

DISCOVERING THE MOLECULAR BASIS OF DNA-PROTEIN INTERACTION SPECIFICITY

A thesis

Submitted in partial fulfilment of the requirements

Of the degree of

Doctor of Philosophy

By

Sanjana Nair

Roll Number: 20142004



Indian Institute of Science Education and Research, Pune
2022

Dedicated to
Santosh Madhavan Nair (Papa)
Darshana Santosh Nair (Mummy)

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



Date: 19/05/2022

Sanjana Nair
Roll Number: 20142004
IISER, Pune

CERTIFICATE

Certified that the work incorporated in the thesis entitled “Discovering the Molecular Basis of DNA-Protein Interaction Specificity”, submitted by Sanjana Nair was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.



डॉ. एम. एस. मधुसुधन / Dr. M. S. Madhusudhan
असोसिएट प्रोफेसर / Associate Professor
भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान
Indian Institute of Science Education & Research
पुणे / Pune - 411 008, India

Date: 19/05/2022

Dr. M. S. Madhusudhan
IISER, Pune

ACKNOWLEDGEMENTS

With this thesis, I present a glimpse of my journey as an integrated Ph. D. student in the Indian Institute of Science Education and Research, Pune. The countless hours of hard work, sleepless nights, exciting discoveries, heart-breaking failures and my development as a researcher are hidden within the depths of these pages. The acknowledgment section is not just a formality but an expression of my appreciation towards everyone who has made this possible.

I would like to break away from tradition and acknowledge my parents first, to whom I have dedicated this thesis. I was highly pampered and lived in a protective bubble. It was challenging for them to let me live independently in Pune and pursue this degree. But as they have always done, they sacrificed their wishes for my betterment. This did not stop them from their fortnightly trips to Pune, when they came on Sunday morning and left in the evening, travelling almost 10 hours to just see me for some time. So much so that the security guards and bus drivers became their friends! They have always ensured that I have everything I want, not just need. They celebrated my tiniest victories and guided me when I was lost. I feel fortunate and blessed to have their constant support and encouragement. I could probably write another thesis to describe their contributions to my life. But I will say that I love you, mummy and papa and will strive to make you happy and proud every day of my life.

Another person who has been an integral and irreplaceable part of this thesis is my supervisor, Dr. M. S. Madhusudhan. His confidence and over-arching optimism convinced me to join his lab and work on this thesis project. Despite having no computational background, he trusted me and created the most constructive learning atmosphere. More than a thesis supervisor, he has been the perfect mentor and an incredible human being. Having insightful discussions on science, sharing his own Ph. D stories to expressing concerns over my eating habits (!), I couldn't have asked for a better guide. We jointly decided that I would also do some wet-lab experiments, and his support throughout this process has been unwavering. Be it ordering chemicals, talking to collaborators, asking his friends for guidance, lobbying for lab space, going through mountains of paperwork- he

has done it all. If there was something that we needed, it was ordered with no compromises or questions asked. His approach towards solving problems is unique, and his passion for science is unparalleled. Apart from an academic standpoint, his management of students is noteworthy. He had faith in me when I didn't. I do not remember a single instance when a mistake was chided, just gently corrected. His jest for perfection and always trying to do better is contagious. I have learned countless skills from him that will stay with me for a lifetime. I sincerely wish to express my heartfelt gratitude to him for everything.

When I joined the lab, I was surprised by the welcoming environment. Having no experience with Linux or coding, I was taught the very basics by the lab members without any judgment or impatience. Neelesh, Neeladri, Yogendra, Sagar Gore, Parichit, Shipra, Abhilesh and Akash were the most helpful colleagues anyone would ever encounter. Neelesh was the go-to person for me to bug with my questions on how to do something, and he never disappointed me. He has provided valuable feedback and input throughout the years on my projects and still continues to do so. Neeladri was my immediate senior in the Integrated Ph. D program and was a pleasure to work with. He was the co-author of my first publication and has been a source of guidance over the years. His presence made the lab lively, and he is the best trip planner you can ever find. He is also the fairest (not complexion-wise!) person I have met when it comes to either giving credits or allocating individual contributions.

Yogendra is always ready to help and has an admirable memory. Sagar was a Masters' project student in IISER who stayed on as a project fellow in our lab. He is extremely hard-working and was always patient in answering doubts. Parichit was the scientific programmer in the lab when I joined and taught me the basics. He was always there to fix any problems without complaining. Shipra, Abhilesh and Akash also contributed a lot during the initial months in the lab. Kuan Pern Tan, Minh Nguyen and Thanh-Binh Nguyen belonged to the Singapore lab Dr. Madhusudhan used to head. Kuan Pern Tan is one of the most intelligent people I have met, and I was awe-struck by his thinking process during lab meetings. Minh developed CLICK, which is the foundation stone of my thesis.

Binh was very inspirational because she was the first one in the lab who combined computational predictions with *in-vitro* experiments.

Since then, there have been many additions to the lab, making me think I have been here forever. I would also like to thank Nida, Anuj, Dayal, Ankit, Swastik, Golding, Vasudha, Saswati and Kaustubh, the Masters students that rotated in the lab. Nida and Neeladri were joint first authors for my first publication, and she is a lovely person. Anuj has been an excellent friend and is a great researcher. Dayal is a prodigy and has shouldered many complicated projects. Ankit did his Masters project in the lab and stayed back as a project assistant. He was very knowledgeable and approachable. He is one of the nicest people I have ever met. Golding was also a Masters student and stayed back as a project assistant. He is very kind and helpful and makes really cool graphs that the rest of us struggle to replicate! Vasudha and Saswati worked with me on semester projects. They are very hard-working people who contributed to certain parts of my thesis. Kaustubh and his banter with Neeladri created a fun atmosphere in the lab.

Tejashree, Atreyi and Mukundan joined as Ph. D students and are delightful juniors. Tejashree contributes to the cheerful atmosphere in the lab. Atreyi joined the lab in 2019 and has helped me tremendously in the wet lab and running molecular dynamics simulations. She is a highly obedient and sincere student. Her helping hand in the lab was one of the reasons I could finish some of the experiments, especially MST. She never said no to any request I made of her and was always there whenever needed. She also proof-read this thesis, so I am very thankful to her. Mukundan was a great help when I needed help with my system while working remotely and he also helped by proof-reading the thesis. Shreyas was a post-doc in the lab for a short while and chauffeured us around in his car for lab lunches and dinners. Finally, Gulzar, who replaced Parichit as our lab scientific programmer, was skilled in hardware and software knowledge and made a fix multiple times when my system gave up on me. There have been many rotation students in the lab whom I have not mentioned. However, every person who has been a part of the lab has contributed to my journey in meaningful ways, and I thank every one of them.

Another two labs that I am grateful to are Dr. Saikrishnan and Dr. Gayathri Pananghat's labs. They trained me in wet-lab techniques, very generously shared their resources and tips and treated me as one of their own. Dr. Saikrishnan was a research advisory committee member and then a collaborator. His feedback and insights on DNA-protein interactions have aided us at every step. He always has a keen interest in my results, and discussions have led to fruitful outcomes. He assigned Sujata and Vinayak from his lab to help me with the experiments. Sujata was a fantastic mentor and explained every step, so I knew exactly what was happening. She taught me everything I know about molecular biology techniques, and I am grateful to her for that. Vinayak assisted me with protein purification experiments in Sujata's absence and solved my doubts patiently anytime I enquired. The same was true of Mahesh, Om and other members in both labs.

Apart from Dr. Saikrishnan, my other research advisory committee members- Dr. Jeetendra Chugh, Dr. Leelavati, and Dr. Deepak Barua have similarly provided invaluable feedback and suggestions on my project and helped enhance the quality of my research. Another faculty member has been a constant in my life at IISER: Dr. Sudha Rajamani. She was the first person I did my semester project with, and I fell in love with the research and her approach. I almost joined her lab, but destiny had other plans. She is a fantastic guide and mentor, and I have been lucky to have known her.

I would like to acknowledge the academic section and the staff, specially Tushar Kurulkar, Sayalee Damle, Swapnil Bhutkar, Dipali Dalvi and Dr. V. S. Rao. They defy the popular notion that government employees do not work efficiently. My life has been infinitely more manageable because of how they do their jobs. I also appreciate the efforts of the biology office staff comprising of Mrinalini Virkar, Mahesh Rote, Piyush Gadekar, Kalpesh, Sandeep Sankpal, Sandeep Beijankki, Shravi and Rupali Jadhav. The IT section has also contributed to sorting issues with the computing systems and allowed for smooth working conditions. I would like to thank Nisha Kurkure, Neeta Deo, the late Sachin Behra and Suresh. When it comes to making life easier, the list would be incomplete without the housekeeping staff who deserve all the praise one can confer.

I would also like to thank CSIR-Shyama Prasad Mukherjee Fellowship for funding my Ph. D between 2016 to 2021. I was financially supported by IISER Pune Fellowship from 2014 to 2016, for which I am grateful. I also thank Scivic Engineering Pvt Ltd. and Innoplexus Consulting Services Pvt Ltd for economic support of INR 39,086. I am very grateful to my mentor Dr. Madhusudhan for financial support through the Zumutor grant from 2021 to 2022.

I wouldn't have known about IISER if it weren't for Ankitha Shetty, Dr. Leena Phadke and Dr. Ravi Phadke. Ankitha had joined IISER Pune as an integrated Ph. D student and came to the microbiology department in Ramnarain Ruia college. It was a tradition for alumni to drop by for a quick chat in the department. I was a third-year B.Sc. student then, and it was while talking to her I came to know of the existence of such an institute. Dr. Leena and Dr. Ravi Phadke were massive believers in research and critical thinking. They encouraged us to pursue research projects in vacations. It was only due to their efforts that I got inspired to pursue research.

I am also grateful to Dr. Rajeev Nerurkar Dr. B.K. Smruti, Dr. Hemant Sant and the staff at AIMS Hospital, Pune and Bombay Hospital, Mumbai. They have ensured my health to continue with the Ph. D program and I will be obliged to them for it.

