non-associated sequences; we further developed prediction models using the input features. We created a hybrid method to combine the strength of alignment-based and alignment-free approaches.

4.2 Materials and Methods

4.2.1 Overall architecture of the study

Figure 4.1 represents the complete workflow of the current study including collection and compilation, feature generation, model development and webserver implementation.

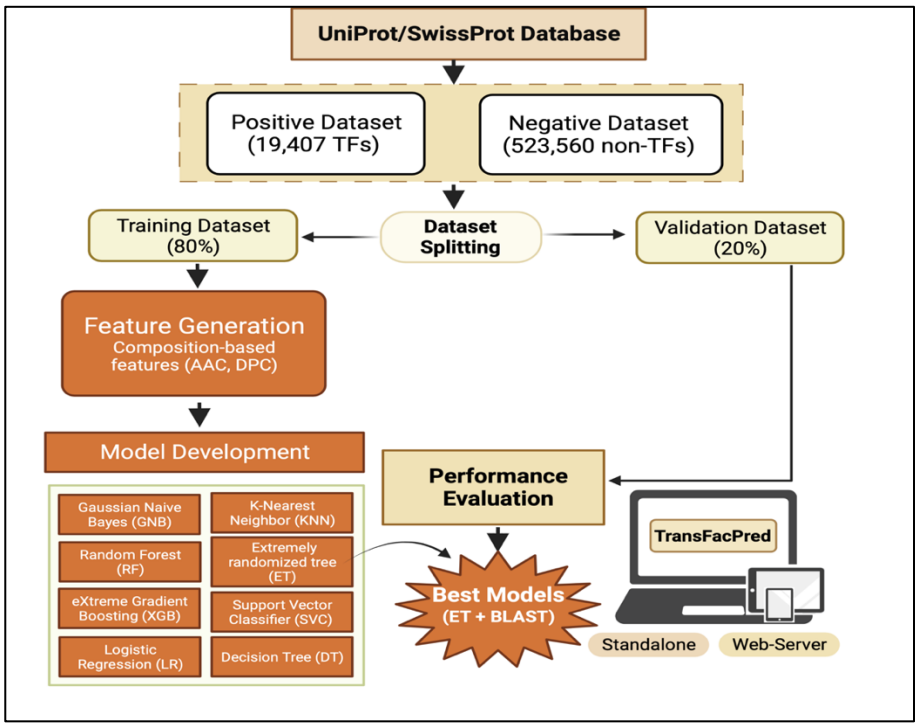


Figure 4.1: Complete pipeline and workflow of the study (Partially adopted from Patiyal, Tiwari, et al., 2022)

4.2.2 Creation of dataset and its pre-processing

We have extracted the dataset from UniProt-KB of the September 2019 release (Boutet, Lieberherr, Tognolli, Schneider, & Bairoch, 2007). The dataset was parsed and processed to segregate the proteins into two classes; transcription-factor (TF) and non-transcription-factor (non-TF), based on the annotation of Gene Ontology. Initially, we fetched 21802 TFs and 539374 non-TFs. Then, we processed the sequences by removing redundant sequences and sequences with non-natural amino acids. Finally, we left with 19406 unique TF sequences and 523560 non-TF sequences. The dataset was then divided into two datasets; training and independent dataset, where the training dataset comprises 80% of the entire dataset, which consists of 15525 TFs and 418848 non-TFs, and the remaining 20% data, i.e., 3882 TFs and 104712 non-TFs, was served as an independent dataset.

4.2.3 Generation of features

In this study, we have used four different kinds of features computed using Pfeature (Pande et al., 2019), like amino acid composition, dipeptide composition, the combination of amino acid and dipeptide composition, and binary profile. Amino acid composition calculates the percent proportion of each residue in the sequence and provides a vector of length 20 for each query. Dipeptide composition computes the percent proportion of all the possible dipeptides generated using 20 natural amino acids, i.e., 20*20=400, and hence generates a vector of length 400 for each query sequence. Then, we combined amino acid and dipeptide composition column-wise to get the vector of size 420. We have also used the binary representation or profile as the input feature in which each amino acid is represented by the binary vector of length 21, where '1' represents the presence, and '0' signifies the absence of the residue. For instance, A is represented by the vector '1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0', where '1' on position one signifies the presence of residue 'A'. The first twenty elements denote the presence/absence of 20 natural amino acids, whereas the last element is for the dummy variable.

4.2.4 Development of model

Various machine learning classifiers were implemented in this study to build the prediction model to classify the transcription factors. Classifiers included Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), eXtreme Gradient Boosting (XGB), K-Nearest Neighbors (KNN), randomized Extra Tree (ET), Support Vector

Classifier (SVC), and Convolutional Neural Network (CNN). We have optimized the parameters for each classifier by using the grid-search algorithm from scikit-learn (Pedregosa et al., 2011). Five-fold cross-validation was implemented to perform the internal validation so the over-fitness and biasness could be avoided.

4.2.5 Performance measures for evaluation

We have used performance measures of two-different categories such as threshold-dependent and threshold-independent. In threshold-dependent we have computed sensitivity, specificity, accuracy, F1-score, Kappa, and Matthews Correlation Coefficient (MCC). Equations 1-5 represents threshold-dependent parameters. Whereas, in threshold-independent parameter we have used Area Under Receiver Operating Characteristics (AUC) to evaluate the models.

$$Sensitivity = \frac{TP}{TP+FN} * 100 \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} * 100 \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (3)$$

$$F1-score = \frac{2TP}{FP+FN} \quad (4)$$

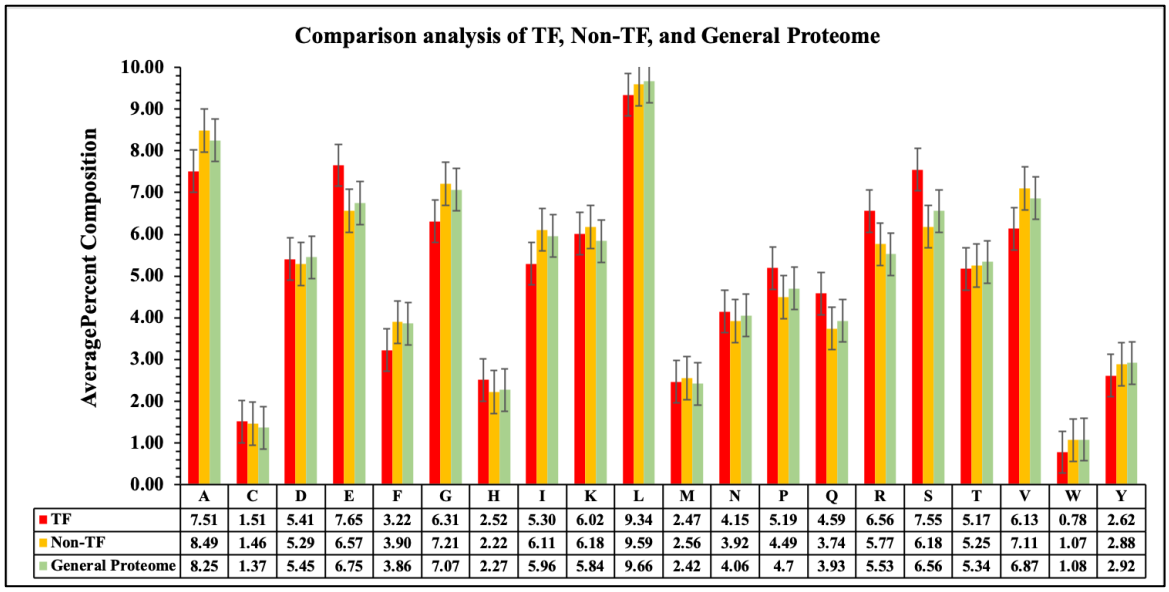
$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Where, FP is false positive, FN is false negative, TP is true positive and TN is true negative.

4.3 Results

4.3.1 Analysis based on composition

In the preliminary analysis, we have computed the amino acid composition of TF sequences and non-TF sequences, and calculated the mean composition for each residue and plotted in Figure 4.2. Compositional analysis showed that TF sequences have higher proportion of residues like E, P, Q, R, and S in comparison to the non-TFs.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
TF	7.51	1.51	5.41	7.65	3.22	6.31	2.52	5.30	6.02	9.34	2.47	4.15	5.19	4.59	6.56	7.55	5.17	6.13	0.78	2.62
Non-TF	8.49	1.46	5.29	6.57	3.90	7.21	2.22	6.11	6.18	9.59	2.56	3.92	4.49	3.74	5.77	6.18	5.25	7.11	1.07	2.88
General Proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.7	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 4.2: Mean percent composition of residues in TF, Non-TF, and General Proteome

4.3.2 Similarity search-based approach

In order to explore the efficiency of similarity search to classify the TFs and non-TFs, we have implemented BLAST by creating the customized database using sequences in the training dataset and then hit the query sequences in independent dataset against it by changing the e-value from 1e-6 to 1e+4. To classify the query sequences in one of the classes, we have considered the top-hit and based on that assign a class to each sequence such that if the top hit belongs to TFs then the assigned class is TF otherwise non-TF. Table 4.1 comprises the information regarding the performance at each e-value. It can be observed that there is an inverse relationship between e-value and probability of correct prediction, i.e., as e-value is increasing the probability of correct prediction is decreasing. Moreover, BLAST alone cannot predict the class for each sequence, hence we have used the machine learning classifiers to classify all sequences.

Table 4.1: Performance on similarity search-based (BLAST) method at different e-values on independent dataset (adopted from Patiyal, Tiwari, et al., 2022)

E-Value	No hits [Positive]	Probability of Correct Prediction
1e-6	68	95.44
1e-5	57	95.40
1e-4	44	95.28
1e-3	39	95.29

1e-2	32	95.30
1e-1	29	95.28
1e+0	22	95.26
1e+1	16	95.14
1e+2	3	94.87
2e+2	3	94.87
1e+e3	3	94.87

4.3.3 Machine learning based approach

In order to develop the prediction models for classifying TFs, we have implemented various classifiers using scikit-learn library. These classifiers included DT, RF, LR, XGB, GNB, ET, KNN, and SVC. Performance on each classifier using amino acid composition as the input feature is reported in Table 4.2, where model developed using ET classifier outperformed the other classifiers with AUC 0.967 and 0.968 on training and independent dataset respectively.

Table 4.2: Performance measures of models developed using AAC as input feature (adopted from Patiyal, Tiwari, et al., 2022)

Classifier	Training Dataset							Independent dataset						
	Sens	**Spec**	**Acc**	**AUC**	**F1**	**K**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**F1**	**K**	**MCC**
DT	52.435	98.192	96.557	0.753	0.523	0.505	0.505	52.486	98.273	96.637	0.754	0.529	0.511	0.511
RF	91.028	89.057	89.127	0.964	0.708	0.698	0.701	91.961	89.195	89.294	0.968	0.721	0.711	0.713
LR	73.895	74.756	74.725	0.814	0.214	0.168	0.215	74.130	75.001	74.970	0.813	0.213	0.169	0.215
XGB	85.592	86.791	86.748	0.940	0.582	0.568	0.574	86.988	86.966	86.967	0.946	0.589	0.575	0.583
KNN	84.684	94.997	94.628	0.913	0.671	0.661	0.669	85.803	95.011	94.682	0.919	0.674	0.663	0.672
GNB	67.835	71.707	71.569	0.772	0.235	0.202	0.206	67.070	72.002	71.825	0.767	0.235	0.201	0.206
ET	90.461	90.866	90.852	0.967	0.733	0.724	0.729	91.033	90.949	90.952	0.968	0.745	0.736	0.740
SVC	80.246	80.812	80.791	0.891	0.488	0.469	0.470	81.474	81.186	81.196	0.897	0.489	0.469	0.470

On the same note, we have also developed models using DPC as the input features by implementing eight different traditional machine learning classifiers. Performance of each model is exhibited in Table 4.3, where ET-based model performed best among the other classifiers with AUC 0.965 and 0.964 on training and independent dataset, respectively.

Table 4.3: Performance measures of models developed using DPC as input feature (adopted from Patiyal, Tiwari, et al., 2022)

Classifier	Training Dataset							Independent dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	52.544	98.180	96.549	0.754	0.522	0.505	0.505	52.409	98.193	96.557	0.753	0.522	0.504	0.504
RF	90.648	89.220	89.271	0.964	0.720	0.710	0.715	90.518	89.444	89.482	0.964	0.728	0.719	0.726
LR	80.343	80.711	80.698	0.876	0.301	0.265	0.303	80.752	80.779	80.778	0.878	0.308	0.272	0.309
XGB	90.113	90.305	90.298	0.965	0.720	0.710	0.715	90.054	90.505	90.488	0.966	0.720	0.711	0.716
KNN	84.117	96.321	95.885	0.913	0.714	0.703	0.704	83.767	96.225	95.780	0.912	0.719	0.709	0.709
GNB	75.866	49.497	50.439	0.694	0.166	0.122	0.135	76.269	49.668	50.619	0.696	0.165	0.122	0.133
ET	90.938	88.708	88.788	0.965	0.757	0.749	0.752	90.673	88.889	88.952	0.964	0.756	0.747	0.753
SVC	88.864	92.169	92.051	0.960	0.781	0.774	0.778	89.307	92.171	92.069	0.964	0.787	0.779	0.782

Further, we combined both the features column-wise such as AAC + DPC to generate a final vector of length 420, and the train and evaluate the models developed using same classifiers as mentioned above. The performance measures exhibiting the efficiency of each classifier is represented in Table 4.4, where XGB-based classifier attained the maximum performance with AUC 0.969 and 0.970 on training and independent dataset, respectively.

Table 4.4: Performance measures of models developed using combination of AAC and DPC as input feature (adopted from Patiyal, Tiwari, et al., 2022)

Classifier	Training Dataset							Independent dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	54.412	98.270	96.703	0.763	0.543	0.526	0.526	53.491	98.287	96.686	0.759	0.537	0.519	0.519
RF	91.885	89.655	89.735	0.969	0.729	0.720	0.723	91.832	89.731	89.806	0.969	0.738	0.729	0.732
LR	80.845	80.190	80.214	0.875	0.295	0.259	0.298	80.881	80.282	80.304	0.878	0.303	0.268	0.303
XGB	90.815	90.592	90.600	0.969	0.735	0.726	0.731	91.007	90.675	90.687	0.970	0.736	0.727	0.731
KNN	85.611	96.317	95.934	0.921	0.719	0.708	0.712	85.442	96.389	95.998	0.920	0.722	0.711	0.711
GNB	76.188	50.433	51.353	0.704	0.180	0.135	0.155	76.424	50.612	51.534	0.706	0.182	0.137	0.156
ET	91.814	88.980	89.082	0.968	0.758	0.750	0.754	91.497	89.178	89.261	0.966	0.759	0.751	0.754
SVC	86.507	84.927	84.984	0.935	0.645	0.633	0.637	86.756	85.212	85.267	0.939	0.650	0.638	0.639

4.3.4 Deep-learning based model

Other than machine-learning models, we also built models using deep-learning approach by implementing CNN using Keras module of TensorFlow library of Python. The results on various features such as AAC, DPC, AAC+DPC, and binary profile is reported in Table 4.5.

Model based on binary profile achieved the maximum AUC of 0.95 on the independent dataset, but still lesser than model developed using ET classifier on amino acid composition. Hence, we proceed further with ET-based model developed on AAC feature.

Table 4.5: Performance measures of models developed using CNN classifier on independent dataset (adopted from Patiyal, Tiwari, et al., 2022)

Feature	Sensitivity	Specificity	Accuracy	AUROC	F1	MCC
AAC	8.00	99.00	96.30	0.54	0.14	0.24
DPC	53.22	99.73	97.92	0.76	0.67	0.68
AAC+DPC	59.34	99.49	97.93	0.79	0.69	0.69
Binary Profile	91.27	98.61	98.32	0.95	0.81	0.81

4.3.5 Alignment free method with similarity search

ET-based model developed on AAC feature performed best among all the other classifiers and features. Hence, we have used the same model and combined with the similarity search approach to improve the performance. Table 4.6 provides the performance of combined model with varying e-values from 1e-6 to 1e+3 on the independent dataset. At e-value 1e+2, model achieved the accuracy of 97.013% with minimum difference between sensitivity and specificity, after this e-value the improvement in accuracy becomes negligible. The same model has been implemented in the backend of the server "TransFacPred".

Table 4.6: Performance of model developed using combination of machine learning and similarity search on independent dataset (adopted from Patiyal, Tiwari, et al., 2022)

E-value	Sensitivity	Specificity	Accuracy	AUC	F1	K	MCC
1.00E-06	95.877	95.406	95.423	0.990	0.936	0.933	0.933
1.00E-05	95.826	95.563	95.572	0.990	0.937	0.934	0.934
1.00E-04	95.697	95.762	95.759	0.990	0.937	0.935	0.935
1.00E-03	95.697	95.931	95.922	0.990	0.938	0.936	0.936
1.00E-02	96.032	96.031	96.031	0.990	0.938	0.936	0.936
1.00E-01	96.161	96.176	96.176	0.990	0.938	0.936	0.936
1.00E+00	96.341	96.413	96.410	0.990	0.938	0.936	0.936
1.00E+01	96.960	96.763	96.770	0.990	0.934	0.931	0.931
1.00E+02	97.063	97.011	97.013	0.990	0.927	0.924	0.924
2.00E+02	97.063	97.150	97.147	0.990	0.926	0.923	0.923
1.00E+03	97.063	97.242	97.236	0.990	0.926	0.923	0.923

4.3.6 Comparison with existing approach

In order to understand the efficiency of our best performing model i.e. hybrid method based model, to classify the transcription factors, we have predicted the classes for proteins using DeepTFactor standalone and our method, and compared the performance as shown in Table 4.7. TransFacPred outperforms the recently published method DeepTFactor in terms of performance measures taken into consideration.

Table 4.7: Comparison of performance of TransFacPred with DeepTFactor on independent dataset (adopted from Patiyal, Tiwari, et al., 2022)

Parameters	TransFacPred	DeepTFactor
Sensitivity	97.06	95.93
Specificity	97.01	95.78
Accuracy	97.01	95.79
AUC	0.99	0.97
F1	0.93	0.85
MCC	0.92	0.85

Additionally, we examined the processing times of DeepTFactor and TransFacPred separately using ML and a hybrid model while submitting a range of sequence counts, and we discovered that DeepTFactor takes longer as the sequence count rises, as shown in Table 4.8. On the other hand, we constructed a hybrid model and a machine learning model based on AAC and compared the results. While hybrid models performed best but took the longest to provide output, ML-based models required less time than DeepTFactor with equal AUC.

Table 4.8: Comparison of the processing time between DeepTFactor and TransFacPred (adopted from Patiyal, Tiwari, et al., 2022)

Number of Sequences	Method	Time (in seconds)		
		Real	User	System
50	**DeepTFactor**	13.285	3.882	1.188
	TransFacPred [ML]	7.666	1.551	0.998
	TransFacPred [Hybrid]	24.111	22.079	1.254
1000	**DeepTFactor**	55.201	51.37	3.954
	TransFacPred [ML]	37.208	2.649	1.157
	TransFacPred [Hybrid]	436.071	429.062	3.157

108594	**DeepTFactor**	6014.113	5629.047	375.138
	TransFacPred [ML]	134.387	130.191	1.945
	TransFacPred [Hybrid]	47932.78	47583.942	304.83

4.4 Web-based services

TransFacPred is a freely available online platform available at URL https://webs.iiitd.edu.in/raghava/transfacpred/ to predict the transcription factors using the best performing ET-based model developed on amino acid composition as well as hybrid model which combined ET-based model developed on amino acid composition and similarity search using BLAST. This user-friendly web-server was developed using PHP, HTML, Python, Perl, and JavaScript.

Several handy tools are incorporated in the web-interface of TransFacPred to facilitate the users to identify the transcription factors using sequence information. It includes three major modules “Predict”, “BLAST Search”, and “Standalone”. The basic prediction module include a approach which accepts multiple protein sequences in the FASTA format and predicts which of these maybe transcription factors. Figure 4.3 is a screenshot of the “Predict” module of the web-server displaying the submission form for the submission of the query sequences in the FASTA format and users are allowed to choose the alignment-free approach i.e. AAC based method which is fast or hybrid approach which is more accurate but slower. Figure 4.4 is an example output page obtained after the submission of the query sequences to the “Predict” module and by choosing hybrid model as the model for making predictions. Other than that, “BLAST scan” module allows to submit the multiple sequences to predict the transcription factors. We have provided the database of the sequences we have used in this study, if the query sequence hit against the transcription factor, the result page show it as transcription factor otherwise non-transcription factor. We have provided the option of varying e-values as per the requirement of the user. Additionally, we have provided the Python-based standalone to predict the transcription factors in the absence of the internet or for making prediction in a huge dataset, for instance, the entire human proteome, as server will take a long time to provide the results. The same standalone is also distributed using GitHub platform which can be cloned using basic git commands and can be used locally.

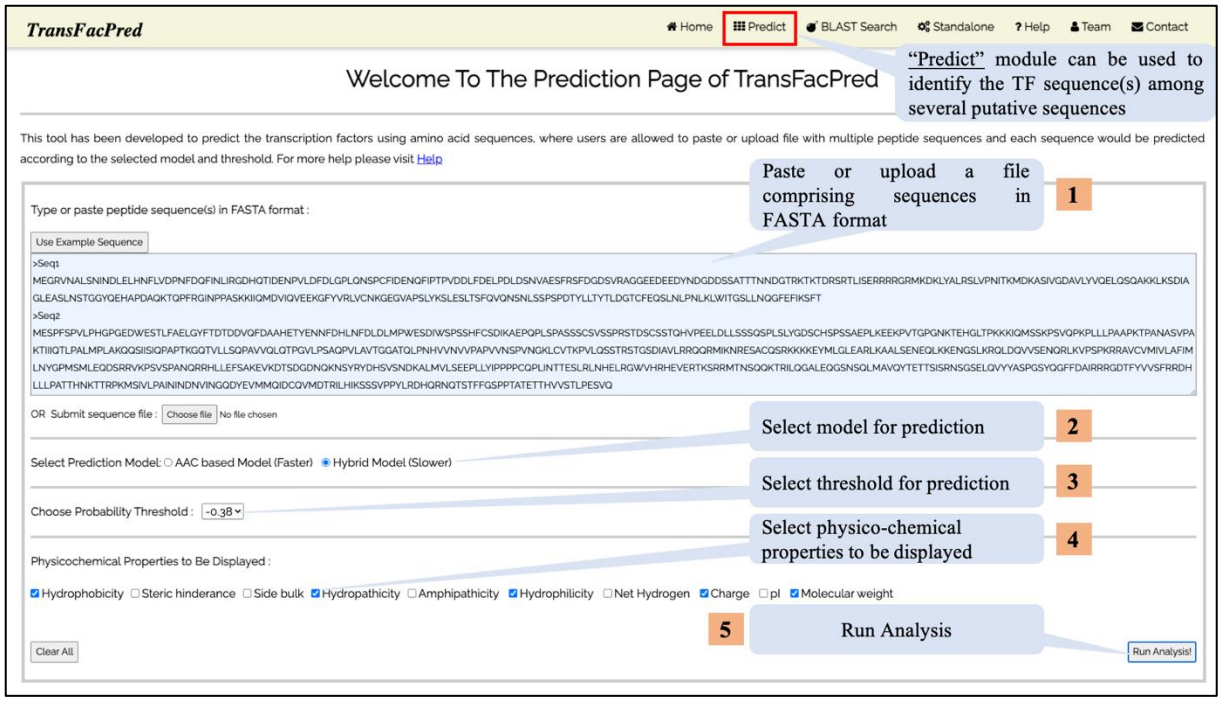


Figure 4.3: Screenshot of "Predict" module of TransFacPred web-server (URL https://webs.iiitd.edu.in/raghava/transfacpred/predict.php)

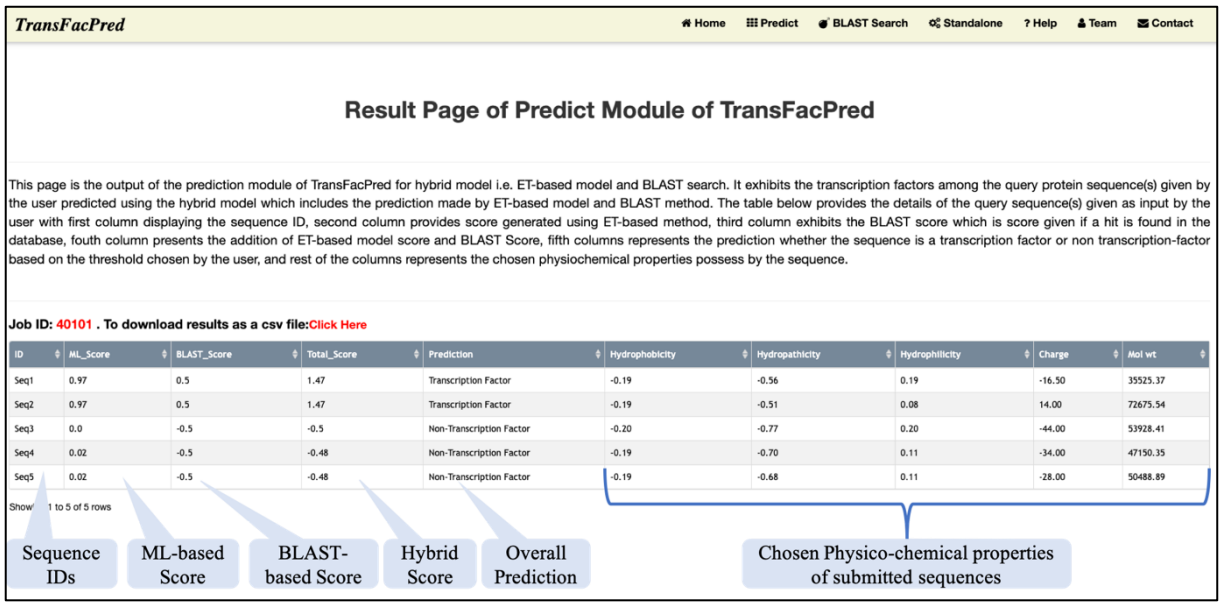


Figure 4.4: Screenshot of the result page of "Predict" module of TransFacPred webserver (URL https://webs.iiitd.edu.in/raghava/transfacpred/predict.php)

4.5 Discussion and Conclusion

Transcription factors play a major role in the initiation of a most important biological process i.e. transcription and hence decides the fact of a cell or the cellular function (Islam et al., 2021; Rhee, Kim, & Tucker, 2017). It is a laborious and expensive effort to determine novel or unknown TFs utilising experimentally based approaches like RNA sequencing (RNA-seq) and Chromatin Immunoprecipitation sequencing (ChIP-seq) (Muhammad, Kong, Akmar Abdullah, & Munusamy, 2019). Several computational approaches have been developed in the past to predict the transcription factors (Eichner et al., 2013; Kim et al., 2021; Zheng et al., 2008). Albeit, several computation method are available but there is still a significant room to improve the performance of methods in terms of accuracy. Therefore, we made an attempt to develop a highly accurate method to predict the transcription factor using primary structure information only.

We have used the sequences from UniProtKB/Swiss-Prot database and classify them based on GO terms into transcription factors and non-transcription factors. Initially, 561176 sequences were obtained out of which 21802 were assigned as transcription factor and rest 539374 were labelled as non- transcription factor, after the preprocessing of the datasets the final dataset was comprised of 19406 transcription factor sequences and 523560 non-transcription factor sequences. The datasets were further divided into 80% training dataset for internal validation using 5-fold cross-validation, and 20% independent dataset for external validation. The primary composition analysis showed that the residues E, P, Q, R, and S are abundant in the transcription factors as compare to non-transcription factors. Further, we have developed a variety of prediction models to classify the transcription factors using protein sequence information, in this study. To differentiate transcription factors from non-transcription-factors, we applied alignment-based approach using BLAST where we built the customized database using training dataset and hit query sequences in independent dataset on it using blastp module at different e-values ranging from 1e-6 to 1e+3. The approach performed well but was not able to classify every sequence in the independent dataset as some of them did not get the hit against the customized database. Then, we implemented alignment-free approach by developing machine learning-based models using diverse features such as amino acid composition, dipeptide composition, combination of amino acid and dipeptide composition, and binary profile. In order to signify the importance of the integration of alignment-based and alignment-free method, we have hit the query sequences in the testing dataset to the database created using

sequences in the training dataset, by implementing blastp algorithm. We have found that there are 34 TF sequences which did not get the hit in the database. We tried to predict the class of these sequences using machine learning based approach and able to classify 30 protein sequences as TF. One of the example is Growth-regulating factor 1 from *Arabidopsis thaliana* with UniProt ID O81001, which did not get the hit using BLAST approach. On the other hand, the same was predicted as TF using alignment-free approach with score of 1.0.

Our models were trained on 80% of the dataset and validated on the remaining uncharacterized 20% dataset. We obtained the AUC of 0.96 on training as well as validation dataset using amino acid composition-based features. Similar performance was obtained for models developed on dipeptide composition and their combination. Since, equivalent performance was attained using less number of features in amino acid composition, we proceeded further with the same. Then, we combined the alignment-based and alignment-free approach to improve the performance and called it as hybrid method and was able to attain the highest AUC of 0.99 on independent dataset with balanced sensitivity and specificity.We have also compared our method with the recently published method DeepTFactor and found out that our proposed method outperformed DeepTFactor at various performance measure. In order to investigate the robustness of the proposed method, we have downloaded the sequences from the Protein Data Bank (PDB) which were reported as TFs. The total number of retrieved entries were 16356. On applying CD-HIT with 90% allowed sequence identity criteria, only 1024 proteins sequences were clustered into different clusters, i.e., these 1024 sequences shared less than 90% sequence identity with the sequences used in this study. TransFacPred method was able to identify 983 (95.99%) TFs sequences accurately. This signifies that the method is robust and can classify the unknown TF sequences with high accuracy. One of the major limitations of this method is the redundancy in the dataset used for the training. In this study, we have followed the same steps as mentioned in the data creation section of DeepTFactor paper to fairly compare the two methods. Therefore, we have not removed the redundancy from the dataset. Ideally, the dataset should be non-redundant. One of the study's main objectives is to aid the scientific community. We developed a user-friendly web server (https://webs.iiitd.edu.in/raghava/transfacpred) that allows users to determine whether or not a particular protein sequence is a transcription factor. We have also provided the Python-based standalone package which can be used to predict the transcription factors in the entire proteome in the absence of internet. We anticipate that the work done here will aid in the annotation of protein sequences.

Chapter 5
Prediction of N-acetylglucosamine interacting residues in a protein

5.1 Introduction

The major challenge in the field of proteins is their annotation at the level of structure and function. The sequencing technology is advancing at an incredible pace, and hence the number of sequences is also increasing in various databases, but their annotation at the same swiftness is still a hurdle. Consequently, the gap between sequence submission and its annotation increases rapidly (Yu et al., 2014). Thus, there is a pressing need to develop computational approaches that can demarcate the protein's function at the residue level. The interaction between the biomolecules and proteins is very critical for many biological processes, which further decide the fate of cells and organisms (Agrawal, Singh, et al., 2019). There were many attempts have been made in the last few decades to determine the ligand-binding residues in the proteins, as reported in the review by Sousa et al. (Sousa, Fernandes, & Ramos, 2006). Initially, the computational approaches designed to predict the binding residues or pockets were non-specific in nature, i.e., the binding or interacting sites were predicted without considering the nature of the ligands it will bind to (Dundas et al., 2006; Le Guilloux, Schmidtke, & Tuffery, 2009). Nevertheless momentarily, it was learned that every ligand retains specific physical and chemical properties, and hence interacting with the protein in a specific manner. Such as several diseases, including breast cancer, arise from mutation in the ligand-binding site. In this respect, identifying potential interacting sites of protein with their ligand within the genome is essential for developing therapeutics. Therefore, new computational methods specific to the ligands were developed (Chauhan et al., 2009b; J. Hu, Li, Zhang, & Yu, 2018; X. Hu, Dong, Yang, & Zhang, 2016; Yu, Hu, Huang, et al., 2013), and these methods performed better in comparison to the non-specific methods (K. Chen et al., 2012; Yu, Hu, Yang, et al., 2013).

The computational approaches involved in predicting the binding site can be broadly classified into structure-based and sequence-based (Agrawal, Raghav, Bhalla, Sharma, & Raghava, 2018; Dukka, 2013). Structure-based are the ones where the interactions can be investigated via docking approaches (Fukunishi & Nakamura, 2011; Heo, Shin, Lee, & Seok, 2014). The three-dimensional (3D) structures of proteins have been used in structure-based drug design. However, these methods fail if the protein structure is unavailable. In order to overwhelm this constraint, sequence-based methods have been developed to predict the interacting residues in the protein with ligand specificities, such as for ATP (Chauhan et al., 2009b; K. Chen et al., 2011), GTP (Chauhan et al., 2010a), NAD (Ansari & Raghava, 2010), and SAM (Agrawal et

al., 2020). The approach mentioned here is the first method for predicting the NAG binding residues in the protein sequence. The NAG or N-acetylglucosamine is an omnipresent monosaccharide responsible for the crucial structure roles at the cell surface area in organisms meandering from bacteria to humans (Naseem et al., 2012). It is present in the peptidoglycan in the bacterial cell wall (Park & Uehara, 2008) and chitin, an important of the fungal cell wall (Gunasekera et al., 2010). The extracellular matrix of the animal cells also possesses glycosaminoglycans (Moussian, 2008). NAG is an essential factor in bacteria and fungi, as it regulates the gene expression of many genes by getting involved in cell signaling. Other than that, plants and animal cells also use NAG for the purpose of cell signaling and also play a significant role in the post-translational modification of proteins by adding O-GlcNAc (Naseem et al., 2012). It has been shown in the literature that NAG may play an essential role in the treatment of autoimmune diseases (Ercolini & Miller, 2009). NAG signaling enables the co-occurrence of a range of bacteria, fungi, and human cells in the human gut (Nicholson et al., 2012).

In this study, we have systematically attempted to predict the NAG interacting residues using the primary sequence information. We have provided a new approach named "NAGbinder," which uses a traditional machine learning algorithm to predict NAG interacting residues in a protein sequence. To serve the scientific community, we have provided a freely-accessible web server at https://webs.iiitd.edu.in/raghava/nagbinder and a standalone package available at https://webs.iiitd.edu.in/raghava/nagbinder/stand.html . Moreover, the same package has been distributed via docker technology in GPSRDocker.

5.2 Materials and Methods

5.2.1 Dataset extraction

We have used Protein Data Bank (PDB) April 2019 release to obtain the PDB IDs for 5736 NAG binding proteins, which comprise 15349 protein chains. Further, to filter the sequences, we have implemented the CD-HIT software (Huang, Niu, Gao, Fu, & Li, 2010) with the standards of 40% sequence identity and acquired 1279 protein chains with no two sequences having more than 40% sequence identity. As shown in the past, for reliable annotation, it is imperative to consider the quality of the protein structure (Chauhan et al., 2010a). Therefore, we have applied the threshold of 3Å, i.e., we have considered only those chains with a resolution equal to or less than 3Å, and we were left with 231 protein chains. Ultimately, to get

the contact information for NAG interacting residues in these protein chains, we have implemented LPC software (Sobolev, Sorokine, Prilusky, Abola, & Edelman, 1999) with the threshold of 4Å, i.e., we have assigned a residue as NAG interacting only if the distance between the atom is less than or equal to 4Å, which is a standard criterion considered by the earlier studies (Chauhan et al., 2010a; N. K. Mishra & Raghava, 2010b). The final dataset comprises 1985 NAG interacting and 74931 non-interacting residues.

Further, the complete dataset was then divided into training and independent datasets based on the protein level rather than the pattern level, as the dataset generated on the pattern- or residue-level lead to the biasness and provide higher performances (Yu et al., 2014). The training dataset contains 80% (186 protein chains) of the entire dataset, whereas the remaining 20% (45 protein chains) were kept for the external validation and termed as a independent dataset. In terms of residues, training data comprises 1335 NAG-interacting and 47198 non-interacting residues, while the independent dataset contains 650 NAG-interacting and 27733 non-interacting residues. Two kinds of datasets were generated for further analysis and to develop the prediction models: a) balanced dataset (1985 interacting and 1985 non-interacting residues), in which an equal number of instances for interacting and non-interacting residues were considered since non-interacting residues were several folds higher, an equal number of instances were randomly picked from non-interacting residues, and b) realistic dataset, which is the original dataset, i.e., 1985 instances of NAG interaction and 74931 of non-interaction.