Acknowledgments also go to my batch mates who shared the joint suffering endowed by the Ph. D Gods- Nilam, Shikha, Punita and Yashaswi. Nilam also gave a lot of inputs for the wet lab experiments. Yashaswi has supported me during dark times, and I am deeply indebted to him.

Finally, some friends last for a reason or for a season. Then there are the ones that last a lifetime. I have been fortunate to have a few forever friends- Aditya Narvekar, Prasenjeet Kawale, Sajyoty Das, Priyanka Yadav, Cleon Baretto and Shubham Singh. Aditya has tolerated my rants, rejoiced in my successes, cheered me up during my failures, and taught me the true meaning of friendship. He is the one constant in my life that I know will never change and I am very thankful for him. Prasenjeet is the best friend that IISER has given me. From staying late nights till my experiment gets over to teaching me what sources to refer to for IISER exams, he started out as the best lab mate. He has

since become my confidant, support system, and biggest cheerleader through time. Sajyoty and I have a friendship that spans 15 years, and she still remains the only person who can talk to me for hours! Priyanka is one of the kindest, most generous and most congenial girls and inspires me every day with the grace with which she leads her life. Cleon has been a wonderful friend and the last bencher who ensured that I would not be too studious! Shubham was my immediate junior in IISER and a friend that has stood by me in good and bad times.

Finally, though my parents played the most significant role in supporting me, I would also like to acknowledge other family members' efforts. I wish to thank my grandmothers- Jyotsna Chandrashekhhar Kanaiya and Komalam Nair. I would also like to thank my aunts, uncles and cousins for their help and support. Many more people have played small but significant parts in my Ph. D journey, and I would like to thank them all.

-Sanjana

Table of Contents

List of Tables.....	15
List of Figures.....	17
Contents of Appendix.....	19
Synopsis.....	20
CHAPTER 1: Introduction.....	26
1.1 Why study the specificity between DNA and proteins?.....	27
1.2 DNA structure.....	27
1.3 Mechanisms of DNA-protein interactions.....	29
1.3.1 Base readout.....	29
1.3.2 Shape readout.....	36
1.4 Previous investigations on the universal recognition code.....	39
1.4.1 The zinc finger recognition code.....	40
1.4.2 The TAL type III effector recognition code.....	42
1.5 Proposed approach for discovering the molecular basis of base readout.....	43
1.6 Introduction to techniques.....	45
1.6.1 CLICK.....	45
1.6.2 Molecular Dynamics (MD) simulations.....	46
1.6.3. Binding free energy calculations.....	47
1.6.4 Protein structure modelling.....	47
1.6.5. Multiple Sequence Alignments (MSA).....	48
1.6.6 Polymerase Chain Reaction (PCR).....	49
1.6.7 Restriction-free (RF) Cloning.....	49
1.6.8 Electrophoretic Mobility Shift Assay (EMSA).....	50
1.6.9 Microscale Thermophoresis.....	51
1.6.10. Nano Differential Scanning Fluorimetry (NanoDSF).....	52
1.7 Chapter summary.....	53
CHAPTER 2: Prediction of amino acids for a DNA sequence.....	54

2.1	Compilation and characterization of DNA-protein complex structure dataset.....	55
2.1.1	Analysis of amino acid composition.....	56
2.1.2	Analysis of molecular interactions.....	60
2.2	Comparing hydrogen bonds across DNA-binding protein families.....	73
2.2.1	Mononucleotide superimpositions.....	76
2.2.2	Dinucleotide superimpositions.....	79
2.2.3	Trinucleotide superimpositions.....	81
2.3	Development of the scoring scheme.....	85
2.4	Accuracy of the predictions on training and testing set.....	88
2.5	Chapter summary.....	92
CHAPTER 3: <i>In silico</i> Validation of Predicting Amino Acids for a DNA Sequence.....		93
3.1	Validation of our method on the zinc finger proteins.....	94
3.1.1	Validation on SELEX selected zinc finger-DNA complexes.....	94
3.1.2	Validation on crystal structures of zinc finger proteins.....	95
3.2	Predicting Multiple Sequence Alignment (MSA) from Position Weight Matrices (PWMs).....	96
3.3	Prediction of the effect of mutation on DNA-protein binding affinity.....	100
3.4	Validation using Molecular Dynamics Simulations.....	105
3.4.1	Zif-268.....	109
3.4.2	Myc.....	115
3.4.3	Ets1.....	118
3.4.4	EcoRV.....	120
3.4.5	Summary of validation with MD simulations.....	123
3.5	Discussions.....	126
3.6	Acknowledgements.....	129
3.7	Chapter summary.....	129
CHAPTER 4: <i>In-vitro</i> Validation of Predicting Amino Acids for a DNA Sequence.....		130
4.1	Changing the specificity of DNA-binding protein.....	131
4.2	Methods for molecular cloning, expression and purification.....	131

4.2.1 Cloning.....	132
4.2.2 Expression and purification.....	136
4.3 Binding Assays for assessing specificity of Ets1 and Myc.....	143
4.3.1 Electrophoretic Mobility Shift Assay (EMSA).....	145
4.3.2 Microscale Thermophoresis (MST).....	153
4.3.3 Nano Differential Scanning Fluorimetry (nanoDSF).....	161
4.4 Comparison of binding with different <i>in-vitro</i> techniques.....	165
4.5 Discussions.....	169
4.6 Acknowledgements.....	171
4.7 Chapter summary.....	172
CHAPTER 5: Predicting DNA-Protein Interface for a Protein Structure.....	173
5.1 Introduction to the problem.....	174
5.1.1 Sequence-based DNA-binding residue prediction.....	175
5.1.2 Structure-based DNA-binding residue prediction.....	175
5.1.3 Sequence and Structure-based DNA-binding residue prediction.....	175
5.2 Methodology.....	176
5.2.1 Compiling datasets used in this study.....	177
5.2.2 Prediction Algorithm.....	178
5.2.3 Comparison of JEDII with GraphBind and HDock on PDNA-285 dataset.....	181
5.2.4 Evaluation Metrics.....	181
5.3 Assessment of the robustness of JEDII in predicting DNA-protein interface.....	183
5.3.1 Assessment of the performance of JEDII on PDNA-285.....	183
5.3.2 Accuracy in the presence of close sequence homologs.....	185
5.3.3 Accuracy using the unbound protein conformation.....	185
5.3.4 Accuracy using protein models as inputs.....	186
5.4 Comparison of JEDII with other methods.....	188
5.4.1 Predicting amino acid residues binding DNA.....	188
5.4.2 Validation of DNA sequence prediction.....	193
5.5 Case Study of Cro Repressor.....	196

5.6 Discussions.....	198
5.7 Chapter summary.....	203
CHAPTER 6: Conclusions and Future Directions.....	205
6.1 Summary of this study.....	206
6.2 Analysis of DNA-protein specificity in non-B-form DNA.....	211
6.3 Predicting binding affinities of the DNA-protein complexes.....	214
6.4 Engineering specific DNA-protein complexes.....	215
6.5 Concluding remarks.....	216
Appendix.....	217
References.....	242
Publications.....	264

List of Tables

Table 1.1: Parameters differentiating A-, B- and Z-DNA.....	38
Table 2.1: Comparison of amino acid composition in DNA-binding, non-binding proteins and DNA-binding domains.....	57
Table 2.2: Normalized amino acid-nucleotide hydrogen bond distribution in the training set.	69
Table 2.3: Statistical propensity of nucleotide-amino acid interaction in the training set....	73
Table 2.4: Mononucleotide superimposition frequency.....	77
Table 2.5: Amino acid-nucleotide mononucleotide superimposition frequency.....	78
Table 2.6: Comparison of di- and trinucleotide preferences.....	84
Table 2.7: The correct predictions (%) of amino acids for a given DNA sequence in the training and testing set.....	90
Table 3.1: Validation of our method on the Liu et al. zinc finger set.....	95
Table 3.2: Validation of our method on the crystal structures of zinc finger proteins.....	96
Table 3.3: Predictions of the effect of mutation on DNA-binding interface.....	102
Table 3.4: Prediction of specificity change in selected DNA-binding proteins.....	107
Table 4.1: The experimental systems selected for changing their DNA-binding specificity	131
Table 4.2: Composition of 10X pfu polymerase buffer (pH 8.3 at 25°C).....	132
Table 4.3: PCR reaction setup for cloning.....	132
Table 4.4: The PCR protocol used for cloning.....	133
Table 4.5: The sequences of primers used for cloning.....	134
Table 4.6: Composition of Myc lysis buffer.....	137
Table 4.7: Composition of Ni-NTA buffers for Myc.....	137
Table 4.8: Composition of 2x SDS dye.....	137
Table 4.9: Composition of 18% SDS-polyacrylamide gel.....	137
Table 4.10: Composition of Myc dialysis buffer.....	138
Table 4.11: Composition of Ets1 lysis buffer.....	140
Table 4.12: Composition of Ni-NTA buffer for Ets1 buffer.....	140
Table 4.13: Composition of Ets1 dialysis buffer.....	140

Table 4.14: Composition of ASB.....	141
Table 4.15: DNA substrates for binding assays.....	144
Table 4.16: DNA annealing protocol.....	144
Table 4.17: Composition of Myc binding buffer (1x).....	144
Table 4.18: Ets1 binding buffer (1x).....	144
Table 4.19: Composition of TBE buffer (1x).....	145
Table 4.20: Composition of 8% polyacrylamide gel (8 mL).....	145
Table 4.21: EMSA results of Myc wild-type and mutant proteins with different substrates.	150
Table 4.22: EMSA results of Ets1 proteins with different substrates.....	153
Table 4.23: Binding affinity values from the MST experiment.....	161
Table 4.24: T _m values measured by nanoDSF.....	162
Table 4.25: Comparison of results from different <i>in-vitro</i> techniques.....	169
Table 5.1: Performance of JEDII on PDNA-285 dataset with redundant and 40% non- redundant template database.....	183
Table 5.2: Comparison of JEDII with existing algorithms on PDNA-62 dataset.....	189
Table 5.3: Comparison of JEDII with existing algorithms on P224 dataset.....	190
Table 5.4: Comparison of JEDII with existing algorithms on PDNA-129 dataset.....	190
Table 5.5: Comparison of performance of JEDII on HOLO-APO 82 dataset with other methods.....	191
Table 5.6: Comparison of the accuracy of JEDII on models built by MODELLER and AlphaFold2.....	192
Table 5.7: Comparison of JEDII with GraphBind and HDock on PDNA-285 dataset.....	193
Table 5.8: Comparison of JEDII with DBD2BS.....	195

List of Figures

Figure 1.1: Nucleotide structure and Watson-Crick base pairing.....	29
Figure 1.2: Distribution of atoms in major and minor grooves.....	32
Figure 1.3: Types of global DNA conformations.....	39
Figure 1.4: Schematic of zinc finger recognition code.....	41
Figure 1.5: Schematic of TAL type III effector recognition code.....	42
Figure 1.6: CLICK superimposition of trinucleotide GAA.....	44
Figure 1.7: CLICK superimposition of proteins with PDB IDs 1xxa (red) and 1tig (blue)....	46
Figure 1.8: Schematic of RF cloning.....	50
Figure 1.9: Schematic representation of an EMSA experiment.....	51
Figure 1.10: Schematic of MST experiment.....	52
Figure 2.1: Amino acid frequency distribution in protein sequences.....	58
Figure 2.2: Types of nucleotide-amino acid interactions.....	60
Figure 2.3: Distribution of π - π interactions.....	63
Figure 2.4: Distribution of cation- π interactions.....	64
Figure 2.5: Hydrogen bond preferences in the training set.....	70
Figure 2.6: Atom-wise frequency of amino acid-DNA backbone <i>hydrogen</i> bonds.....	71
Figure 2.7: Merging amino acids with similar hydrogen bond forming functional groups..	72
Figure 2.8: Definition of hydrogen bond overlap.....	74
Figure 2.9: Fold-wise analysis of RMSD of superimposed nucleotides.....	76
Figure 2.10: Nucleotide-amino acid interaction preferences in the superimposed mononucleotides.....	77
Figure 2.11: Nucleotide-amino acid preferences of superimposed dinucleotides.....	80
Figure 2.13: Schematic of prediction of amino acids for a given DNA.....	88
Figure 2.14: Distribution of hydrogen bonds in the training and testing sets.....	90
Figure 2.15: Correlating top 3 predictions with the MSA constituents.....	91
Figure 3.1: Rank distribution of the predicted amino acids that matched with the MSA.....	97
Figure 3.2: Amino acid distribution in the cases where the prediction matched the MSA...	98
Figure 3.3: Distribution of MCC values for predicting the top 3 amino acids in the MSA....	99
Figure 3.4: Comparison of the contribution of different components toward MCC.....	100

Figure 3.5: Distribution of correct and incorrect predictions of binding affinity change.....	105
Figure 3.6: Schematic for changing the specificity of a DNA-binding protein.....	106
Figure 3.7: Systems for validating the change in specificity of DNA-binding proteins.....	107
Figure 3.8: MD results of Zif-268 wild-type and mutant proteins.....	113
Figure 3.9: MD results of Zif-268 wild-type and mutant protein models.....	114
Figure 3.10: MD results of Myc wild-type and mutant proteins.....	117
Figure 3.11: MD results of Ets1 wild-type and mutant proteins.....	120
Figure 3.12: MD results of EcoRV wild-type and mutant proteins.....	122
Figure 4.1: SDS-PAGE gel after HisTrap column elution for Myc.....	138
Figure 4.2: SDS PAGE gel after HisTrap column elution of Ets1.....	140
Figure 4.3: Expression check gel for EcoRV.....	142
Figure 4.4: Protocol for competent cell preparation containing methyltransferase gene..	143
Figure 4.5: EMSA of Myc wild-type protein.....	147
Figure 4.6: EMSA of Myc R20Q single mutant protein.....	148
Figure 4.7: EMSA of Myc QY double mutant protein.....	149
Figure 4.8: EMSA of Ets1 wild-type protein.....	151
Figure 4.9: EMSA of Ets1 Y395R mutant protein.....	152
Figure 4.10: MST data for Myc wild-type protein.....	156
Figure 4.11: MST data for Myc R20Q single mutant protein.....	157
Figure 4.12: MST data for Myc QY double mutant protein.....	158
Figure 4.13: MST data for Ets1 wild-type protein.....	159
Figure 4.14: MST data for Ets1 Y395R protein.....	160
Figure 4.15: NanoDSF data for Myc.....	163
Figure 4.16: NanoDSF data for Ets1.....	164
Figure 5.1: Flowchart of the steps in the JEDII protocol.....	180
Figure 5.2: Performance of JEDII on the PDNA-285 dataset.....	184
Figure 5.3: Distribution of motif similarity score in ChIP-seq dataset.....	194
Figure 5.4: Comparison of JEDII with Farrel and Guo's study.....	196
Figure 5.5: The prediction of binding pose of cro repressor.....	197
Figure 6.1: Comparison of binding preferences in different DNA conformations.....	213

Contents of Appendix

1. List of PDBs removed by manual curation to retain only B-form sequence-specific DNA-protein complex structures.....	217
2. Amino acid (three-letter representation) composition (%) in training and testing sets. .	217
3. Hydrogen bond donor-acceptor list used in this study.....	218
4. Superimposition frequency when the first nucleotide of the dinucleotide is hydrogen-bonded.....	220
5. Superimposition frequency when the second nucleotide of the dinucleotide is hydrogen-bonded.....	221
6. Superimposition frequency of trinucleotides.....	222
7. Statistically significant preferences common between Di- and Trinucleotides.....	224
8. Weights optimized by grid-search for the prediction of an amino acid specific for a given DNA sequence.....	228
9. List of chains selected for MSA prediction from PWM.....	230
10. System-size for MD simulations of DNA-protein complexes.....	232
11. RMSD of wild-type and mutant proteins with DNA substrates (Å).....	232
12. Centroid distance between wild-type and mutant DNA-protein complexes (Å).....	233
13. Total hydrogen bonds between different wild-type and mutant DNA-protein complexes.	234
14. Nucleotide-specific hydrogen bonds between wild-type and mutant DNA-protein complexes.....	235
15. Binding affinities of wild-type and mutant proteins with DNA substrates (kcal/mol)....	236
16. Total energy of the Zif-268 wild-type protein with wild-type DNA.....	237
17. Temperature of the system with Zif-268 wild-type DNA-protein complex.....	238
18. Retention of hydrogen bonds present in the crystal structure of the proteins.....	239
19. Distance of N and O atoms with DNA in the PDNA-285 datasets.....	239
20. Accuracy of prediction of amino acid specific for a sequence after three-fold cross-validation.....	240
21. Comparison of intra and intermolecular potential energy.....	240

Synopsis

The recognition of specific sites on the DNA by proteins mediates crucial biological processes such as replication [1–4], DNA repair [5–7], recombination [8–11] and gene regulation [12–14], to name a few. This recognition of cognate sequences that drives these processes needs to bypass billions of non-cognate base pairs. The specificity preferences that drive this process are not clear. One factor that hinders insights into the specificity preferences is the structural diversity of DNA-binding proteins [15]. Another factor is that the binding of proteins to their cognate sites could be mediated by the cooperative action, as in the case of transcription factors [16–18]. Insights into the preferences of nucleotide-amino acid would help delineate the contribution of individual proteins to the functioning of biological systems.

This insight may prove useful because mutations of these proteins are responsible for many diseases such as cancer, neurodegenerative, metabolic, infectious and inherited diseases. One of the most viable options for the treatment of these diseases is gene therapy [19–24]. With the advent of the CRISPR-Cas9 system, replacing defective genes is possible [25,26]. However, off-target effects remain a potential concern, and efforts are underway to reduce them [27–29]. Thus, investigating the mechanism by which DNA-protein complexes show specificity could allow rational modification of the system and assist in developing appropriate therapeutics.

The earliest studies on the DNA-protein specificity investigation showed the requirement for more than two hydrogen bonds for specific recognition of nucleotides [30]. The study predicted that the hydrogen bond between arginine and guanine would be specific without the crystal structures of DNA-protein complexes. Since then this prediction has been proven and is observed most frequently in the X-ray crystal structures of DNA-protein complexes. The statistical analysis of these DNA-protein crystal structures has led to knowledge about nucleotide-amino acid interaction preferences but not a generalized recognition code [31–34]. However, analyses of DNA-binding protein families have resulted in specific recognition preferences for zinc finger and Transcription Activation factor-Like effector (TAL Type III effector) protein families [35–39].

Due to the crucial role of DNA-protein specificity, we hypothesized the conservation of molecular mechanisms across DNA-binding protein structural families. We wanted to investigate this hypothesis and had two objectives. The first was to predict the amino acids specifically binding a DNA sequence. The second was predicting the DNA sequence that a protein would recognize. These objectives ensured that we adopted a holistic approach and studied the specificity problem from both the DNA and protein standpoints.

To address these objectives, we analysed the structures of a 40% non-redundant (NR) set of DNA-protein complexes. We found hydrogen bonds between nucleotides and amino acids to be the predominant interactions among the directional interactions. Like our predecessors in the field, we focussed on finding their role in imparting specificity to the DNA-protein interface. We hypothesized that the common molecular mechanism of specificity could relate to a similar arrangement of hydrogen bond donor and acceptor atoms at the DNA-protein interface. Hydrogen bonds have a limited number of the hydrogen bond donor and acceptor pair orientations. Thus, the amino acids in different protein structural folds would have to adopt these orientations to recognize nucleotides specifically. Hence, hydrogen bond donors and acceptors for superimposition make our approach independent of the structural fold.

To highlight the significance of hydrogen bonds, we superimposed hydrogen bond donors and acceptors across structural folds in the 40% NR set [40,41]. For the first objective, we analysed superimpositions of the hydrogen-bonded nucleotide and its neighbours on either side. Therefore, we performed mono-, di- and trinucleotide superimpositions across structural folds and found some statistically significant nucleotide-amino acid interaction preferences. The superimposition of structures boosts our confidence in the significance of preferences, especially when data is sparse. We developed a probabilistic scoring scheme to predict the amino acids that would specifically bind a given DNA sequence based on these analyses.