5.2.2 Size of the pattern

To generate the feature vector of equal size, we have generated the overlapping patterns of different window sizes with odd numbers (5-23) for each protein sequence. The residue in the middle is the overall representative of the pattern, such as, if the central residue is interacting, the while pattern is assigned as interacting, otherwise non-interacting. To handle the residues in each terminus, (n-1)/2 dummy variable "X" is added on both ends, where n is the window/pattern size.

5.2.3 Binary profile

The binary profile or one-hot encoding represents each amino acid with a vector size of 21, where each element of the vector signifies the presence or absence of particular amino acid, such as residue ‘A’ is represented by ‘1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0’, similarly

dummy variable 'X' is represented by ‘0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1’. Hence, if a pattern has a length of K amino acids, the resulting vector would be of length K*21 (Agrawal & Raghava, 2018).

5.2.4 PSSM profile

Besides the binary profile, we have also calculated each sequence PSSM (Position-Specific Scoring Matrix) or evolutionary profile (K. Chen et al., 2012). The PSSM profile was calculated using PSI-BLAST (Altschul et al., 1997) by searching against the SwissProt database (Bairoch & Apweiler, 2000). To run the PSI-BLAST, the number of iterations was set to three with an e-value of 0.001. Further, the profile was normalized between 0 and 1, using equation 1. Finally, the normalized PSSM profile has the dimension of N X 21, where N is the length of the sequence. Further, the vector against each amino acid is concatenated that comes in the pattern of size K; hence resulting vector size for each pattern would be K*21.

$$\boldsymbol{Norm_{PSSM}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{x}}} \quad [1]$$

Where, x is the PSSM score and $Norm_{PSSM}$ is the normalized PSSM value.

5.2.5 Model building

In order to build the prediction model, we have implemented various machine learning classifiers using the scikit-learn (Pedregosa et al., 2011) library of Python. We have also hyper-tuned the parameter using the grid search module of sklearn. We have implemented five-fold cross-validation for the purpose of internal validation. We have used various classifiers such as RF (Random Forest), ET (Extra Tree), MLP (MultiLayer Perceptron), SVC (Support Vector Classifier), KNN (K-Nearest Neighbour), and Ridge classifier.

5.2.6 Performance measures

To compare and evaluate the performance of the generated models, we have calculated various measures, which can be broadly classified into threshold-dependent and threshold-independent. In threshold-dependent measures, we have calculated Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), and Matthews Correlation Coefficient (MCC); on the other hand, in threshold-independent measures, we have considered Area-Under the Receiver

Operating Curve (AUROC), which signifies the relation between the True Positive Rate (TPR) and False-Positive Rate (FPR). The pROC package of R (Sachs, 2017) was implemented to calculate and plot the AUROC. The equation for threshold-dependent parameters is provided in equations 1-5 of section 4.2.5 of Chapter 4.

5.3 Results

5.3.1 Overall workflow

The overall workflow of NAGbinder is depicted in Figure 5.1

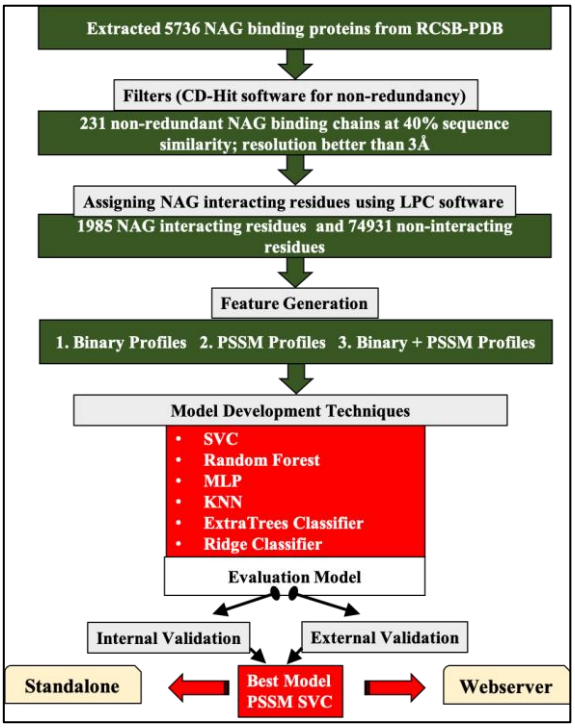
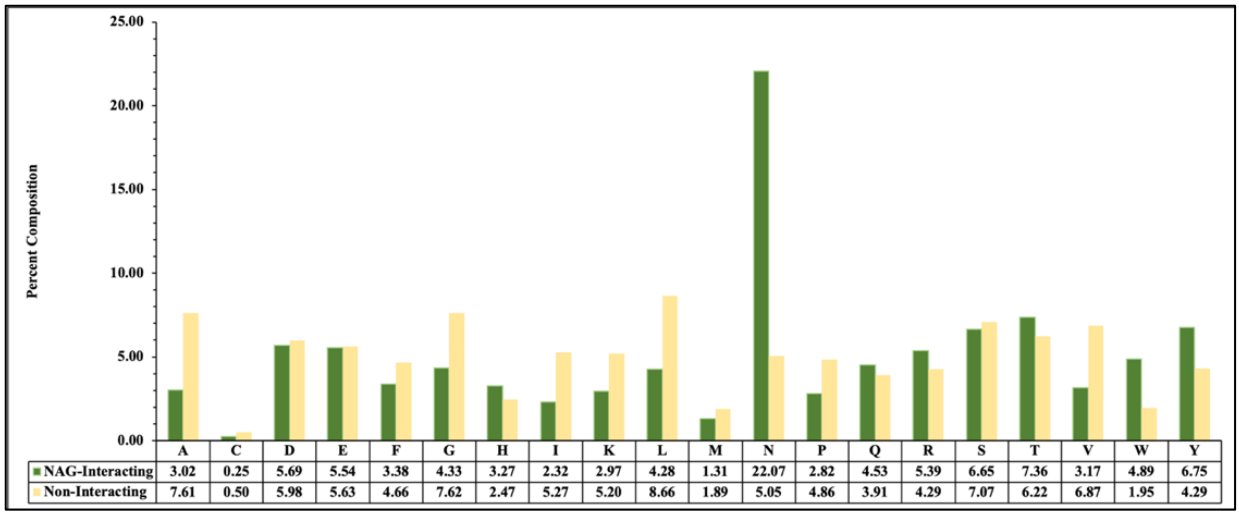


Figure 5.1: Complete workflow of the study; including data collection, model generation and webserver development

5.3.2 Composition based analysis

In the primary analysis, we have calculated and compared the amino acid composition of NAG-interacting and non-interacting residues in the NAG binding proteins and plotted the graph as shown in Figure 5.2. Figure 5.2 depicts that residues N, Q, R, T, W, Y, and H are more preferred in NAG interacting residues compared to the non-interacting residues. A similar fact is reported in the study by Ramakrishnan et al. (Ramakrishnan, Boeggeman, & Qasba, 2012).



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
NAG-Interacting	3.02	0.25	5.69	5.54	3.38	4.33	3.27	2.32	2.97	4.28	1.31	22.07	2.82	4.53	5.39	6.65	7.36	3.17	4.89	6.75
Non-Interacting	7.61	0.50	5.98	5.63	4.66	7.62	2.47	5.27	5.20	8.66	1.89	5.05	4.86	3.91	4.29	7.07	6.22	6.87	1.95	4.29

Figure 5.2: Composition of NAG interacting residues and non-interacting residues for each type of residue

5.3.3 Propensity based analysis

To explore the preference of amino acids in the NAG-interacting sites, we have calculated the propensity score for each residue using equation 1 (H. Singh, Srivastava, & Raghava, 2016) and plotted the same in Figure 5.3. As exhibited in the Figure 5.3, residues N and W are preferred in NAG-interacting sites as compared to other amino acids.

$$\boldsymbol{Propensity_{Score_i}} = \frac{\boldsymbol{R_i}}{\boldsymbol{T_i}}\boldsymbol{X\ 100} \qquad [\mathbf{1}]$$

Where, $Propensity_{Scorei}$ is the propensity score of residue of type i, R_i is the number of residue of type i, and T_i is the total number of residue (interacting and non-interacting) of type i.

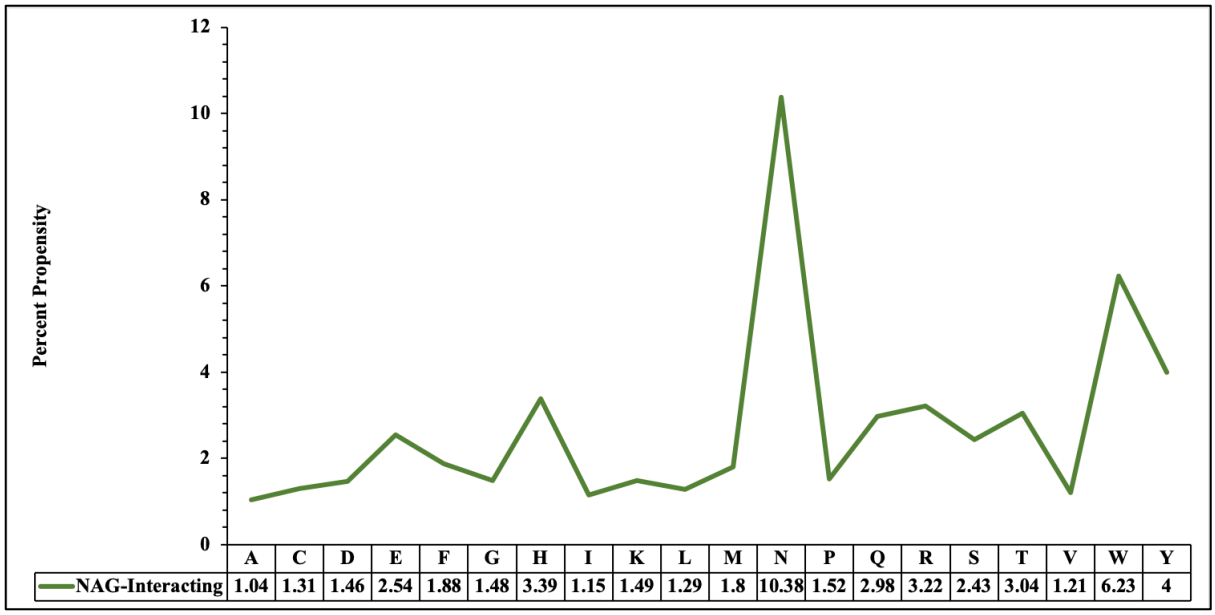


Figure 5.3. Percent propensity of NAG interaction of each type of residue

5.3.4 Physicochemical properties based analysis

The nature of residues plays a significant role in the interacting with ligands, hence we have calculated the composition of eight different physicochemical properties such as, nature of amino acid (acidic or basic), aromaticity of side chain (aliphatic or aromatic), size of side chain (small or large), and polarity (polar or non-polar). Figure 5.4 represents the physicochemical properties based composition, which signifies that NAG-interacting residues are rich in small, polar and aromatic amino acids.

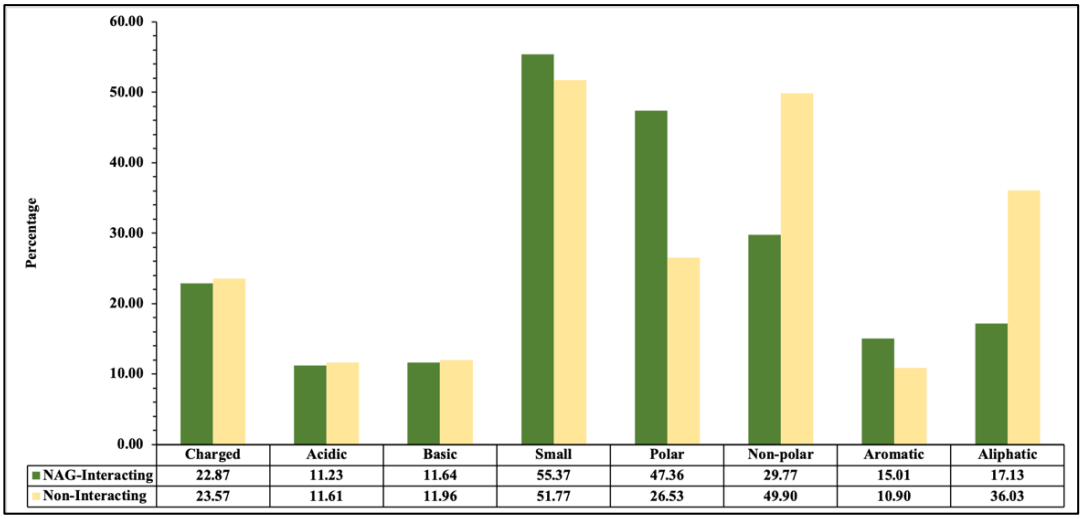


Figure 5.4. Percentage composition of physicochemical properties possess NAG interacting and non-interacting residues

5.3.5 Binary profile based models

Binary profile is able to capture the compositional and spatial information in a sequence (Agrawal & Raghava, 2018; Chauhan et al., 2010a). We have generated the patterns of different window sizes [5-23] and calculated the binary profile for balanced dataset which comprises of 1985 NAG-interacting and 1985 non-interacting patterns. The performance of the best classifiers among all the used classifiers for each window size is reported in Table 5.1. On observing the results, we have found that model developed using RF classifier on window size 9 outperforms all the other classifiers and window sizes, with AUROC 0.73 and MCC 0.31 on training dataset, and AUROC 0.70 and MCC 0.25 on the independent dataset, with balanced sensitivity and specificity.

Table 5.1: The performance of best model developed using binary pattern for each window size on balanced dataset (adopted from Patiyal et al., 2020)

Pattern (Classifier)	**Training Dataset**					Independent **Dataset**				
	Sens	**Spec**	**Accu**	**MCC**	**AUROC**	**Sens**	**Spec**	**Accu**	**MCC**	**AUROC**
Pat5(SVC)	67.42	60.9	64.16	0.28	0.71	67.18	55.69	61.43	0.23	0.68
Pat7(SVC)	66.52	64.12	65.32	0.31	0.72	66.26	60.00	63.13	0.26	0.70
Pat9(RF)	65.39	65.77	65.58	0.31	0.73	65.69	59.69	62.69	0.25	0.70
Pat11(RF)	66.07	65.17	65.62	0.31	0.72	69.08	60.62	64.85	0.30	0.71
Pat13(RF)	65.62	65.77	65.69	0.31	0.72	69.69	62.31	66.00	0.32	0.71
Pat15(RF)	66.52	65.24	65.88	0.32	0.72	68.00	59.23	63.62	0.27	0.71
Pat17(RF)	67.64	61.12	64.38	0.29	0.71	68.15	58.92	63.54	0.27	0.69
Pat19(RF)	66.37	62.47	64.42	0.29	0.71	67.69	59.54	63.62	0.27	0.70
Pat21(RF)	67.87	61.57	64.72	0.29	0.71	67.38	60.15	63.77	0.28	0.70
Pat23(RF)	67.57	62.02	64.79	0.30	0.71	66.00	59.85	62.92	0.26	0.69

5.3.6 PSSM profile based models

Evolutionary or PSSM profile comprises more information than a single sequences as it is based on the alignment with the sequences the non-redundant database (H. Kaur & Raghava, 2003; Kuznetsov, Gou, Li, & Hwang, 2006). Hence, we have used the PSSM profile as input feature to predict the NAG-interaction for different window sizes. Table 5.2 comprises of the performance measures of best performing classifiers in each window size. As shown by Table 5.2, RF-based model with pattern size 9 outperformed other window size with AUROC 0.69

and MCC 0.24 on training dataset, and AUROC 0.66 and MCC 0.22 on the independent dataset.

Table 5.2: The performance of best model developed using PSSM pattern for each window size on balanced dataset (adopted from Patiyal et al., 2020)

Pattern (Classifier)	Training Dataset					Independent Dataset				
	Sens	Spec	Accu	MCC	AUROC	Sens	Spec	Accu	MCC	AUROC
Pat5(RF)	61.27	61.57	61.42	0.23	0.67	52.00	64.46	58.23	0.17	0.64
Pat7(RF)	61.27	61.87	61.57	0.23	0.68	56.46	66.92	61.69	0.24	0.66
Pat9(RF)	62.47	61.87	62.17	0.24	0.69	55.38	66.92	61.15	0.22	0.66
Pat11(RF)	62.92	62.55	62.73	0.25	0.68	56.15	66.31	61.23	0.23	0.66
Pat13(RF)	64.27	62.17	63.22	0.26	0.68	56.92	66.46	61.69	0.23	0.66
Pat15(RF)	62.47	62.17	62.32	0.25	0.68	56.62	64.92	60.77	0.22	0.66
Pat17(RF)	63.67	61.8	62.73	0.25	0.68	54.15	63.23	58.69	0.17	0.65
Pat19(ETree)	64.04	62.77	63.41	0.27	0.68	53.85	65.38	59.62	0.19	0.65
Pat21(ETree)	65.02	62.25	63.63	0.27	0.69	54.31	65.69	60.00	0.20	0.66
Pat23(ETree)	63.45	63.00	63.22	0.26	0.68	54.62	66.77	60.69	0.22	0.65

5.3.7 Performance on realistic dataset

On analysing the results on each window size, we conclude that pattern size 9 in the optimal one since, performance measures of models developed using binary and PSSM profile exhibited that best performing models are the one with window size 9. Therefore, we developed models on realistic dataset using binary profile for pattern size 9 and performance measures for each classifier is reported in Table 5.3. According to Table 5.3, RF-based model achieved the maximum MCC of 0.26 with AUROC of 0.70 on training dataset, and MCC of 0.27 and AUROC 0.71 on the independent dataset. Figure 5.5 exhibits the AUROC plots for each classifier on window size 9 on training and independent dataset.

Table 5.3. The performance of binary pattern-based models developed for window size 9 on realistic dataset (adopted from Patiyal et al., 2020)

Classifiers	Training Dataset					Independent Dataset				
	Sens	Spec	Acc	MCC	AUROC	Sens	Spec	Acc	MCC	AUROC
SVC	14.91	99.47	97.15	0.25	0.71	18.95	99.43	97.59	0.28	0.72
RF	16.70	99.41	97.14	0.26	0.70	19.69	99.35	97.53	0.27	0.71
ET	17.30	99.22	96.97	0.25	0.70	19.69	99.26	97.44	0.26	0.70
KNN	8.61	98.88	96.40	0.11	0.61	10.92	98.99	96.97	0.13	0.63
MLP	13.78	98.94	96.60	0.18	0.71	17.85	98.78	96.92	0.20	0.72
Ridge	13.11	99.11	96.74	0.18	0.70	16.62	99.2	97.31	0.22	0.71

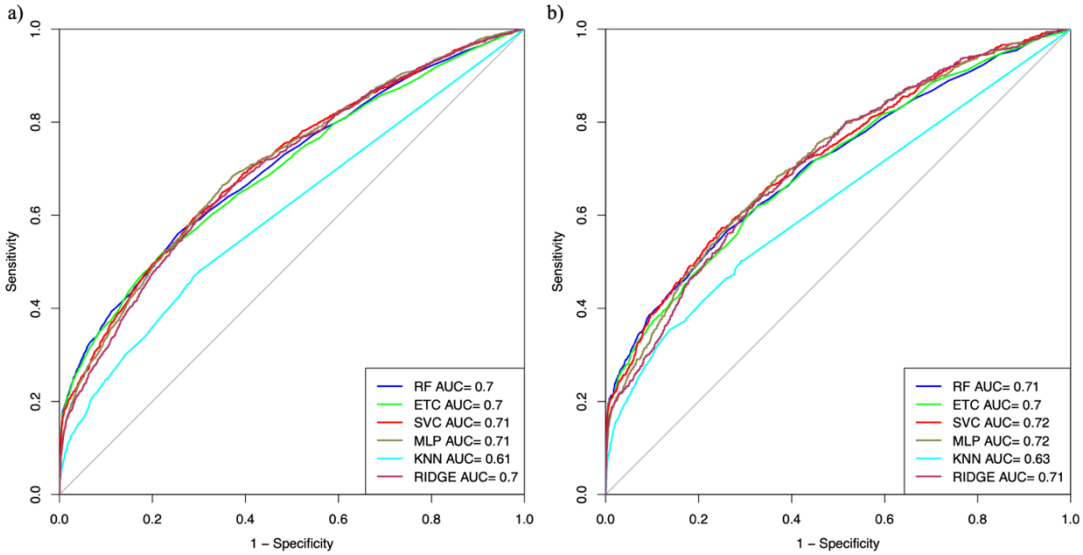


Figure 5.5. AUROC curve for window length 9 developed using binary profiles on realistic dataset for (a) training dataset and (b) independent dataset

5.4 Web-based services

NAGbinder is a freely available web-interface available at URL https://webs.iiitd.edu.in/raghava/nagbinder to predict the NAG interacting residues in a protein using the best performing models developed on binary profile of pattern length 9. This user-friendly web-server was developed using PHP, HTML, Python, Perl, and JavaScript.

Several handy tools are incorporated in the web-server of NAGbinder to facilitate the users to determine the NAG interacting residues using sequence information. It includes four major modules "Sequence", "PSSM Profile", "Standalone", and "Download". The basic sequence

module include a approach which accepts multiple protein sequences in the FASTA format and predicts NAG interacting residues in each sequence. In each prediction module of NAGbinder, overlapping patterns of length 9 is generated and then these patterns are used to make predictions. Users are allowed to provide their email, so that for the long processes they do not have to wait and the results will be send to their provided email IDs, once the process gets completed. Figure 5.6 is a screenshot of the "Sequence" module of the web-server displaying the submission form for the submission of the query sequences in the FASTA format and users are allowed to choose the classifiers such as Random Forest, SVM, ANN, or KNN, by default best performing classifier Random Forest is chosen. Figure 5.7 is an example output page obtained after the submission of the query sequences to the "Sequence" module, by choosing Random Forest based model for making predictions. The output page exhibits sequences with interacting residues highlighted in red color and bigger font size, whereas non-interacting residues are shown in black color. The result page is downloadable in the .txt, .pdf, and .png format. Other than that, "PSSM Profile" module allows to submit the multiple sequences to predict the NAG interacting residues by generating PSSM profile by implementing PSI-BLAST in the backend, followed by generation of overlapping patterns for window length 9. The result page show the submitted sequences with interacting residues highlighted in red color and bigger font size. Additionally, to predict the NAG interacting residues in a protein in the absence of the internet or for prediction in a large dataset, such as the complete human proteome, since server will take a long time to produce the results, we have developed the Python- and Perl-based standalone. The same standalone is also released through the GitHub platform, which may be locally downloaded using standard git instructions.

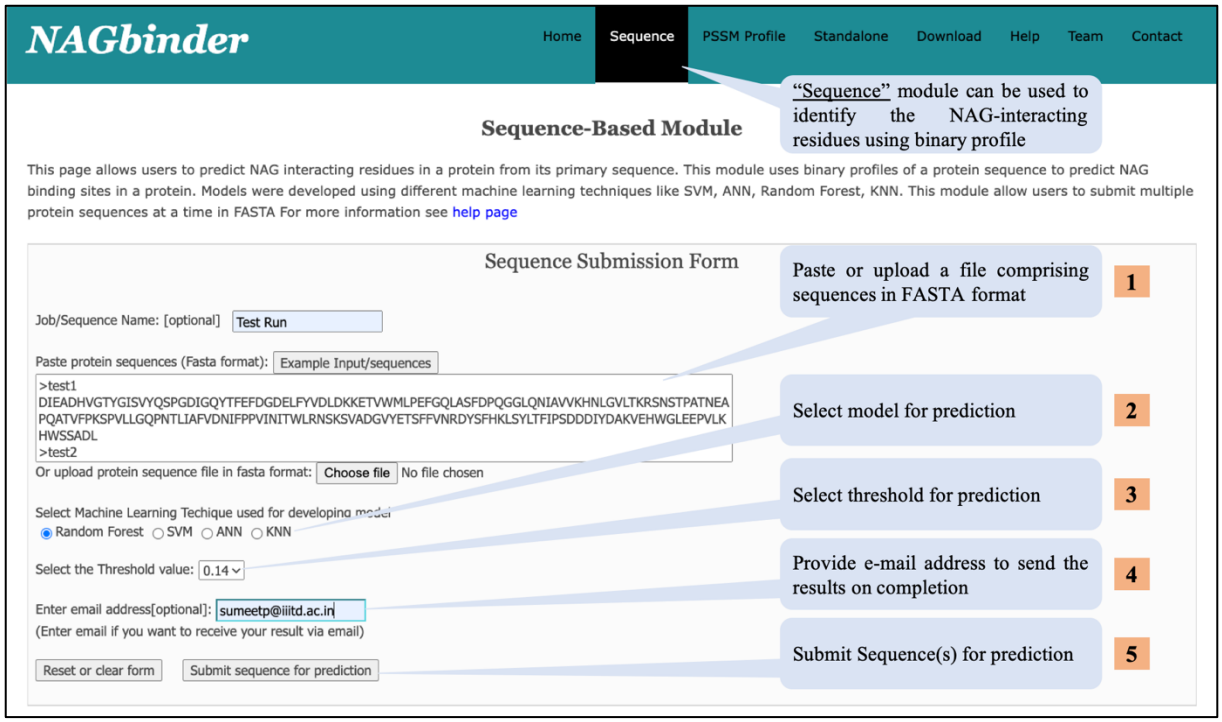


Figure 5.6: Screenshot of "Sequence" module of NAGbinder web-server

(URL https://webs.iiitd.edu.in/raghava/nagbinder/batch.html)

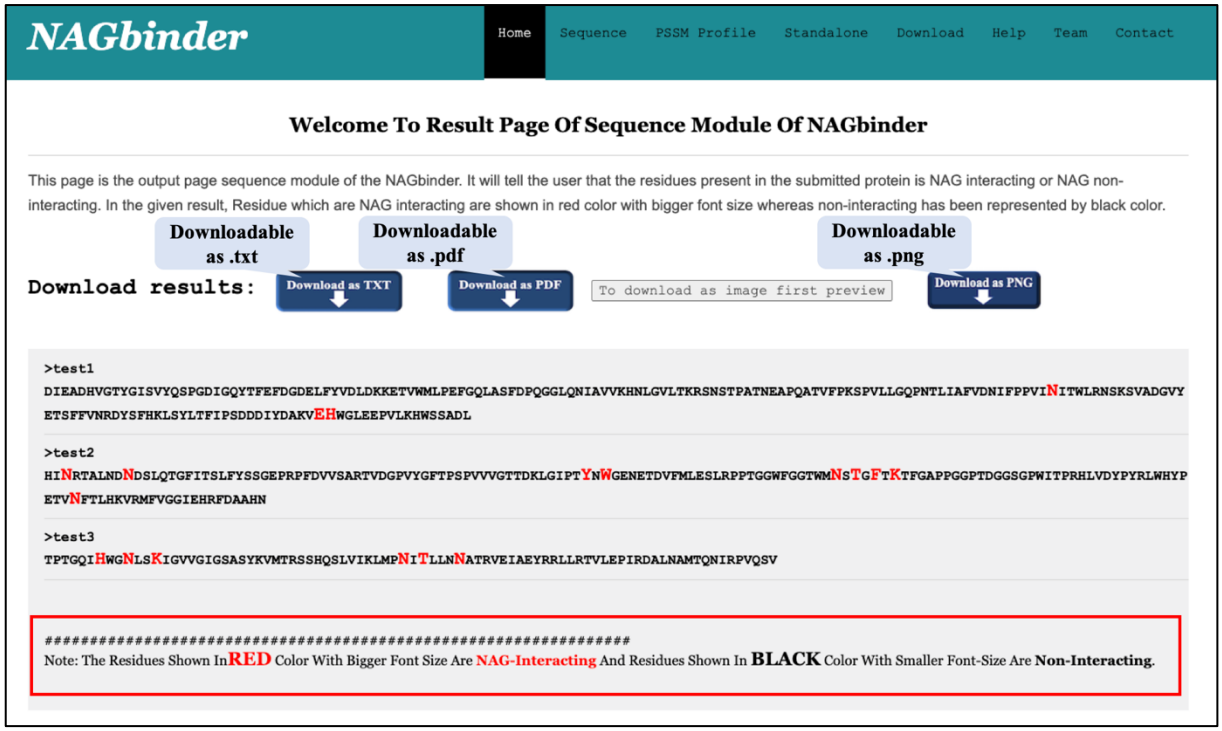


Figure 5.7: Screenshot of the result page of "Sequence" module of NAGbinder webserver (URL https://webs.iiitd.edu.in/raghava/nagbinder/batch.html)

5.5 Discussion and Conclusion

N-acetylglucosamine plays a very crucial role in the maintenance of several biological systems. NAG is one of the eight crucial sugars required to maintain the optimal health and functioning of the human body and also plays a significant role in the communication between the cells. As shown in the clinical trials, essential saccharides play a vital role in reducing allergies and reduce symptoms in chronic diseases such as arthritis, diabetes, lupus, and kidney disease. NAG actively participates in the numerous process in the human body as repairing of cartilage, decreases inflammation with bones joints, tissue rebuilding, the functioning of the digestive tract and nervous system, molecule transportation such as thyroglobulin. The presence of NAG in the liver controls the secretion of insulin. This sugar is also known to possess antiviral and anti-tumor activity. Because NAG is so pervasive in the environment, it is clear how important it is to upkeep and coordination of several systems, from microbes to people. The determination of the structure, which is a highly difficult task, is a requirement for understanding the mechanisms behind the interactions. Therefore, there is an urgent need to develop sequence-based computational approaches to anticipate the NAG interacting residues in proteins due to the intricacy of structure determination and limits of current technologies.

In order to predict the NAG interacting residues in the uncharacterized proteins using their sequence information, we investigated various properties of the NAG interacting protein chains, including compositional analysis, propensity, and physiochemical properties. The compositional and propensity analysis showed that residue “N” is most abundant in the NAG-interacting residues. N-glycans are covalently attached to protein at “N” residues by an N-glycosidic bond, and N-acetylglucosamine to asparagine (GlcNAcβ1-Asn) is the most common linkage among the five N-glycans. We then developed various prediction models using machine learning techniques using different kinds of input features like binary and PSSM profile using two type of datasets such as balanced and realistic dataset. The models were first created on balanced data using various window widths. Performance measures of models developed on different window size signified that pattern size 9 is the optimum one with binary profile as the input feature. Hence, the final model was developed on realistic dataset using binary profile with window size 9. The performances of the models were confirmed using the independent datasets. We built a user-friendly prediction web server that incorporates various modules to predict the NAG interacting residues in a protein sequences to support the scientific community. The study's prediction models are implemented in the backend of the web server

named as “NAGbinder” available at https://webs.iiitd.edu.in/raghava/nagbinder. Sequence, PSSM Profile, Standalone, and Download are the four major modules in the “NAGbinder”. In the guise of docker technology, we've also offered a stand-alone facility. This standalone is included in our GPSRDocker package (Agrawal, Kumar, et al., 2019), which may be obtained from https://webs.iiitd.edu.in/gpsrdocker. Figure 5.8 depicts the utility of NAGbinder. One of the limitations of the proposed method is that the model was trained on the limited dataset, due to which it was not possible to divide the dataset into N-acetylglucosamine and O-GlcNAcylation. In the future, when a sufficient amount of data will be available this method can further be trained to classify the residues in two-layers, such as, in the first layers it will be predicted if the residues is NAG-interacting or not, which can further be distinguish the N-acetylglucosamine from O-linked-N- acetylglucosaminylation.

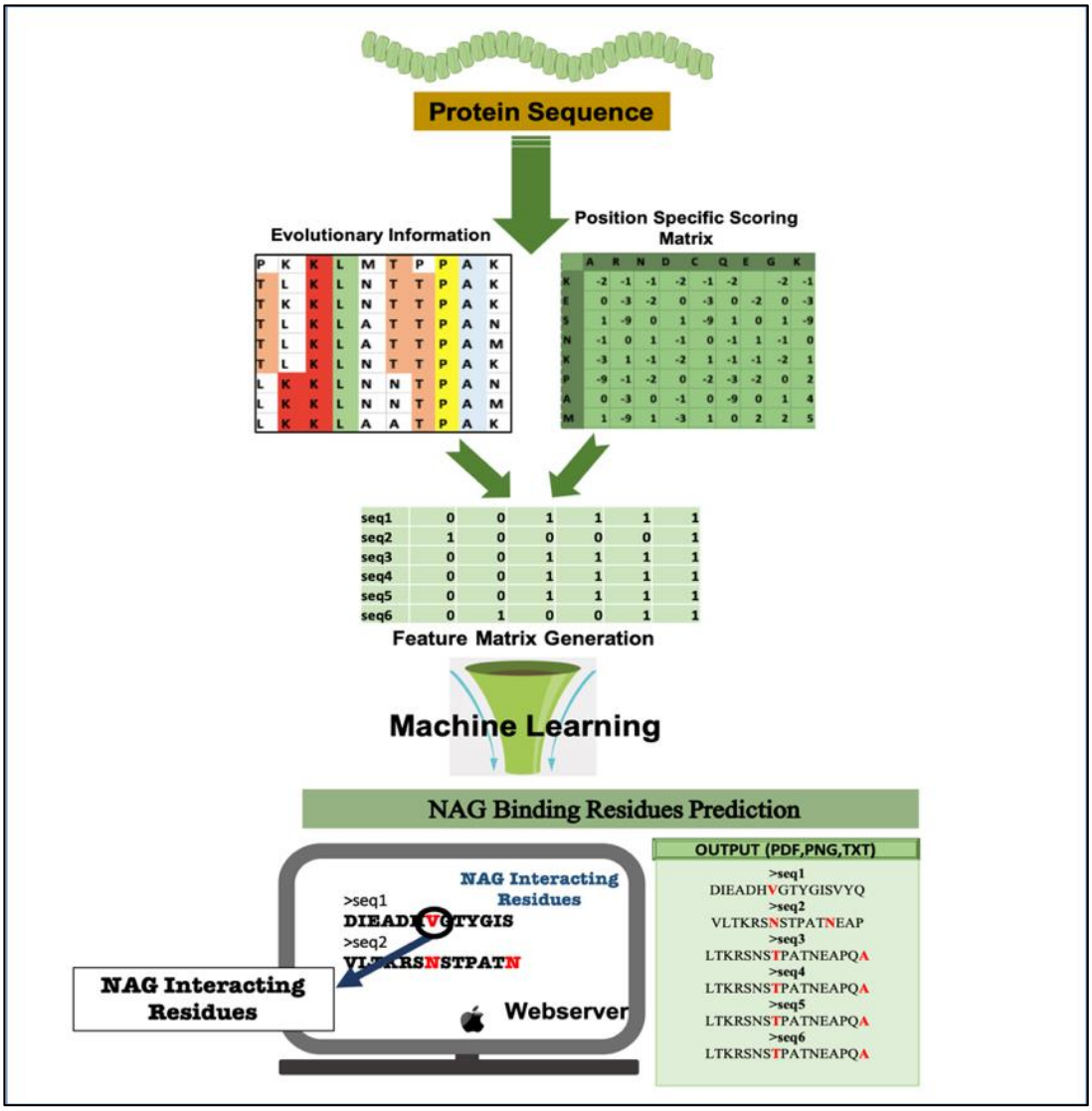


Figure 5.8: Utility of NAGbinder webserver

Chapter 6

Identification of DNA-binding residues in a protein

6.1 Introduction

Molecular interactions such as DNA-protein, RNA-protein, and protein-protein interactions are crucial for a living organism to perform various biological functions (Emamjomeh, Choobineh, Hajieghrari, MahdiNezhad, & Khodavirdipour, 2019). The interactions between DNA and proteins play a significant role in many biological processes such as transcription, regulation of gene expression, and splicing (Aeling et al., 2007; Choi & Han, 2011; Si et al., 2015; Wong et al., 2016). There are many experimental approaches to determine the DNA-protein interactions (Collas, 2010; Furlan-Magaril et al., 2009; Jayaram et al., 2002). The three-dimensional structure of DNA-protein complexes aids researchers in understanding the essential aspects of interactions, such as DNA confirmations, stability via hydrogen bonds, nature of amino acids, interactions like electrostatics, Vander Waals forces, etc. (Berger et al., 2006; Ho et al., 2006; S Jones et al., 1999; Lejeune et al., 2005; Nadassy et al., 1999; Nagarajan et al., 2013; Ponting et al., 1999). The advancement in sequencing technology is responsible for the exponential increase in the sequences of DNA binding proteins in the respective databases. However, structure determination techniques could not cope up with the pace of sequencing technology; hence only a limited number of structures have been deposited in the Protein Data Bank (PDB) (Rose et al., 2015). On the contrary, betterment in the machine learning algorithms allowed us to predict the protein structures with high accuracies; AlphaFold and AlphaFold2 (Jumper et al., 2021) are the products of the same advancement. But the computational requirements of these tools are costly in terms of space and memory.