We used multiple *in-silico* and *in-vitro* strategies to validate our predicted preferences. Apart from the testing set, we validated our predictions on the zinc finger proteins, since they have well-characterized specificity preferences. Another validation was

to reconcile the multiple sequence alignment (MSA) of a protein with the variation observed in its recognition site. While our method needs improvement, currently, it is the only one that can predict these specific interactions to some degree of accuracy. The validation of the preferences by predicting the effect of a mutation in the DNA-binding interface had comparatively more accuracy.

We also attempted to validate the method using the probabilistic interaction propensities to change the specificity of DNA-binding proteins using Molecular Dynamics (MD) simulations and *in-vitro* assays.

For MD simulations, we made predictions of the amino acids mutations for recognizing a non-consensus DNA for Ets1, Myc, EcoRV. Ets1 and Myc are transcription factors, whereas EcoRV is a restriction enzyme. We generated the models of mutant protein-non-consensus DNA, wild-type protein-non-consensus DNA and mutant protein-consensus DNA. Using MD simulation, we compared these complexes to the wild-type protein-consensus DNA.

We used Zif-268 wild-type protein-consensus DNA and crystal structures of the mutant proteins with their preferred substrate as a control system. We modelled the Zif-268 wild-type and mutant proteins to check the effect of modelling on the binding energy of the complex. However, in the control system, the wild-type protein had a better binding to a different substrate than its consensus. This reduced our confidence in the results, where we observed that the Myc mutant did not bind to the non-consensus DNA preferentially while Ets1 did. There was no difference in binding when the EcoRV mutant protein interacted with either the consensus or non-consensus DNA.

We predicted three mutations in Ets1 and Myc that would disrupt binding to the original recognition site and be specific to our desired site for the *in-vitro* assay validation. We cloned and purified the wild-type and mutant proteins and performed binding assays to derive the specificity experimentally. We used Electrophoretic Mobility Shift Assay (EMSA), MicroScale Thermophoresis (MST) and nano Differential Scanning Fluorimetry (nanoDSF) to assess the binding. We could show that the desired specificity change was present in one of the three predicted mutations. In one other case the wild-type itself had a broad

specificity and changing one amino acid was possibly not sufficient to change the specificity.

The validation provided confidence in using hydrogen bond donors and acceptors to study specificity. Thus, we employed the same strategy for the second objective of predicting the DNA sequence from a protein structure. We developed a template-dependent methodology using CLICK superimpositions, with hydrogen bond donors and acceptors in proteins as representative atoms.

We generated putative models of the given protein with DNA. Then we selected the model with the highest donor/acceptor match between the given and template proteins among the different models. This provides information about the DNA-binding residue and the binding pose. We relied on the chemical complementarity of the binding interface. The hydrogen bond donor/acceptor atoms in the template protein matched those in the given protein. Thus the template DNA, which would complement this protein interface, would also match. Our method outperforms all existing algorithms that predict either the DNA-binding residues or the DNA sequence of the protein. Our strategy integrated the prediction of both sides of the DNA-protein interface and to the best of our knowledge, it is the only one that does so.

Our findings provide new insights into the specificity determining factors in DNA-protein complexes. The applications of the results from our study extend to various spheres. We described some applications in our thesis and proposed others, which are a direct extension of the results from this thesis. The amino acid prediction from the DNA sequence may be employed to study the evolution of transcription factors. The current methodology can predict the effect of the mutation on the DNA-protein interface and therefore, could help rationalize diseases that occur as a result of this. An extension of this would be to predict the binding affinity of the DNA-protein complexes

We have shown that the predicted DNA sequence correlates with the binding site obtained through Chromatin Immunoprecipitation sequencing (ChIP-seq). In the future, we can engineer proteins with desired specificity using the hydrogen bond donor/acceptor derived interface, without regard for conforming to a specific fold. A similar approach may

be adopted to analyse the specificity in non-B form DNA and RNA-protein complexes or any interface that relies on hydrogen bonds to recognize the binding partner. Given the central role of DNA-protein interactions in cellular processes, this study would interest the scientific community across different research areas.

Thesis Organization

The thesis has 6 chapters and an appendix section. Each of them is summarized below-

■ Chapter 1: Introduction

We explain the methodology for predicting amino acids specifically binding to a given DNA sequence. It describes the amino acid composition and the molecular interactions in the dataset of specific DNA-protein complexes. Also analysed are the hydrogen bond preferences of mono-, di- and trinucleotides superimposed across DNA-binding protein families. It then focuses on the scoring scheme for predicting the amino acids for a DNA sequence and reports the results on the training and testing sets.

■ Chapter 2: Predicting Amino Acids Specific for a Given DNA Sequence

This chapter begins with the methodology adopted for solving the first objective of the project, which was to find the amino acids specific for a given DNA sequence. It describes the amino acid composition and the molecular interactions in the dataset of specific DNA-protein complexes. Also analysed are the hydrogen bond preferences of mono-, di- and trinucleotides superimposed across DNA-binding protein families. It then focuses on the scoring scheme for predicting the amino acids for a DNA sequence and reports the results on the training and testing sets.

■ Chapter 3: *In-silico* Validation of Predicting Amino Acids Specific for a DNA Sequence

Chapter 3 discusses the four *in-silico* approaches for validating the amino acid prediction for a DNA sequence. The chapter first describes validation on zinc finger proteins, which have a known recognition code. It then focuses on the second approach of predicting the multiple sequence alignment of a protein from the

position weight matrix of its recognition site. The following section outlines predicting the effect of a mutation at the DNA-protein interface on the binding affinity of the complex. Finally, the chapter deals with changing the specificity of DNA-binding proteins and assessing it using molecular dynamics simulations.

■ **Chapter 4: *In-vitro* Validation of Predicting Amino Acids Specific for a DNA Sequence**

Here we discuss the *in-vitro* assays to validate the change in specificity of DNA-binding proteins experimentally.. The methods section describes the protocols followed for cloning, expression and purification of three DNA-binding proteins. The chapter then outlines the three binding assays to determine the DNA-protein specificities- Electrophoretic Mobility Shift Assay, MicroScale Thermophoresis and nano Differential Scanning Fluorimetry. The last section reconciles the inferences from the different techniques to conclude whether the predictions were validated.

■ **Chapter 5: Predicting DNA-Protein Interface for a Given Protein Structure**

This chapter deals with the second objective of predicting the DNA sequence for a given protein. It describes the currently existing tools and the lacunae that exist in them. The following section proposes a template-dependent methodology to predict the DNA-binding residues, the binding pose and the DNA sequence for a protein structure. The chapter then illustrates the technique's robustness in making predictions using unbound protein conformations, using protein models and in the absence of close homologs. Then the method is compared with the existing tools and validated using different benchmark datasets.

■ **Chapter 6; Conclusions and Future Directions**

The final chapter summarizes the findings and discusses the key takeaways from this study. It then proposed the extensions of the project, such as obtaining specificity preferences for non-B-form DNA, predicting the binding affinity of DNA-protein complexes and engineering specific DNA-protein complexes.

■ **Appendix**

This section has additional data from the chapters for reference.



CHAPTER 1: Introduction

OVERVIEW

- Why study the specificity between DNA and proteins?
 - DNA structure
 - Mechanisms of DNA-protein interactions
 - Base readout
 - Shape readout
 - Previous investigations on universal recognition code
 - Case study of TAL-effector family
 - Case study of zinc finger family
 - Proposed approach for discovering the molecular basis of base readout
 - Introduction to techniques used in this study
 - Chapter summary
-

1.1 Why study the specificity between DNA and proteins?

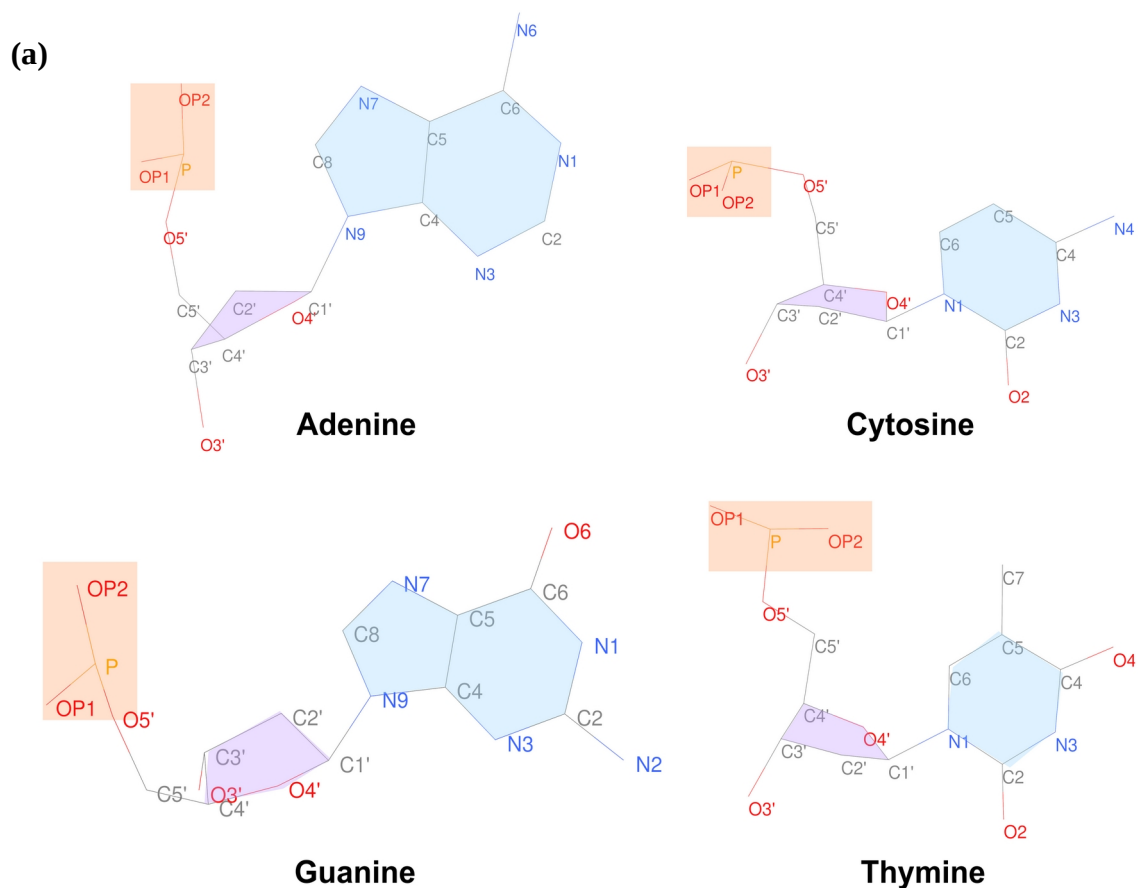
The recognition of specific sites on the DNA by proteins mediates crucial biological processes such as replication [1–4], DNA repair [5–7], recombination [8–11] and gene regulation [12–14], to name a few. It is not clear how these processes bypass billions of non-cognate base pairs and recognize their cognate partner.

One of the factors that make this a difficult problem to solve is the structural diversity of DNA-binding proteins [15]. Many transcription factors bind in tandem or have co-operative action in regulation of gene networks and their mutations lead to diseased conditions [16–20]. Understanding the DNA-protein specificity preferences can bring new insights into the functioning of biological systems. There are many diseases such as cancer, neurodegenerative diseases, metabolic diseases, infectious diseases and inherited diseases for which gene therapy is one of the most viable options [21–24]. Understanding the DNA-protein specificity can help us develop targeted therapeutics. With the advent of the CRISPR-Cas9 system, replacing defective genes is possible [25,26]. However, off-target effects still remain a potential concern and efforts are underway to reduce them [27–29]. Thus, investigating the mechanism by which DNA-protein complexes show specificity could allow rational modification of the system and assist in developing appropriate therapeutics.

1.2 DNA structure

DNA is composed of nucleotides held together by phosphodiester bonds. Each nucleotide, in turn, has three components- nucleoside base or nitrogenous base, deoxyribose sugar and phosphate backbone. A double-stranded (ds) DNA has two strands of these linked nucleotides. There are four commonly found nucleotides in biological systems with the nucleosides-adenine (A), cytosine (C), guanine (G) and thymine (T) (Figure 1.1 (a)). The two strands in a ds-DNA interact primarily through hydrogen bonds between A-T and G-C. These hydrogen bonds between the bases may be formed in multiple permutations and combinations of the interacting atoms. The atomic contacts between the complementary bases result in different types of base pairing, such as the

Watson-Crick and Hoogsteen. The kind of base pairing dictates the conformation of the DNA. Watson-Crick base pairing forms B-DNA, the most commonly occurring conformation of DNA in living systems (Figure 1.1(b)). B-DNA forms a right-handed double-helical structure that has grooves named major and minor grooves based on their sizes. The major groove interactions are frequently associated with the sequence-specific recognition between protein and DNA (base readout). This sequence specificity is conferred by the nucleoside base since all the four commonly found nucleotides have the same deoxyribose and phosphate backbone composition and structure [42]. In contrast, the specific recognition that depends on the DNA shape (shape readout) uses a combination of all components of the nucleotides to achieve specificity.



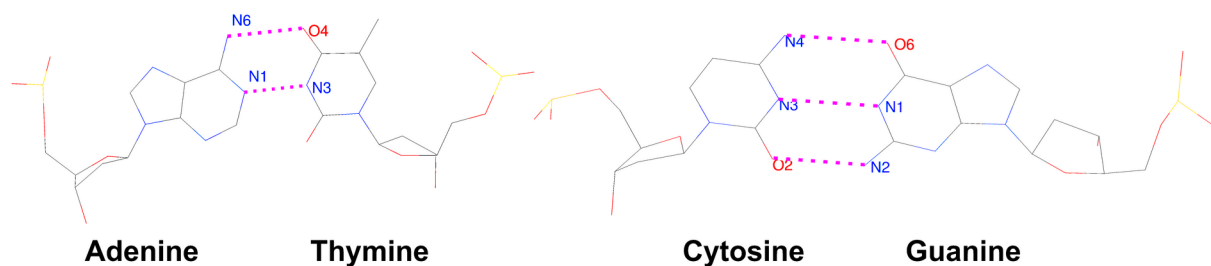


Figure 1.1: Nucleotide structure and Watson-Crick base pairing.

(a) Structure of adenine, cytosine, guanine and thymine nucleotides. The phosphate backbone, deoxyribose sugar and the nucleoside base are represented with orange, purple and blue planes, respectively. (b) Watson-Crick hydrogen bond partners. The hydrogen bonds are shown in purple dotted lines.

1.3 Mechanisms of DNA-protein interactions

There are two modes by which the proteins and DNA recognize their cognate partners- base readout or direct readout and shape readout or indirect readout. This thesis deals with the mechanism that is associated with base readout.

1.3.1 Base readout

- (b) The base readout or direct readout involves recognizing a specific sequence of nucleotides. The recognition is independent of the DNA-shape and relies on the chemical bonds formed between the protein and the DNA, each of which contributes towards the binding affinity of the DNA-protein complex. Free energy calculations show that the cognate sites have a higher binding affinity compared to non-cognate sites. Binding sites exhibit favourable interactions when the interface between the nucleotides and proteins is complementary. These favourable interactions may either be electrostatic or hydrophobic in nature.

However, in non-binding sites, there may be mismatched pairs of amino acids and nucleotides that may have for example similar charges interacting with each other. These unfavourable interactions heavily influence the binding affinity compared to the favourable interactions. This dependence is because the energetic cost of an unfavourable interaction pose is higher than the interaction energy of a favourable nucleotide-amino acid interaction. Hence, unfavourable binding partners between the DNA and protein moieties dictate the probability of a binding event [46]. This implies that the total binding energy of the complex is dependent on the specific and non-specific amino acid-nucleotide interactions. Thus, a combination of specific and non-specific interactions between proteins and DNA is essential for stabilizing the association and therefore, recognizing the target site. These chemical bonds are predominantly hydrogen bonds, electrostatic interactions (salt bridges), π - π stacking, cation- π stacking, van der Waals interactions and hydrophobic contacts [33,43–45]. These interactions will be described briefly in the following sections.

1.3.1.1 Hydrogen bonds

A hydrogen bond is a directional interaction between two partially charged electronegative atoms. An atom with covalently bound hydrogen atom acts as a donor and shares this hydrogen with the second atom, which acts as an acceptor. Owing to their directionality, hydrogen bonds are regarded as the major contributors to the specificity of DNA-protein interactions [47]. Another aspect of hydrogen bonds that makes them attractive investigative targets is that the energy required for their formation is around 10-40 kJ/mol. This energy is sufficient to account for the specificity while simultaneously being low enough to allow conformational changes occurring during binding, unbinding events and conformational sampling [48]. The hydrogen bond distribution in the major groove is partly responsible for the higher specificity of this groove. The hydrogen bond pattern in the major groove has an asymmetric hydrogen bond donor-acceptor pattern that is unique for a base-pair. In other words, the hydrogen bond donor/acceptor orientation is unique for A-T, T-A, G-C and C-G base pairs. However, in the minor groove, A-T and T-A have the same hydrogen bond donor-acceptor composition, as do G-C and C-G (Figure 1.2).

Hydrogen bonds between amino acids and bases may be classified into three types- single, bidentate and complex. Single hydrogen bonds refer to the existence of one hydrogen bond between an amino acid and the nitrogenous base. The interaction is termed bidentate when more than one hydrogen bond exists between an amino acid and nitrogenous base. Complex hydrogen bonds constitute those where an amino acid forms hydrogen bonds with more than one base [33]. A previous study has observed that nucleotide-amino acid single hydrogen bond interactions did not exhibit preferences except arginine, serine, and histidine. Bidentate interactions are the dominant force behind specific recognition of DNA, partly by increasing the binding energy due to multiple favourable interactions.

The recognition of guanine by arginine can be explained by bidentate interaction. Since arginine has donor atoms, the nucleobase must have an acceptor. If the structures of adenine and guanine are considered, they differ at the 6th position containing N6 and O6, respectively. If only O6 is used for recognition, arginine can distinguish between guanine and adenine but not thymine. On the other hand, if N7 alone is used for differentiation, then guanine and thymine can be distinguished but not guanine and adenine. This conundrum can only be solved in the scenario where a hydrogen bond is formed between N7 as well as O6. This interaction mode was proposed by Seeman et al. even before the structure for a DNA-protein complex was solved [49]. Complex hydrogen bonds have an obvious advantage in specific recognition as they form hydrogen bonds with adjacent nitrogenous bases and can therefore recognize the target site more effectively.

The hydrogen bonds between the nucleotides and amino acids may not necessarily be direct. They could be mediated via a water molecule. Since water molecules have both a donor and an acceptor, they could act as a bridge between the nucleotides and amino acids that are spatially too far to form a direct hydrogen bond. The TrpR operator complex is one such example of specific site recognition using water molecules. The trp repressor lacks sufficient direct contacts between the base and amino acids to provide sequence dependence to the binding. The crystal structure revealed the presence of three ordered water moieties at the interface that could assist recognition by forming hydrogen bonds

with the protein and DNA [50,51]. Homeodomains provide another example of the significance of water in specifying binding [52]. These proteins bind AT-rich regions in the genome that have a spine of water to provide stability to the conformation and mediate DNA-protein complex formation.