Myriads of computational approaches have been designed in the last few decades to predict the DNA-interacting (Miao & Westhof, 2015; Schmidtke & Barril, 2010; Si et al., 2015; Liangjiang Wang & Brown, 2006). These tools can be vastly classified into four categories such as sequence-based, structure-based, evolutionary-based, and hybrid approaches which combine sequence- and structure-based approaches (Chowdhury et al., 2017; Ferguson & Allen, 1988; Hwang et al., 2007; Susan Jones et al., 2003; B.-Q. Li et al., 2014; R. Liu & Hu, 2013; Tjong & Zhou, 2007; Yuan et al., 2022). Methods developed in the earlier times are developed on a limited number of complexes; some of the examples are BindN (Liangjiang Wang & Brown, 2006), BindN+ (Liangjiang Wang et al., 2010), NucBind (Su, Liu, Sun, Peng, & Yang, 2019), and DP-Bind (Hwang et al., 2007). At the same time, recent methods have considered a large number of DNA-protein complexes to train their models (Ferguson & Allen, 1988; Qiu et al., 2020; J. Yan & Kurgan, 2017; Yuan et al., 2022; Jian Zhang et al., 2019). In

spite of tremendous advancements in the prediction algorithms during the course, there is still enough room for improvements in the performance. Therefore, it is the need of the hour to develop an accurate method to predict the DNA binding residues using sequence information.

In this study, we have made a systematic attempt to develop a new method with the capability of classifying DNA-interacting residues in a protein sequence using a deep-learning approach. Internal and external validation was performed to develop an unbiased method. To serve the scientific community, we have provided a freely-accessible web server "DBPred" at https://webs.iiitd.edu.in/raghava/dbpred and a Perl- and Python-based standalone package is available at . Moreover, the same package has been deployed via docker technology in GPSRDocker (Agrawal, Kumar, et al., 2019).

6.2 Materials and Methods

6.2.1 Overall architecture of the study

Figure 6.1 represents the complete workflow of the current study including collection and compilation, feature generation, model development and webserver implementation.

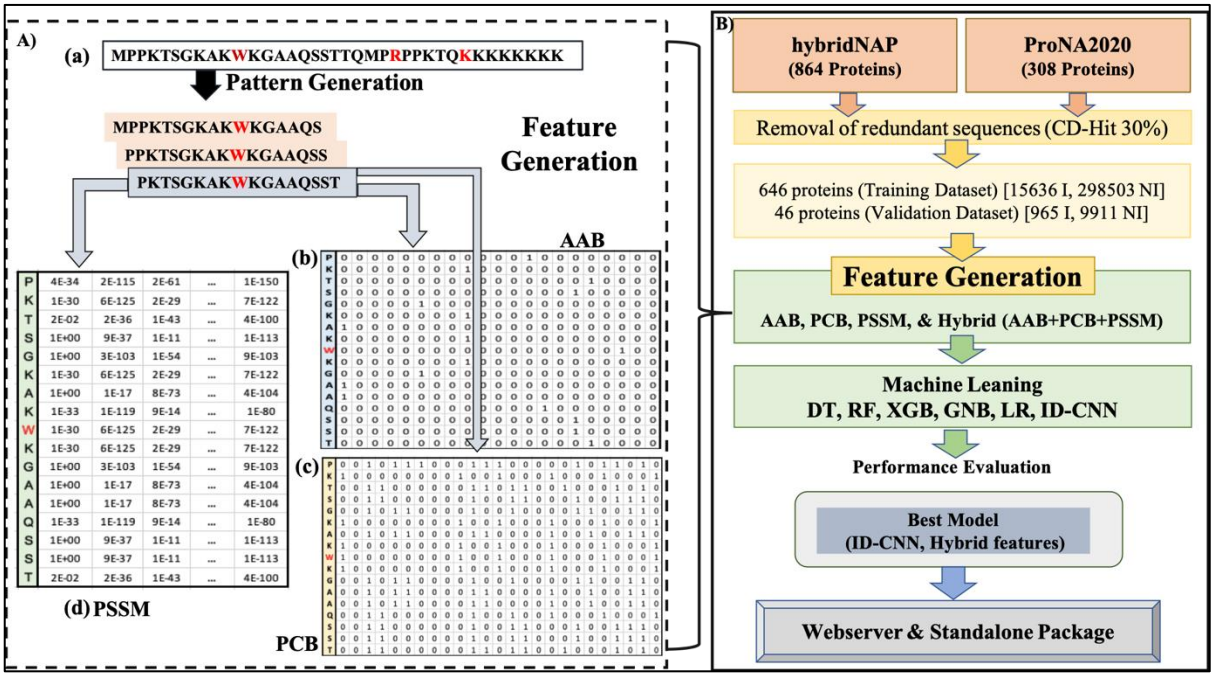


Figure 6.1: A Comprehensive workflow for feature generation(A) and model development(B). Following steps were taken to generate of different profiles from sequence; a) generation of fixed length patterns from a sequence, b) binary profile from

pattern, c) generation of physicochemical properties profile and d) PSSM profile. Overall algorithm for predicting DNA binding residues is shown in Figure 6.1B. (adopted from Patiyal, Dhall, & Raghava, 2021)

6.2.2 Training and testing dataset

The dataset for this study was obtained from two different methods, hybridNAP (Jian Zhang et al., 2019) and ProNA2020 (Qiu et al., 2020), which were trained and evaluated on 864 (817 in training and 47 in the independent dataset) and 308 (199 in training and 109 in the independent dataset) proteins, respectively. We have implemented the CD-HIT (Huang et al., 2010) software with 30% sequence identity criteria to form the non-redundant datasets. CD-HIT results in the 646 proteins in the training dataset and 46 protein sequences in the independent dataset, with no sequences in the training datasets having more than 30% sequence identity with sequences in the validation dataset. The final dataset comprises 15636 DNA-interacting and 298503 non-interacting residues in the training dataset, whereas the independent dataset contains 965 DNA-interacting and 9911 non-interacting residues.

6.2.3 Generation of patterns

Using the protein sequences, overlapping patterns of length 17 were generated, where the central residue, i.e., 9^{th} residue represents the entire pattern. If the central residue is DNA-interacting then the pattern is assigned as DNA-interacting, otherwise non-interacting. To tackle the residues in both ends, eight dummy variable "X" is added on both ends so that all residues can be in the middle of each pattern.

6.2.4 Patterns profile

We have calculated two types of profiles, i.e., one-hot encoding and physicochemical properties profile, to represent the pattern. These profiles were calculated using the binary profile module of Pfeature (Pande et al., 2019). The term one-hot encoding and binary profile is used interchangeably in literature due to the different terminology in different fields, such as, one-hot encoding is used in the field of computer science, whereas, researchers in the field of bioinformatics use binary profile term to represent the same feature. In the binary profile, each amino acid is represented as a vector of size 21, where the elements in the vector are either 1 or 0. 1 exhibits the presence of amino acid, whereas 0 represents the absence. For instance,

amino acid 'A' is represented as '1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0' where first 20 elements corresponds to the natural amino acids and the last element corresponds to dummy variable 'X', whereas 'X' is denoted as '0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1'. Hence, each pattern resulted in a vector size of 357 (17*21). On the other hand, in the physicochemical properties profile, each amino acid is represented by the vector size of 25, where each position in the vector is responsible for a particular property; for example, residue 'A' is represented as '0,0,1,0,1,1,0,0,0,0,1,1,0,0,0,0,1,0,0,1,0,0,1,1,0' and 'X' is denoted by the zero vector of length 25. Thus, the final vector size for pattern size 17 is 425 (17*25).

6.2.5 Evolutionary information

Besides the aforementioned profiles, we have also tried to capture the evolutionary information in terms of PSSM (Position-Specific Scoring Matrix) profile (K. Chen et al., 2012) for each sequence. The PSSM profile was calculated using PSI-BLAST (Altschul et al., 1997) by searching against the SwissProt database (Bairoch & Apweiler, 2000). To run the PSI-BLAST, the number of iterations was set to three with an e-value of 0.001. Further, the profile was normalized between 0 and 1, using equation 1. Finally, the normalized PSSM profile has the dimension of N X 21, where N is the length of the sequence. Further, the vector against each amino acid is concatenated that comes in the pattern of size K; hence resulting vector size for each pattern would be 357 (17*21), as value of K here is 17.

$$\boldsymbol{Norm_{PSSM}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{x}}} \quad [1]$$

Where, x is the PSSM score and $Norm_{PSSM}$ is the normalized PSSM value.

6.2.6 Machine learning classifiers

To develop the prediction models, we have implemented various traditional classifiers using scikit-learn (Pedregosa et al., 2011) and one deep-learning classifier (One-dimensional convolutional neural network (1D-CNN)) using TensorFlow. The traditional machine learning classifiers included Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), eXtreme Gradient Boosting (XGB), and Gaussian Naïve Bayes (GNB). The parameters were hyper-tuned using grid-search algorithm. Internal validation was implemented using five-fold cross validation to avoid overfitting and biasness.

6.2.7 Model architecture for 1D-CNN

The detailed architecture of 1D-CNN implemented in this study is represented in Figure 6.2. In this work, we used typical CNN architecture to build prediction models. It was created with the Python package Keras, which is based on TensorFlow. Each branch has four convolutional layers with the first layer using 256 filters. This implies that the input characteristics are represented by 256 filters in the first layer, and these features are decreased to half in each layer. The final or fourth layer will have 32 characteristics. Finally, we concatenated and flattened all of these vectors to create a feature vector. Instead of transmitting the complete vector directly for classification, we sent it via the densely connected neural network layers to capture the relevance of each feature for the classification job. Because of its simplicity and efficacy, we employed the ReLU activation function for each hidden layer. In the last layer, we utilised the sigmoid function to generate values between 0 and 1, which were then used to determine the ideal threshold that provides a good mix of sensitivity and specificity.

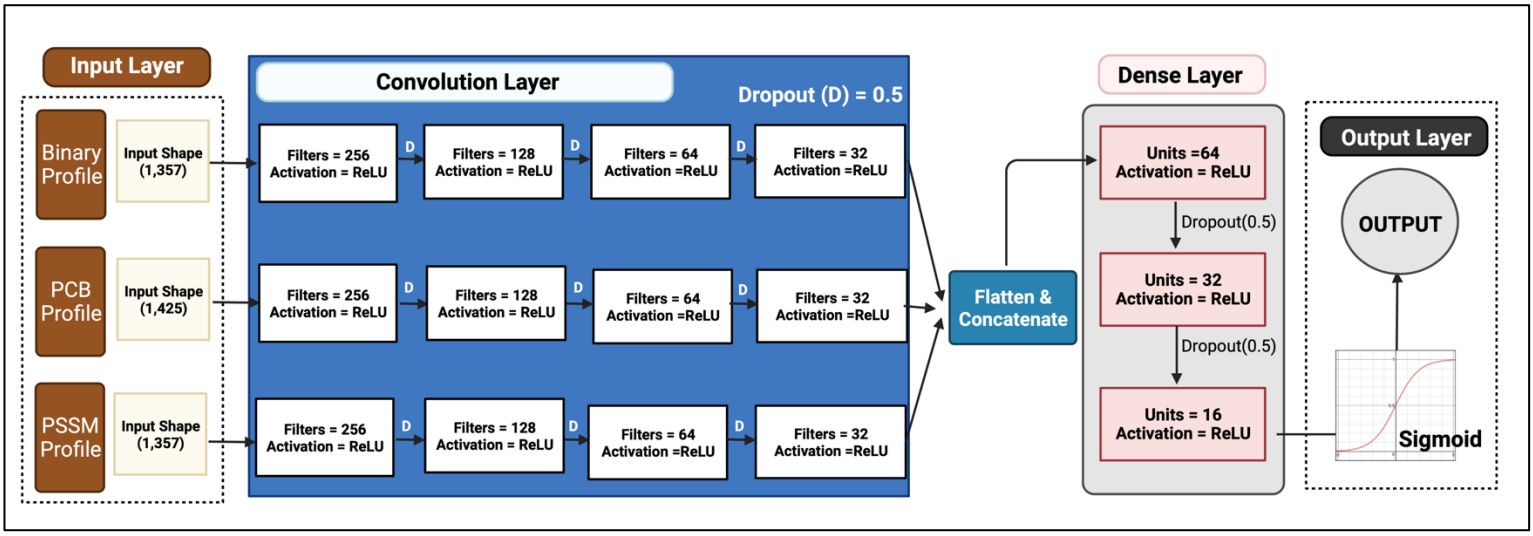


Figure 6.2: Detailed architecture of 1D-CNN implemented in this study

6.2.8 Measures to evaluate performance

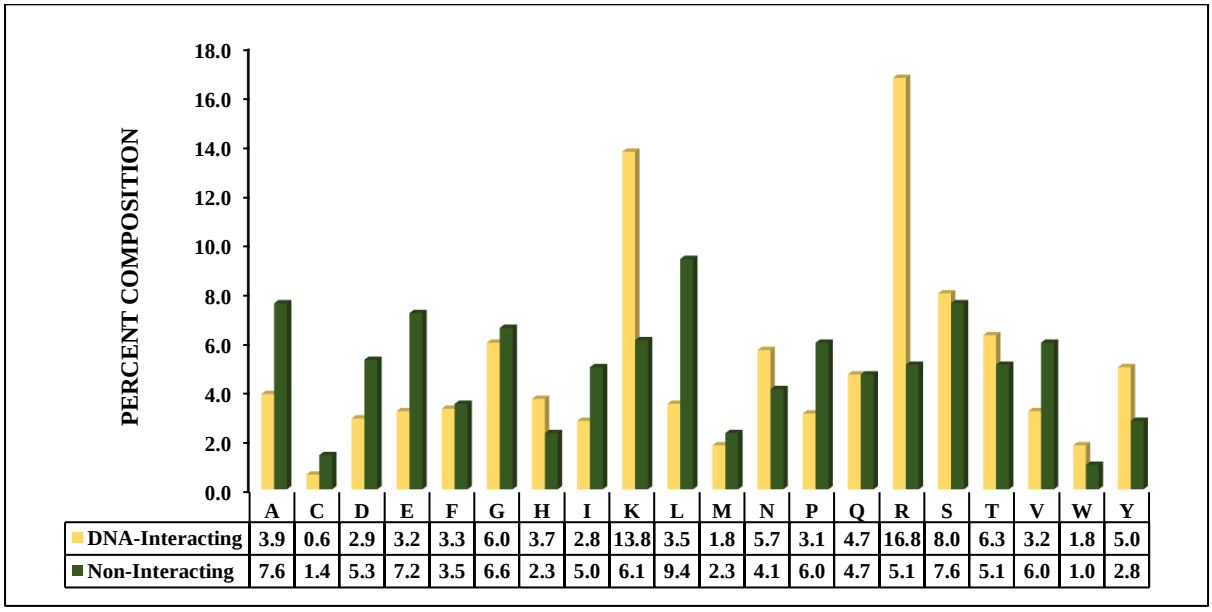
To compare and evaluate the performance of the generated models, we have calculated various measures, which can be broadly classified into threshold-dependent and threshold-independent. In threshold-dependent measures, we have calculated Sensitivity, Specificity, Accuracy, and Matthews Correlation Coefficient (MCC); on the other hand, in threshold-

independent measures, we have considered Area-Under the Receiver Operating Curve (AUROC), which signifies the relation between the True Positive Rate (TPR) and False-Positive Rate (FPR). The pROC package of R (Sachs, 2017) was implemented to calculate and plot the AUROC. The equation for threshold-dependent parameters is provided in equations 1-5 of section 4.2.5 of Chapter 4.

6.3 Results

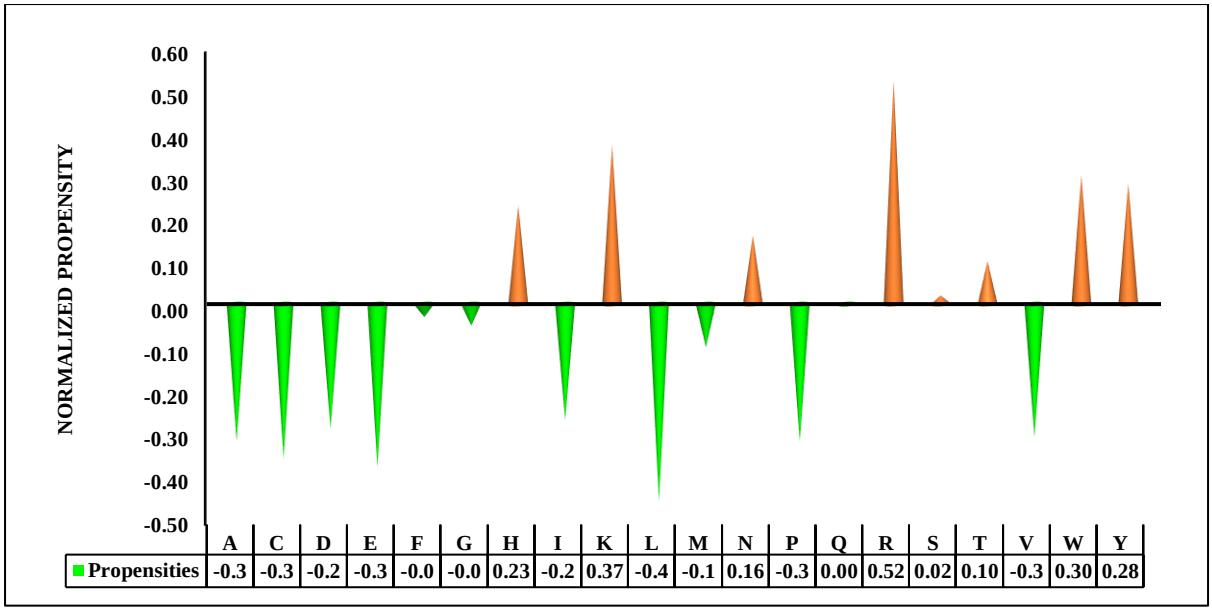
6.3.1 Preliminary analysis

In the primary analysis, we have performed three different analysis, such as, amino acid composition based analysis, propensity analysis, and physicochemical properties based analysis. Composition analysis revealed that DNA-interacting residues are abundant in H, R, K, N, and Y as shown in Figure 6.3. Similar trend was shown in propensity analysis, which showed that residues R, K, W, and Y are most preferred in the DNA-interacting sites as shown in Figure 6.4. Figure 6.5 depicts the composition of physicochemical properties of residues involved in the DNA-interaction and it signifies that residues with properties like positive charge, basic, hydrophilic, helix secondary structure, and large side chain are abundant in the DNA-interaction sites.



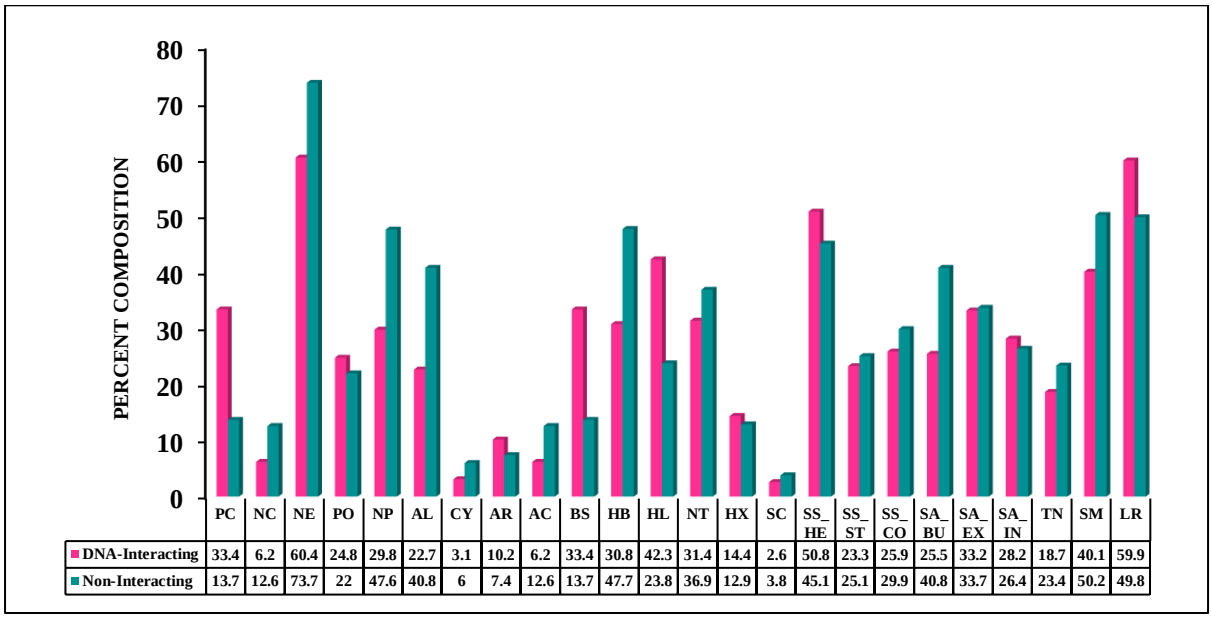
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
DNA-Interacting	3.9	0.6	2.9	3.2	3.3	6.0	3.7	2.8	13.8	3.5	1.8	5.7	3.1	4.7	16.8	8.0	6.3	3.2	1.8	5.0
Non-Interacting	7.6	1.4	5.3	7.2	3.5	6.6	2.3	5.0	6.1	9.4	2.3	4.1	6.0	4.7	5.1	7.6	5.1	6.0	1.0	2.8

Figure 6.3: Compositional analysis of DNA-interacting residues



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Propensities	-0.3	-0.3	-0.2	-0.3	-0.0	-0.0	0.23	-0.2	0.37	-0.4	-0.1	0.16	-0.3	0.00	0.52	0.02	0.10	-0.3	0.30	0.28

Figure 6.4: Propensity-based analysis of DNA-interacting residues



	PC	NC	NE	PO	NP	AL	CY	AR	AC	BS	HB	HL	NT	HX	SC	SS_HE	SS_ST	SS_CO	SA_BU	SA_EX	SA_IN	TN	SM	LR
DNA-Interacting	33.4	6.2	60.4	24.8	29.8	22.7	3.1	10.2	6.2	33.4	30.8	42.3	31.4	14.4	2.6	50.8	23.3	25.9	25.5	33.2	28.2	18.7	40.1	59.9
Non-Interacting	13.7	12.6	73.7	22	47.6	40.8	6	7.4	12.6	13.7	47.7	23.8	36.9	12.9	3.8	45.1	25.1	29.9	40.8	33.7	26.4	23.4	50.2	49.8

Figure 6.5: Physico-chemical properties-based analysis of DNA-interacting residues

6.3.2 Models based on one-hot encoding profile

We have developed various models based on traditional machine learning classifier and one-deep-learning model using one-hot encoding profile as the input features. The performance measures of the models developed on each classifier is reported in Table 6.1. As shown in the

table, LR-based and 1D-CNN based model achieved the same AUROC of 0.74 on the independent dataset.

Table 6.1: Performance measures of various models developed using one-hot encoding on independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	12.62	92.81	85.61	0.53	0.06
RF	67.05	65.29	65.45	0.72	0.19
LR	68.19	66.59	66.73	0.74	0.21
XGB	67.15	68.17	68.08	0.73	0.21
GNB	66.22	63.19	63.46	0.70	0.17
1D-CNN	70.67	66.54	66.00	0.74	0.21

6.3.3 Physicochemical properties profile based models

Similarly, the models were developed using physicochemical properties as the input features and their performance is reported in Table 6.2. Similar trend as per the binary profile was observed here, LR and 1D-CNN based models outperformed the other classifiers with AUROC of 0.73 on the independent dataset.

Table 6.2: Performance measures of various models developed using physicochemical properties profile on independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	09.32	94.82	87.24	0.52	0.05
RF	63.11	63.67	63.62	0.69	0.16
LR	68.39	66.50	66.67	0.73	0.21
XGB	63.32	68.98	68.48	0.72	0.19
GNB	67.46	58.87	59.64	0.68	0.15
1D-CNN	67.08	67.86	67.79	0.73	0.20

6.3.4 Models based on evolutionary information

Further, the models were developed using evolutionary (PSSM) profile as the input features and their performance is reported in Table 6.3. XGB-based model performed best among the other classifiers with AUROC of 0.77 on the independent dataset.

Table 6.3: Performance measures of various models developed using evolutionary information on independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	13.26	94.47	87.27	0.54	0.09
RF	73.06	62.46	63.41	0.74	0.21
LR	69.33	67.98	68.10	0.75	0.22
XGB	72.12	67.51	67.92	0.77	0.24
GNB	64.87	56.24	57.91	0.63	0.12
1D-CNN	64.89	69.97	69.43	0.74	0.21

6.3.5 Models based on combined profile

Finally, the features were combined by concatenating all the profiles, i.e. binary-, physicochemical properties-, and PSSM-profile, and used as the input feature. The performance of each model was calculated and reported in Table 6.4. 1D-CNN based model was able to outperforms all the other classifiers with AUROC of 0.79 on the independent dataset.

Table 6.4: Performance measures of various models developed using combined profile on independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

Classifier	Sensitivity	Specificity	Accuracy	AUROC	MCC
DT	15.34	94.91	87.84	0.55	0.18
RF	70.98	63.02	63.73	0.75	0.20
LR	70.88	69.18	69.33	0.77	0.29
XGB	69.95	62.62	63.27	0.72	0.19
GNB	66.84	62.70	63.06	0.70	0.17

1D-CNN	70.78	78.40	77.72	0.79	0.32

6.3.6 Comparison with existing approaches

It is important to compare the newly developed approach to the old methods in order to concede it. The comparison shows the advantages and disadvantages of the newly discovered approach. Because there are several approaches for predicting DNA-binding residues in proteins (Qiu et al., 2020; Liangjiang Wang et al., 2010, 2009; Jian Zhang et al., 2019), a thorough comparison is required to comprehend the advantages of the newly created method "DBPred." We compare the performance of existing approaches and the proposed method using an independent dataset of 46 proteins utilised in this study to give an unbiased comparison. Table 6.5 summarises the results of all approaches in terms of sensitivity, specificity, AUROC, accuracy, and MCC. Among known methods, DRNAPred (J. Yan & Kurgan, 2017) produced the highest AUROC of 0.75 and MCC of 0.22, whereas SVMnuc, NucBind (Su et al., 2019), DNAPred (Zhu, Hu, Song, & Yu, 2019), DNAbindR (C. Yan et al., 2006), and ProNA2020 (Qiu et al., 2020) achieved comparable MCC and AUROC. On the independent dataset, our technique DBPred outperformed the existing methods with an AUROC of 0.79 and MCC of 0.32, as shown in Table 6.5. The ROC curve depicts a comparison of the AUROC of the available approaches (Figure 6.6). We are unable to compare our solutions with a few others since they are either non-functional or lack webserver/standalone software.

Table 6.5: The performance of existing methods and proposed method on the independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

Method	Year	Sensitivity	Specificity	AUROC	Accuracy	MCC
DNABindR	2006	52.16	78.09	0.71	75.80%	0.20
DP-Bind	2007	47.41	71.14	0.56	69.06%	0.11
DRNAPred	2017	67.67	69.19	0.75	69.06%	0.22
TargetDNA	2017	48.71	77.52	0.69	74.98%	0.17
HybridNAP	2017	38.79	79.58	0.66	75.99%	0.13
funDNApred	2018	62.93	63.70	0.69	63.70%	0.16
DNAPred	2019	67.10	65.50	0.73	65.64%	0.19
SVMnuc	2019	66.81	66.57	0.72	66.59%	0.20
NucBind	2019	62.50	64.86	0.72	64.66%	0.16

iProDNA-CapsNet	2019	63.79	61.28	0.68	61.28%	0.14
ProNA2020[#]	2020	42.22	76.28	0.72	74.31%	0.22
NCBRPred[#]	2021	67.67	67.44	0.71	67.46%	0.21
DBPred[#]	2022	70.78	78.40	0.79	77.72%	0.32

[#]**Standalones are also available**

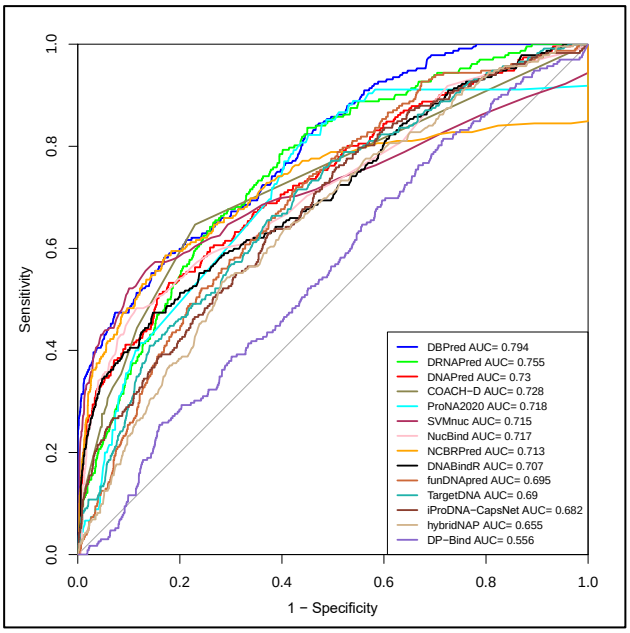


Figure 6.6: AUROC plots obtained for existing methods using independent dataset (adopted from Patiyal, Dhall, & Raghava, 2021)

6.4 Web-based services

DBPred is a freely available online platform to determine the DNA interacting residues in a protein using the best performing models developed on different features like binary profile, physico-chemical profile, PSSM profile, and their combination called as 'hybrid feature'. In this method we have considered pattern length 17, i.e. we generate overlapping patterns of length 17 from the submitting sequences and use them to make predictions. This user-friendly web-server was developed using PHP, HTML, Python, Perl, and JavaScript.

Several handy tools are incorporated in the web-server of DBPred to facilitate the users to determine the DNA interacting residues using sequence information. It includes four major modules "Sequence", "PSSM Profile", "Hybrid", and "Standalone". The basic sequence module include a approach which accepts multiple protein sequences in the FASTA format and predicts DNA interacting residues in each sequence. In this module, amino acid binary

profile and physico-chemical properties based profile is calculated as the input feature for making the predictions. Similarly, in “PSSM profile” module, PSSM profile is calculated using PSI-BLAST by hitting the query sequences against the Swiss-Prot database, and use the same as the input feature to predict the DNA interacting residues in a protein sequences. “Hybrid” module calculates all the three profiles such as binary, physico-chemical properties, and PSSM profile and provide it the 1D-CNN model implemented in the backend. Users are allowed to provide their email, so that for the long processes they do not have to wait and the results will be send to their provided email IDs, once the process gets completed. Figure 6.7 is a screenshot of the “Hybrid” module of the web-server displaying the submission form for the submission of the query sequences in the FASTA format. Figure 6.8 is an example output page obtained after the submission of the query sequences to the “Hybrid” module for making predictions for the DNA interacting residues in the protein sequences. The output page exhibits the sequences in which DNA interacting residues are highlighted in red color with bigger font size, whereas non-interacting residues are shown in black color. The result page is downloadable in the .txt, .pdf, and .png format. Additionally, to predict the DNA interacting residues in a protein in the absence of the internet or for prediction in a large dataset, such as the complete human proteome, since server will take a long time to produce the results, we have developed the Python- and Perl-based standalone.