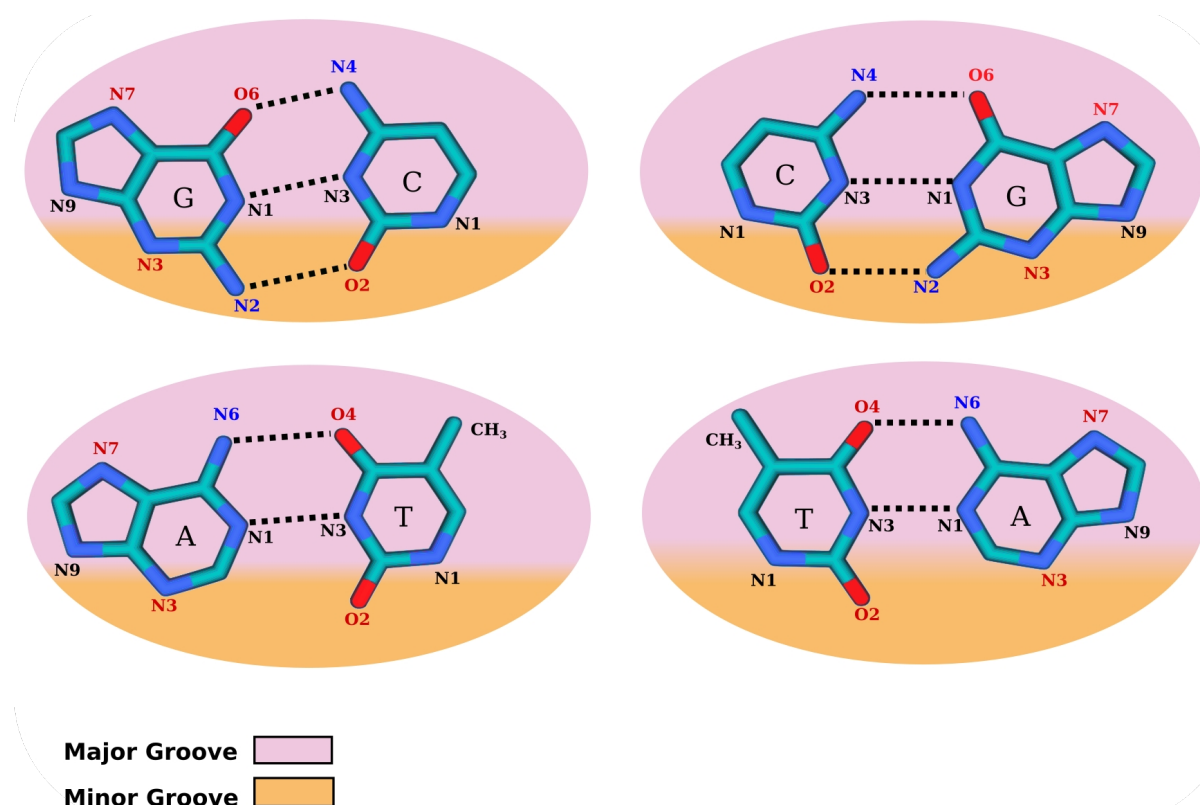


Figure 1.2: Distribution of atoms in major and minor grooves.

The nucleotides are in cyan with the oxygen atoms in red and nitrogen atoms in blue. The red and blue coloured atom labels indicate hydrogen bond acceptors and donors, respectively. The black atom labels specify the atoms that are unavailable for hydrogen bonding. The dotted line between the nucleotides indicates hydrogen bonds. The major groove region is coloured pink and the minor groove is coloured cream.

Even though there are wide-ranging examples that highlight the role of water in assisting binding, certain binding sites are entirely hydrophobic. This occurs in lac repressors where the non-specific binding sites have water-mediated interactions and the specific sites lack water [50,53]. As there is a spine of water around the DNA, the bound water molecules have to be displaced for binding to occur. Since this is an energetically

cumbersome process, the energy from non-specific binding would not be able to displace these water molecules. Therefore, it ensures specific binding indirectly.

Another class of non-classical hydrogen bonds that has been gaining momentum as an essential player in mediating DNA-protein interaction is the CH...O interaction. This interaction involves C5/C5-met of cytosine/thymine and an oxygen from the amino acid. CH...O bonds are weak attractive bonds and their role in gene regulation has been proposed because strong bonds between protein and DNA would not allow rapid changes in gene expression [54].

1.3.1.2 Salt bridges or electrostatic interactions

Salt bridges or electrostatic interactions are non-covalent interactions between two oppositely charged moieties. These interactions are indirectly responsible for the enhanced stability of the DNA-protein complex via their enthalpy contribution to the free energy. The formation of salt bridges is accompanied by the displacement of water molecules and ions near the DNA, providing entropic contribution to the free energy [55,56]. The salt bridges are commonly formed between the negatively charged phosphate backbone and the cationic amino acids: arginine, lysine and histidine (generally at pH<7, though it is dependent on the local environment). Studies have shown that the formation of salt bridges induces bending of the DNA and thereby creates a complementary interface for DNA-protein interaction. This bending results from backbone charge neutralization by the positively charged residues of the protein [44]. Salt bridges have also been implicated to be a part of specific recognition of target DNA sites. Salt bridges assist in recognition by contributing to a network of favourable interactions or by restricting the size of the DNA recognition site through DNA wrapping [57–59].

1.3.1.3 π - π stacking interactions

π - π stacking is an attractive non-covalent interaction between molecules containing π -bonds. These interactions have been shown to occur in cyclic and non-cyclic molecules with π bonds with an interaction strength in the range of -0.5 to 2 kcal/mol [60]. The

proximity of π electron-rich systems does not ensure a π - π stacking interaction and the angle between the planes of the two molecules is crucial. Stacking may occur with the planes perpendicular or parallel to each other. The energy of a stacking interaction is equivalent to that of a hydrogen-bonding interaction [61]. Pyrimidines contribute to 62% of stacking interactions. Among the amino acids, phenylalanine (44%) and tyrosine (33%) have a higher proportion of stacking interactions as compared to tryptophan (13%) and histidine (11%). Tryptophan is under-represented, which could be due to its rare occurrence in proteins. However, the natural frequency of occurrence of histidine is similar to that of phenylalanine and tyrosine. Therefore fewer histidine mediated π - π stacking interactions could indicate lesser propensity to form these interactions [61]. The aromatic amino acids intercalate between bases and cause the deformation of DNA, which is one of the mechanisms by which the protein finds its cognate site on the DNA [33]. A previous study has shown that π - π stacking interactions increase in the hydrogen bonding capacity of the DNA bases. The hypothesis proposed was that stacking promotes repulsion between the bases and increases the hydrogen bonding capacity [62].

1.3.1.4 Cation- π stacking interactions

The non-covalent attractive interactions involving π electron-rich aromatic ring systems and a proximal cation are termed as cation- π interactions [63]. They play a significant role in biologically relevant processes, since it has an energetic contribution of 2.6 kcal/mol [63–67]. These stabilizing interactions can be quantified using thresholds of distance, van der Waals forces and electrostatic forces. Since π electron-rich systems prefer a hydrophobic environment and cationic groups prefer a hydrophilic environment, there was a debate about the location of this interaction. However, they have been proven to occur on the solvent-exposed surface [66]. Between lysine and arginine, arginine is a dominant participant in this interaction even when the total number of both residues in the dataset is almost the same [63]. Purines are more frequently involved in cation- π interaction than pyrimidines [63]. An example of cation- π mediated recognition is the glycosylase AlkA, which has a hydrophobic pocket with an abundance of aromatic residues in the active site of the enzyme. This enzyme processes the DNA bases that have undergone alkylation and therefore have a positive charge. A positively charged

substrate combined with the aromatic rings in the enzyme active site implies the role of cation- π interaction for enzyme function [68].

1.3.1.5 Van der Waals interactions and hydrophobic contacts

Van der Waals forces are “the attractive or repulsive forces between molecular entities (or between groups within the same molecular entity) other than those due to bond formation or to the electrostatic interaction of ions or of ionic groups with one another or with neutral molecules” as quoted from the IUPAC Gold book [69]. These forces are present between all entities within a specified distance. Therefore, their role in specific DNA-protein interaction was considered negligible. However, through the study of three-dimensional structures of DNA-protein complexes, it was revealed that two-thirds of all interactions in DNA protein complexes are van der Waals interactions [33].

They influence the recognition process by limiting the permissive steric conformations of the interacting partners in the DNA-protein complex [33,70]. These interactions participate in maintaining shape complementarity between the DNA-protein interface and are also directly involved in recognizing methyl groups of thymine bases [12,71,72]. It has also been suggested that alternating van der Waals and hydrogen bonds are markers for the specific recognition site [73]. The energy provided by van der Waals interactions, with each interaction providing -0.5 to 1Kcal/mol, positively correlates with the stability of the DNA-protein complex [47,70,74]. Increased stability at the target sites due to a larger number of van der Waals interactions could contribute towards specificity. Van der Waals interactions also assist in imparting specificity in the minor groove. In the minor groove, hydrogen bonding pattern alone is insufficient to recognise the exact DNA sequence (section 1.3.1.1).

Therefore a combination of hydrogen bonds, hydrophobic contacts and van der Waals interactions promote specific recognition in the minor groove [47]. The intercalation of hydrophobic side chains between nucleobases is also associated with specific recognition in DNA-protein complexes such as those involving purR, TBP, Sso7d and

Sac7d. These hydrophobic interactions often stabilize the kinks introduced by the conformational change to A-DNA [75–79].

1.3.2 Shape readout

Shape readout, also called indirect readout, involves the recognition of cognate DNA based on features other than the nucleotide sequence. These characteristics include strandedness, local geometric variations, global changes in the type of DNA (A-form, B-form, or Z-form), or the ease of deformation.

1.3.2.1 DNA strandedness

DNA can exist in the single-stranded (ss-DNA) or ds-DNA form inside the cell. ss-DNA is transient and is mainly formed during the processes of replication [80,81], repair [82–84] and recombination [85–87]. Some insights into the differentiation between ss-DNA and ds-DNA-binding to proteins have been provided by three-dimensional (3D) features of protein bound to ss-DNA [88–92]. In particular, while the ds-DNA-binding proteins have diverse binding interfaces, many ss-DNA-binding proteins possess a structural fold known as Oligonucleotide Binding (OB) fold, which can be used to differentiate between ss-DNA and ds-DNA binders [93]. Other methods to predict the strandedness of DNA bound by proteins include sequence-based and machine learning algorithms, which combine structural and sequential features [94,95].

1.3.2.2 Local geometric variations

The local geometric variations may arise from DNA sequences, such as AT-rich stretches, GC-rich stretches, A- or T-tracts, CpG methylation sites, CpA steps, RY alternating bases, TATA box, RpY, and YpR steps [96,97,106,98–105]. The presence of these bases change the helical parameters of the double helix locally, often altering the width of the minor groove and changing the electrostatic potential. Proteins recognize these changes, most notably the ones in the electrostatic potential. Some of the examples of proteins that use this mechanism include TBP-binding to TATA-box during transcription

initiation, wrapping DNA around a nucleosome, binding of homeodomain proteins such as Ultrabithorax (Ubx) to its promoter and p53 protein binding to cognate sites [105,107,108]. Most of these proteins employ arginine residues, which form short linear motifs that are functionally important for recognition [109,110].

1.3.2.3 Global conformational changes

The deviation of the global conformation from the classic B-form has been associated with binding certain classes of proteins such as endonucleases and proteins involved in innate immunity [111,112]. The A- and Z-DNA forms differ from the B-form with regards to the width and depth of the two grooves and the orientation of the deoxyribose sugar along with many other features that are discussed in detail elsewhere [113](Figure 1.3, Table 1.1).

While both A and B-DNA have a right-handed helix, Z-DNA possesses a left-handed helix. It has been hypothesized that the left-handed characteristic of the Z-DNA is a result of B-DNA underwinding by helicases and polymerases. Though A-DNA nucleotide bases form contacts with their cognate proteins, Z-DNA recognition is exclusively mediated by conformational changes. In A-DNA, the deoxyribose sugar is more exposed to the solvent, involving more non-polar interactions compared to B-DNA. An example of the recognition of the A-DNA conformation is by endonucleases, where the exposed phosphates in A-DNA facilitate cleavage of the DNA [114]. Z-DNA has a distinctive zig-zag positioning of phosphates, which acts as a signature for Z-DNA-binding proteins and a structural code has been proposed for the same [115].

Table 1.1: Parameters differentiating A-, B- and Z-DNA.

Parameters	A-DNA	B-DNA	Z-DNA
Helix sense	Right	Right	Left
Residue per turn	11	10 (10.5)	12
Axial rise (Å)	2.55	3.4	3.7
Helix pitch (°)	28	34	45
Base pair tilt (°)	20	-6	7
Rotation per residue (°)	33	36 (34.3)	-30
Diameter of helix (Å)	23	20	18
Glycosidic bond configuration			
dA, dt, dC	anti	anti	anti
dG	anti	anti	syn
Sugar Pucker			
dA, dt, dC	C3' endo	C2' endo	C2' endo
dG	C3' endo	C2' endo	C3' endo
Intrastrand phosphate-phosphate distance			
dA, dT,	5.9	7	7
dG	5.9	7	5.9

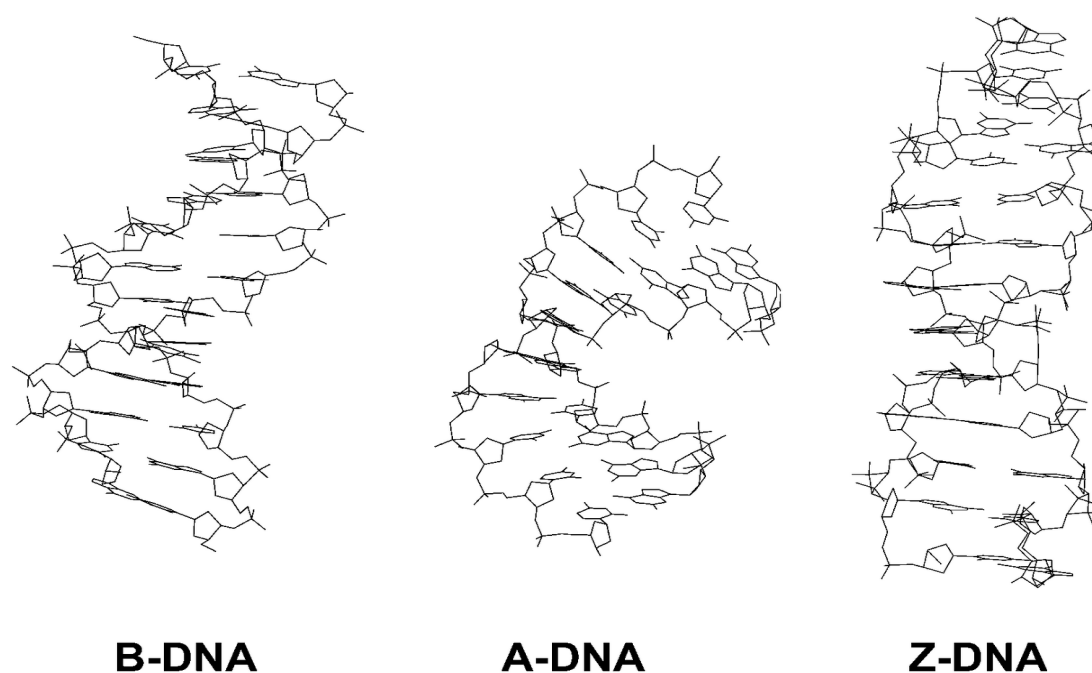


Figure 1.3: Types of global DNA conformations.

The figure was generated using PDB files 5eyq (B-DNA), 4izq (A-DNA) and 2ei1 (Z-DNA).

1.4 Previous investigations on the universal recognition code

Numerous features about DNA-protein binding have been established as we have seen in the previous sections. However, the guidelines that dictate binding between protein and DNA molecules are yet to be substantiated. Seeman et al.'s [30] study in 1976 showed the requirement for more than two hydrogen bonds for specific recognition of nucleotides. The study predicted that the hydrogen bond between arginine and guanine would be specific without having structural data of DNA-protein complexes. The structures of DNA-protein complexes solved later have validated his prediction. This hinted towards a definite pattern for DNA recognition and shifted the research focus towards analysing crystal structures to explain the molecular basis for DNA-protein interaction.

Pabo and Sauer proposed the term “Recognition Code” for representing the interactions governing the formation of the DNA-protein complex [116]. Though attempts to find the recognition code were made, information from crystal structures revealed diverse

DNA-protein interaction modes. ***It was observed that more than one amino acid recognized a single base, and conformational changes in the protein and DNA facilitated complex formation. This observation pointed towards the absence of a universal recognition code [117].*** However, due to the increasing availability of structural information, studies were still carried out to find specific recognition patterns between DNA and proteins. These studies focused on hydrogen bonds in conjunction with other interactions [32,72,118–121]. But these efforts only focussed on interactions of a single nucleotide or a particular family of DNA-binding proteins. Statistical analysis of DNA-protein crystal structures has led to knowledge about amino acid nucleotide preferences but not to a generalized recognition code [31–34]. However, analyses of families of DNA-binding proteins have resulted in specific nucleotide-amino acid association preferences that are only applicable to that family. The zinc finger protein and the Transcription Activation factor-Like effector (TAL Type III effector) families are the only two DNA-binding protein families for which the recognition code is known yet.

1.4.1 The zinc finger recognition code

The zinc finger code delineates a 1:1 relationship between amino acids and nucleotides. It was discovered in 1994 through phage display experiments when a library of zinc finger proteins specifically bound DNA triplets [35–37]. This was further compiled and validated through the rational design of zinc fingers that would bind a given DNA sequence [122].

The code states that the amino acids at positions +6, +3, -1 and +2 would recognize three nucleotides on the forward strand and the last nucleotide on the reverse strand, respectively (Figure 1.4). In the illustrative example the DNA sequence AGTC is recognized by glutamine, serine, arginine and serine at the +6, +3, -1 and +2 positions. Each unit of zinc finger can bind a four nucleotide DNA sequence. Since each unit is independent of the other, multiple zinc fingers can theoretically be combined to recognise a larger sequence if they have overlapping binding sites. However, In practice, not all DNA sequences can be specifically recognised by zinc fingers [23]. In the case of the TATA box, not all of the 1:1 interactions predicted by the code are conserved [123]. This deviation

from the code may be attributed to the shape of the TATA-box site, which deviates from the classical B-DNA form.

Using the recognition code, custom zinc finger nucleases were designed, which could be used for gene editing. It involved attaching a zinc finger protein to FokI nuclease, which is a non-specific nuclease, to cut the target site of the zinc finger protein [124,125]. It has been successfully used for target gene mutation and gene replacement in *Drosophila melanogaster* [126,127]. However, off-target effects remain a challenging problem to address [126,128].

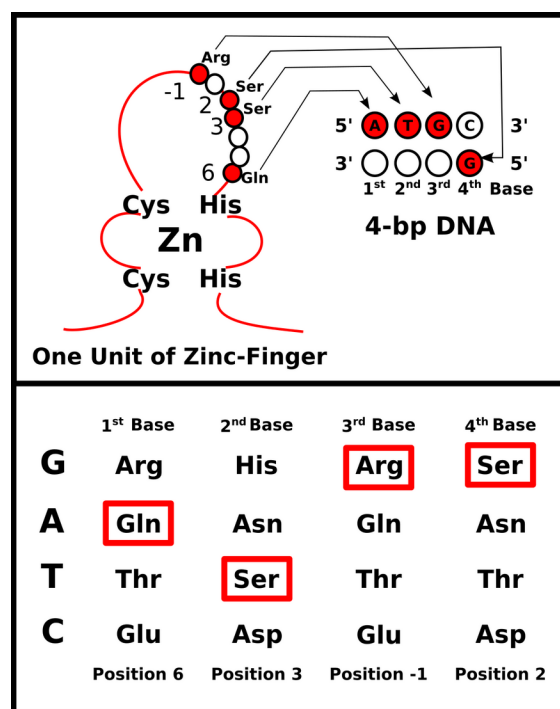


Figure 1.4: Schematic of zinc finger recognition code (Recreated from [122]). Amino acids interacting with DNA are shown as red circles on the zinc finger. The arrows from each amino acid point to the nucleotide it recognizes. The table at the bottom describes the 1:1 correlation between the nucleotide at each position (rows, single-letter representation) and the amino acid (columns, three-letter representation) that it would bind specifically. Red rectangles highlight the amino acids involved in ATGC binding.