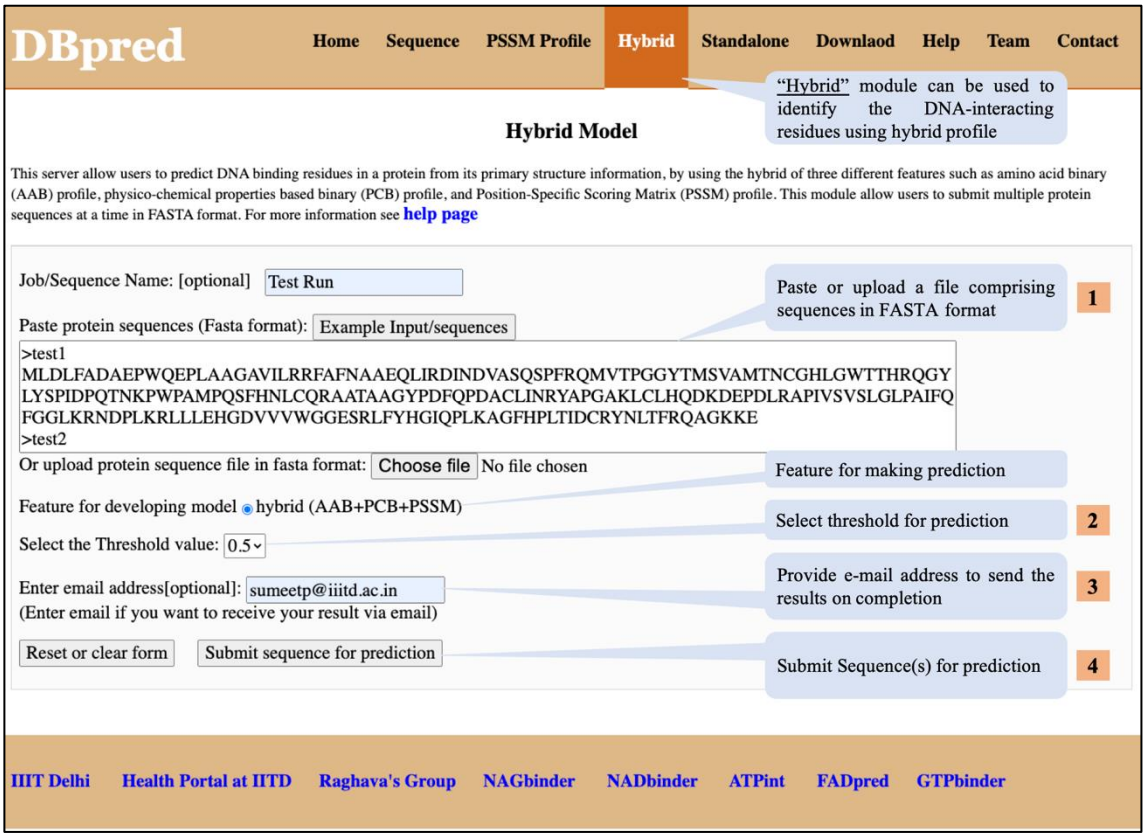


Figure 6.7: Screenshot of "Hybrid" module of DBPred web-server
(URL https://webs.iiitd.edu.in/raghava/dbpred/hybrid.php)

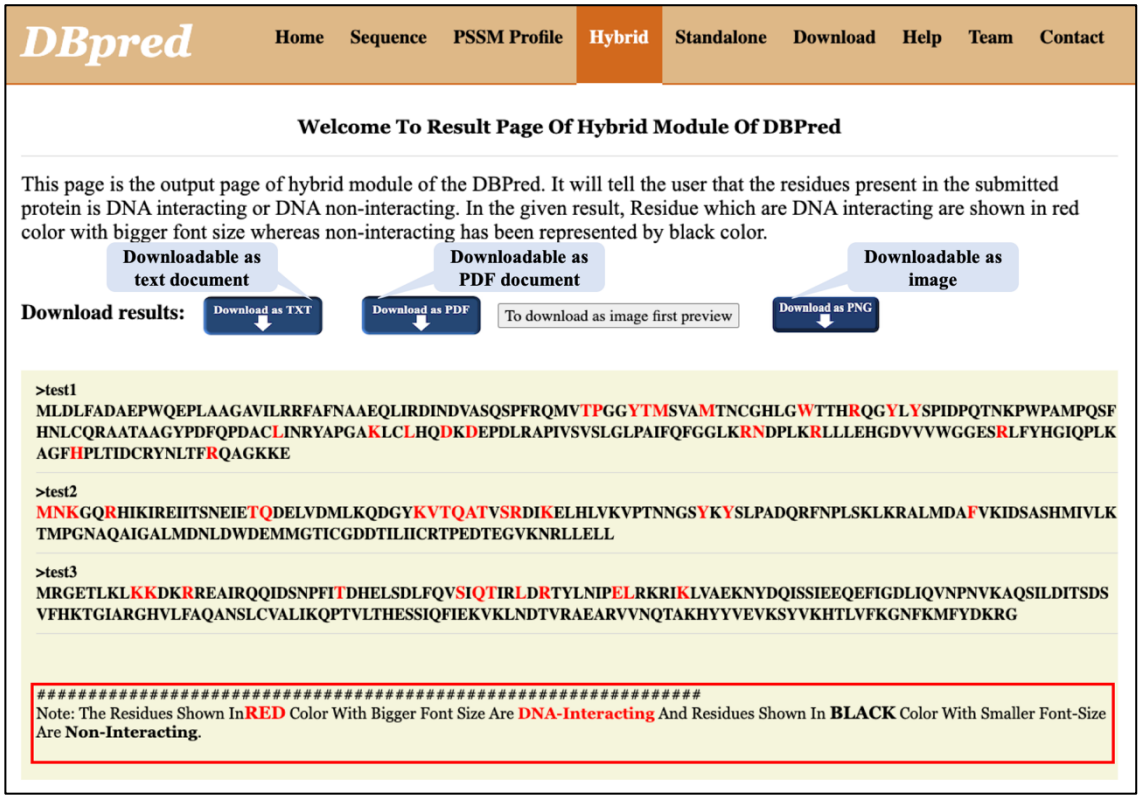


Figure 6.8: Screenshot of the result page of "Hybrid" module of DBPred webserver
(URL https://webs.iiitd.edu.in/raghava/dbpred/hybrid.php)

6.5 Discussion and Conclusion

The interactions between the biological molecules is a very crucial phenomenon, which is responsible for the well-being of an organism. The interaction between the DNA and protein is one of the most important factor which in turn regulates the crucial biological processes such as transcription, replication, and expression of genes (Ofran, Mysore, & Rost, 2007). Therefore, better knowledge of these interaction is crucial for processes like drug-designing, where these interactions can be exploited in order to design novel therapeutics for diseases (Csermely, Korcsmaros, Kiss, London, & Nussinov, 2013; Hopkins, 2008). It should be noted that the three-dimensional information of the protein is the sole way to extract the interacting residues. The PDB database provides a large number of protein structures that have been experimentally validated and were discovered using NMR and X-ray crystallography (Berman et al., 2002; Burley et al., 2021). According to research, 3D knowledge of protein binding residues is useful in structure-based drug design because it allows one to understand the interaction of drug molecules with DNA-interacting residues (Anderson, 2003; Goodwin, Long, & Georgiadis, 2005; Moravek, Neidle, & Schneider, 2002; Pradhan, Das, & Mattaparthi, 2018). As a result, during the last several decades, a lot of academicians have worked tirelessly to comprehend the physical interaction between DNA and protein molecules. There are ample of computational approaches have been designed in the past which are able to predict the DNA-interacting sites based on different features such as structural motifs, functions, structural classes, DNA-type specific interaction etc. DNAgenie (Zhang et. al, 2021) is an accurate method for the prediction of DNA-specific binding residues in a protein sequence which predict the residues that interact with A-DNA, B-DNA and single-stranded DNA. Structural motif templates were used to detect the proteins with DNA binding function, such as, method proposed by Jones et al, in 2003, which achieved the accuracy of 88.4% to predict the DNA-binding helix–turn–helix (HTH) structural motifs (Jones et al, 2003). Another method proposed by Pellegrini-Calace et al., in 2005 which utilizes the sequence and structural information of the proteins to predict the DNA-binding (HTH) structural motifs (Pellegrini-Calace, M., & Thornton, J. M., 2005). In 2009, another method LCV was introduced to predict the DNA-binding (HTH) motif using sequence information as local combinatorial variables, that attained the accuracy of 93.29%. Researchers have also created a number of computational methods for predicting DNA-interacting sites on proteins, which may be divided into three major categories: sequence-based, structure-based, and hybrid approaches (Chowdhury et al., 2017; A. Mishra, Pokhrel, & Hoque, 2019). However, the fundamental disadvantage of structure-

based or hybrid techniques is their reliance on protein structural information, which limits their use because protein structure identification is an expensive, time-consuming, and difficult procedure (Patiyal et al., 2020b). On the other hand, the amount of sequence information in various databases is increasing dramatically, strengthening the use of sequence-based approaches with consistent performance. A variety of computer approaches for predicting DNA-interacting residues have been developed in recent years. However, most studies employed small datasets, and performance on independent datasets is poor. As a result, it is important to create a new approach for predicting DNA-interacting residues utilising protein sequences using the largest dataset available. This study was aimed to develop a computational method for identifying the DNA-interacting residues using the sequence information. We have used the latest benchmark dataset of ProNA2020 and hybridNAP. We preprocessed the dataset by applying CD-HIT software at 40% criteria to create the non-redundant dataset. Finally, we left with 646 proteins (15636 DNA-interacting and 298503 non-interacting residues) in the training dataset, and 46 proteins (965 interacting and 9911 non-interacting residues) in the independent dataset. We have performed various analyses such as composition, propensity, and physicochemical properties based analysis, to understand the nature of DNA-interacting residues. It was found that certain residues like lysine, arginine, and tyrosine are more frequent in DNA-interaction. Most of DNA-interacting residues possess positive charged, basic, hydrophilic residues and helix secondary structure properties. Finally, we have developed various machine learning models using different kinds of input features like binary, physicochemical properties, PSSM profile and their combination which was referred as hybrid features. Model developed by implementing 1D-CNN based classifier using all the three profiles outperformed all the other classifiers and features types with AUROC of 0.79 on the independent dataset. Further, we compared the performance of our method with the existing methods using independent dataset, and found out that our method has outperformed all the existing approaches. We built a user-friendly prediction web server that incorporates various modules to predict the DNA-interacting residues in a protein sequences to support the scientific community. The study's prediction models are implemented in the backend of the web server named as “DBPred” available at https://webs.iiitd.edu.in/raghava/dbpred. Sequence, PSSM Profile, Hybrid, Standalone, and Download are the five major modules in the “DBPred”.In the guise of docker technology, we've also offered a stand-alone facility. This standalone is included in our GPSRDocker package (Aggarwal et al., 2019), which may be obtained from https://webs.iiitd.edu.in/gpsrdocker.

Chapter 7

Determination of RNA-binding sites in a protein

7.1 Introduction

Being a part of many biological machineries like the ribosome and spliceosome, RNA is one of the essential components of the living cell. Hence, it is involved in various biological functions (S Jones, Daley, Luscombe, Berman, & Thornton, 2001). The interaction between RNA and protein is crucial for many biological processes, such as regulating gene expression, viral assembly and replication, post-translation modification, synthesis of protein, and splicing (Gangloff et al., 2000; B. Lin & Pang, 2019; Pattnaik et al., 2018; Payne et al., 2018; Standart & Jackson, 1994; Turner & Diaz-Munoz, 2018). Besides, RNA has also shown the involvement in the development of cancer and neurological diseases such as ALS (Amyotrophic Lateral Sclerosis) and Alzheimer's (Carey & Wickramasinghe, 2018; Idda et al., 2018; Kwiatkowski et al., 2009; Tsai et al., 2011; M. Zhou et al., 2019). Other than that, the interaction between RNA and protein is responsible for cellular homeostasis, and disturbance in these interactions may result in abnormal cellular processes and diseases (Allerson, Cazzola, & Rouault, 1999; Batista & Chang, 2013; Khalil & Rinn, 2011; Ramanathan, Porter, & Khavari, 2019). To understand and exploit the interaction between RNA and proteins, it is imperative to identify the residues that take part in the interaction. One can develop an RNA-based treatment to treat or diagnose various RNA-related illnesses with a better understanding of the RNA-protein interaction residues.

With the refinement in the experimental methods like nuclear magnetic resonance (NMR) and X-ray crystallography, numerous structures of protein-RNA complex have been discovered and documented in Protein Data Bank (Berman et al., 2002). However, these experimental techniques are highly time-consuming and expensive. Determination of RNA-interacting residues holds great importance, but at the same time, it is highly complex. On the contrary, the likely identification of RNA-binding residues is achievable and cost-effective by computer-based methods using sequence data. Various computational approaches, i.e., sequence-based and structure-based (Luo, Liu, Venkateswaran, Song, & Zhou, 2017; Pan, Rijnbeek, Yan, & Shen, 2018; Poursheikhali Asghari & Abdolmaleki, 2019; Sanchez de Groot et al., 2019; Zhao, Yang, & Zhou, 2011), have been created during the past several years to predict the RNA-interacting residues. However, the limitation of the structure-based methods is the dependency on the availability of the RNA-protein complex structure; on the other hand, the increase of sequences in the respective databases is exponential due to the advancement in sequencing technology. Therefore, myriads of computational methods based on sequence information have been developed in the last few decades, such as PPRInt (M. Kumar, Gromiha, et al., 2008),

PredPRBA (Deng, Yang, & Liu, 2019), RPiRLS (Shen, Cui, Chen, Zhang, & Xu, 2018), hybridNAP (Jian Zhang et al., 2019), and ProNA2020 (Qiu et al., 2020).

Despite the availability of numerous computational-based approaches, there is still some room for improvement in the performance. Therefore, to complement the existing methods, we have developed the update of our older version of PPRInt (M. Kumar, Gromiha, et al., 2008), named "PPRInt2". PPRInt2 is a methodical attempt to predict the RNA-interacting residues in a protein sequence using an evolutionary profile and 1D-CNN. To aid the scientific community, we have provided our approach in the form of a web server at https://webs.iiitd.edu.in/raghava/pprint2/ and a Python-based standalone available at https://webs.iiitd.edu.in/raghava/pprint2/stand.php. We have also provided the source code and other files on GitHub (https://github.com/raghavagps/pprint2).

7.2 Materials and Methods

7.2.1 Overall architecture

Figure 7.1 depicts the overall workflow that we have adapted for this study.

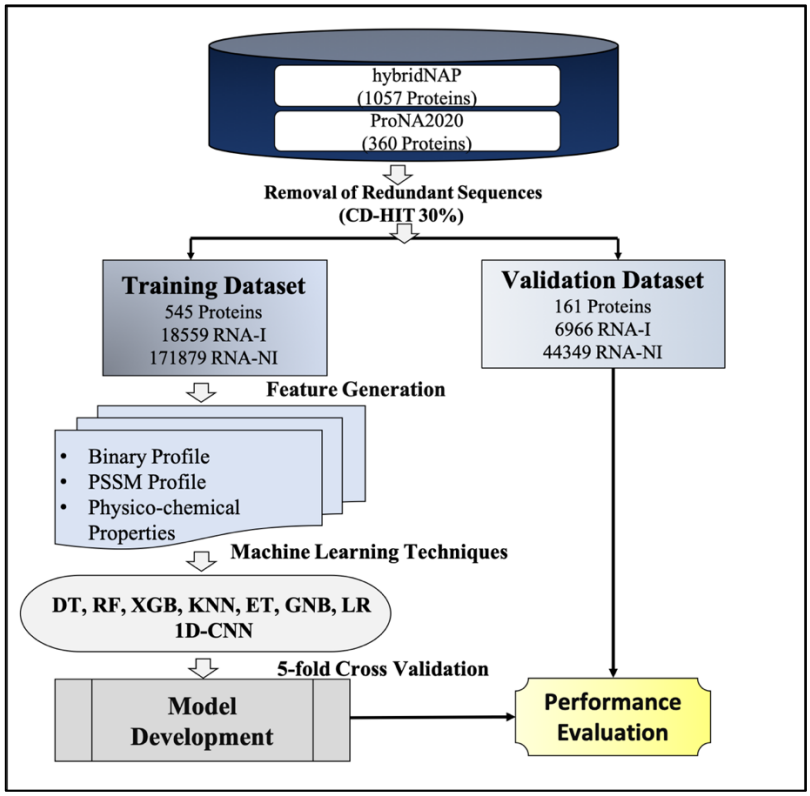


Figure 7.1: Overall architecture of the present study

7.2.2 Dataset collection

We have obtained the annotated sequences from two studies, hybridNAP (Jian Zhang et al., 2019) and ProNA2020 (Qiu et al., 2020), which comprise 1057 and 360 protein sequences, respectively. To handle the redundancy issue, we have applied CD-HIT software (Huang et al., 2010) with the 30% sequence identity criteria, which means no two sequences in two different datasets will share more than 30% sequence identity. Further, we left with 545 protein sequences in the training dataset and 161 protein sequences in the independent dataset. Further, overlapping patterns of window size 17 were generated for each sequence, where the labels are assigned to each pattern based on their central residues, i.e., if the central residue is annotated as RNA-interacting, then the pattern is assigned as RNA-interacting, otherwise non-interacting. In order to handle the terminal residues, we have added the eight dummy 'X' residues on both ends. Finally, the training dataset comprises 18559 RNA-interacting and 171879 non-interacting residues/patterns, whereas the independent dataset contains 6966 RNA-interacting and 44349 non-interacting residues.

7.2.3 Generation of features

In this study, we have calculated three types of features: binary profile, physicochemical properties profile, and PSSM profile using Pfeature (Pande et al., 2019). In binary profile, each amino acid is represented with a vector of length 21, having elements as 1 or 0, where 1 signifies the presence and 0 exhibits the absence. The first twenty positions in the vector represent the 20 natural amino acids, and the last element presents the dummy variable 'X'. Similarly, in the physicochemical properties profile, each amino acid is represented by the vector of size 25, where each position is responsible for a particular property, and the zero vector of length 25 represents 'X'. Moreover, the PSSM profile is generated using the PSI-BLAST (Altschul et al., 1997) module, where the sequences were hit against the non-redundant database Swiss-Prot (Bairoch & Apweiler, 2000) with the number of iterations equal to three and e-value 1e-3. Further, the PSSM profile was normalized between 0 and 1, and patterns were generated as per the respective sequences.

7.2.4 Building of prediction model

We have implemented seven traditional machine learning classifiers and one deep-learning classifier to achieve the classification of RNA-interacting residues using sequence information.

Traditional machine learning classifiers included Decision Tree (DT), Random Forest (RF), Support Vector Classifier (SVC), eXtreme Gradient Boosting (XGB), Gaussian Naive Bayes (GNB), Logistic regression (LR), and K-Nearest Neighbour (KNN). In contrast, we have implemented One-Dimensional Convolutional Neural Network (1D-CNN) as the deep-learning classifier. We have performed the internal validation on the training dataset using five-fold cross-validation to avoid overfitting and biasness. The parameters for each classifier were hyper-tuned using the grid-search module in python.

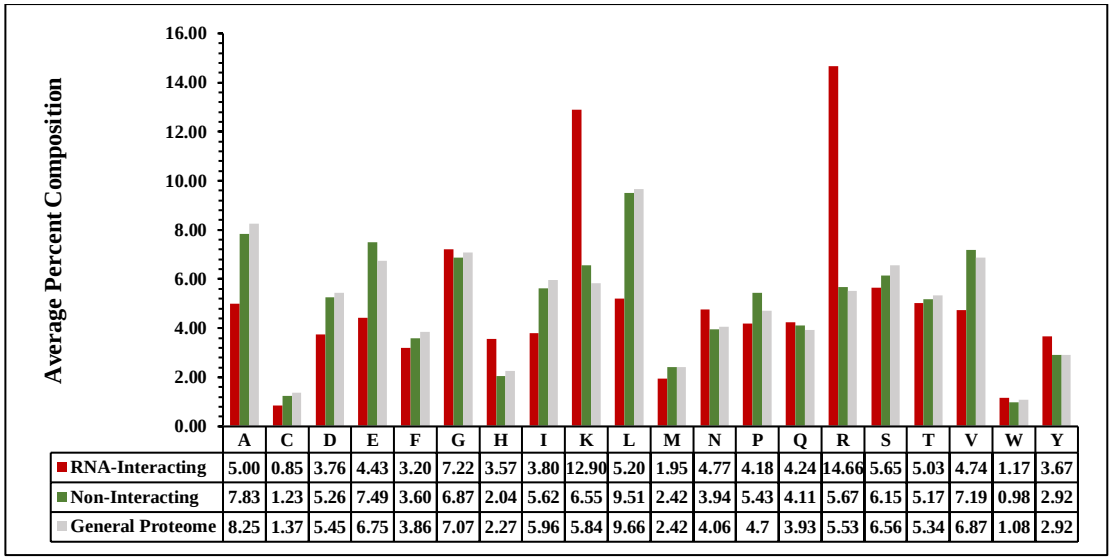
7.2.5 Evaluation measures

In order to evaluate and compare the performance of the generated models, we have used threshold-dependent and threshold-independent parameters. We have measured the performance of each model in terms of Sensitivity, Specificity, Accuracy, F1-Score, and Matthews Correlation Coefficient (MCC) as threshold-dependent measures and Area Under the Receiver Operating Characteristics (AUROC) curve as the threshold-independent parameter. The equation for threshold-dependent parameters is provided in equations 1-5 of section 4.2.5 of Chapter 4.

7.3 Results

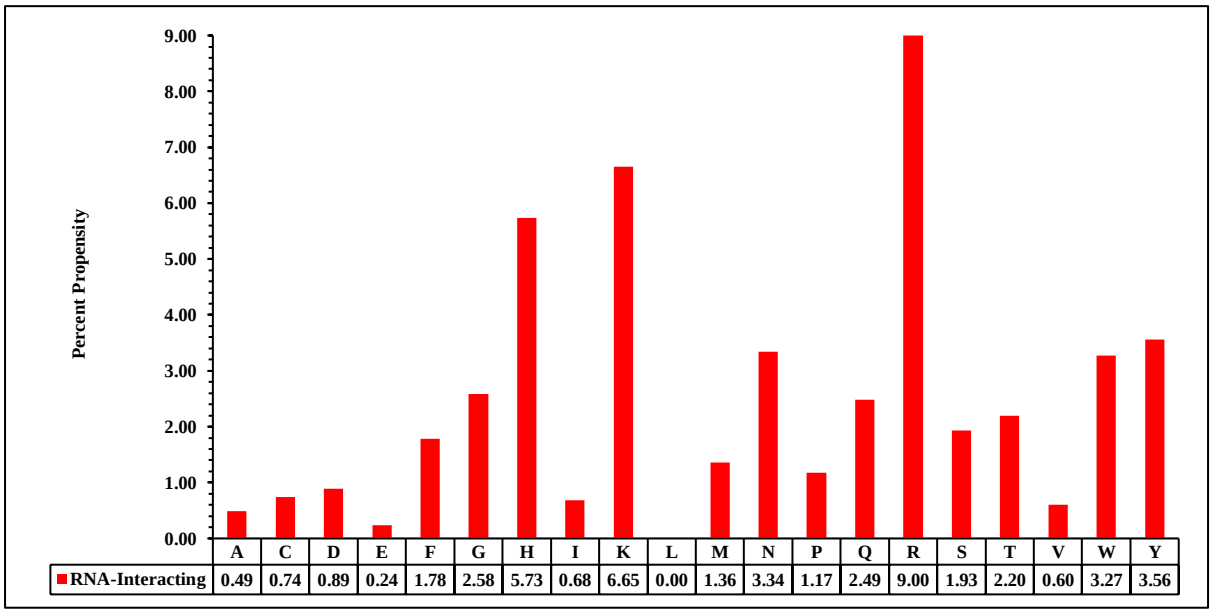
7.3.1 Preliminary analysis

In the preliminary analysis, we have performed the compositional analysis and physicochemical properties-based composition analysis, to understand the nature of the amino acid abundant in RNA-interacting sites. Figure 7.2 represents the percent amino acid composition of RNA-interacting, non-interacting, and general proteome, which signifies that positively charged residues H, R, and K are more abundant in the RNA-interacting sites. The similar trend was seen for the propensity analysis for RNA-interacting residues. Where, physicochemical properties based analysis showed that RNA-interacting residues are rich in positively charged, hydrophilic, and basic residues as shown in Figure 7.4.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
RNA-Interacting	5.00	0.85	3.76	4.43	3.20	7.22	3.57	3.80	12.90	5.20	1.95	4.77	4.18	4.24	14.66	5.65	5.03	4.74	1.17	3.67
Non-Interacting	7.83	1.23	5.26	7.49	3.60	6.87	2.04	5.62	6.55	9.51	2.42	3.94	5.43	4.11	5.67	6.15	5.17	7.19	0.98	2.92
General Proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.7	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 7.2: Composition of amino-acid residues in RNA-interacting, non-interacting, and general proteome



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
RNA-Interacting	0.49	0.74	0.89	0.24	1.78	2.58	5.73	0.68	6.65	0.00	1.36	3.34	1.17	2.49	9.00	1.93	2.20	0.60	3.27	3.56

Figure 7.3: Propensity analysis for RNA-interacting, non-interacting residues

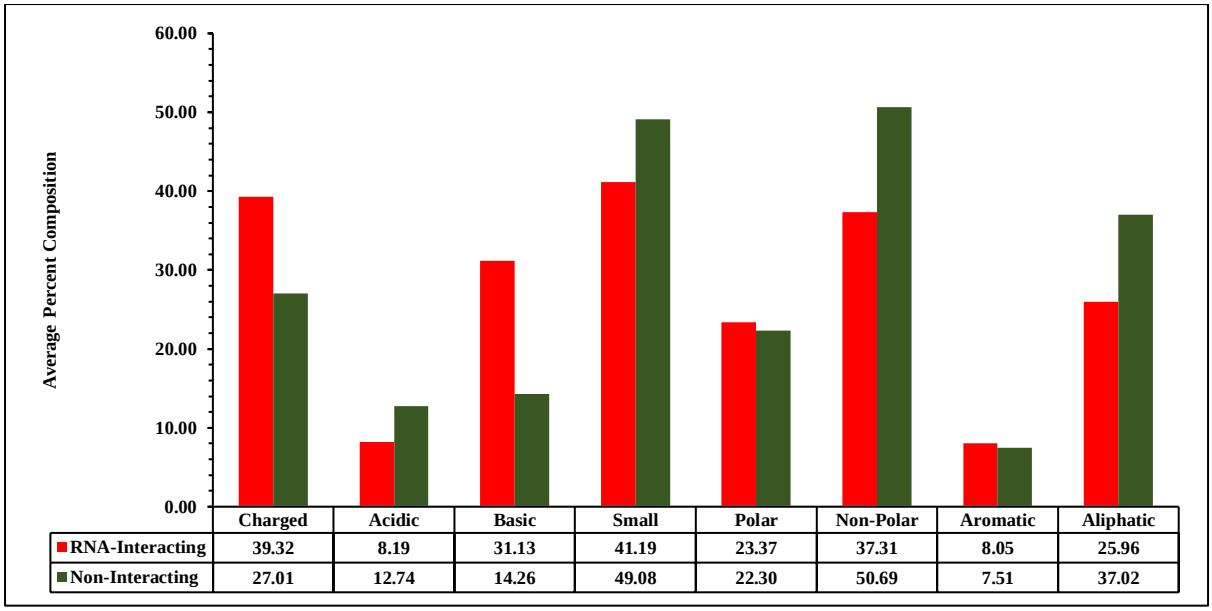


Figure 7.4: Physico-chemical properties based composition of RNA-interacting, non-interacting residues

7.3.2 Performance using position-based profile

We have implemented the position-based profile or binary profile as the input feature to develop the prediction models using various classifiers. Table 7.1 provides the performance measures for each classifier on training and independent dataset. The 1D-CNN-based model achieved the highest AUROC of 0.81 on the training dataset and 0.68 on the independent dataset. Whereas XGB and LR-based models also achieved the AUROC of 0.68 on the independent dataset.

Table 7.1: The performance binary profile based on models developed using different classifiers (adopted from Patiyal, Dhall, Bajaj, Sahu, & Raghava, 2022)

Classifier	Training Dataset							Validation Dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	19.14	88.93	82.13	0.54	0.17	0.07	0.07	15.94	89.07	79.14	0.53	0.17	0.05	0.05
RF	66.57	67.53	67.43	0.73	0.29	0.16	0.21	55.77	68.54	66.81	0.67	0.31	0.15	0.18
LR	67.20	68.16	68.07	0.74	0.29	0.16	0.22	55.79	70.37	68.39	0.68	0.32	0.16	0.19
XGB	66.88	67.54	67.48	0.74	0.29	0.16	0.21	55.86	69.65	67.78	0.68	0.32	0.16	0.19
KNN	60.32	63.75	63.41	0.64	0.24	0.10	0.15	52.60	64.66	63.02	0.60	0.28	0.10	0.12
GNB	66.63	65.14	65.28	0.71	0.27	0.14	0.19	56.30	67.11	65.64	0.65	0.31	0.14	0.17
ET	68.25	65.72	65.97	0.73	0.28	0.15	0.21	58.38	66.64	65.52	0.67	0.32	0.15	0.18
1D-CNN	73.91	73.76	73.78	0.81	0.36	0.24	0.31	50.66	74.63	71.38	0.68	0.33	0.17	0.19

7.3.3 Performance using physicochemical properties profile

Other than binary profile, we have also develop models on physicochemical properties profile which represents each pattern with the vector size of 425. Table 7.2 comprises the performance for each classifier and it exhibits that 1D-CNN based model attained the maximum AUROC 0.79 and 0.68 on the training and validation dataset, respectively. In terms of independent dataset, physicochemical properties profile and binary profile-based model achieved the same performance.

Table 7.2: Performance measures for models developed by implementing various classifiers using physicochemical properties profile as the input feature (adopted from Patiyal, Dhall, Bajaj, Sahu, & Raghava, 2022)

Classifier	Training Dataset							Validation Dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	16.15	91.03	83.73	0.54	0.16	0.07	0.07	14.00	91.46	80.95	0.53	0.17	0.06	0.06
RF	67.74	64.56	64.87	0.72	0.27	0.14	0.20	56.88	66.37	65.08	0.66	0.31	0.14	0.17
LR	67.30	68.19	68.10	0.74	0.29	0.16	0.22	55.70	70.48	68.47	0.68	0.32	0.16	0.19
XGB	65.41	68.60	68.29	0.73	0.29	0.16	0.21	54.22	70.50	68.29	0.68	0.32	0.16	0.18
KN	62.43	63.32	63.23	0.65	0.25	0.11	0.16	53.22	63.81	62.37	0.60	0.28	0.10	0.12
GNB	65.44	66.15	66.08	0.71	0.27	0.14	0.19	54.19	68.58	66.63	0.66	0.31	0.14	0.16
ET	65.81	67.25	67.11	0.72	0.28	0.15	0.20	54.65	68.74	66.83	0.66	0.31	0.14	0.17
1D-CNN	69.84	69.60	69.62	0.77	0.31	0.19	0.25	56.42	71.14	69.14	0.69	0.33	0.17	0.20

7.3.4 Performance using evolutionary profile

Further, we have used evolutionary profile in terms of PSSM profile as the input features to build various classifier-based prediction models. In Table 7.3, we have provided the performance measures for each classifier on training and independent dataset. As shown in Table 7.3, PSSM profile based models have performed better in comparison to the binary and physicochemical properties profile. 1D-CNN based model outperformed all the other classifiers with AUROC 0.91 and 0.82 on the training and validation dataset.

Table 7.3: Performance of various classifiers using PSSM profile as input feature for training and validation dataset (adopted from Patiyal, Dhall, Bajaj, Sahu, & Raghava, 2022)

Classifier	Training Dataset							Validation Dataset						
	Sens	Spec	Acc	AUC	F1	K	MCC	Sens	Spec	Acc	AUC	F1	K	MCC
DT	30.60	92.02	86.03	0.61	0.30	0.22	0.22	22.08	92.42	82.87	0.57	0.26	0.17	0.17
RF	74.43	70.64	71.01	0.80	0.33	0.22	0.28	62.26	70.90	69.73	0.73	0.36	0.20	0.24
LR	71.93	71.63	71.66	0.79	0.33	0.21	0.28	61.71	74.12	72.44	0.75	0.38	0.23	0.27
XGB	74.23	73.65	73.71	0.82	0.36	0.24	0.31	61.47	75.83	73.88	0.76	0.39	0.25	0.28
KN	74.19	66.17	66.96	0.75	0.30	0.18	0.25	64.36	66.71	66.39	0.69	0.34	0.18	0.22
GNB	67.64	68.82	68.70	0.75	0.30	0.17	0.23	54.78	73.28	70.77	0.70	0.34	0.18	0.21
ET	74.28	70.24	70.64	0.80	0.33	0.21	0.28	62.82	70.12	69.13	0.72	0.36	0.20	0.24
1D-CNN	83.08	83.37	83.34	0.91	0.49	0.41	0.47	80.19	82.35	82.05	0.82	0.55	0.45	0.49

7.3.5 Performance of existing tools

To understand the efficiency of a newly proposed method, it is of utmost importance to compare its performance with the existing tools. There are several tools available for the prediction of RNA-interacting residues in a protein but there is still enough possibility of improvement in the performance of the existing methods. We have used validation dataset created in this study, and predict the RNA-interacting residues by existing tools. Further, we compared their performance using different performance measures such as sensitivity, specificity, accuracy, AUC, F1, kappa, and MCC, and reported in Table 7.4. Our proposed method Pprint2, outperformed all the existing methods by attaining the highest AUC of 0.82 on the validation dataset. Figure 7.5 exhibits the comparison between the existing tools and Pprint2 in terms of AUC plots.

Table 7.4: Performance comparison between proposed method and existing tools on validation dataset (adopted from Patiyal, Dhall, Bajaj, Sahu, & Raghava, 2022)

Methods	Sensitivity	Specificity	Accuracy	AUC	F1	K	MCC
DRNAPred	45.40	51.87	51.59	0.52	0.08	0.00	0.01
HybridNAP	56.90	56.05	56.09	0.59	0.10	0.02	0.05
Pprint	60.15	60.29	60.28	0.63	0.12	0.04	0.08
ProNA2020	62.07	62.44	62.43	0.68	0.13	0.05	0.10

RNABindRPlus	67.24	68.36	68.31	0.73	0.16	0.09	0.15
Pprint2	80.19	82.34	82.05	0.82	0.55	0.45	0.49

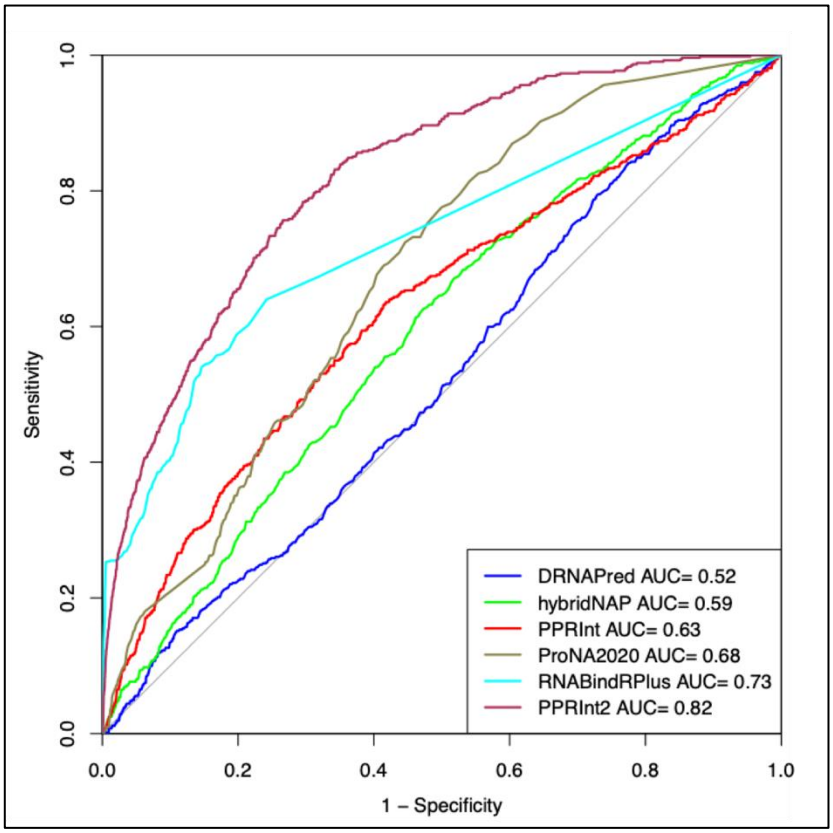


Figure 7.5: Performance comparison between Pprint2 and existing RNA-interacting residues prediction tools (adopted from Patiyal, Dhall, Bajaj, Sahu, & Raghava, 2022)

7.4 Web-based services

Pprint2 is an updated version of 'Pprint' which was published in year 2008. Pprint2 is available as a web-server (https://webs.iiitd.edu.in/raghava/pprint2/), and Python-based standalone (https://webs.iiitd.edu.in/raghava/pprint2/stand.php). The same tool is distributed via GitHub platform at https://github.com/raghavagps/pprint2. The web-interface was developed using HTML, PHP, CSS, JavaScript, Perl, and Python.

The web-interface of Pprint2 provides two major modules, "Predict" and "Standalone". The basic predict module include an approach which accepts multiple protein sequences in the FASTA format and predicts RNA interacting residues in each sequence. In this module, users are allowed to choose from four different types of features such as amino acid binary profile,

physico-chemical properties based profile, PSSM Profile, and hybrid profile for making the predictions. We have implemented ID-CNN based model developed on each feature type in the backend to predict the RNA-interacting residues in a protein sequence. Users are allowed to provide their email, so that for the long processes they do not have to wait and the results will be send to their provided email IDs, once the process gets completed. Figure 7.6 is a screenshot of the "Predict" module of the web-server displaying the submission form for the submission of the query sequences in the FASTA format. Figure 7.7 is an example output page obtained after the submission of the query sequences for making predictions for the RNA interacting residues in the protein sequences. The output page exhibits the sequences, in which RNA interacting residues are highlighted in red color with bigger font size, whereas non-interacting residues are shown in black color. The result page is downloadable in the .txt, .pdf, and .png format. Additionally, to predict the RNA interacting residues in a protein, in the absence of the internet or for prediction in a large dataset, such as the complete human proteome, since server will take a long time to produce the results, we have developed the Python-based standalone. The links for the same is available in the "Standalone" module of the Pprint2 webserver.

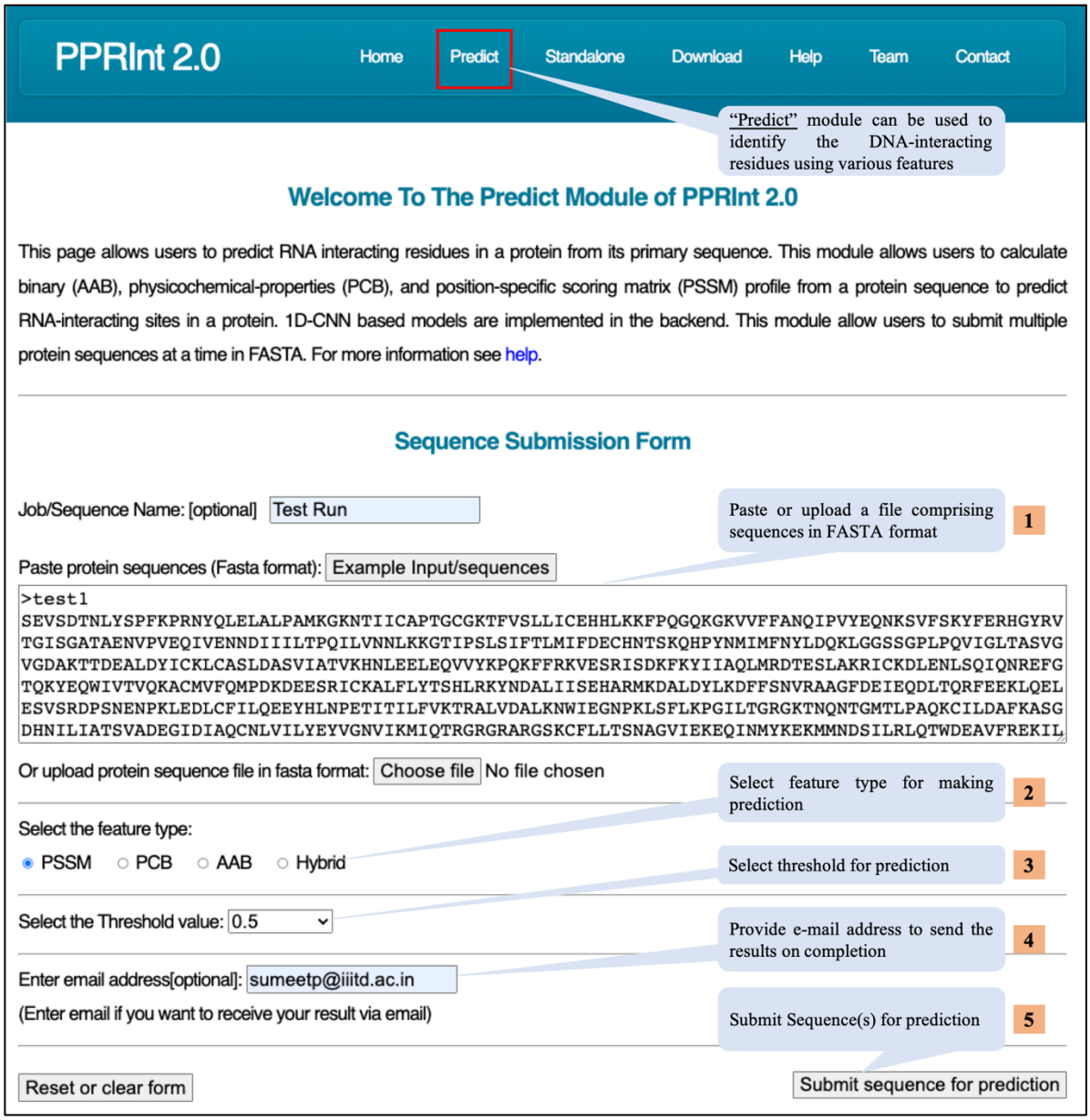


Figure 7.6: Screenshot of "Predict" module of Pprint2 web-server

(URL https://webs.iiitd.edu.in/raghava/pprint2/predict.php)

Figure 7.7: Screenshot of the result page of "Predict" module of Pprint2 webserver

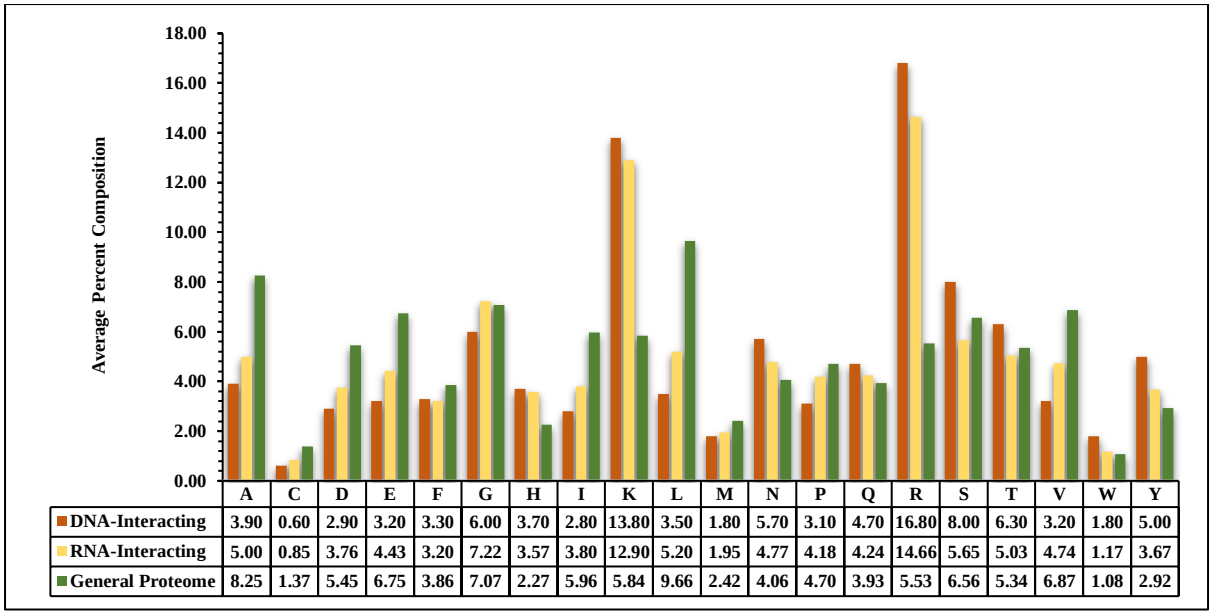
(URL https://webs.iiitd.edu.in/raghava/pprint2/predict.php)

7.5 Discussion and Conclusion

Many essential functions, like as splicing, translation, transport, and silencing, are dependent on the interaction between RNA and protein complexes (Cozzolino et al., 2021; Re, Joshi, Kulberkyte, Morris, & Workman, 2014). The accurate determination of RNA-interacting residues is necessary to understand or exploit the various biological processes (Y. Chen &

Varani, 2013; Jain, Gupte, & Aduri, 2018). The structure of RNA-protein complexes is required for accurate identification of RNA-interacting residues in proteins. Unfortunately, due to the limitations of experimental methods such as crystallography and NMR, crystallisation of all RNA-protein complexes is not attainable. Furthermore, experimental approaches are both expensive and time demanding. Many approaches for predicting RNA interacting residues have been developed to aid researchers in the field of RNA biology (W. Chen, Zhang, Cheng, & Pan, 2011; Y. C. Chen et al., 2014; Xiong, Zeng, & Gong, 2015). One of the key limitations of prior techniques was that they were trained and tested on a small number of RNA binding proteins. For instance, in our prior method Pprint (M. Kumar, Gromiha, et al., 2008), for example, we trained and tested our models on just 86 RNA binding proteins. Due to a paucity of the data, a lenient cut-off threshold of 70% was utilised to remove the redundant proteins. This is true for most of the older approaches that employed a restricted number of RNA binding proteins for training and assessment. Furthermore, the proteins in the collection have a high degree of similarity with one another. The PDB structure of RNA-protein complexes has increased dramatically over the years (Y. Chen & Varani, 2013; Velankar, Burley, Kurisu, Hoch, & Markley, 2021). As a result, a novel technique based on a large number of RNA binding proteins whose structures are accessible in the PDB is required. Furthermore, distinct datasets for training and validation are required. Proteins in training and validation should not have a minimal similarity to avoid over-optimization of machine learning models.

In order to understand the preference of amino acid residues in the DNA- and RNA-interacting sites, we have performed the compositional analysis of the DNA- and RNA- interacting residues and compared the same with the general proteome average percent composition. As shown in Figure 7.8, the DNA- and RNA-interacting residues have shown the similar trends in comparison with general proteome composition but differ in the magnitude. The positively charged residues “R” and “K” are higher in abundance in DNA-interacting sites as compared to RNA-interacting sites, whereas residue “H” has equivalent preference in both the sites. On the other hand, residues “S”, “W”, and “T” are higher in abundance in DNA-interacting sites as compared to the general proteome composition, whereas, in case of RNA-interacting sites, the abundance of these residues are either less or equivalent to the general proteome composition.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
DNA-Interacting	3.90	0.60	2.90	3.20	3.30	6.00	3.70	2.80	13.80	3.50	1.80	5.70	3.10	4.70	16.80	8.00	6.30	3.20	1.80	5.00
RNA-Interacting	5.00	0.85	3.76	4.43	3.20	7.22	3.57	3.80	12.90	5.20	1.95	4.77	4.18	4.24	14.66	5.65	5.03	4.74	1.17	3.67
General Proteome	8.25	1.37	5.45	6.75	3.86	7.07	2.27	5.96	5.84	9.66	2.42	4.06	4.70	3.93	5.53	6.56	5.34	6.87	1.08	2.92

Figure 7.8 Compositional analysis comparison between DNA and RNA-interacting residues

We aim to train our models on a broad set of proteins in this work. We first got 1057 and 360 protein sequences from recently published studies hybridNAP (Jian Zhang et al., 2019) and proNA2020 (Qiu et al., 2020), respectively. We used CD-HIT at 30% to remove duplicate sequences from the dataset of redundant RNA-binding protein. This resulted in a training dataset of 545 proteins and a validation dataset of 161 sequences. This is one of the largest training and validation datasets. Furthermore, protein sequences in the training and validation datasets have a similarity of 30% or less. We have developed a variety of prediction models to predict the RNA-interacting residues in a protein sequence, in this study. To differentiate interacting residues from non-interacting residues, we developed machine learning-based models using diverse features such as binary-profile based features (binary- and physicochemical properties profile) and evolutionary information-based features (PSSM). One of the study's main objectives is to provide assistance to the scientific community. We developed a user-friendly web server (http://webs.iiitd.edu.in/raghava/pprint2) that allows users to determine whether or not a particular residue is a RNA-interacting or non-interacting in a protein sequence (See Figure 7.9). We have also provided the Python-based standalone package which can be used to predict the RNA-interacting residues in bulk proteins or proteome in the absence of internet. We anticipate that the work done here will aid in the annotation of protein sequences.

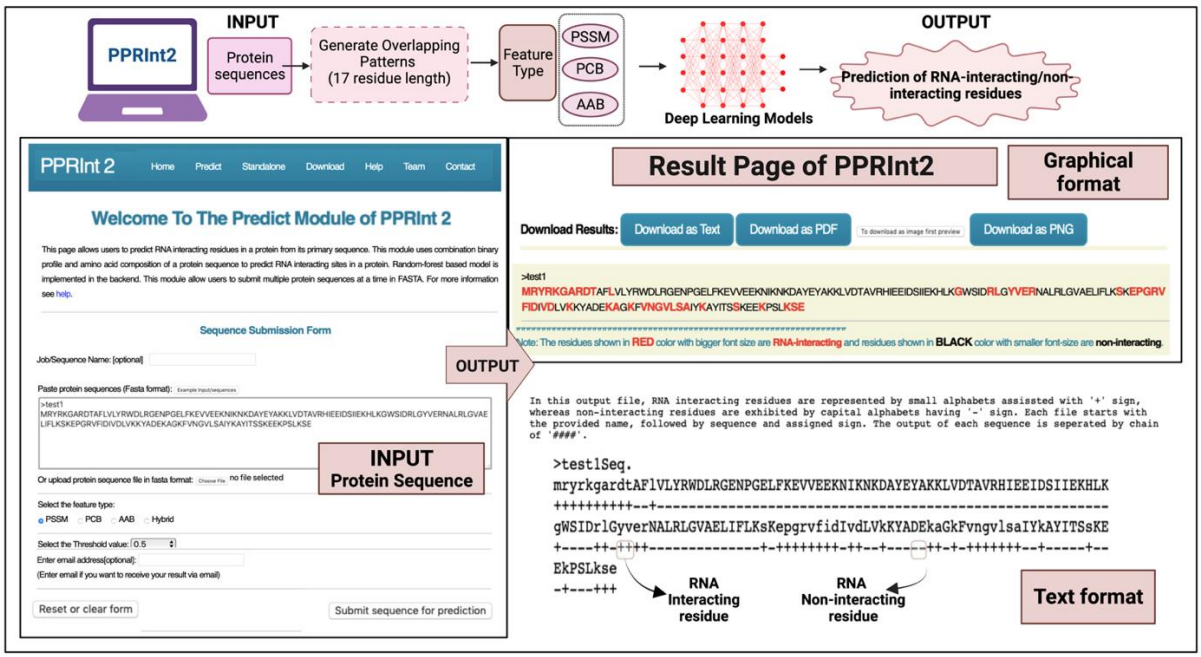


Figure 7.9: Utility of webserver where different steps represents processing of data, generation of features and prediction of RNA-interacting and non-interacting residues

Chapter 8

Benchmarking of mutation calling techniques and identification of cancer biomarkers based on mutation

8.1 Introduction

The world health organisation states that cancer is the primary cause of death worldwide and a potentially fatal condition. According to statistics on cancer worldwide, there will be 10.3 million cancer-related deaths and 19.3 million new cases estimated in 2020 (Sung et al., 2021). Due to the tremendous heterogeneity of cancer, patients with comparable types of cancer cannot benefit from the same treatment plan. There is currently no one, effective treatment for all known kinds of cancers is available. A number of targeted medicines are provided for the treatment of cancer, with a primary focus on the identification of genetic abnormalities (Gerlinger et al., 2012). A number of cancer medicines have been developed in recent years based on mutant genes. As an illustration, Sorafenib, a BRAF inhibitor, is used in melanoma treatment associated with the V600E mutation (Ascierto et al., 2012; S. S. Taylor, 1987). Non-small-cell lung cancer is treated with medications that target the EGFR mutation, such as afatinib and erlotinib (Hirsch et al., 2017; Lynch et al., 2004). Additionally, olaparib, a poly (ADP-ribose) polymerase (PARP) inhibitor, has been used to treat ovarian cancer patients with BRCA1/BRCA2 gene mutations. Notably, understanding the proper disease mechanism requires research on the genetic mutations found in cancer patients (Audeh et al., 2010). With a lot of improvements in whole-genome, whole-exome, and mutation identification techniques used in next-generation sequencing, it is now possible to use sequencing data to identify over 98 percent of the disease-related mutation (LaDuca et al., 2017; Lelieveld, Spielmann, Mundlos, Veltman, & Gilissen, 2015). Next-generation sequencing techniques are widely accessible, affordable, and allow researchers to conduct tests on big cohorts of cancer patients (Hartley et al., 2018).

Single nucleotide variants (SNV), structural variants (SV) and insertion/deletion (indel) are the three basic categories for genomic variations. Numerous somatic mutation calling algorithms have been created in recent years to find genetic mutations using sequencing data, including Varscan2, Mutect2, MuSE, SomaticSniper, etc. (Alioto et al., 2015; Cibulskis et al., 2013; do Valle et al., 2016; Fan et al., 2016; S. Kim et al., 2018; Koboldt et al., 2012; Larson et al., 2012). Among the 36 malignancies listed by Global Cancer Statistics 2020, liver cancer is the fifth most prevalent and one of the deadliest diseases (Sung et al., 2021). Although numerous therapeutic options have been discovered in the past, the survival rate of people with liver cancer is still quite low, contributing to a high mortality rate (Revathidevi & Munirajan, 2019). The Cancer Genome Atlas (TCGA), the most comprehensive repository for information on

cancer-related research, offers two types of file formats for mutation data, including MAF and VCF file format. The genomic sequence variants that directly resulted from the various automated variant calling methods are stored and reported in VCF files, which are the raw mutation files. The processed version of the VCF files, known as MAF files, are curated by deleting false positives or by retrieving known calls that the automated pipelines could have missed. VCF files report all mutations, regardless of their significance, whereas MAF files only describe the most significant mutations by excluding the lesser-quality mutations. Both types of files are available in the Genomic Data Commons (GDC) site and were produced utilising the four main mutation calling methods, MuTect2, MuSE, Varscan2, and SomaticSniper. Even though there are numerous methodologies, it can be challenging to decide which approach and data set is best for examining the function of mutations in cancer.

The goal of the current study was to identify substantially mutated genes linked with patients at high risk for liver cancer by carefully evaluating the four mutation calling techniques that are often used in TCGA. For all the mutation calling strategies, we have extracted the VCF and MAF files from 418 individuals with liver cancer. ANNOtate VARiation (ANNOVAR) and Maftools are used for the identification of gene-based annotations (Mayakonda, Lin, Assenov, Plass, & Koeffler, 2018; K. Wang, Li, & Hakonarson, 2010). We performed correlation and survival analysis for the identification of highly significant genes which have major impact on the outcome of patients. In conclusion, we selected top-10 risk-associated genes and developed survival prediction/classification models using various ML techniques. Based on the conclusions, we benchmarked various methodologies that can serve as a useful guide and reference for researchers for identifying the mutated genes significantly influencing cancer patients' survival.

8.2 Materials and Methods

8.2.1 Construction of dataset and overall workflow

We have used the GDC data portal (Grossman et al., 2016) to download the mutation data in the form of VCF and MAF files for liver cancer (TCGA-LIHC and TCGA-CHOL) patients. We considered the data of mutation profiles generated using four widely used mutation calling techniques such as Mutect2, MuSE, Varscan2, and SomaticSniper. The total number of patients we have assessed is 418, and we downloaded the mutation profiles for each patient generated using each technique. Additionally, to download the clinical characteristics of the patients, we

have used the TCGA assembler 2 module (Wei et al., 2018). The final matrix is divided into two datasets, such as, training dataset which captures 80% of the instances, where rest of the data was used for the purpose of external validation and called it as independent dataset. The overall workflow that we have adapted in this study is shown in Figure 8.1.

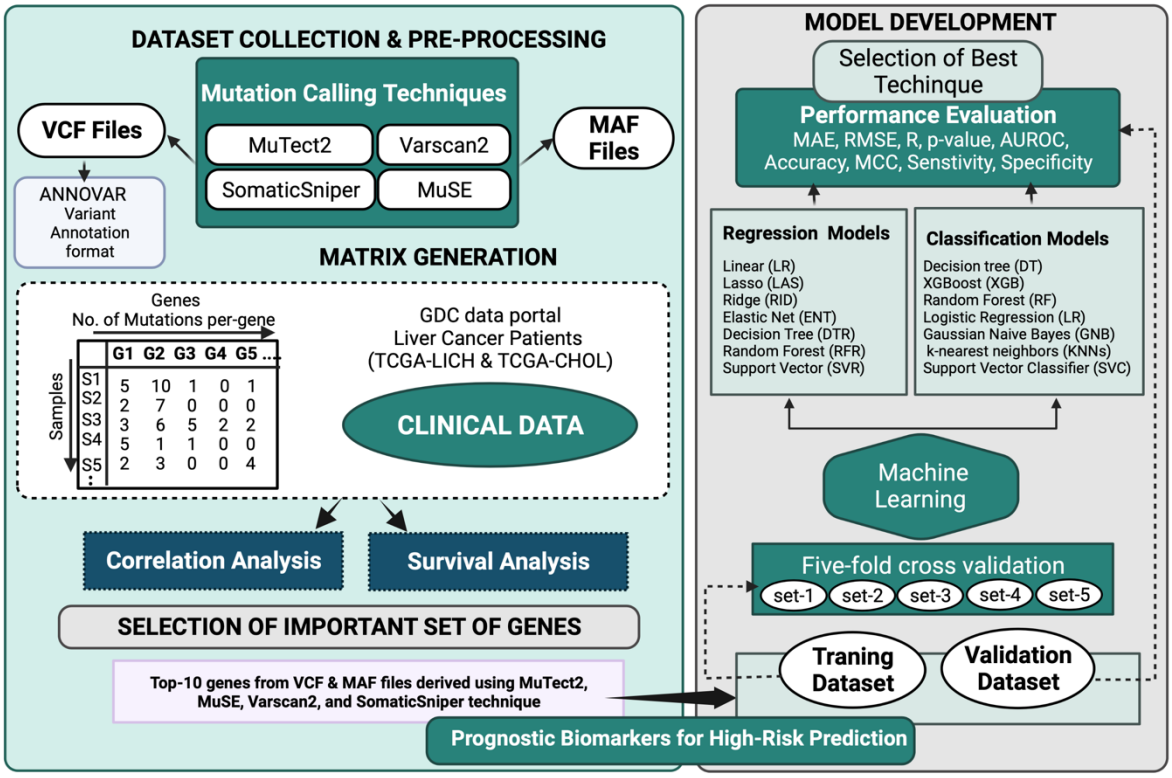


Figure 8.1: Overall workflow acquired in this study

8.2.2 Annotation of mutations

We have implemented the ANNOVAR software to annotate the mutations in the VCF files derived from each technique. First, we have converted the VCF file to ANNOVAR genetic variant file, which is five columns containing file such as chromosome number, start position, end position, reference nucleotide, and altered nucleotides. We have used the gene-based annotated file in this study. Finally, we have created a matrix that contains the information on the number of mutations in each gene in each sample. We have generated this information for VCF as well as the MAF file.

8.2.3 Statistical analysis

In statistical analysis, we have performed correlation and survival analysis. We have computed the correlation between the number of mutations in each gene and the overall survival time of the patients. Further, we have removed the genes with a non-significant p-value for the correlation test and sorted the rest based on their coefficient to choose the top negatively correlated genes for further analysis for each file type and technique. Moreover, we have also performed the univariate survival analysis using the cox proportional hazard (Cox PH) model using the 'survival' package of R, to understand the role of mutation frequency on the survival of the patients. The log-rank test was used to explain the significance of the distribution of patients into high- and low-risk groups. We have calculated different measures like hazard ratio, p-value, 95% confidence interval, and concordance. Also plotted the Kaplan–Meier to represent the survival curves.

8.2.4 Prediction models

We have developed the classification and regression models to classify the patients into high- and low-risk groups and, to predict their survival time using the number of mutations in the risk-associated genes. To develop prediction models, we have implemented various classifiers and regressors using Python's scikit-learn library (Pedregosa et al., 2011). Classifiers that we have implemented include decision tree, random forest, logistic regression, extreme gradient boosting, Gaussian naive Bayes, extra tree, support vector classifier, and k-nearest neighbor. To train and develop the classification models, we have segregated the patients into high- and low-risk groups based on the median overall survival time, i.e., if a patient has an overall survival time less than the median then the high-risk group is assigned to that patient, otherwise low-risk. To predict the overall survival time using the number of mutations in risk-associated genes, we have implemented various regressors such as decision tree regressor, random forest regressor, linear regressor, lasso, ridge, elastic net, and support vector regressor.

8.2.5 Performance measures

We have implemented various performance measures used in the previous studies (Bhalla, Kaur, Dhall, & Raghava, 2019; Dhall, Patiyal, Kaur, et al., 2020; Schemper, 1993) to compare and evaluate the models. For the sake of classification, we have implemented parameters like sensitivity, specificity, accuracy, F1-score, kappa, and Matthews correlation coefficient (MCC)

as the threshold-dependent parameters, whereas area-under the receiver operating characteristics curve (AUROC) is the threshold-independent parameter. To evaluate the regression models, we have used mean absolute error (MAE), root-mean-square error (RMSE), correlation coefficient (R), and *P*-value as the performance measures.

8.3 Results

8.3.1 Preliminary analysis

In the preliminary analysis, we have analyzed the number of genes and mutations by each mutation calling technique for VCF and MAF files. We have observed that VCF files from different methods reported a higher number of mutations than the MAF files, as reported in Figure 8.2A. Where Mutect2 and SomaticSniper reported the maximum number of mutations and genes in VCF files. Figure 8.2B and 8.2C exhibit the UpSet Plot (Lex, Gehlenborg, Strobelt, Vuillemot, & Pfister, 2014) for VCF and MAF files, respectively, which is a graphical representation to understand the distribution of genes in each technique. As per the figures, 18758 genes are common in all the VCF files derived using all the four mutation calling methods, whereas 182, 5, 2, and 630 genes are unique to MuTect2, MuSE, Varscan2, and SomaticSniper techniques, respectively. On the same note, 14585 genes are common in all the MAF files derived using all the four mutation calling methods, whereas 461, 73, 115, and 41 genes are unique to MuTect2, MuSE, Varscan2, and SomaticSniper techniques, respectively.

A)

File Type	Technique	Number of Genes	Number of Mutations
VCF	MuTect2	25366	5237093
	MuSE	19425	379368
	Varscan2	19422	576231
	SomaticSniper	25785	5003969
MAF	MuTect2	16474	59741
	MuSE	15712	51184
	Varscan2	15950	54877
	SomaticSniper	14979	44102

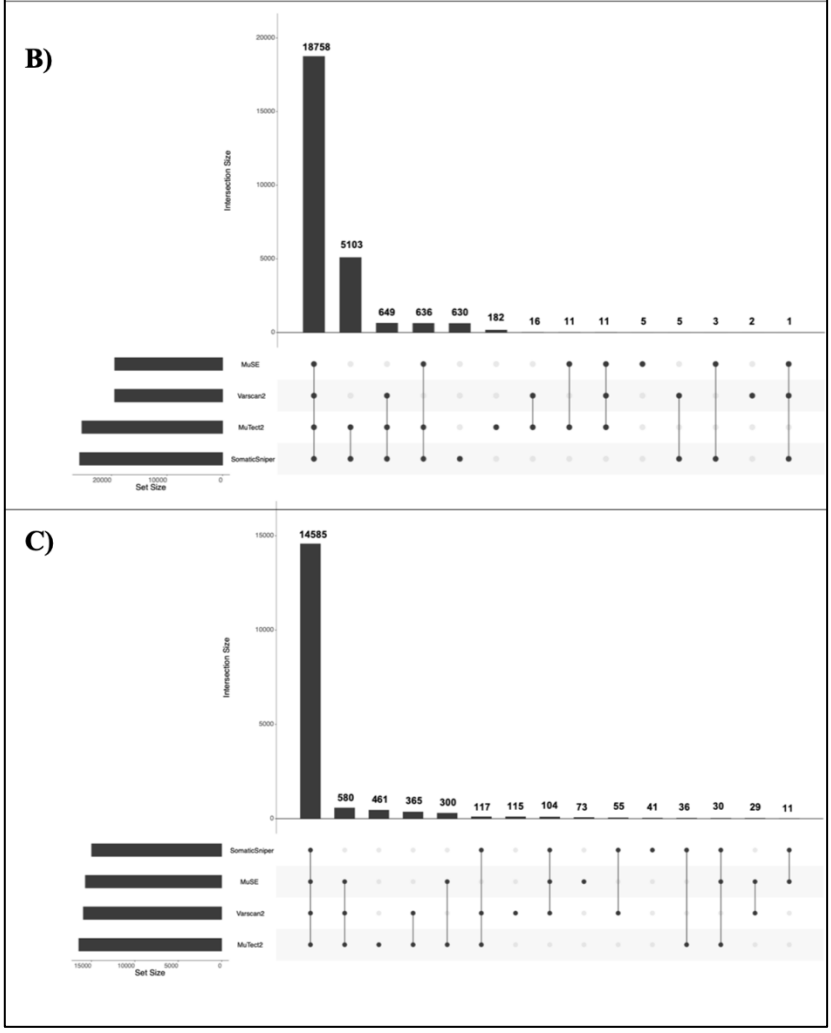


Figure 8.2: Preliminary analysis exhibiting A) technique-wise frequency distribution of mutations and genes B) UpSet plot for gene distribution in VCF files derive using Mutect2, MuSE, Varscan2, and SomaticSniper C) UpSet plot for gene distribution in MAF files derive using Mutect2, MuSE, Varscan2, and SomaticSniper (adopted from Patiyal, Dhall, & Raghava, 2022)

8.3.2 MAF file comparison

We used processed and annotated MAF data from the TCGA to test alternative mutation calling strategies. We used the Maftools package to analyse the somatic variations retrieved from MuSE, Mutect2, Varscan2, and the SomaticSniper mutation calling approach in depth. The investigation revealed minor differences in mutation calling approaches for the same cohort of samples. MuSE and SomaticSniper MAF files, for example, solely provide SNPs whereas Varscan2 and MuTect2 (Fig. 8.3) depict SNPs, INS, and DEL under the variant type.

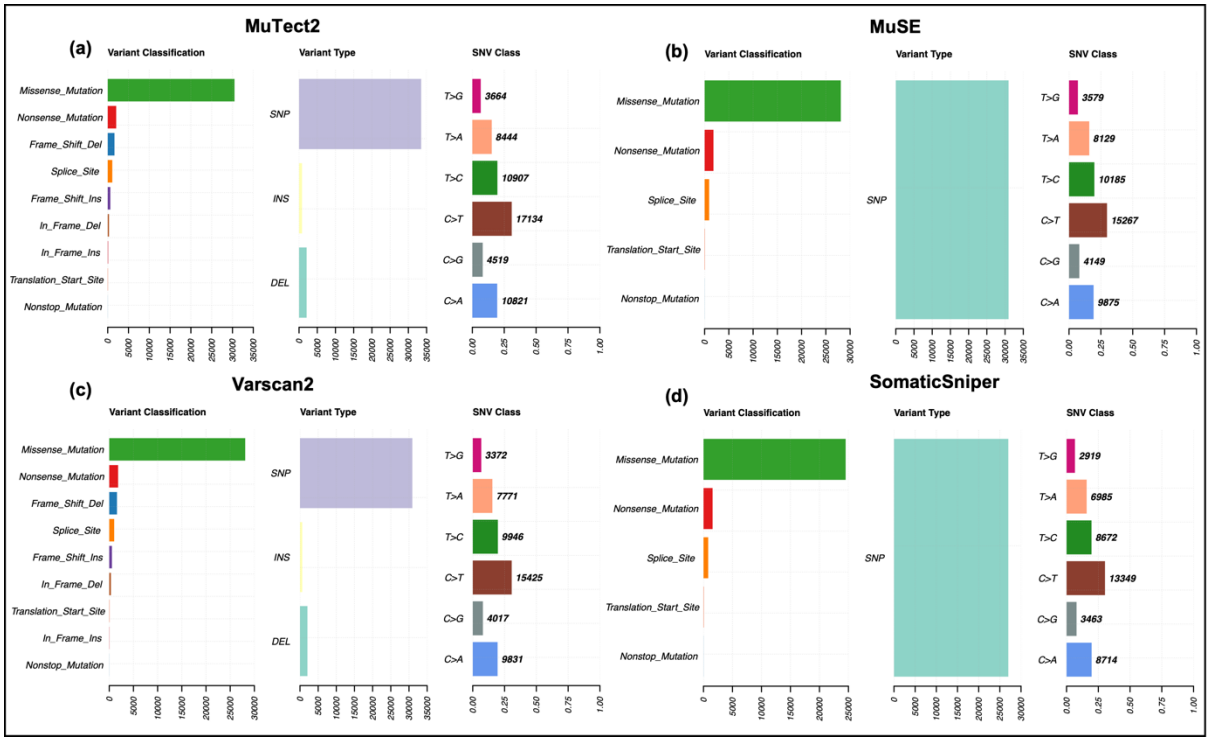


Figure 8.3: Exhibition of mutation summary (variants classification, type and SNVs) for (A) MuTect2, (B) MuSE, (C) Varscan2 and (D) SomaticSniper MAF files (adopted from Patiyal, Dhall, & Raghava, 2022)

The variant classification distribution in Varscan2 and MuTect2 represents nine types of mutations: Missense Mutation, Nonsense Mutation, Splice Site, Translational Start Site, Frame Shift Insertions, Frame Shift Deletion, In Frame Insertion, In Frame Deletion, and Nonstop Mutations, whereas MuSE and SomaticSniper MAF files include Missense Mutation, Nonsense Mutation, Splice Site, Translational Start Site and Nonstop Mutations. The SNV class represents single-nucleotide variations in the TCGA cohort; we discovered that all

approaches exhibit a heterogeneous distribution of SNV. Maftools visualisation module Oncoplots depicting the somatic landscape of cancer patients for Varscan2, MuTect2, MuSE, and SomaticSniper MAF files. In Figure 8.4, we showed the top mutated genes together with their mutation proportion (5%) in the total number of samples. According to the findings, TP53 is a highly altered gene with around 20% or more mutations among different methodologies.

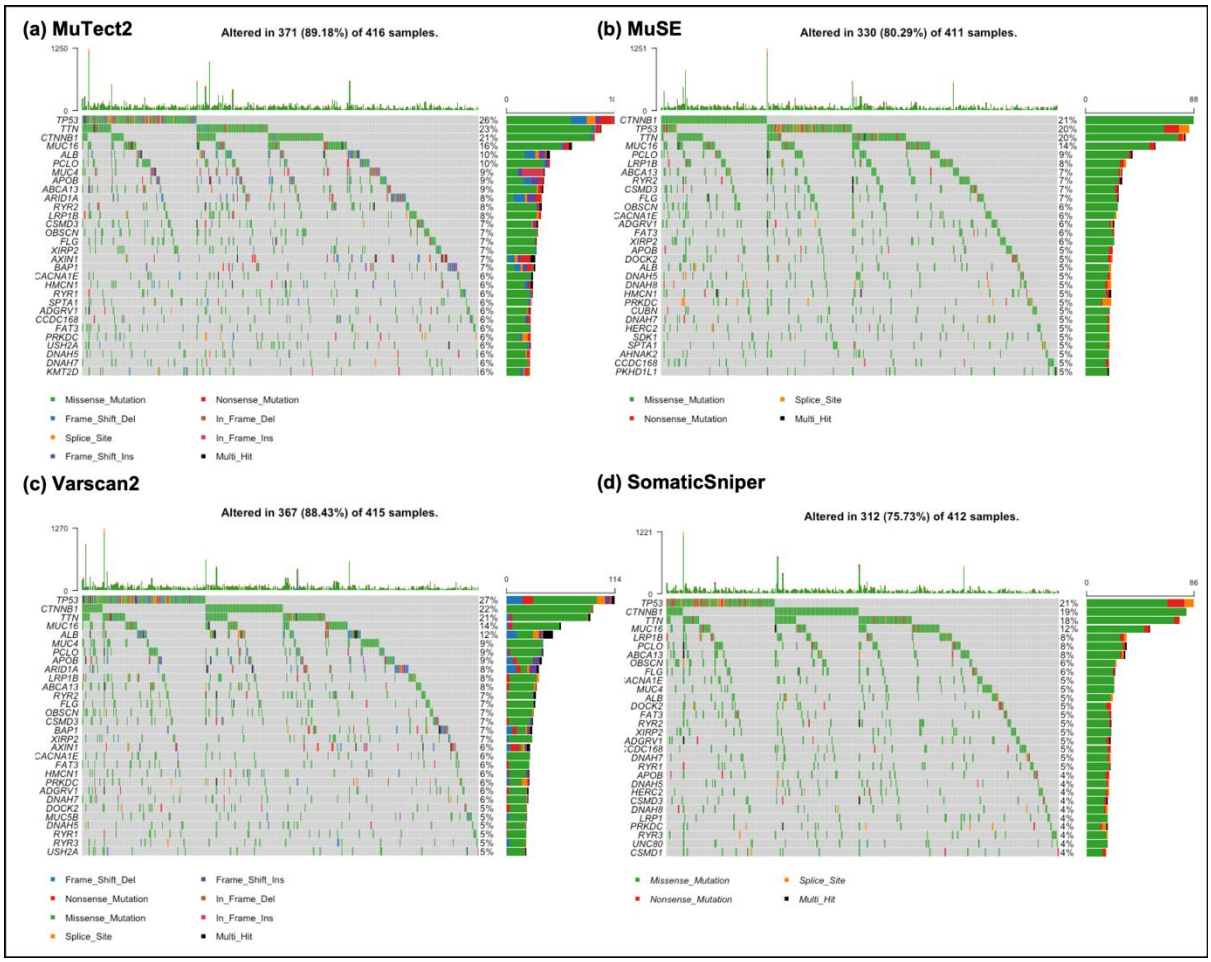


Figure 8.4: Oncoplot representation of the top-most mutated genes' mutation frequency. The rows indicated the genes with the highest percentage of mutations, while the columns represented the samples. (A) Shows the oncoplot of the MuTect2 approach, which shows that 89.18 percent of samples had altered genes. (B) Shows the oncoplot of the MuSE approach and reveals that 80.29 percent of samples had altered genes. (C) Displays the oncoplot of the Varscan2 technique, revealing that 88.43 percent of samples had altered genes. (D) Shows the oncoplot of the SomaticSniper approach, which shows that 75.73 percent of samples had alerted/mutated genes. (adopted from Patiyal, Dhall, & Raghava, 2022)

8.3.3 Correlation analysis

Using the correlation test, we ranked the genes and chose the top ten that had a significant negative correlation with overall survival time. The approach was performed for all four techniques utilising MAF and VCF data from liver cancer patients, yielding a total of 80 genes.

8.3.4 Prediction of biomarkers based on single gene

To explore the gene-wise impact on the survival of liver cancer patients, we have performed the univariate survival analysis on high-risk associated genes derived using correlation analysis for each mutation calling technique. The results on each mutation calling approach and file type, i.e., VCF and MAF, are reported in Table 8.1 and Table 8.2, respectively. We have compared the performance of each gene in terms of HR, P-value, 95% CI, and concordance index (C-index). SomaticSniper technique attained the maximum HR of 7.06 with a p-value of 6.62e-07 on the gene CLDN20, followed by Varscan2, Mutect2, and MuSE performed the least with an HR 3.01 (p-value 1.67e-05) on CLMP gene, for VCF files.

Table 8.1: Univariate survival analysis results for top-10 genes from VCF files derived using MuTect2, MuSE, Varscan2, and SomaticSniper technique

MuTect2					MuSE				
Gene	**HR**	**P-value**	**95% CI**	**C-index**	**Gene**	**HR**	**P-value**	**95% CI**	**C-index**
SNHG10	5.49	3.94E-06	2.66 - 11.31	0.53	**CLMP**	3.01	1.67E-05	1.82 - 4.97	0.54
WIZ	2.69	9.71E-07	1.81 - 4.00	0.56	**BIRC6**	2.80	4.46E-04	1.58 - 4.99	0.54
MGAT4EP	2.49	4.46E-04	1.50 - 4.15	0.54	**LINC02210-CRHR1**	2.03	6.42E-03	1.22 - 3.39	0.53
LINC00304	2.39	7.40E-05	1.55 - 3.67	0.55	**DHX8**	2.00	2.90E-02	1.07 - 3.74	0.52
CACNG7	1.93	5.72E-04	1.33 - 2.81	0.56	**LINC00972**	1.91	9.31E-03	1.17 - 3.10	0.54
OR52B6	1.83	1.12E-03	1.27 - 2.63	0.56	**PAX7**	1.90	8.29E-04	1.30 - 2.76	0.56
TYK2	1.80	2.21E-03	1.24 - 2.63	0.56	**TAS1R2**	1.61	2.63E-02	1.06 - 2.44	0.53
PIGO	1.79	1.66E-02	1.11 - 2.88	0.52	**SNTG1**	1.53	3.37E-02	1.03 - 2.27	0.54
S100A12	1.71	1.10E-02	1.13 - 2.59	0.54	**CNTN5**	1.34	2.25E-01	0.83 - 2.16	0.51
DNAJC9-AS1	1.08	6.51E-01	0.77 - 1.51	0.52	**ZNF521**	1.26	2.63E-01	0.84 - 1.91	0.52
Varscan2					**SomaticSniper**				
Gene	**HR**	**P-value**	**95% CI**	**C-index**	**Gene**	**HR**	**P-value**	**95% CI**	**C-index**
FAM160A2	6.81	4.01E-05	2.73 - 17.02	0.52	**CLDN20**	7.06	6.62E-07	3.27 - 15.2	0.53

LOC100420587	5.45	1.31E-07	2.90 - 10.22	0.54	**NR2C2AP**	5.17	3.16E-05	2.38 - 11.2	0.52
SPDYA	3.08	7.70E-04	1.60 - 5.94	0.53	**ATG9B**	3.34	2.59E-04	1.75 - 6.37	0.53
BRSK2	2.55	1.01E-03	1.46 - 4.46	0.54	**HAUS5**	2.79	2.22E-05	1.74 - 4.48	0.55
ADGRF4	2.21	1.23E-02	1.19 - 4.10	0.53	**LOC100287329**	2.58	8.23E-04	1.48 - 4.49	0.53
LINC00972	2.11	2.18E-03	1.31 - 3.41	0.55	**P4HTM**	2.18	2.43E-02	1.11 - 4.31	0.52
TM4SF18	2.07	1.40E-02	1.16 - 3.70	0.53	**OR6C76**	2.12	1.18E-03	1.35 - 3.35	0.54
OR5AS1	1.86	1.43E-02	1.13 - 3.06	0.54	**CLK2**	1.94	3.58E-02	1.05 - 3.61	0.52
PDE11A	1.72	2.74E-03	1.21 - 2.46	0.55	**FAM187B**	1.64	1.51E-02	1.10 - 2.43	0.55
LOC101929073	1.29	2.98E-01	0.80 - 2.11	0.52	**NOMO3**	1.34	1.45E-01	0.90 - 1.98	0.52

Similar, analysis was done on genes derived from MAF files generated using MuTect2, MuSE, Varscan2, and SomaticSniper technique. Mutect2 technique attained the highest performance followed by Varscan2, MuSE and SomaticSniper for genes LAMC3, SYDE1, ITGB8 and CAD, respectively, as shown in Table 8.2.

Table 8.2: Univariate survival analysis results for top-10 genes from MAF files derived using MuTect2, MuSE, Varscan2, and SomaticSniper technique

MuTect2					**MuSE**				
Gene	**HR**	**P-value**	**95% CI**	**C-index**	**Gene**	**HR**	**P-value**	**95% CI**	**C-index**
LAMC3	9.25	1.78E-06	3.71 - 23.05	0.52	**ITGB8**	8.37	5.69E-07	3.64 - 19.24	0.52
EVC2	4.30	8.66E-05	2.08 - 8.91	0.53	**TBX3**	8.10	6.06E-05	2.91 - 22.53	0.52
NYNRIN	3.94	1.22E-03	1.72 - 9.05	0.52	**SIPA1L3**	4.90	5.54E-05	2.26 - 10.61	0.52
KIAA2026	3.85	1.49E-03	1.68 - 8.86	0.52	**CAD**	4.45	3.58E-03	1.63 - 12.14	0.52
SUPT20H	3.41	7.53E-03	1.39 - 8.40	0.51	**EVC2**	4.16	2.97E-04	1.92 - 9.01	0.52
BRINP2	2.83	2.43E-02	1.14 - 6.98	0.52	**ARHGEF11**	3.17	2.37E-02	1.17 - 8.64	0.51
LRP1B	1.93	7.81E-03	1.19 - 3.14	0.54	**BRINP2**	2.80	2.56E-02	1.13 - 6.92	0.52
TP53	1.48	3.60E-02	1.03 - 2.14	0.55	**PCDH15**	1.72	1.20E-01	0.87 - 3.39	0.51
TG	1.46	4.53E-01	0.54 - 3.97	0.51	**TG**	1.46	4.55E-01	0.54 - 3.97	0.51
PCDH15	1.43	3.30E-01	0.70 - 2.93	0.51	**CSMD3**	1.24	4.54E-01	0.71 - 2.15	0.51
Varscan2					**SomaticSniper**				
Gene	**HR**	**P-value**	**95% CI**	**C-index**	**Gene**	**HR**	**P-value**	**95% CI**	**C-index**
SYDE1	8.46	3.71E-05	3.07 - 23.35	0.52	**CAD**	5.56	8.10E-04	2.04 - 15.17	0.52
ALPP	4.33	1.44E-03	1.76 - 10.66	0.52	**TOP2A**	4.63	2.73E-03	1.70 - 12.62	0.52
KIAA2026	3.85	1.49E-03	1.68 - 8.86	0.52	**KIAA2026**	4.01	2.62E-03	1.62 - 9.93	0.52
CAD	3.32	1.91E-02	1.22 - 9.04	0.51	**EVC2**	4.00	1.04E-03	1.75 - 9.17	0.52
BRINP2	2.83	2.43E-02	1.14 - 6.98	0.52	**KTN1**	2.56	1.09E-01	0.81 - 8.10	0.51
TP53	1.60	9.85E-03	1.12 - 2.30	0.56	**EPHA3**	2.25	1.67E-01	0.71 - 7.13	0.51
PCDH15	1.48	2.81E-01	0.72 - 3.05	0.51	**KIF26B**	2.03	1.66E-01	0.74 - 5.55	0.51
TG	1.46	4.53E-01	0.54 - 3.97	0.51	**PCDH15**	1.76	1.78E-01	0.77 - 4.02	0.51

PLCB1	1.25	7.00E-01	0.40 - 3.96	0.50	**TP53**	1.63	1.20E-02	1.11 - 2.38	0.55
XIRP2	1.11	7.55E-01	0.58 - 2.12	0.51	**TG**	1.18	8.17E-01	0.29 - 4.79	0.50

8.3.5 Prediction of biomarkers based on multiple gene

We have projected the survival time to determine the high-risk group in liver cancer patients in order to investigate the overall effects of mutations in the chosen genes. For each approach that corresponds to each file type, the Cox PH model was used to calculate the HR and P-value using the anticipated OS time. For the VCF files produced using the MuTect2 approach, we obtained the highest HR = 4.50 with a very significant P-value of 3.83E-15. But when it came to MAF files, the MuSE approach outperformed with HR = 2.47 and P-value = 9.64E-07. Additionally, the separation of high- and low-risk groups is evident in KM survival plots; Figure 8.5 compares several mutation calling methods based on two file formats.

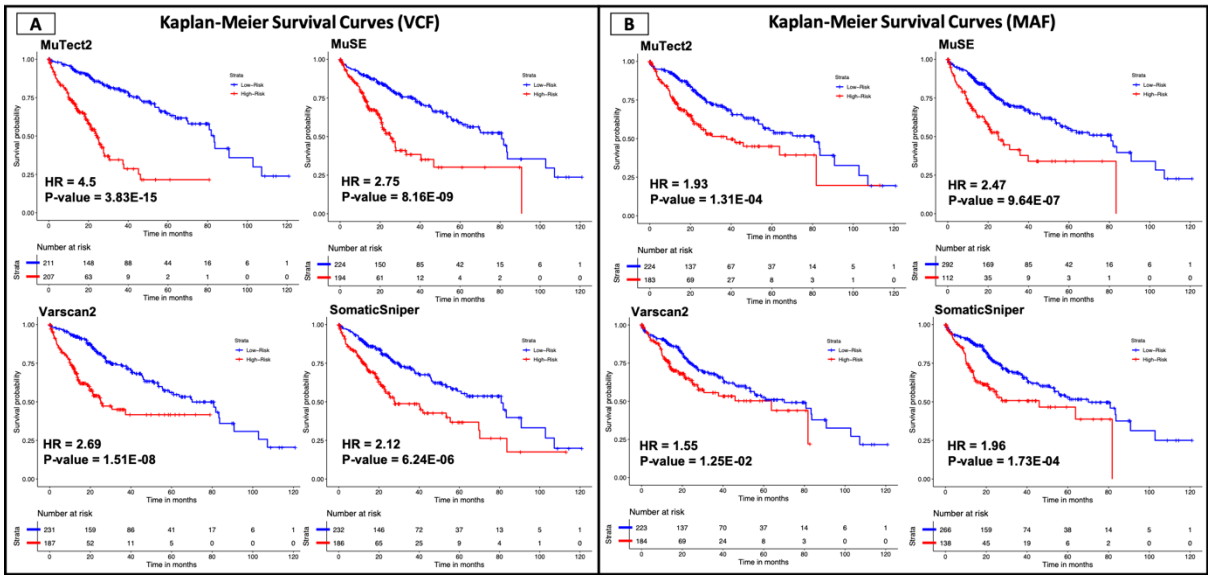


Figure 8.5: Kaplan Meier survival plots for the risk-estimation using multiple genes (adopted from Patiyal, Dhall, & Raghava, 2022)

8.3.6 Overall survival time prediction

In order to develop the technique-wise regression models using top-10 risk-associated genes from VCF and MAF file types, we have implemented various regressors such as decision tree, random forest, linear, lasso, ridge, elastic net, and support vector regressor. Table 8.3 represents the performance of the best performing regressors in each technique and file type. In the files belonging to the VCF format, the Mutect2 technique attained the minimum error (MAE 12.52

month and R 0.57); similarly, in the case of the MAF file format too, Mutetct2 performed best by achieving the lowest error and highest correlation value with significant p-value (MAE 16.47 months and R 0.37). In a nutshell, Mutect2-based files performed best among the other techniques in the case of VCF and MAF file formats.

Table 8.3: Performance of best regressors on VCF and MAF files extracted using different techniques (adopted from Patiyal, Dhall, & Raghava, 2022)

Technique	File Type	MAE	RMSE	R	p-value
MuTect2	**VCF**	12.52	19.58	0.57	7.00E-37
	MAF	16.47	22.16	0.37	1.31E-14
MuSE	**VCF**	13.88	20.38	0.51	1.38E-29
	MAF	16.89	22.48	0.34	1.68E-12
Varscan2	**VCF**	14.57	20.78	0.48	4.77E-26
	MAF	16.53	22.26	0.36	9.11E-14
SomaticSniper	**VCF**	15.76	21.82	0.40	3.31E-17
	MAF	16.72	22.26	0.33	8.46E-12

8.3.7 Prediction of Risk-group

To develop the prediction model to classify the patients into high- and low-risk groups, w have used diverse classifiers using top-10 genes from VCF and MAF files derived using different mutation calling techniques. We provided the class to each patient as high- or low-risk based on the median overall survival time, such as patients with overall survival time less than the median are assigned as the high-risk group, and the patients having overall survival time higher than the median overall survival time are designated as a low-risk group. Table 8.4 comprises the performance measures for the logistic regression-based model on VCF and MAF file types for each technique, as the logistic regression-based model outperforms other classifiers in most of the methods. As shown in Table 8.4, Mutect2 based files performed best with AUROC 0.765 and 0.659 on VCF and MAF files, respectively. In terms of mean values, VCF file-based models have higher performance in comparison to the models generated on MAF files with an AUROC of 0.699 ± 0.061 on validation dataset. In conclusion, for VCF and MAF files, the MuTect2 technique performed best among other approaches in terms of AUROC, F1, Kappa and MCC values.

Table 8.4: Performance of logistic regression based models developed on VCF and MAF files extracted using different techniques (adopted from Patiyal, Dhall, & Raghava, 2022)

Technique	File Type	AUROC	F1	Kappa	MCC
MuTect2	**VCF**	0.765	0.767	0.421	0.442
	MAF	0.659	0.661	0.259	0.335
MuSE	**VCF**	0.735	0.737	0.400	0.421
	MAF	0.621	0.667	0.225	0.277
Varscan2	**VCF**	0.656	0.661	0.250	0.348
	MAF	0.653	0.661	0.308	0.309
SomaticSniper	**VCF**	0.638	0.672	0.276	0.277
	MAF	0.617	0.667	0.225	0.243
Average	**VCF**	0.699 ± 0.061	0.709 ± 0.051	0.337 ± 0.086	0.372 ± 0.075
	MAF	0.638 ± 0.022	0.664 ± 0.003	0.254 ± 0.039	0.291 ± 0.040

8.4 Important Discoveries

We reviewed the results to better understand the limits and to make some suggestions. We discovered that classification and regression models produced using the MuTect2 technique's VCF/MAF file performed better than models developed using alternative mutation calling strategies. Notably, we may infer that MuTect2 outperforms the other mutation calling approaches used in this study. Furthermore, our findings show that models based on VCF files outperform models based on MAF files for the majority of mutation calling approaches except Varscan2. Because the VCF file contains information in its raw form, it is larger in size than the MAF file, which is a processed version. Hence, when we convert the VCF to MAF format, the number of mutations decreases dramatically, but performance decreases as well, i.e., useful and efficient variants may be discarded during the conversion process. As a result, an effective technique for converting VCF to MAF format without losing essential information is required. Furthermore, we discover that gene-based prognostic indicators change as per the mutation calling technique, as well as VCF and MAF format. Ideally, these variant calling algorithms should show the same mutations and biomarkers in a particular gene. It demonstrates that the list of mutations in a particular gene differs depending on the mutation calling techniques used. Therefore, improved variant calling methods or identifying consensus mutations are required.

A recent research (M. Wang et al., 2020) found that consensus mutations outperformed mixed models.

8.5 Discussion and Conclusion

Liver cancer is one of most deadly disease which occurs after chronic liver diseases which are associated with progressive genetic mutation (Davis et al., 2008; Farazi & DePinho, 2006; Muller, Bird, & Nault, 2020). Liver cancer is one of the most frequent cancer types and is associated with a poor prognosis and a high mortality rate (Balogh et al., 2016; L. Lin et al., 2020). To detect the mutation landscape in tumor/normal patients, numerous mutation calling approaches are now accessible. Until far, there has been no adequate comparison of mutation detection technologies for predictive and prognostic analysis. Using the TCGA liver cancer cohort, we examined the performance of four commonly used mutation calling techniques: MuTect2, MuSE, Varscan2, and SomaticSniper. We used correlation and survival analyses to identify prognostic biomarkers (i.e., risk-associated genes) in patients with liver cancer. Furthermore, we used a variety of machine learning approaches to compare all of the strategies for predicting high-risk liver cancer patients. First, we utilised VCF and MAF files created by the various mutation calling mechanisms. To detect gene-associated mutations in liver cancer samples, we employed the most common software (ANNOVAR and Maftools). Based on our findings, the VCF files of Mutect2 and SomaticSniper have the greatest number of altered genes and encompass more than 5 million mutations. MAF files, on the other hand, report less altered genes for each approach. Then, in order to understand the influence of mutations on the survival of liver cancer patients, we used correlation analysis. The univariate survival analysis indicated that risk-associated genes such as LncRNA SNGH10, CLMP, FAM160A2, and CLDN20 attained the highest HR values in the MuTect2, MuSE, Varscan2, and SomaticSniper techniques. A study by Lan et al. supported our findings by revealing that the oncogenic lncRNA SNGH10 is related with poor survival in patients with liver cancer (Lan et al., 2019). Furthermore, SNGH10 down-regulation is related with poor survival in non-small cell lung cancer patients, with HR 2.09 and p-value 0.02 (Liang, Wang, Cao, Song, & Wu, 2020). Our findings are consistent with prior research, demonstrating that mutations in the SNGH10 gene are related with a poor prognosis in patients with liver cancer, with an HR of 5.49 and a p-value of 3.94E-06. The differential expression of the CLMP gene, on the other hand, has been linked to the advancement of breast cancer (Nilchian et al., 2019). Yang et al. also found the CLDN20 gene to be important in the survival of breast cancer patients, with an HR of 1.38 and

a p-value of 0.047 (G. Yang, Jian, & Chen, 2021). Our research also highlighted the importance of the CLMP and CLDN20 genes in the survival of individuals with liver cancer. In the case of MAF files, univariate survival analysis demonstrates that the genes SYDE1, LAMC3, ITGB8, CAD, EVC2, NYNRIN, BRSK2, and TP53 substantially lower the overall survival. According to a recent study, the overexpressed SYDE1 oncogene acts as an essential diagnostic and prognostic biomarker in patients with glioma (Han et al., 2021). Furthermore, down-regulation of LAMC3 in ovarian cancer patients is associated with a poor prognosis and metastases (Lei et al., 2021). A research also found that mutations in the LAMC3 genes might induce PNH (a rare condition of clonal stem cells in the foetus), which can lead to infection and preterm delivery (De Angelis et al., 2021; Qian et al., 2021).

We also discovered that mutations in LAMC3 significantly impair patient survival, with HR = 9.25 and p-value 1.78E-06. Furthermore, ITGB8 has been demonstrated to be substantially elevated in high-grade ovarian cancer patients, resulting in a shorter OS with a significant HR 1.42 (He, Liu, Zhang, & Zhang, 2018). According to Paul et al., the EVC2 gene is highly mutated in breast cancer patients and dysregulates pathways such as mTOR, CDK/RB, cAMP/PKA, WNT, and others (Paul et al., 2020). Our findings suggest that mutations in the EVC2 genes diminish overall patient survival, with HR = 4.3 and p-value 8.66E-05. Overexpression of the BRSK2 gene has been found to be associated with patient survival and prognosis in pancreatic cancer. A lot of studies have found that TP53 is the most frequently altered gene in most human malignancies, affecting cancer patients' survival (Monti et al., 2020; Olivier, Hollstein, & Hainaut, 2010; Petitjean, Achatz, Borresen-Dale, Hainaut, & Olivier, 2007; Rosenberg, Okamura, Kato, Soussi, & Kurzrock, 2020). In our investigation, we also discovered that the number of mutations related with the TP53 gene is particularly high among liver cancer patients, accounting for over 20% of all mutations. Correlation and survival analyses revealed that the TP53 mutation significantly lowers overall survival, with HR = 1.63 and P = 1.20E-02. When the overall effect of the chosen genes in each file was considered, MuTect2 beat all other approaches in VCF files with HR = 4.50 (P = 3.83E-15), but MuSE outperformed other mutation calling methods in MAF files with HR = 2.47 (P = 9.64E-07). Furthermore, we create multiple survival prediction and classification models utilising the top ten risk-associated genes in order to compare the different mutation calling methodologies. The MuTect2 technique's logistic regression-based model created on 10 chosen genes from the VCF file performed best among the other approaches in stratifying patients into high- and low-risk groups, with an AUROC of 0.765 on the validation dataset. Furthermore, MuSE performs

pretty well on the validation dataset, with an AUROC of 0.735, but Varscan2 and SomaticSniper-based models do not perform well on both VCF and MAF files. We examined models developed using various machine learning techniques, and the results show that the error is not due to machine learning techniques because the performance measure AUROC was similar on the training and validation datasets, indicating that these models are reliable and no overfitting was observed. Our findings show that the VCF file created using the MuTect2 mutation calling approach has extensive information that may be utilised to estimate the risk of a liver cancer cohort. Furthermore, this has to be validated in other cancer cohorts in order to investigate the prognostic potential of mutations in various types of cancer. We created a comprehensive Python-based end-to-end pipeline (https://github.com/raghavagps/mutationbench) to support the scientific community working in this era. Users can compare the performances of various prediction models built using different mutation calling techniques by simply providing VCF/MAF files.

Chapter 9

Summary

Biological macromolecules and molecules contribute significantly to the well-being of an organism. Major biological macromolecules include carbohydrates, lipids, proteins, and nucleic acid. Each of them is the essential component of the cell and performs a wide variety of critical biological functions. In the biological system, major molecules coordinate with each other via interactions, for instance, the interaction between a protein, say transcription factor, and nucleic acid, such as DNA, is responsible for decoding the genetic information into RNA via the process called transcription, which further translated to the functional unit of life, i.e., proteins via translation. The flawless coordination between these molecules is responsible for the complete functionality of the organism; on the other hand, a minor glitch in this coordination may lead to various life-threatening disorders. Several studies revealed that protein-RNA interactions are majorly involved in the development of human cancers, neurological disorders like Alzheimer's and sclerosis, and genetic disorders. One of the major flaws that disrupt or disturb the coordination between the important molecules is mutations which further alter the system's supposed functionality and lead to many disorders such as cancer. Cancer is associated with genetic mutations and is the second leading cause of death worldwide. Several attempts are in the pipeline to develop the target therapies for cancer as well as other diseases, that mainly target the mutated genes in order to develop new therapeutics.

Protein is one of the most important biological macromolecules and also works as the functional unit of the cell, we have explored various areas related to a protein's functional annotation in this study. The entire study is divided into three parts: 1) Functional annotation of protein; 2) Identification of Protein-molecules interaction; and 3) Prediction of cancer-associated mutations. In the functional annotation of protein, we endeavored to understand the various features associated with the protein sequences and structures, and used them to functionally annotate the proteins into transcription factors. This section is further subdivided into two categories: i) generation of features from the protein structure and sequences, and ii) identification of transcription factors from the primary structure. This section is briefly explained in Chapter 3 and 4. In the first section, we first tried to gather knowledge from the literature about the features related to protein sequences and structures. Besides the features reported in the literature, we attempted to come up with some novel features based on the trends amino acid sequences follow in protein with different functionalities. We developed a platform called "Pfeature" to compute all those features by providing the sequence or structure of a protein. We have provided this tool in the form of a web-server

(https://webs.iiitd.edu.in/raghava/pfeature/), Python-based standalone, Python library, and more than 180 Python scripts that users can modify as per their desire. Entire source codes and standalone are available at GitHub platform (https://github.com/raghavagps/Pfeature).

We divided the features into five major categories: composition, binary profile, evolutionary information, structure, and pattern. Pfeature can calculate more than two lakhs features for a single sequence which can further be used to develop models for functional annotation at the level of sequence or residues. The composition category further subdivided into five different sub-categories, such as: i) simple which included amino acid composition, dipeptide composition, tripeptide composition, and atom & bond composition; ii) physico-chemical properties based composition which calculates composition for standard, amino acid index, advanced, and structure based physico-chemical properties; iii) repeats & distribution that captures the repeat information in a sequence and calculate composition for residue repeats, property repeats, and distance distribution of residues; iv) Shannon entropy that computes features at the level of protein, residues, and properties; and v) miscellaneous which provide the option to calculated features like, autocorrelation, conjoint triad distribution, composition enhanced-transition and distribution, pseudo-amino acid composition, amphiphilic pseudo-amino acid composition, quasi-sequence order, and sequence order coupling number. Composition features provides feature vector of fixed length for variable length proteins and hence fail to capture the position specific features. For calculation residues position specific features, binary profile category provides features at the level of amino acid, dipeptide, atom & bond, physico-chemical properties, and amino acid index. Further, in the category of evolutionary information, we provided the facility to generate the position-specific scoring matrix (PSSM) profile by implementing PSI-BLAST in the back-end and used Swiss-Prot database to hit the query sequences. In the same module we have provided four different normalization method, which normalizes the raw PSSM profile. The structure category compute features like fingerprints by implementing PaDEL software, SMILES by implementing open-babel software, secondary structure by implementing DSSP software, and solvent accessibility by using NACCESS software in the back-end. Last category, pattern generates patterns of desired length for the input sequences or profile, which can further be used as the feature to generate models. In addition, Pfeature also provides facility to generate models in the module of model building, which further subdivided into sub-categories like merging features to merge two different feature matrices column-wise into one matrix, feature relevance to provide importance of each feature based on mean, classification to develop

classification models, and regression to develop regression models based on the input matrices. In the classification and regression modules, users are allowed to perform various operation on the input matrix such as dimension reduction, feature selection, normalization, and clustering. Other than that, Pfeature provides facilities like parameter optimization, k-fold cross validation, and options of five different classifiers and regressors.

In the second section of this part of study, an in-silico method "TransFacPred" was developed to predict the transcription factors using the primary structure information. We have downloaded the sequences from the UniProt/Swiss-Prot database and classify the sequences based on the GO terms, and obtained 19406 sequences as transcription factors, and 523560 sequences as non-transcription factors. We have split the dataset in 80:20 ratio, where 80% data used for training and referred as training dataset; and remaining 20% data is used for external validation, referred as independent dataset. First, we tried to classify the sequences using similarity search approach using BLAST and called it as alignment-based method, this method works fine but was not able to results for every sequence in the independent dataset. Then, we applied alignment-free method in which we implemented several machine-learning and one deep-learning classifier was implemented to develop the prediction model using features like amino acid composition, dipeptide composition, their combination, and one-hot encoding. ET-based method developed on amino acid composition performed best with AUROC of 0.97 in training and independent dataset, respectively. To improve the performance of the method, we further combined alignment-based (BLAST) and alignment-free (ET) methods to develop hybrid approach, and attained the highest AUROC of 0.99 on the independent dataset. We have provided this method as web-server (https://webs.iiitd.edu.in/raghava/transfacpred/) and Python-based standalone, which is also available on GitHub platform (https://github.com/raghavagps/transfacpred).

In the second part of the study, i.e., identification of protein-molecules interaction, we developed methods to understand the interactions between the protein and different molecules. This section is briefly explained in Chapter 5, 6, and 7. We have considered three important molecules such as NAG, DNA, and RNA, and developed bioinformatic-ware for each. First, we developed computational tool "NAGbinder", for predicting the N-acetylglucosamine interacting residues in a protein using sequence information. In order to make the dataset, we have downloaded the NAG-interacting protein complexes from the PDB April 2019 release, which were 5736 in number. In the preprocessing, we have removed the redundant chains by

applying CD-HIT software with 40% criteria and left with 231 protein chains. The contact information was retrieved by applying LPC software. Finally, we left with 231 protein chains which constituted 1985 NAG-interacting and 74931 non-interacting residues, and called it as realistic dataset. Further, we generated balanced dataset out of it by randomly selecting 1985 non-interacting residues out of 74931 non-interacting residues, which further subdivided into 80:20 ratio where 80% dataset was used for internal validation and remaining 20% was used for external validation, referred as training and independent dataset, respectively. This was the first method proposed for NAG-interaction study, hence there was no prior information was available regarding the pattern size to be considered. Hence, we generated the overlapping patterns of window size ranging from 5 to 23 using the sequences, where central residue represented the whole pattern as NAG-interacting or non-interacting. We applied different classifiers to develop the models on various features like binary profile, PSSM profile, and hybrid profile which is the combination of binary- and PSSM-profile. We further found out that RF-based model developed on binary profile of pattern length 9 performed best among the other classifiers and feature types on balanced dataset with AUROC of 0.73 and 0.70 on training and independent dataset, respectively. Hence, we concluded that pattern length 9 is the optimum length. We generated the model on realistic dataset and achieved AUROC of 0.70 on training and independent dataset, respectively. We have provided this tool as web-server (https://webs.iiitd.edu.in/raghava/nagbinder/), Python- and Perl-based standalone (https://webs.iiitd.edu.in/raghava/nagbinder/stand.html) available at web-site, and GitHub (https://github.com/raghavagps/nagbinder). Also, distributed the same using docker technology via GPSRdocker (https://webs.iiitd.edu.in/gpsrdocker/).

In the next section of the second part of the study, we made a systematic attempt to develop the method to predict the DNA interacting residues on a protein sequence. We have used the largest benchmarked dataset provided by the recently published studies hybridNAP and ProNA2020. We removed the redundant sequences by applying CD-HIT software with 30% criteria, where no sequences in training and independent dataset shared more the 30% sequence identity. Training dataset included 646 proteins with 15636 DNA-interacting, and 298503 non-interacting residues, whereas independent dataset comprised 46 protein with 965 DNA-interacting, and 9911 non-interacting residues. We have generated the pattern size 17 for generating the overlapping patterns and calculated four different type of features such as binary profile, physico-chemical properties profile, PSSM profile, and hybrid profile which is combination of binary, physico-chemical properties, and PSSM profiles. We have

implemented several machine learning and one deep-learning classifier to develop models on each feature type. Our results indicated that 1D-CNN based model developed on hybrid profile attained the highest AUROC of 0.79 on the independent dataset. We have used the independent dataset to compare the performance of the proposed method with existing approaches and found out that our method outperformed all the functional existing method. We have called this method as "DBPred" and made it available to the scientific community via web-interface (https://webs.iiitd.edu.in/raghava/dbpred/), Python- and Perl-based standalone (https://webs.iiitd.edu.in/raghava/dbpred/stand.html). We have also made available via GPSRdocker.

In the last section of the second part of the study, we updated the already existing tool 'Pprint' developed in 2008 to predict the RNA interacting residues on a protein sequence. We called this method as "Pprint2". We have used the largest benchmarked dataset provided by the recently published studies hybridNAP and ProNA2020. We removed the redundant sequences by applying CD-HIT software with 30% criteria, where no sequences in training and independent dataset shared more the 30% sequence identity. Training dataset included 545 protein sequences with 18559 RNA-interacting, and 171879 non-interacting residues, whereas independent dataset comprised 161 protein sequences with 6966 RNA-interacting, and 44349 non-interacting residues. We have generated the pattern size 17 for generating the overlapping patterns and calculated three different type of features such as binary profile, physico-chemical properties profile, and PSSM profile. We have implemented several machine learning classifiers like DT, RF, LR, XGB, KNN, ET, and GNB, and one deep-learning classifier such as 1D-CNN, to develop models on each feature type. Our results indicated that 1D-CNN based model developed on PSSM profile attained the highest AUROC of 0.82 on the independent dataset. We have used the independent dataset to compare the performance of the proposed method with existing approaches and found out that our method outperformed all the functional existing method. This method is available as web-interface (https://webs.iiitd.edu.in/raghava/pprint2/), Python-based standalone (https://webs.iiitd.edu.in/raghava/pprint2/stand.php) via web-server and GitHub platform (https://github.com/raghavagps/pprint2).

In the last part of the study i.e. third part, we made a systematic attempt to understand the role of mutations in liver cancer patients, and used their mutation profiles to identify the diagnostic and prognostic biomarkers. We have described the method in details in Chapter 8. At first, we

benchmarked the four widely used mutation calling techniques to understand the detection of mutations in genome, as there are several ways available, but there is no scientific consensus on the optimal variant calling pathway at the moment. Mutation calling techniques that we have considered are Mutect2, MuSE, Varscan2, and Somaticsniper. In order to acquire the mutation profiles of the liver cancer patients, we have downloaded the mutations in two formats, i.e., VCF and MAF format from TCGA. In addition, we obtained the clinical characteristics of each sample of TCGA cohort of liver cancer patients. We had four VCF and four MAF files derived from each mutation calling technique. VCF files were converted to gene annotation format by implementing ANNOVAR tool. The VCF and MAF files were used to generate a matrix with number of mutations per gene per sample. Finally, we have eight matrices belong to each mutation calling technique and file type. Further, to explore the influence of number of mutations on overall survival time of the patients, we have implemented correlation analysis in which we have taken top-10 highly negatively correlation genes with significant p-value for each technique for VCF and MAF file, and referred them as high-risk associated genes. Then, we performed the univariate survival analysis for each selected gene and calculated hazard ratio and p-value along with other parameters, to understand the impact of each gene on the survival of the patients. Using these top-10 high-risk associated genes from each technique and file type, we have developed classification and regression models to predict the risk-group and overall survival time of the patients, respectively. Our results indicated that Mutect2 based VCF and MAF file performed best among the other techniques in classifying the patients in high- and low-risk group, and predicted the overall survival time of the patients with minimum error and high correlation between predicted and actual overall survival time. In VCF and MAF file, models developed on Mutect2 derived VCF file outperformed models developed on MAF file. Hence, we concluded that Mutect2 worked better than the other mutation calling techniques to explore the diagnostic and prognostic role of mutation profile in liver cancer patients. Furthermore, this has to be validated in other cancer cohorts in order to investigate the prognostic potential of mutations in various types of cancer. To assist the scientific community working in this period, we created a comprehensive Python-based end-to-end pipeline (https://github.com/raghavagps/mutation_bench), where users may evaluate the performances of multiple prediction models generated using different mutation calling approaches. Overall, the study done in this thesis addresses various aspects of the functional annotation and use of mutation profiles to identify the diagnostic and prognostic biomarkers, and provides useful insights. We anticipate that clinicians and researchers will use the findings of our investigations to develop advanced treatment approaches.

Bibliography

Abarca-Cabrera, L., Fraga-Garcia, P., & Berensmeier, S. (2021). Bio-nano interactions: binding proteins, polysaccharides, lipids and nucleic acids onto magnetic nanoparticles. *Biomaterials Research*, *25*(1), 12. https://doi.org/10.1186/s40824-021-00212-y

Aeling, K. A., Steffen, N. R., Johnson, M., Hatfield, G. W., Lathrop, R. H., & Senear, D. F. (2007). DNA deformation energy as an indirect recognition mechanism in protein- DNA interactions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *4*(1), 117–125. https://doi.org/10.1109/TCBB.2007.1000

Aggarwal, M., & Mondal, A. K. (2006). Role of N-terminal hydrophobic region in modulating the subcellular localization and enzyme activity of the bisphosphate nucleotidase from Debaryomyces hansenii. *Eukaryotic Cell*, *5*(2), 262–271. https://doi.org/10.1128/EC.5.2.262-271.2006

Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., & Raghava, G. P. S. (2021). AntiCP 2.0: an updated model for predicting anticancer peptides. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa153

Agrawal, P., Bhalla, S., Chaudhary, K., Kumar, R., Sharma, M., & Raghava, G. P. S. (2018). In Silico Approach for Prediction of Antifungal Peptides. *Frontiers in Microbiology*, *9*, 323. https://doi.org/10.3389/fmicb.2018.00323

Agrawal, P., Kumar, R., Usmani, S. S., Dhall, A., Patiyal, S., Sharma, N., Jain, S. (2019). GPSRdocker: a Docker-based resource for genomics, proteomics and systems biology. *BioRxiv*, 827766.

Agrawal, P., Mishra, G., & Raghava, G. P. S. (2020). SAMbinder: A Web Server for Predicting S-Adenosyl-L-Methionine Binding Residues of a Protein From Its Amino Acid Sequence . *Frontiers in Pharmacology* . Retrieved from https://www.frontiersin.org/articles/10.3389/fphar.2019.01690

Agrawal, P., Raghav, P. K., Bhalla, S., Sharma, N., & Raghava, G. P. S. (2018). Overview of Free Software Developed for Designing Drugs Based on Protein-Small Molecules Interaction. *Current Topics in Medicinal Chemistry*, *18*(13), 1146–1167. https://doi.org/10.2174/1568026618666180816155131

Agrawal, P., & Raghava, G. P. S. (2018). Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure. *Frontiers in Microbiology*, *9*, 2551.

https://doi.org/10.3389/fmicb.2018.02551

Agrawal, P., Singh, H., Srivastava, H. K., Singh, S., Kishore, G., & Raghava, G. P. S. (2019). Benchmarking of different molecular docking methods for protein-peptide docking. *BMC Bioinformatics*, *19*(Suppl 13), 426. https://doi.org/10.1186/s12859-018-2449-y

Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics (Oxford, England)*, *20*(4), 477–486. https://doi.org/10.1093/bioinformatics/btg432

Aksentijevich, I., & Zhou, Q. (2017). NF-kappaB Pathway in Autoinflammatory Diseases: Dysregulation of Protein Modifications by Ubiquitin Defines a New Category of Autoinflammatory Diseases. *Frontiers in Immunology*, *8*, 399. https://doi.org/10.3389/fimmu.2017.00399

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*, 10001. https://doi.org/10.1038/ncomms10001

Allerson, C. R., Cazzola, M., & Rouault, T. A. (1999). Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. *The Journal of Biological Chemistry*, *274*(37), 26439–26447. https://doi.org/10.1074/jbc.274.37.26439

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amirkhani, A., Kolahdoozi, M., Wang, C., & Kurgan, L. A. (2020). Prediction of DNA-Binding Residues in Local Segments of Protein Sequences with Fuzzy Cognitive Maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(4), 1372–1382. https://doi.org/10.1109/TCBB.2018.2890261

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & Biology*, *10*(9), 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002

Ansari, H. R., & Raghava, G. P. S. (2010). Identification of NAD interacting residues in

proteins. *BMC Bioinformatics*, *11*(1), 160. https://doi.org/10.1186/1471-2105-11-160

Ascierto, P. A., Kirkwood, J. M., Grob, J.-J., Simeone, E., Grimaldi, A. M., Maio, M., Mozzillo, N. (2012, July). The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine*. England. https://doi.org/10.1186/1479-5876-10-85

Audeh, M. W., Carmichael, J., Penson, R. T., Friedlander, M., Powell, B., Bell-McGuinn, K. M., Tutt, A. (2010). Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet (London, England)*, *376*(9737), 245–251. https://doi.org/10.1016/S0140-6736(10)60893-8

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. https://doi.org/10.1093/nar/28.1.45

Balogh, J., Victor, D. 3rd, Asham, E. H., Burroughs, S. G., Boktour, M., Saharia, A., Monsour, H. P. J. (2016). Hepatocellular carcinoma: a review. *Journal of Hepatocellular Carcinoma*, *3*, 41–53. https://doi.org/10.2147/JHC.S61146

Batista, P. J., & Chang, H. Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. *Cell*, *152*(6), 1298–1307. https://doi.org/10.1016/j.cell.2013.02.012

Batool, M., Ahmad, B., & Choi, S. (2019). A Structure-Based Drug Discovery Paradigm. *International Journal of Molecular Sciences*, *20*(11). https://doi.org/10.3390/ijms20112783

Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K. D., Giles, H., Zaugg, J. B. (2019). Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. *Cell Reports*, *29*(10), 3147-3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W. 3rd, & Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, *24*(11), 1429–1435. https://doi.org/10.1038/nbt1246

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallographica. Section D, Biological*

Crystallography, *58*(Pt 6 No 1), 899–907. https://doi.org/10.1107/s0907444902003451

Bhalla, S., Kaur, H., Dhall, A., & Raghava, G. P. S. (2019). Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Scientific Reports*, *9*(1), 15790. https://doi.org/10.1038/s41598-019-52134-4

Bhasin, M., & Raghava, G. P. S. (2004a). ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, *32*(Web Server), W414–W419. https://doi.org/10.1093/nar/gkh350

Bhasin, M., & Raghava, G. P. S. (2004b). GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Research*, *32*(Web Server), W383–W389. https://doi.org/10.1093/nar/gkh416

Blum, T., Briesemeister, S., & Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, *10*(1), 274. https://doi.org/10.1186/1471-2105-10-274

Booth, W. T., Schlachter, C. R., Pote, S., Ussin, N., Mank, N. J., Klapper, V., Chruszcz, M. (2018). Impact of an N-terminal Polyhistidine Tag on Protein Thermal Stability. *ACS Omega*, *3*(1), 760–768. https://doi.org/10.1021/acsomega.7b01598

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)*, *406*, 89–112. https://doi.org/10.1007/978-1-59745-535-0_4

Brown, A.-L., Li, M., Goncearenco, A., & Panchenko, A. R. (2019). Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLoS Computational Biology*, *15*(4), e1006981. https://doi.org/10.1371/journal.pcbi.1006981

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V, Zhuravleva, M. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, *49*(D1), D437–D451. https://doi.org/10.1093/nar/gkaa1038

Bushweller, J. H. (2019). Targeting transcription factors in cancer - from undruggable to reality. *Nature Reviews. Cancer*, *19*(11), 611–624. https://doi.org/10.1038/s41568-019-0196-7

Cai, S. F., & Levine, R. L. (2019). Genetic and epigenetic determinants of AML pathogenesis. *Seminars in Hematology*, *56*(2), 84–89. https://doi.org/10.1053/j.seminhematol.2018.08.001

Cai, Y.-H., & Huang, H. (2012). Advances in the study of protein-DNA interaction. *Amino Acids*, *43*(3), 1141–1146. https://doi.org/10.1007/s00726-012-1377-9

Cao, D.-S., Liang, Y.-Z., Yan, J., Tan, G.-S., Xu, Q.-S., & Liu, S. (2013). PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *Journal of Chemical Information and Modeling*, *53*(11), 3086–3096. https://doi.org/10.1021/ci400127q

Cao, D.-S., Xiao, N., Xu, Q.-S., & Chen, A. F. (2015). Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics (Oxford, England)*, *31*(2), 279–281. https://doi.org/10.1093/bioinformatics/btu624

Cao, D.-S., Xu, Q.-S., & Liang, Y.-Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics (Oxford, England)*, *29*(7), 960–962. https://doi.org/10.1093/bioinformatics/btt072

Carey, K. T., & Wickramasinghe, V. O. (2018). Regulatory Potential of the RNA Processing Machinery: Implications for Human Disease. *Trends in Genetics : TIG*, *34*(4), 279–290. https://doi.org/10.1016/j.tig.2017.12.012

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Mathelier, A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *50*(D1), D165–D173. https://doi.org/10.1093/nar/gkab1113

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Taylor, B. S. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*, *34*(2), 155–163. https://doi.org/10.1038/nbt.3391

Chauhan, J. S., Mishra, N. K., & Raghava, G. P. (2009a). Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, *10*(1), 434. https://doi.org/10.1186/1471-2105-10-434

Chauhan, J. S., Mishra, N. K., & Raghava, G. P. (2010a). Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC*

Bioinformatics, *11*(1), 301. https://doi.org/10.1186/1471-2105-11-301

Chauhan, J. S., Mishra, N. K., & Raghava, G. P. S. (2009b). Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, *10*(1), 434. https://doi.org/10.1186/1471-2105-10-434

Chauhan, J. S., Mishra, N. K., & Raghava, G. P. S. (2010b). Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics*, *11*(1), 301. https://doi.org/10.1186/1471-2105-11-301

Chen, J.-K., Shen, C.-R., & Liu, C.-L. (2010). N-acetylglucosamine: production and applications. *Marine Drugs*, *8*(9), 2493–2516. https://doi.org/10.3390/md8092493

Chen, K., Mizianty, M. J., & Kurgan, L. (2011). ATPsite: sequence-based prediction of ATP-binding residues. *Proteome Science*, *9 Suppl 1*(Suppl 1), S4. https://doi.org/10.1186/1477-5956-9-S1-S4

Chen, K., Mizianty, M. J., & Kurgan, L. (2012). Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics (Oxford, England)*, *28*(3), 331–341. https://doi.org/10.1093/bioinformatics/btr657

Chen, W., Zhang, S.-W., Cheng, Y.-M., & Pan, Q. (2011). Identification of protein-RNA interaction sites using the information of spatial adjacent residues. *Proteome Science*, *9 Suppl 1*, S16. https://doi.org/10.1186/1477-5956-9-S1-S16

Chen, X.-F., Zhang, Y., Xu, H., & Bu, G. (2013). Transcriptional regulation and its misregulation in Alzheimer's disease. *Molecular Brain*, *6*, 44. https://doi.org/10.1186/1756-6606-6-44

Chen, Y. C., Sargsyan, K., Wright, J. D., Huang, Y.-S., & Lim, C. (2014). Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Research*, *42*(3), e15. https://doi.org/10.1093/nar/gkt1299

Chen, Y., & Varani, G. (2013). Engineering RNA-binding proteins for biology. *The FEBS Journal*, *280*(16), 3734–3754. https://doi.org/10.1111/febs.12375

Chen, Z., Liu, X., Zhao, P., Li, C., Wang, Y., Li, F., Song, J. (2022). iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac351

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Zhao, Q. (2021). iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Research*, *49*(10), e60–e60.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Song, J. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics (Oxford, England)*, *34*(14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Song, J. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, *21*(3), 1047–1057. https://doi.org/10.1093/bib/bbz041

Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Ballester, B. (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Research*, *48*(D1), D180–D188. https://doi.org/10.1093/nar/gkz945

Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., Wei, X. (2019). Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy*, *4*, 62. https://doi.org/10.1038/s41392-019-0095-0

Cho, H.-J., Lee, S., Ji, Y. G., & Lee, D. H. (2018). Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PloS One, 13(11), e0207204. https://doi.org/10.1371/journal.pone.0207204

Choi, S., & Han, K. (2011). Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics*, *12 Suppl 1*, S7. https://doi.org/10.1186/1471-2105-12-S13-S7

Chow, C.-N., Lee, T.-Y., Hung, Y.-C., Li, G.-Z., Tseng, K.-C., Liu, Y.-H., Chang, W.-C. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Research*, *47*(D1), D1155–D1163. https://doi.org/10.1093/nar/gky1081

Chowdhury, S. Y., Shatabda, S., & Dehzangi, A. (2017). iDNAProt-ES: Identification of DNA-binding Proteins Using Evolutionary and Structural Features. *Scientific Reports*, *7*(1), 14938. https://doi.org/10.1038/s41598-017-14945-1

Chu, W.-Y., Huang, Y.-F., Huang, C.-C., Cheng, Y.-S., Huang, C.-K., & Oyang, Y.-J. (2009). ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Research*, *37*(Web Server issue), W396-401. https://doi.org/10.1093/nar/gkp449

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. https://doi.org/10.1038/nbt.2514

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5

Colak, S., & Ten Dijke, P. (2017). Targeting TGF-beta Signaling in Cancer. *Trends in Cancer*, *3*(1), 56–71. https://doi.org/10.1016/j.trecan.2016.11.008

Collas, P. (2010). The current state of chromatin immunoprecipitation. *Molecular Biotechnology*, *45*(1), 87–100. https://doi.org/10.1007/s12033-009-9239-8

Coppede, F., Lopomo, A., Spisni, R., & Migliore, L. (2014). Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World Journal of Gastroenterology*, *20*(4), 943–956. https://doi.org/10.3748/wjg.v20.i4.943

Cox, D. B. T., Platt, R. J., & Zhang, F. (2015). Therapeutic genome editing: prospects and challenges. *Nature Medicine*, *21*(2), 121–131. https://doi.org/10.1038/nm.3793

Cozzolino, F., Iacobucci, I., Monaco, V., & Monti, M. (2021). Protein-DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era. *Journal of Proteome Research*, *20*(6), 3018–3030. https://doi.org/10.1021/acs.jproteome.1c00074

Craik, D. J., Fairlie, D. P., Liras, S., & Price, D. (2013). The future of peptide-based drugs. *Chemical Biology & Drug Design*, *81*(1), 136–147. https://doi.org/10.1111/cbdd.12055

Csermely, P., Korcsmaros, T., Kiss, H. J. M., London, G., & Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, *138*(3), 333–408. https://doi.org/10.1016/j.pharmthera.2013.01.016

Dasberg, H. (1991). Why we were silent--an Israeli psychiatrist speaks to Germans on psychic pain and past persecution. *The Israel Journal of Psychiatry and Related Sciences*, *28*(2),

29–38.

Davis, G. L., Dempster, J., Meler, J. D., Orr, D. W., Walberg, M. W., Brown, B., Goldstein, R. M. (2008). Hepatocellular carcinoma: management of an increasingly common problem. *Proceedings (Baylor University. Medical Center)*, *21*(3), 266–280. https://doi.org/10.1080/08998280.2008.11928410

De Angelis, C., Byrne, A. B., Morrow, R., Feng, J., Ha, T., Wang, P., Barnett, C. (2021). Compound heterozygous variants in LAMC3 in association with posterior periventricular nodular heterotopia. *BMC Medical Genomics*, *14*(1), 64. https://doi.org/10.1186/s12920-021-00911-4

de la Cruz, X., Hutchinson, E. G., Shepherd, A., & Thornton, J. M. (2002). Toward predicting protein topology: an approach to identifying beta hairpins. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11157–11162. https://doi.org/10.1073/pnas.162376199

Deng, L., Yang, W., & Liu, H. (2019). PredPRBA: Prediction of Protein-RNA Binding Affinity Using Gradient Boosted Regression Trees. *Frontiers in Genetics*, *10*, 637. https://doi.org/10.3389/fgene.2019.00637

Dhall, A., Patiyal, S., Kaur, H., Bhalla, S., Arora, C., & Raghava, G. P. S. (2020). Computing Skin Cutaneous Melanoma Outcome From the HLA-Alleles and Clinical Characteristics. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00221

Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. S. (2020). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Briefings in Bioinformatics*, *2020*(00), 1–10. https://doi.org/10.1093/bib/bbaa259

Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. S. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Briefings in Bioinformatics*, *22*(2), 936–945. https://doi.org/10.1093/bib/bbaa259

Dhanda, S. K., Gupta, S., Vir, P., & Raghava, G. P. S. (2013). Prediction of IL4 Inducing Peptides. *Clinical and Developmental Immunology*, *2013*. https://doi.org/10.1155/2013/263952

Dhanda, S. K., Vir, P., & Raghava, G. P. S. (2013). Designing of interferon-gamma inducing MHC class-II binders. *Biology Direct*, *8*(1). https://doi.org/10.1186/1745-6150-8-30

Didiasova, M., Schaefer, L., & Wygrecka, M. (2018). Targeting GLI Transcription Factors in Cancer. *Molecules (Basel, Switzerland)*, *23*(5). https://doi.org/10.3390/molecules23051003

do Valle, I. F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., Castellani, G. (2016). Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics*, *17*(Suppl 12), 341. https://doi.org/10.1186/s12859-016-1190-7

Dong, J., Yao, Z.-J., Wen, M., Zhu, M.-F., Wang, N.-N., Miao, H.-Y., Cao, D.-S. (2016). BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions. *Journal of Cheminformatics*, *8*, 34. https://doi.org/10.1186/s13321-016-0146-2

Dong, J., Yao, Z.-J., Zhang, L., Luo, F., Lin, Q., Lu, A.-P., Cao, D.-S. (2018). PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *Journal of Cheminformatics*, *10*(1), 16. https://doi.org/10.1186/s13321-018-0270-2

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Liu, S.-Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *International Journal of Molecular Sciences*, *17*(2). https://doi.org/10.3390/ijms17020144

Dukka, B. K. (2013). Structure-based Methods for Computational Protein Functional Site Prediction. *Computational and Structural Biotechnology Journal*, *8*, e201308005. https://doi.org/10.5936/csbj.201308005

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research*, *34*(Web Server issue), W116-8. https://doi.org/10.1093/nar/gkl282

Eichner, J., Topf, F., Drager, A., Wrzodek, C., Wanke, D., & Zell, A. (2013). TFpredict and SABINE: sequence-based prediction of structural and functional characteristics of transcription factors. *PloS One*, *8*(12), e82238. https://doi.org/10.1371/journal.pone.0082238

Emamjomeh, A., Choobineh, D., Hajieghrari, B., MahdiNezhad, N., & Khodavirdipour, A. (2019). DNA-protein interaction: identification, prediction and data analysis. *Molecular*

Biology Reports, *46*(3), 3571–3596. https://doi.org/10.1007/s11033-019-04763-1

Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *Journal of Molecular Biology*, *300*(4), 1005–1016. https://doi.org/10.1006/jmbi.2000.3903

Ercolini, A. M., & Miller, S. D. (2009). The role of infections in autoimmune disease. *Clinical and Experimental Immunology*, *155*(1), 1–15. https://doi.org/10.1111/j.1365-2249.2008.03834.x

Fan, Y., Xi, L., Hughes, D. S. T., Zhang, J., Zhang, J., Futreal, P. A., Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, *17*(1), 178. https://doi.org/10.1186/s13059-016-1029-6

Farazi, P. A., & DePinho, R. A. (2006). Hepatocellular carcinoma pathogenesis: from genes to environment. *Nature Reviews. Cancer*, *6*(9), 674–687. https://doi.org/10.1038/nrc1934

Ferguson, R. L., & Allen, B. L. J. (1988). Considerations in the treatment of cerebral palsy patients with spinal deformities. *The Orthopedic Clinics of North America*, *19*(2), 419–425.

Finn, R. D., Miller, B. L., Clements, J., & Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*, *42*(Database issue), D364-73. https://doi.org/10.1093/nar/gkt1210

Fong, A. P., & Tapscott, S. J. (2013). Skeletal muscle programming and re-programming. *Current Opinion in Genetics & Development*, *23*(5), 568–573. https://doi.org/10.1016/j.gde.2013.05.002

Freeman, T. C., & Wimley, W. C. (2012). TMBB-DB: a transmembrane β-barrel proteome database. *Bioinformatics (Oxford, England)*, *28*(19), 2425–2430. https://doi.org/10.1093/bioinformatics/bts478

Fuchs, P. F. J., & Alix, A. J. P. (2005). High accuracy prediction of beta-turns and their types using propensities and multiple alignments. *Proteins*, *59*(4), 828–839. https://doi.org/10.1002/prot.20461

Fukunishi, Y., & Nakamura, H. (2011). Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Science : A*

Publication of the Protein Society, *20*(1), 95–106. https://doi.org/10.1002/pro.540

Furlan-Magaril, M., Rincon-Arano, H., & Recillas-Targa, F. (2009). Sequential chromatin immunoprecipitation protocol: ChIP-reChIP. *Methods in Molecular Biology (Clifton, N.J.)*, *543*, 253–266. https://doi.org/10.1007/978-1-60327-015-1_17

Gallina, A. M., Bork, P., & Bordo, D. (2014). Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs. *Journal of Molecular Recognition : JMR*, *27*(2), 65–72. https://doi.org/10.1002/jmr.2332

Gangloff, S., Soustelle, C., & Fabre, F. (2000). Homologous recombination is responsible for cell death in the absence of the Sgs1 and Srs2 helicases. *Nature Genetics*, *25*(2), 192–194. https://doi.org/10.1038/76055

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., Luo, J. (2006). DRTF: a database of rice transcription factors. *Bioinformatics (Oxford, England)*, *22*(10), 1286–1287. https://doi.org/10.1093/bioinformatics/btl107

Gao, Y., Wells, L., Comer, F. I., Parker, G. J., & Hart, G. W. (2001). Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic beta-N-acetylglucosaminidase from human brain. *The Journal of Biological Chemistry*, *276*(13), 9838–9845. https://doi.org/10.1074/jbc.M010420200

Garg, A., Bhasin, M., & Raghava, G. P. S. (2005). Support Vector Machine-based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order, and Similarity Search. *Journal of Biological Chemistry*, *280*(15), 14427–14432. https://doi.org/10.1074/jbc.M411789200

Garg, A., & Raghava, G. P. S. (2008). ESLpred2: Improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-503

Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., & Raghava, G. P. S. (2013). In silico approaches for designing highly effective cell penetrating peptides. *Journal of Translational Medicine*, *11*(1), 74. https://doi.org/10.1186/1479-5876-11-74

Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, *366*(10), 883–892. https://doi.org/10.1056/NEJMoa1113205

Goodwin, K. D., Long, E. C., & Georgiadis, M. M. (2005). A host-guest approach for determining drug-DNA interactions: an example using netropsin. *Nucleic Acids Research*, *33*(13), 4106–4116. https://doi.org/10.1093/nar/gki717

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *The New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Gunasekera, A., Alvarez, F. J., Douglas, L. M., Wang, H. X., Rosebrock, A. P., & Konopka, J. B. (2010). Identification of GIG1, a GlcNAc-induced gene in Candida albicans needed for normal sensitivity to the chitin synthase inhibitor nikkomycin Z. *Eukaryotic Cell*, *9*(10), 1476–1483. https://doi.org/10.1128/EC.00178-10

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Open Source Drug Discovery Consortium, G. P. S., & Raghava, G. P. S. (2013). In silico approach for predicting toxicity of peptides and proteins. *PloS One*, *8*(9), e73957. https://doi.org/10.1371/journal.pone.0073957

Guruprasad, K., & Rajkumar, S. (2000). Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *Journal of Biosciences*, *25*(2), 143–156.

Han, Z., Zhuang, X., Yang, B., Jin, L., Hong, P., Xue, J., Tian, Z. (2021). SYDE1 Acts as an Oncogene in Glioma and has Diagnostic and Prognostic Values. *Frontiers in Molecular Biosciences*, *8*, 714203. https://doi.org/10.3389/fmolb.2021.714203

Hart, G. W., Slawson, C., Ramirez-Correa, G., & Lagerlof, O. (2011). Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annual Review of Biochemistry*, *80*, 825–858. https://doi.org/10.1146/annurev-biochem-060608-102511

Hartley, T., Wagner, J. D., Warman-Chardon, J., Tetreault, M., Brady, L., Baker, S., Boycott, K. M. (2018). Whole-exome sequencing is a valuable diagnostic tool for inherited peripheral neuropathies: Outcomes from a cohort of 50 families. *Clinical Genetics*, *93*(2), 301–309. https://doi.org/10.1111/cge.13101

Hayden, M. S., & Ghosh, S. (2012). NF-kappaB, the first quarter-century: remarkable progress and outstanding questions. *Genes & Development*, *26*(3), 203–234.

https://doi.org/10.1101/gad.183434.111

Hayes, C. S., Alarcon-Hernandez, E., & Setlow, P. (2001). N-terminal amino acid residues mediate protein-protein interactions between DNA-bound alpha /beta -type small, acid-soluble spore proteins from Bacillus species. *The Journal of Biological Chemistry*, *276*(3), 2267–2275. https://doi.org/10.1074/jbc.M007858200

He, J., Liu, Y., Zhang, L., & Zhang, H. (2018). Integrin Subunit beta 8 (ITGB8) Upregulation Is an Independent Predictor of Unfavorable Survival of High-Grade Serous Ovarian Carcinoma Patients. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *24*, 8933–8940. https://doi.org/10.12659/MSM.911518

Heo, L., Shin, W.-H., Lee, M. S., & Seok, C. (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Research*, *42*(Web Server issue), W210-4. https://doi.org/10.1093/nar/gku321

Hirsch, F. R., Scagliotti, G. V, Mulshine, J. L., Kwon, R., Curran, W. J. J., Wu, Y.-L., & Paz-Ares, L. (2017). Lung cancer: current therapies and new targeted treatments. *Lancet (London, England)*, *389*(10066), 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8

Ho, S.-W., Jona, G., Chen, C. T. L., Johnston, M., & Snyder, M. (2006). Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9940–9945. https://doi.org/10.1073/pnas.0509185103

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, *4*(11), 682–690. https://doi.org/10.1038/nchembio.118

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, *35*(Web Server), W585–W587. https://doi.org/10.1093/nar/gkm259

Hu, J., Li, Y., Zhang, Y., & Yu, D.-J. (2018). ATPbind: Accurate Protein-ATP Binding Site Prediction by Combining Sequence-Profiling and Structure-Based Comparisons. *Journal of Chemical Information and Modeling*, *58*(2), 501–510. https://doi.org/10.1021/acs.jcim.7b00397

Hu, J., Zheng, L.-L., Bai, Y.-S., Zhang, K.-W., Yu, D.-J., & Zhang, G.-J. (2021). Accurate

prediction of protein-ATP binding residues using position- specific frequency matrix. *Analytical Biochemistry*, *626*, 114241. https://doi.org/10.1016/j.ab.2021.114241

Hu, X., Dong, Q., Yang, J., & Zhang, Y. (2016). Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics (Oxford, England)*, *32*(21), 3260–3269. https://doi.org/10.1093/bioinformatics/btw396

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics (Oxford, England)*, *26*(5), 680–682. https://doi.org/10.1093/bioinformatics/btq003

Huh, H. D., Kim, D. H., Jeong, H.-S., & Park, H. W. (2019). Regulation of TEAD Transcription Factors in Cancer Biology. *Cells*, *8*(6). https://doi.org/10.3390/cells8060600

Hwang, S., Gou, Z., & Kuznetsov, I. B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics (Oxford, England)*, *23*(5), 634–636. https://doi.org/10.1093/bioinformatics/btl672

Idda, M. L., Munk, R., Abdelmohsen, K., & Gorospe, M. (2018). Noncoding RNAs in Alzheimer's disease. *Wiley Interdisciplinary Reviews. RNA*, *9*(2). https://doi.org/10.1002/wrna.1463

Islam, Z., Ali, A. M., Naik, A., Eldaw, M., Decock, J., & Kolatkar, P. R. (2021). Transcription Factors: The Fulcrum Between Cell Development and Carcinogenesis. *Frontiers in Oncology*, *11*, 681377. https://doi.org/10.3389/fonc.2021.681377

Jahandideh, S., Sarvestani, A. S., Abdolmaleki, P., Jahandideh, M., & Barfeie, M. (2007). gamma-Turn types prediction in proteins using the support vector machines. *Journal of Theoretical Biology*, *249*(4), 785–790. https://doi.org/10.1016/j.jtbi.2007.09.002

Jain, D. S., Gupte, S. R., & Aduri, R. (2018). A Data Driven Model for Predicting RNA-Protein Interactions based on Gradient Boosting Machine. *Scientific Reports*, *8*(1), 9552. https://doi.org/10.1038/s41598-018-27814-2

Jayaram, B., McConnell, K., Dixit, S. B., Das, A., & Beveridge, D. L. (2002). Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. *Journal of Computational Chemistry*, *23*(1), 1–14. https://doi.org/10.1002/jcc.10009

Jenner, R. G., Townsend, M. J., Jackson, I., Sun, K., Bouwman, R. D., Young, R. A., Lord, G. M. (2009). The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17876–17881. https://doi.org/10.1073/pnas.0909357106

Jin, J., Wu, X., Yin, J., Li, M., Shen, J., Li, J., Qu, L. (2019). Identification of Genetic Mutations in Cancer: Challenge and Opportunity in the New Era of Targeted Therapy. *Frontiers in Oncology*, *9*, 263. https://doi.org/10.3389/fonc.2019.00263

Jiramongkol, Y., & Lam, E. W.-F. (2020). FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Reviews*, *39*(3), 681–709. https://doi.org/10.1007/s10555-020-09883-w

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), W5-9. https://doi.org/10.1093/nar/gkn201

Jones, S, Daley, D. T., Luscombe, N. M., Berman, H. M., & Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Research*, *29*(4), 943–954. https://doi.org/10.1093/nar/29.4.943

Jones, S, van Heyningen, P., Berman, H. M., & Thornton, J. M. (1999). Protein-DNA interactions: A structural analysis. *Journal of Molecular Biology*, *287*(5), 877–896. https://doi.org/10.1006/jmbi.1999.2659

Jones, Susan, Barker, J. A., Nobeli, I., & Thornton, J. M. (2003). Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Research*, *31*(11), 2811–2823. https://doi.org/10.1093/nar/gkg386

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalita, M. K., Nandal, U. K., Pattnaik, A., Sivalingam, A., Ramasamy, G., Kumar, M., Gupta, D. (2008). CyclinPred: A SVM-Based Method for Predicting Cyclin Protein Sequences. *PLoS ONE*, *3*(7), e2605. https://doi.org/10.1371/journal.pone.0002605

Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., & Suzuki, H. (2004). A genome-wide and nonredundant mouse transcription factor database. *Biochemical and*

Biophysical Research Communications, *322*(3), 787–793. https://doi.org/10.1016/j.bbrc.2004.07.179

Kaubryte, J., & Lai, A. G. (2022). Pan-cancer prognostic genetic mutations and clinicopathological factors associated with survival outcomes: a systematic review. *NPJ Precision Oncology*, *6*(1), 27. https://doi.org/10.1038/s41698-022-00269-5

Kaundal, R., & Raghava, G. P. S. (2009). RSLpred: An integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. *Proteomics*, *9*(9), 2324–2342. https://doi.org/10.1002/pmic.200700597

Kaur, D., Arora, C., & Raghava, G. P. S. (2020). A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.00071

Kaur, H., & Raghava, G. P. S. (2003). Prediction of beta-turns in proteins from multiple alignment using neural network. *Protein Science*, *12*(3), 627–634. https://doi.org/10.1110/ps.0228903

Kaur, Harpreet, & Raghava, G. P. S. (2004). Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins*, *55*(1), 83–90. https://doi.org/10.1002/prot.10569

Kemmeren, P., Sameith, K., van de Pasch, L. A. L., Benschop, J. J., Lenstra, T. L., Margaritis, T., Holstege, F. C. P. (2014). Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, *157*(3), 740–752. https://doi.org/10.1016/j.cell.2014.02.054

Khalil, A. M., & Rinn, J. L. (2011). RNA-protein interactions in human health and disease. *Seminars in Cell & Developmental Biology*, *22*(4), 359–365. https://doi.org/10.1016/j.semcdb.2011.02.016

Kilic, S., White, E. R., Sagitova, D. M., Cornish, J. P., & Erill, I. (2014). CollecTF: a database of experimentally validated transcription factor- binding sites in Bacteria. *Nucleic Acids Research*, *42*(Database issue), D156-60. https://doi.org/10.1093/nar/gkt1123

Kim, G. B., Gao, Y., Palsson, B. O., & Lee, S. Y. (2021). DeepTFactor: A deep learning-based tool for the prediction of transcription factors. *Proceedings of the National Academy of Sciences*, *118*(2). https://doi.org/10.1073/PNAS.2021171118

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Kallberg, M., Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, *10*(1), 3583. https://doi.org/10.1038/s41467-019-11526-w

Kishtagari, A., Levine, R. L., & Viny, A. D. (2020). Driver mutations in acute myeloid leukemia. *Current Opinion in Hematology*, *27*(2), 49–57. https://doi.org/10.1097/MOH.0000000000000567

Kleinjan, D. A., & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, *76*(1), 8–32. https://doi.org/10.1086/426833

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Kountouris, P., & Hirst, J. D. (2010). Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics*, *11*(1), 407. https://doi.org/10.1186/1471-2105-11-407

Kumar, K. K., Pugalenthi, G., & Suganthan, P. N. (2009). DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. *Journal of Biomolecular Structure & Dynamics*, *26*(6), 679–686. https://doi.org/10.1080/07391102.2009.10507281

Kumar, M., Gromiha, M. M., & Raghava, G. P. S. (2007). Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, *8*(1), 463. https://doi.org/10.1186/1471-2105-8-463

Kumar, M., Gromiha, M. M., & Raghava, G. P. S. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, *71*(1), 189–194. https://doi.org/10.1002/prot.21677

Kumar, M., Thakur, V., & Raghava, G. P. S. (2008). COPid: composition based protein

identification. *In Silico Biology*, *8*(2), 121–128.

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., & Raghava, G. P. S. (2015). An in silico platform for predicting, screening and designing of antihypertensive peptides. *Scientific Reports*, *5*(1), 12512. https://doi.org/10.1038/srep12512

Kumar, V., Agrawal, P., Kumar, R., Bhalla, S., Usmani, S. S., Varshney, G. C., & Raghava, G. P. S. (2018). Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. *Frontiers in Microbiology*, *9*(APR). https://doi.org/10.3389/fmicb.2018.00725

Kuznetsov, I. B., Gou, Z., Li, R., & Hwang, S. (2006). Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins: Structure, Function, and Bioinformatics*, *64*(1), 19–27. https://doi.org/10.1002/prot.20977

Kwiatkowski, T. J. J., Bosco, D. A., Leclerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Brown, R. H. J. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science (New York, N.Y.)*, *323*(5918), 1205–1208. https://doi.org/10.1126/science.1166066

LaDuca, H., Farwell, K. D., Vuong, H., Lu, H.-M., Mu, W., Shahmirzadi, L., Chao, E. C. (2017). Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. *PloS One*, *12*(2), e0170843. https://doi.org/10.1371/journal.pone.0170843

Lambert, M., Jambon, S., Depauw, S., & David-Cordonnier, M.-H. (2018). Targeting Transcription Factors for Cancer Treatment. *Molecules (Basel, Switzerland)*, *23*(6). https://doi.org/10.3390/molecules23061479

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Lan, T., Yuan, K., Yan, X., Xu, L., Liao, H., Hao, X., Wu, H. (2019). LncRNA SNHG10 Facilitates Hepatocarcinogenesis and Metastasis by Modulating Its Homolog SCARNA13 via a Positive Feedback Loop. *Cancer Research*, *79*(13), 3220–3234. https://doi.org/10.1158/0008-5472.CAN-18-4044

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ding, L.

(2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, *28*(3), 311–317. https://doi.org/10.1093/bioinformatics/btr665

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, *10*(1), 168. https://doi.org/10.1186/1471-2105-10-168

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Lei, S. M., Liu, X., Xia, L. P., Ke, Y., Wei, L. W., Li, L., & Yin, F. J. (2021). [Relationships between decreased LAMC3 and poor prognosis in ovarian cancer]. *Zhonghua fu chan ke za zhi*, *56*(7), 489–497. https://doi.org/10.3760/cma.j.cn112141-20210426-00230

Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A., & Brasseur, R. (2005). Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, *61*(2), 258–271. https://doi.org/10.1002/prot.20607

Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., & Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, *36*(8), 815–822. https://doi.org/10.1002/humu.22813

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, B.-Q., Feng, K.-Y., Ding, J., & Cai, Y.-D. (2014). Predicting DNA-binding sites of proteins based on sequential and 3D structural information. *Molecular Genetics and Genomics : MGG*, *289*(3), 489–499. https://doi.org/10.1007/s00438-014-0812-x

Li, H., Yang, Y., Hong, W., Huang, M., Wu, M., & Zhao, X. (2020). Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, *5*(1), 1. https://doi.org/10.1038/s41392-019-0089-y

Li, P., Spolski, R., Liao, W., & Leonard, W. J. (2014). Complex interactions of transcription factors in mediating cytokine biology in T cells. *Immunological Reviews*, *261*(1), 141–156. https://doi.org/10.1111/imr.12199

Li, Z. R., Lin, H. H., Han, L. Y., Jiang, L., Chen, X., & Chen, Y. Z. (2006). PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *34*(Web Server issue), W32-7. https://doi.org/10.1093/nar/gkl305

Liang, M., Wang, L., Cao, C., Song, S., & Wu, F. (2020). LncRNA SNHG10 is downregulated in non-small cell lung cancer and predicts poor survival. *BMC Pulmonary Medicine*, *20*(1), 273. https://doi.org/10.1186/s12890-020-01281-w

Lim, S. M., Westover, K. D., Ficarro, S. B., Harrison, R. A., Choi, H. G., Pacold, M. E., Gray, N. S. (2014). Therapeutic targeting of oncogenic K-Ras by a covalent catalytic site inhibitor. *Angewandte Chemie (International Ed. in English)*, *53*(1), 199–204. https://doi.org/10.1002/anie.201307387

Lin, B., & Pang, Z. (2019). Stability of methods for differential expression analysis of RNA-seq data. *BMC Genomics*, *20*(1), 35. https://doi.org/10.1186/s12864-018-5390-6

Lin, L., Yan, L., Liu, Y., Qu, C., Ni, J., & Li, H. (2020). The Burden and Trends of Primary Liver Cancer Caused by Specific Etiologies from 1990 to 2017 at the Global, Regional, National, Age, and Sex Level Results from the Global Burden of Disease Study 2017. *Liver Cancer*, *9*(5), 563–582. https://doi.org/10.1159/000508568

Lin, W.-Z., Fang, J.-A., Xiao, X., & Chou, K.-C. (2011). iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PloS One*, *6*(9), e24756. https://doi.org/10.1371/journal.pone.0024756

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings in Bioinformatics*, *20*(4), 1280–1294. https://doi.org/10.1093/bib/bbx165

Liu, B., Gao, X., & Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Research*, *47*(20), e127. https://doi.org/10.1093/nar/gkz740

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., & Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research*, *43*(W1), W65-71. https://doi.org/10.1093/nar/gkv458

Liu, B. , Wu, H. and Chou, K. (2017) Pse-in-One 2.0: An improved package of web servers for

generating various modes of pseudo components of DNA, RNA, and protein sequences. Natural Science, 9, 67-91. https://10.4236/ns.2017.94007

Liu, B., Wang, S., & Wang, X. (2015). DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Scientific Reports*, *5*, 15479. https://doi.org/10.1038/srep15479

Liu, R., & Hu, J. (2013). DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins*, *81*(11), 1885–1899. https://doi.org/10.1002/prot.24330

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., & Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics (Oxford, England)*, *26*(13), 1616–1622. https://doi.org/10.1093/bioinformatics/btq253

Loeb, L. A., Loeb, K. R., & Anderson, J. P. (2003). Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 776–781. https://doi.org/10.1073/pnas.0334858100

Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., & Zhang, H. (2014). Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. *PloS One*, *9*(1), e86703. https://doi.org/10.1371/journal.pone.0086703

Lu, W.-C., Xie, H., Yuan, C., Li, J.-J., Li, Z.-Y., & Wu, A.-H. (2020). Identification of potential biomarkers and candidate small molecule drugs in glioblastoma. *Cancer Cell International*, *20*, 419. https://doi.org/10.1186/s12935-020-01515-1

Luo, J., Liu, L., Venkateswaran, S., Song, Q., & Zhou, X. (2017). RPI-Bind: a structure-based method for accurate identification of RNA- protein binding sites. *Scientific Reports*, *7*(1), 614. https://doi.org/10.1038/s41598-017-00795-4

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, *350*(21), 2129–2139. https://doi.org/10.1056/NEJMoa040938

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Micklem, G. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biology*, *8*(7), R129. https://doi.org/10.1186/gb-2007-8-7-r129

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., & Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, *8*(44), 77121–77136. https://doi.org/10.18632/oncotarget.20365

Manavalan, B., Basith, S., Shin, T. H., Wei, L., & Lee, G. (2018). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics (Oxford, England)*. https://doi.org/10.1093/bioinformatics/bty1047

Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M. O., & Lee, G. (2018). iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. *Frontiers in Immunology*, *9*, 1695. https://doi.org/10.3389/fimmu.2018.01695

Manavalan, B., & Patra, M. C. (2022). MLCPP 2.0: An Updated Cell-penetrating Peptides and Their Uptake Efficiency Predictor. *Journal of Molecular Biology*, *434*(11), 167604. https://doi.org/10.1016/j.jmb.2022.167604

Manavalan, B., Shin, T. H., Kim, M. O., & Lee, G. (2018a). AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. *Frontiers in Pharmacology*, *9*, 276. https://doi.org/10.3389/fphar.2018.00276

Manavalan, B., Shin, T. H., Kim, M. O., & Lee, G. (2018b). PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. *Frontiers in Immunology*, *9*, 1783. https://doi.org/10.3389/fimmu.2018.01783

Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., & Lee, G. (2018). Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *Journal of Proteome Research*, *17*(8), 2715–2726. https://doi.org/10.1021/acs.jproteome.8b00148

Mariani, L., Lohning, M., Radbruch, A., & Hofer, T. (2004). Transcriptional control networks of cell differentiation: insights from helper T lymphocytes. *Progress in Biophysics and Molecular Biology*, *86*(1), 45–76. https://doi.org/10.1016/j.pbiomolbio.2004.02.007

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

Miao, Z., & Westhof, E. (2015). A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs. *PLoS Computational Biology*, *11*(12), e1004639.

https://doi.org/10.1371/journal.pcbi.1004639

Miao, Z., & 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), W562-7. https://doi.org/10.1093/nar/gkw251

Miles, B., & Tadi, P. (2022). Genetics, Somatic Mutation. Treasure Island (FL).

Miraghazadeh, B., & Cook, M. C. (2018). Nuclear Factor-kappaB in Autoimmunity: Man and Mouse. *Frontiers in Immunology*, *9*, 613. https://doi.org/10.3389/fimmu.2018.00613

Mishra, A., Pokhrel, P., & Hoque, M. T. (2019). StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics (Oxford, England)*, *35*(3), 433–441. https://doi.org/10.1093/bioinformatics/bty653

Mishra, N. K., & Raghava, G. P. S. (2010a). Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. *BMC Bioinformatics*, *11 Suppl 1*(Suppl 1), S48. https://doi.org/10.1186/1471-2105-11-S1-S48

Mishra, N. K., & Raghava, G. P. S. (2010b). Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. *BMC Bioinformatics*, *11 Suppl 1*(S1), S48. https://doi.org/10.1186/1471-2105-11-S1-S48

Miyazaki, K., & Miyazaki, M. (2021). The Interplay Between Chromatin Architecture and Lineage-Specific Transcription Factors and the Regulation of Rag Gene Expression. *Frontiers in Immunology*, *12*, 659761. https://doi.org/10.3389/fimmu.2021.659761

Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., Teixeira, M. C. (2020). YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Research*, *48*(D1), D642–D649. https://doi.org/10.1093/nar/gkz859

Monti, P., Menichini, P., Speciale, A., Cutrona, G., Fais, F., Taiana, E., Fronza, G. (2020). Heterogeneity of TP53 Mutations and P53 Protein Residual Function in Cancer: Does It Matter? *Frontiers in Oncology*, *10*, 593383. https://doi.org/10.3389/fonc.2020.593383

Mooney, C., Wang, Y.-H., & Pollastri, G. (2011). SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics (Oxford, England)*, *27*(20), 2812–2819. https://doi.org/10.1093/bioinformatics/btr494

Moravek, Z., Neidle, S., & Schneider, B. (2002). Protein and drug interactions in the minor

groove of DNA. *Nucleic Acids Research*, *30*(5), 1182–1191. https://doi.org/10.1093/nar/30.5.1182

Moussian, B. (2008). The role of GlcNAc in formation and function of extracellular matrices. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *149*(2), 215–226. https://doi.org/10.1016/j.cbpb.2007.10.009

Muhammad, I. I., Kong, S. L., Akmar Abdullah, S. N., & Munusamy, U. (2019). RNA-seq and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional Regulatory Mechanism. *International Journal of Molecular Sciences*, *21*(1). https://doi.org/10.3390/ijms21010167

Muhammod, R., Ahmed, S., Md Farid, D., Shatabda, S., Sharma, A., & Dehzangi, A. (2019). PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. *Bioinformatics (Oxford, England)*, *35*(19), 3831–3833. https://doi.org/10.1093/bioinformatics/btz165

Muller, M., Bird, T. G., & Nault, J.-C. (2020). The landscape of gene mutations in cirrhosis and hepatocellular carcinoma. *Journal of Hepatology*, *72*(5), 990–1002. https://doi.org/10.1016/j.jhep.2020.01.019

Munro, S., & Pelham, H. R. (1987). A C-terminal signal prevents secretion of luminal ER proteins. *Cell*, *48*(5), 899–907. https://doi.org/10.1016/0092-8674(87)90086-9

Munsky, B., Neuert, G., & van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science (New York, N.Y.)*, *336*(6078), 183–187. https://doi.org/10.1126/science.1216379

Muppirala, U. K., Honavar, V. G., & Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, *12*, 489. https://doi.org/10.1186/1471-2105-12-489

Muthukrishnan, S., Garg, A., & Raghava, G. P. S. (2007). Oxypred: prediction and classification of oxygen-binding proteins. *Genomics, Proteomics & Bioinformatics*, *5*(3–4), 250–252. https://doi.org/10.1016/S1672-0229(08)60012-1

Nadassy, K., Wodak, S. J., & Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry*, *38*(7), 1999–2017. https://doi.org/10.1021/bi982362d

Nagarajan, R., Ahmad, S., & Gromiha, M. M. (2013). Novel approach for selecting the best

predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Research*, *41*(16), 7606–7614. https://doi.org/10.1093/nar/gkt544

Nagarajan, R., Archana, A., Thangakani, A. M., Jemimah, S., Velmurugan, D., & Gromiha, M. M. (2016). PDBparam: Online Resource for Computing Structural Parameters of Proteins. *Bioinformatics and Biology Insights*, *10*, 73–80. https://doi.org/10.4137/BBI.S38423

Nagel, K., Jimeno-Yepes, A., & Rebholz-Schuhmann, D. (2009). Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb. *BMC Bioinformatics*, *10 Suppl 8*(S8), S4. https://doi.org/10.1186/1471-2105-10-S8-S4

Nagpal, G., Usmani, S. S., Dhanda, S. K., Kaur, H., Singh, S., Sharma, M., & Raghava, G. P. S. (2017). Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Scientific Reports*, *7*(1), 42851. https://doi.org/10.1038/srep42851

Nalejska, E., Maczynska, E., & Lewandowska, M. A. (2014). Prognostic and predictive biomarkers: tools in personalized oncology. *Molecular Diagnosis & Therapy*, *18*(3), 273–284. https://doi.org/10.1007/s40291-013-0077-9

Naseem, S., Parrino, S. M., Buenten, D. M., & Konopka, J. B. (2012). Novel roles for GlcNAc in cell signaling. *Communicative & Integrative Biology*, *5*(2), 156–159. https://doi.org/10.4161/cib.19034

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012). Host-gut microbiota metabolic interactions. *Science (New York, N.Y.)*, *336*(6086), 1262–1267. https://doi.org/10.1126/science.1223813

Nilchian, A., Johansson, J., Ghalali, A., Asanin, S. T., Santiago, A., Rosencrantz, O., Fuxe, J. (2019). CXADR-Mediated Formation of an AKT Inhibitory Signalosome at Tight Junctions Controls Epithelial-Mesenchymal Plasticity in Breast Cancer. *Cancer Research*, *79*(1), 47–60. https://doi.org/10.1158/0008-5472.CAN-18-1742

Nimrod, G., Schushan, M., Szilagyi, A., Leslie, C., & Ben-Tal, N. (2010). iDBPs: a web server for the identification of DNA binding proteins. *Bioinformatics (Oxford, England)*, *26*(5), 692–693. https://doi.org/10.1093/bioinformatics/btq019

Odom, D. T. (2011). Identification of Transcription Factor-DNA Interactions In Vivo. *Sub-Cellular Biochemistry*, *52*, 175–191. https://doi.org/10.1007/978-90-481-9069-0_8

Ofran, Y., Mysore, V., & Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics (Oxford, England)*, *23*(13), i347-53. https://doi.org/10.1093/bioinformatics/btm174

Oliveira Monteiro, L. M., Saraiva, J. P., Brizola Toscan, R., Stadler, P. F., Silva-Rocha, R., & Nunes da Rocha, U. (2022). PredicTF: prediction of bacterial transcription factors in complex microbial communities using deep learning. *Environmental Microbiome*, *17*(1), 7. https://doi.org/10.1186/s40793-021-00394-x

Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor Perspectives in Biology*, *2*(1), a001008. https://doi.org/10.1101/cshperspect.a001008

Ortet, P., De Luca, G., Whitworth, D. E., & Barakat, M. (2012). P2TF: a comprehensive resource for analysis of prokaryotic transcription factors. *BMC Genomics*, *13*, 628. https://doi.org/10.1186/1471-2164-13-628

Ozdemir, Y., Cag, M., Colak, E., Coskun, N., Basgoz, N., Sarici, H., Ozkul, Y. (2021). The Effect of Gene Mutations on Metastasis and Overall Survival in Metastatic and Nonmetastatic Colon Cancers. *Asian Pacific Journal of Cancer Prevention : APJCP*, *22*(12), 3839–3846. https://doi.org/10.31557/APJCP.2021.22.12.3839

Pan, X., Rijnbeek, P., Yan, J., & Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics*, *19*(1), 511. https://doi.org/10.1186/s12864-018-4889-1

Pande, A., Patiyal S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Raghava, G. P. S. (2019). Computing wide range of protein/peptide features from their sequence and. *BioRxiv*, 599126. https://doi.org/10.1101/599126

Panwar, B., Gupta, S., & Raghava, G. P. S. (2013). Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinformatics*, *14*(1), 44. https://doi.org/10.1186/1471-2105-14-44

Park, J. T., & Uehara, T. (2008). How bacteria consume their own exoskeletons (turnover and recycling of cell wall peptidoglycan). *Microbiology and Molecular Biology Reviews : MMBR*, *72*(2), 211–227, table of contents. https://doi.org/10.1128/MMBR.00027-07

Patiyal, S., Agrawal, P., Kumar, V., Dhall, A., Kumar, R., Mishra, G., & Raghava, G. P. S. (2020). NAGbinder: An approach for identifying N-acetylglucosamine interacting

residues of a protein from its primary sequence. *Protein Science*, *29*(1), 201–210. https://doi.org/10.1002/pro.3761

Patiyal, S., Dhall, A., & Raghava, G. P. S. (2021). DBpred: A deep learning method for the prediction of DNA interacting residues in protein sequences. BioRxiv. https://doi.org/10.1101/2021.08.05.455224

Patiyal, S., Dhall, A., Bajaj, K., Sahu, H., & Raghava, G. P. S. (2022). Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. BioRxiv, 2022.06.03.494705. https://doi.org/10.1101/2022.06.03.494705

Patiyal, S., Dhall, A., & Raghava, G. P. S. (2022). Prediction of risk-associated genes and high-risk liver cancer patients from their mutation profile: Benchmarking of mutation calling techniques. Biology Methods and Protocols.

Patiyal, S., Tiwari, P., Ghai, M., Dhapola, A., Dhall, A., & Raghava, G. P. S. (2022). A hybrid approach for predicting transcription factors. BioRxiv, 2022.07.13.499865. https://doi.org/10.1101/2022.07.13.499865

Pattnaik, A., Palermo, N., Sahoo, B. R., Yuan, Z., Hu, D., Annamalai, A. S., Xiang, S.-H. (2018). Discovery of a non-nucleoside RNA polymerase inhibitor for blocking Zika virus replication through in silico screening. *Antiviral Research*, *151*, 78–86. https://doi.org/10.1016/j.antiviral.2017.12.016

Paul, M. R., Pan, T.-C., Pant, D. K., Shih, N. N., Chen, Y., Harvey, K. L., Chodosh, L. A. (2020). Genomic landscape of metastatic breast cancer identifies preferentially dysregulated pathways and targets. *The Journal of Clinical Investigation*, *130*(8), 4252–4265. https://doi.org/10.1172/JCI129941

Payne, J. L., Khalid, F., & Wagner, A. (2018). RNA-mediated gene regulation is less evolvable than transcriptional regulation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), E3481–E3490. https://doi.org/10.1073/pnas.1719138115

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830. Retrieved from http://www.jmlr.org/papers/v12/pedregosa11a.html

Pellegrini-Calace, M., & Thornton, J. M. (2005). Detecting DNA-binding helix-turn-helix

structural motifs using sequence and structure information. Nucleic acids research, 33(7), 2129–2140. https://doi.org/10.1093/nar/gki349

Peng, Z., Wang, C., Uversky, V. N., & Kurgan, L. (2017). Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. *Methods in Molecular Biology (Clifton, N.J.)*, *1484*, 187–203. https://doi.org/10.1007/978-1-4939-6406-2_14

Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*, *9*(1). https://doi.org/10.3390/jcm9010132

Petitjean, A., Achatz, M. I. W., Borresen-Dale, A. L., Hainaut, P., & Olivier, M. (2007). TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, *26*(15), 2157–2165. https://doi.org/10.1038/sj.onc.1210302

Pizzino, G., Irrera, N., Cucinotta, M., Pallio, G., Mannino, F., Arcoraci, V., Bitto, A. (2017). Oxidative Stress: Harms and Benefits for Human Health. *Oxidative Medicine and Cellular Longevity*, *2017*, 8416763. https://doi.org/10.1155/2017/8416763

Ponting, C. P., Schultz, J., Milpetz, F., & Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research*, *27*(1), 229–232. https://doi.org/10.1093/nar/27.1.229

Poursheikhali Asghari, M., & Abdolmaleki, P. (2019). Prediction of RNA- and DNA-Binding Proteins Using Various Machine Learning Classifiers. *Avicenna Journal of Medical Biotechnology*, *11*(1), 104–111.

Pradhan, S., Das, P., & Mattaparthi, V. S. K. (2018). Characterizing the Binding Interactions between DNA-Binding Proteins, XPA and XPE: A Molecular Dynamics Approach. *ACS Omega*, *3*(11), 15442–15454. https://doi.org/10.1021/acsomega.8b01793

Qian, X., Liu, X., Zhu, Z., Wang, S., Song, X., Chen, G., Cao, L. (2021). Variants in LAMC3 Causes Occipital Cortical Malformation. *Frontiers in Genetics*, *12*, 616761. https://doi.org/10.3389/fgene.2021.616761

Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F., & Rost, B. (2020). ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *Journal of Molecular Biology*, *432*(7), 2428–2443. https://doi.org/10.1016/j.jmb.2020.02.026

Ramakrishnan, B., Boeggeman, E., & Qasba, P. K. (2012). Binding of N-acetylglucosamine (GlcNAc) β1-6-branched oligosaccharide acceptors to β4-galactosyltransferase I reveals a new ligand binding mode. *The Journal of Biological Chemistry*, *287*(34), 28666–28674. https://doi.org/10.1074/jbc.M112.373514

Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA-protein interactions. *Nature Methods*, *16*(3), 225–234. https://doi.org/10.1038/s41592-019-0330-1

Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., & Chen, Y. Z. (2011). Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *39*(Web Server issue), W385-90. https://doi.org/10.1093/nar/gkr284

Re, A., Joshi, T., Kulberkyte, E., Morris, Q., & Workman, C. T. (2014). RNA-protein interactions: an overview. *Methods in Molecular Biology (Clifton, N.J.)*, *1097*, 491–521. https://doi.org/10.1007/978-1-62703-709-9_23

Revathidevi, S., & Munirajan, A. K. (2019). Akt in cancer: Mediator and more. *Seminars in Cancer Biology*, *59*, 80–91. https://doi.org/10.1016/j.semcancer.2019.06.002

Rhee, C., Kim, J., & Tucker, H. O. (2017). Transcriptional Regulation of the First Cell Fate Decision. *Journal of Developmental Biology & Regenerative Medicine*, *1*(1).

Rivera, J., Keranen, S. V. E., Gallo, S. M., & Halfon, M. S. (2019). REDfly: the transcriptional regulatory element database for Drosophila. *Nucleic Acids Research*, *47*(D1), D828–D834. https://doi.org/10.1093/nar/gky957

Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Burley, S. K. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Research*, *43*(Database issue), D345-56. https://doi.org/10.1093/nar/gku1214

Rosenberg, S., Okamura, R., Kato, S., Soussi, T., & Kurzrock, R. (2020). Survival Implications of the Relationship between Tissue versus Circulating Tumor DNA TP53 Mutations-A Perspective from a Real-World Precision Medicine Cohort. *Molecular Cancer Therapeutics*, *19*(12), 2612–2620. https://doi.org/10.1158/1535-7163.MCT-20-0097

Rybak-Wolf, A., & Plass, M. (2021). RNA Dynamics in Alzheimer's Disease. *Molecules (Basel, Switzerland)*, *26*(17). https://doi.org/10.3390/molecules26175113

S, L., & GP, R. (2008). CytoPred: A Server for Prediction and Classification of Cytokines. *Protein Engineering, Design & Selection : PEDS*, *21*(4). https://doi.org/10.1093/PROTEIN/GZN006

Sachs, M. C. (2017). plotROC: A Tool for Plotting ROC Curves. *Journal of Statistical Software*, *79*(Code Snippet 2). https://doi.org/10.18637/jss.v079.c02

Saha, S., & Raghava, G. P. S. (2006). VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics, Proteomics & Bioinformatics*, *4*(1), 42–47. https://doi.org/10.1016/S1672-0229(06)60015-6

Sanchez de Groot, N., Armaos, A., Grana-Montes, R., Alriquet, M., Calloni, G., Vabulas, R. M., & Tartaglia, G. G. (2019). RNA structure drives interaction with proteins. *Nature Communications*, *10*(1), 3246. https://doi.org/10.1038/s41467-019-10923-5

Saravanan, V., & Lakshmi, P. T. V. (2013). APSLAP: an adaptive boosting technique for predicting subcellular localization of apoptosis protein. *Acta Biotheoretica*, *61*(4), 481–497. https://doi.org/10.1007/s10441-013-9197-1

Schemper, M. (1993). The relative importance of prognostic factors in studies of survival. *Statistics in Medicine*, *12*(24), 2377–2382. https://doi.org/10.1002/sim.4780122413

Schmidtke, P., & Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, *53*(15), 5858–5867. https://doi.org/10.1021/jm100574m

Schonbach, C., Tan, T. W., Kelso, J., Rost, B., Nathan, S., & Ranganathan, S. (2011, November). InCoB celebrates its tenth anniversary as first joint conference with ISCB-Asia. *BMC Genomics*. England. https://doi.org/10.1186/1471-2164-12-S3-S1

Schuschel, K., Helwig, M., Huttelmaier, S., Heckl, D., Klusmann, J.-H., & Hoell, J. I. (2020). RNA-Binding Proteins in Acute Leukemias. *International Journal of Molecular Sciences*, *21*(10). https://doi.org/10.3390/ijms21103409

Shanahan, H. P., Garcia, M. A., Jones, S., & Thornton, J. M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Research*, *32*(16), 4732–4741. https://doi.org/10.1093/nar/gkh803

Sharma, A., Kapoor, P., Gautam, A., Chaudhary, K., Kumar, R., Chauhan, J. S., Raghava, G.

P. S. (2013). Computational approach for designing tumor homing peptides. *Scientific Reports*, *3*(1), 1607. https://doi.org/10.1038/srep01607

Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. P. S. (2020). AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbaa294

Shen, W.-J., Cui, W., Chen, D., Zhang, J., & Xu, J. (2018). RPiRLS: Quantitative Predictions of RNA Interacting with Any Protein of Known Sequence. *Molecules (Basel, Switzerland)*, *23*(3). https://doi.org/10.3390/molecules23030540

Shima, F., Matsumoto, S., Yoshikawa, Y., Kawamura, T., Isa, M., & Kataoka, T. (2015). Current status of the development of Ras inhibitors. *Journal of Biochemistry*, *158*(2), 91–99. https://doi.org/10.1093/jb/mvv060

Shimizu, M., Shibuya, H., & Tanaka, N. (2022). Enhanced O-GlcNAc modification induced by the RAS/MAPK/CDK1 pathway is required for SOX2 protein expression and generation of cancer stem cells. *Scientific Reports*, *12*(1), 2910. https://doi.org/10.1038/s41598-022-06916-y

Si, J., Zhao, R., & Wu, R. (2015). An overview of the prediction of protein DNA-binding sites. *International Journal of Molecular Sciences*, *16*(3), 5194–5215. https://doi.org/10.3390/ijms16035194

Sim, J. C. H., White, S. M., & Lockhart, P. J. (2015). ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable & Rare Diseases Research*, *4*(1), 17–23. https://doi.org/10.5582/irdr.2014.01021

Singh, H., Khan, A. A., & Dinner, A. R. (2014). Gene regulatory networks in the immune system. *Trends in Immunology*, *35*(5), 211–218. https://doi.org/10.1016/j.it.2014.03.006

Singh, H., Singh, S., & Raghava, G. P. S. (2015). In silico platform for predicting and initiating β-turns in a protein at desired locations. *Proteins: Structure, Function and Bioinformatics*, *83*(5), 910–921. https://doi.org/10.1002/prot.24783

Singh, H., Srivastava, H. K., & Raghava, G. P. S. (2016). A web server for analysis, comparison and prediction of protein ligand binding sites. *Biology Direct*, *11*(1), 14. https://doi.org/10.1186/s13062-016-0118-5

Singh, S., Singh, H., Tuknait, A., Chaudhary, K., Singh, B., Kumaran, S., & Raghava, G. P. S.

(2015). PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues. *Biology Direct*, *10*(1), 73. https://doi.org/10.1186/s13062-015-0103-4

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics (Oxford, England)*, *15*(4), 327–332. https://doi.org/10.1093/bioinformatics/15.4.327

Sousa, S. F., Fernandes, P. A., & Ramos, M. J. (2006). Protein-ligand docking: current status and future challenges. *Proteins*, *65*(1), 15–26. https://doi.org/10.1002/prot.21082

Standart, N., & Jackson, R. J. (1994). Regulation of translation by specific protein/mRNA interactions. *Biochimie*, *76*(9), 867–879. https://doi.org/10.1016/0300-9084(94)90189-9

Su, H., Liu, M., Sun, S., Peng, Z., & Yang, J. (2019). Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics (Oxford, England)*, *35*(6), 930–936. https://doi.org/10.1093/bioinformatics/bty756

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Syriopoulou, A., Markopoulos, I., Tzakos, A. G., & Mavromoustakos, T. (2021). Ligand-Receptor Interactions and Drug Design. *Methods in Molecular Biology (Clifton, N.J.)*, *2266*, 89–104. https://doi.org/10.1007/978-1-0716-1209-5_5

Szilagyi, A., & Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. *Journal of Molecular Biology*, *358*(3), 922–933. https://doi.org/10.1016/j.jmb.2006.02.053

Taylor, J. P., Brown, R. H. J., & Cleveland, D. W. (2016). Decoding ALS: from genes to mechanism. *Nature*, *539*(7628), 197–206. https://doi.org/10.1038/nature20413

Taylor, S. S. (1987). Protein kinases: a diverse family of related proteins. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *7*(1), 24–29.

https://doi.org/10.1002/bies.950070106

Teng, Y., Xu, Z., Zhao, K., Zhong, Y., Wang, J., Zhao, L., Xia, Y. (2021). Novel function of SART1 in HNF4alpha transcriptional regulation contributes to its antiviral role during HBV infection. *Journal of Hepatology*, *75*(5), 1072–1082. https://doi.org/10.1016/j.jhep.2021.06.038

Terribilini, M., Sander, J. D., Lee, J.-H., Zaback, P., Jernigan, R. L., Honavar, V., & Dobbs, D. (2007). RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Research*, *35*(Web Server issue), W578-84. https://doi.org/10.1093/nar/gkm294

Tjong, H., & Zhou, H.-X. (2007). DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Research*, *35*(5), 1465–1477. https://doi.org/10.1093/nar/gkm008

Tomlinson, I. P., Novelli, M. R., & Bodmer, W. F. (1996). The mutation rate and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14800–14803. https://doi.org/10.1073/pnas.93.25.14800

Tomlinson, I., Sasieni, P., & Bodmer, W. (2002). How many mutations in a cancer? *The American Journal of Pathology*, *160*(3), 755–758. https://doi.org/10.1016/S0002-9440(10)64896-1

Tsai, M.-C., Spitale, R. C., & Chang, H. Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Research*, *71*(1), 3–7. https://doi.org/10.1158/0008-5472.CAN-10-2483

Turner, M., & Diaz-Munoz, M. D. (2018). RNA-binding proteins control gene expression and cell fate in the immune system. *Nature Immunology*, *19*(2), 120–129. https://doi.org/10.1038/s41590-017-0028-4

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., & Raghava, G. P. S. (2013). In silico models for designing and discovering novel anticancer peptides. *Scientific Reports*, *3*(1), 2984. https://doi.org/10.1038/srep02984

Usmani, S. S., Bhalla, S., & Raghava, G. P. S. (2018). Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. *Frontiers in Pharmacology*, *9*, 954. https://doi.org/10.3389/fphar.2018.00954

Velankar, S., Burley, S. K., Kurisu, G., Hoch, J. C., & Markley, J. L. (2021). The Protein Data Bank Archive. *Methods in Molecular Biology (Clifton, N.J.)*, *2305*, 3–21. https://doi.org/10.1007/978-1-0716-1406-8_1

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *The New England Journal of Medicine*, *319*(9), 525–532. https://doi.org/10.1056/NEJM198809013190901

Voon, P. J., & Kong, H. L. (2011). Tumour genetics and genomics to personalise cancer treatment. *Annals of the Academy of Medicine, Singapore*, *40*(8), 362–368.

Walia, R. R., Xue, L. C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., & Honavar, V. (2014). RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA- binding residues in proteins. *PloS One*, *9*(5), e97725. https://doi.org/10.1371/journal.pone.0097725

Wang, C., Zhang, Z., Yao, H., Zhao, F., Wang, L., Wang, X., Xu, S. (2014). Effects of atrazine and chlorpyrifos on DNA methylation in the liver, kidney and gill of the common carp (Cyprinus carpio L.). *Ecotoxicology and Environmental Safety*, *108*, 142–151. https://doi.org/10.1016/j.ecoenv.2014.06.011

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., Lithgow, T. (2017). POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics (Oxford, England)*, *33*(17), 2756–2758. https://doi.org/10.1093/bioinformatics/btx302

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, L., & 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*(Web Server issue), W243-8. https://doi.org/10.1093/nar/gkl298

Wang, L., Huang, C., Yang, M. Q., & Yang, J. Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Systems Biology*, *4 Suppl 1*, S3. https://doi.org/10.1186/1752-0509-4-S1-S3

Wang, L., Yang, M. Q., & Yang, J. Y. (2009). Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*, *10 Suppl 1*, S1.

https://doi.org/10.1186/1471-2164-10-S1-S1

Wang, L., Yan, K., Zhou, J., Zhang, N., Wang, M., Song, J., Wang, L. (2019). Relationship of liver cancer with LRP1B or TP53 mutation and tumor mutation burden and survival. *Journal of Clinical Oncology*, *37*(15_suppl), 1573. https://doi.org/10.1200/JCO.2019.37.15_suppl.1573

Wang, M., Luo, W., Jones, K., Bian, X., Williams, R., Higson, H., Zhu, B. (2020). SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Scientific Reports*, *10*(1), 12898. https://doi.org/10.1038/s41598-020-69772-8

Wang, Y.-X., Zuo, X., Wang, J., Yu, P., & Butcher, S. E. (2010). Rapid global structure determination of large RNA and RNA complexes using NMR and small-angle X-ray scattering. *Methods (San Diego, Calif.)*, *52*(2), 180–191. https://doi.org/10.1016/j.ymeth.2010.06.009

Wang, Y, Xue, Z., Shen, G., & Xu, J. (2008). PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, *35*(2), 295–302. https://doi.org/10.1007/s00726-007-0634-9

Wang, Yan, Xue, Z., & Xu, J. (2006). Better prediction of the location of alpha-turns in proteins with support vector machine. *Proteins*, *65*(1), 49–54. https://doi.org/10.1002/prot.21062

Wang, Zhenjia, Civelek, M., Miller, C. L., Sheffield, N. C., Guertin, M. J., & Zang, C. (2018). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics (Oxford, England)*, *34*(16), 2867–2869. https://doi.org/10.1093/bioinformatics/bty194

Wang, Zhining, Jensen, M. A., & Zenklusen, J. C. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA). *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 111–141. https://doi.org/10.1007/978-1-4939-3578-9_6

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., & Ji, Y. (2018). TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics (Oxford, England)*, *34*(9), 1615–1617. https://doi.org/10.1093/bioinformatics/btx812

Weinstein, M. J., Blanchard, R., Moake, J. L., Vosburgh, E., & Moise, K. (1989). Fetal and neonatal von Willebrand factor (vWF) is unusually large and similar to the vWF in patients with thrombotic thrombocytopenic purpura. *British Journal of Haematology*,

72(1), 68–72. https://doi.org/10.1111/j.1365-2141.1989.tb07654.x

Wingender, E., Dietze, P., Karas, H., & Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, *24*(1), 238–241. https://doi.org/10.1093/nar/24.1.238

Wong, K.-C., Li, Y., Peng, C., & Wong, H.-S. (2016). A Comparison Study for DNA Motif Modeling on Protein Binding Microarray. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(2), 261–271. https://doi.org/10.1109/TCBB.2015.2443782

Wyrick, J. J., & Roberts, S. A. (2015). Genomic approaches to DNA repair and mutagenesis. *DNA Repair*, *36*, 146–155. https://doi.org/10.1016/j.dnarep.2015.09.018

Xiao, N., Cao, D.-S., Zhu, M.-F., & Xu, Q.-S. (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics (Oxford, England)*, *31*(11), 1857–1859. https://doi.org/10.1093/bioinformatics/btv042

Xiong, D., Zeng, J., & Gong, H. (2015). RBRIdent: An algorithm for improved identification of RNA-binding residues in proteins from primary sequences. *Proteins*, *83*(6), 1068–1077. https://doi.org/10.1002/prot.24806

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, *16*, 15–24. https://doi.org/10.1016/j.csbj.2018.01.003

Yan, C., Terribilini, M., Wu, F., Jernigan, R. L., Dobbs, D., & Honavar, V. (2006). Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, *7*, 262. https://doi.org/10.1186/1471-2105-7-262

Yan, J., & Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Research*, *45*(10), e84. https://doi.org/10.1093/nar/gkx059

Yang, G., Jian, L., & Chen, Q. (2021). Comprehensive analysis of expression and prognostic value of the claudin family in human breast cancer. *Aging*, *13*(6), 8777–8796. https://doi.org/10.18632/aging.202687

Yang, X., Wang, J., Sun, J., & Liu, R. (2015). SNBRFinder: A Sequence-Based Hybrid

Algorithm for Enhanced Prediction of Nucleic Acid-Binding Residues. *PloS One*, *10*(7), e0133260. https://doi.org/10.1371/journal.pone.0133260

Yang, Y., Liu, L., Naik, I., Braunstein, Z., Zhong, J., & Ren, B. (2017). Transcription Factor C/EBP Homologous Protein in Health and Diseases. *Frontiers in Immunology*, *8*, 1612. https://doi.org/10.3389/fimmu.2017.01612

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. https://doi.org/10.1002/jcc.21707

Yu, D.-J., Hu, J., Huang, Y., Shen, H.-B., Qi, Y., Tang, Z.-M., & 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. https://doi.org/10.1002/jcc.23219

Yu, D.-J., Hu, J., Yan, H., Yang, X.-B., Yang, J.-Y., & Shen, H.-B. (2014). Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble. *BMC Bioinformatics*, *15*(1), 297. https://doi.org/10.1186/1471-2105-15-297

Yu, D.-J., Hu, J., Yang, J., Shen, H.-B., Tang, J., & Yang, J.-Y. (2013). Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(4), 994–1008. https://doi.org/10.1109/TCBB.2013.104

Yuan, Q., Chen, S., Rao, J., Zheng, S., Zhao, H., & Yang, Y. (2022). AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Briefings in Bioinformatics*, *23*(2). https://doi.org/10.1093/bib/bbab564

Zaorsky, N. G., Churilla, T. M., Egleston, B. L., Fisher, S. G., Ridge, J. A., Horwitz, E. M., & Meyer, J. E. (2017). Causes of death among cancer patients. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *28*(2), 400–407. https://doi.org/10.1093/annonc/mdw604

Zaretsky, J. Z., & Wreschner, D. H. (2008). Protein multifunctionality: principles and mechanisms. *Translational Oncogenomics*, *3*, 99–136.

Zhang, Jian, Ma, Z., & Kurgan, L. (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Briefings in Bioinformatics*, *20*(4), 1250–1268. https://doi.org/10.1093/bib/bbx168

Zhang, Jian, Ghadermarzi, S., Katuwawala, A., Kurgan, L. (2021). DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences, *Briefings in Bioinformatics*, 22(6), bbab336. https://doi.org/10.1093/bib/bbab336

Zhang, Jianming, Yang, P. L., & Gray, N. S. (2009). Targeting cancer with small molecule kinase inhibitors. *Nature Reviews. Cancer*, *9*(1), 28–39. https://doi.org/10.1038/nrc2559

Zhang, Jinghui, Walsh, M. F., Wu, G., Edmonson, M. N., Gruber, T. A., Easton, J., Downing, J. R. (2015). Germline Mutations in Predisposition Genes in Pediatric Cancer. *The New England Journal of Medicine*, *373*(24), 2336–2346. https://doi.org/10.1056/NEJMoa1508054

Zhang, Junjun, Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database : The Journal of Biological Databases and Curation*, *2011*, bar026. https://doi.org/10.1093/database/bar026

Zhang, P., Tao, L., Zeng, X., Qin, C., Chen, S. Y., Zhu, F., Chen, Y. Z. (2017). PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *Journal of Molecular Biology*, *429*(3), 416–425. https://doi.org/10.1016/j.jmb.2016.10.013

Zhang, S., Zeng, X., Lin, S., Liang, M., & Huang, H. (2022). Identification of seven-gene marker to predict the survival of patients with lung adenocarcinoma using integrated multi-omics data analysis. *Journal of Clinical Laboratory Analysis*, *36*(2), e24190. https://doi.org/10.1002/jcla.24190

Zhao, H., Yang, Y., & Zhou, Y. (2011). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Research*, *39*(8), 3017–3025. https://doi.org/10.1093/nar/gkq1266

Zheng, G., Qian, Z., Yang, Q., Wei, C., Xie, L., Zhu, Y., & Li, Y. (2008). The combination approach of SVM and ECOC for powerful identification and classification of transcription factor. *BMC Bioinformatics*, *9*, 282. https://doi.org/10.1186/1471-2105-9-282

Zhou, J., Lu, Q., Xu, R., He, Y., & Wang, H. (2017). EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM Relation Transformation. *BMC Bioinformatics*, *18*(1), 379. https://doi.org/10.1186/s12859-017-1792-8

Zhou, M., Zhao, H., Wang, X., Sun, J., & Su, J. (2019). Analysis of long noncoding RNAs

highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. *Briefings in Bioinformatics*, *20*(2), 598–608. https://doi.org/10.1093/bib/bby021

Zhou, Y., Cui, C., Ma, X., Luo, W., Zheng, S. G., & Qiu, W. (2020). Nuclear Factor kappaB (NF-kappaB)-Mediated Inflammation in Multiple Sclerosis. *Frontiers in Immunology*, *11*, 391. https://doi.org/10.3389/fimmu.2020.00391

Zhu, Y.-H., Hu, J., Song, X.-N., & Yu, D.-J. (2019). DNAPred: Accurate Identification of DNA-Binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines. *Journal of Chemical Information and Modeling*, *59*(6), 3057–3071. https://doi.org/10.1021/acs.jcim.8b00749