1.4.2 The TAL type III effector recognition code

The TAL type III effector proteins are secreted by the pathogen *Xanthomonas* when it infects plants. It activates transcription in the plant nucleus by binding to the host DNA via its central repeat domain. The TAL effector has 17.5 repeats of a 33-35 amino acid sequence (17 complete sequences and one-half sequence). In this repeat sequence, the amino acids at positions 12 and 13 were variable while the rest of the region was conserved. In 2009 it was discovered that the variable amino acids at positions 12 and 13 bind to one base specifically, forming the TAL type III effector recognition code [38,39]. Therefore each repeat binds to one nucleotide (Figure 1.5). Though only two residues in this repeated sequence are involved in the process of recognition, all of them are essential.

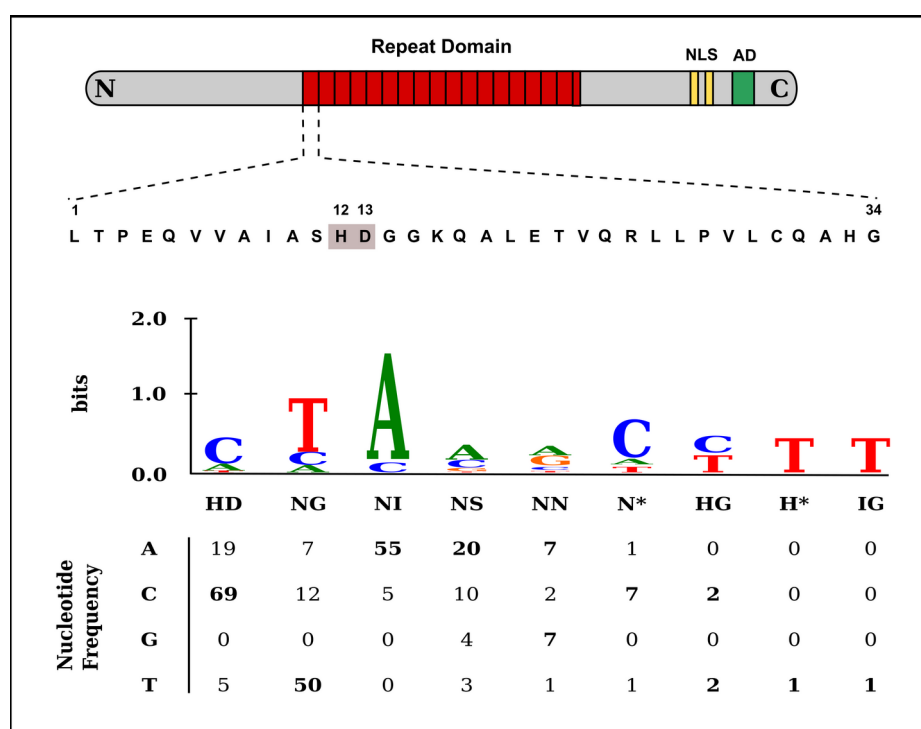


Figure 1.5: Schematic of TAL type III effector recognition code

(Recreated from [38]). TAL architecture is at the top of the figure with red, yellow and green regions depicting the repeat region, nuclear localization signals and the activation domain, respectively. The sequence of the repeat region is shown in single amino acid letter codes with positions 12 and 13 highlighted in grey. The table at the bottom is the frequency of interaction of each pair of amino acids at positions 12 and 13 with the four

nucleotides shown in single letter codes. Adenine, cytosine, guanine and thymine are depicted in green, blue, orange and red, respectively.

Similar to Zinc Finger nucleases, artificial nucleases were also engineered for TAL effectors using FokI called the TALENs. They find applications in genome editing just as zinc finger nucleases and possess the same challenge of off-target effects [129–132].

1.5 Proposed approach for discovering the molecular basis of base readout

The previous approaches had certain limitations, which we wish to address using our method. The first problem was that of structural diversity among DNA-binding proteins. This structural diversity makes it difficult to compare different DNA-binding proteins to find common recognition patterns. To overcome this problem, we made use of CLICK, a topology independent structural alignment software [40,41] (refer to Chapter 1, section 1.6.1).

While CLICK can superimpose structures with different folds, another feature that is apt for the current study is the flexibility to select any atom/(s) as representatives to guide the alignment. Since we were interested in looking at interactions, we used the nucleotide hydrogen bond donor and atoms to guide the alignment, resulting in the superimposition of the DNA-protein interface. This superimposition allowed us to compare hydrogen bond similarities between different DNA-binding proteins. We also analysed the influence of neighbouring nucleotides on the specificity of interactions, which has not been studied yet. An example alignment of the regions of an endonuclease (PDB ID: 1mow) and a transcription repressor (PDB ID: 2drp) bound to the same trinucleotide sequence GAA shows similar orientations of hydrogen bonds (Figure 1.6). The region that shows similar hydrogen bonds belongs to α helix in the transcription repressor and β sheet in the endonuclease.

Therefore our CLICK-based hydrogen bond similarity approach is one of the ways in which we can decipher protein family-independent DNA-protein interaction specificity. We used

the data obtained from CLICK superimpositions to develop a model that predicts amino acids that would specifically recognize a given sequence of DNA. We validated the model using various *in-silico* and *in-vitro* methods.

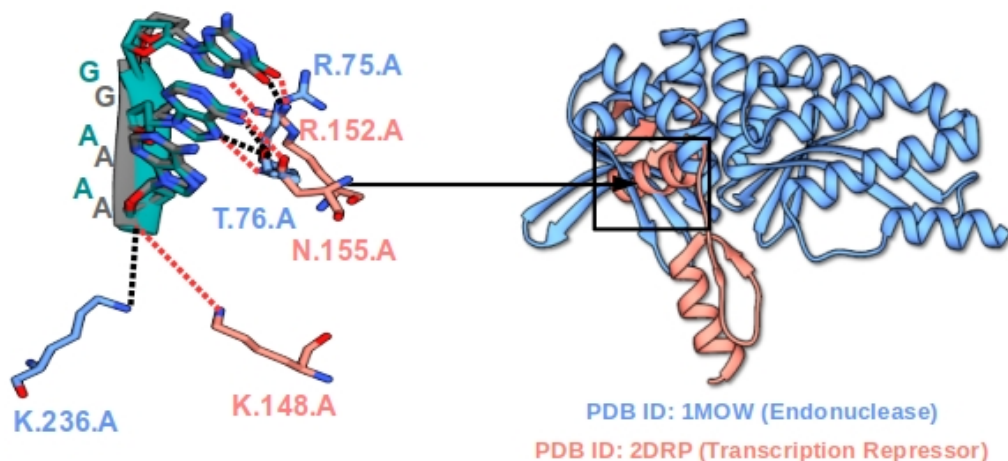


Figure 1.6: CLICK superimposition of trinucleotide GAA.

Nucleotides are shown in grey and cyan with single-letter codes for 2drp and 1mow, respectively. Amino acids are coloured blue (1mow) and salmon (2drp) and represented in single-letter codes. The hydrogen bonds are depicted as black and red dotted lines. The superimposed structure on the left belongs to the region of the protein shown in the black rectangle on the right.

We also addressed the reverse problem of recognizing the DNA sequence, given a protein. For this problem, we also utilized the same hypothesis that the hydrogen bonding pattern is conserved across protein families. We performed CLICK alignments between a template dataset of DNA-protein complexes and the query protein guided by the hydrogen bond donor and acceptor atoms. We developed a method to predict the putative pose of the DNA and the nucleotide sequence that would bind a given protein. It was validated using *in-silico* approaches.

This study is based on structures of 235 proteins in the training set, which is the biggest limitation of our method. Many of the interactions are not captured in this set and therefore, it affects the accuracy of our approach for predicting specific DNA-protein interactions. With an increase in the available structural data, this method can be refined to obtain the protein family-independent nucleotide-amino acid specificity preferences.

1.6 Introduction to techniques

This section describes the principles of the techniques used to implement the proposed approach and refers to the sources to obtain the latest advances in the field.

1.6.1 CLICK

CLICK, is a topology independent structural alignment software [40,41]. CLICK interprets biomolecular structures in terms of a collection of representative atoms, called cliques. The clique size varies between 3-7 points. The algorithm initially matches three atoms and then progressively increases the clique size to obtain regions with the highest similarity. The least-squares fit is then performed with cliques from the two biomolecules disregarding the atom connectivity and thereby bypassing secondary structure differences. The structural similarity of the aligned structures is evaluated by two measures- Root Mean Square Deviation (RMSD) and % Structural Overlap (% SO). RMSD is the average distance between the points from the two structures after superimposition by CLICK. %SO is the percentage of atoms that have matched between the two structures.

$$RMSD(v, w) = \sqrt{\frac{1}{n} \times \sum_{i=1}^n \|v_i - w_i\|^2}$$

where i to n are the superimposed atoms in structures v and w .

$$\%SO = 100 \times \frac{N}{\min(M1, M2)}$$

where N is the number of atoms with RMSD < user-defined threshold, $M1$ is the number of atoms in structure v and $M2$ is the number of atoms in structure w .

Though there are many more structural alignment programs [133–136], CLICK allows structures from different folds to get superimposed (Figure 1.7). In the example alignment, arginine repressor (1xxa) is superimposed with translation initiation factor 3 (1tig) using C α atoms as representatives. Though they look similar, the β sheet is parallel

in 1tig and anti-parallel in 1xxa. They also have different CATH [137,138] (protein structure classification database) IDs with 3.30.110.10 for 1xxa and 3.30.1360.40 for 1tig.

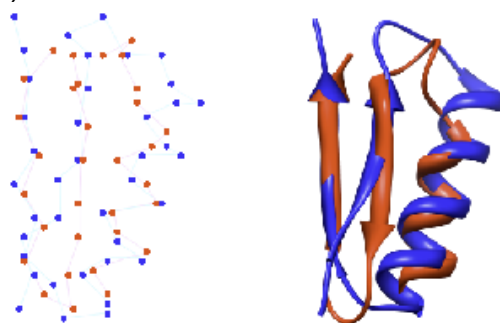


Figure 1.7: CLICK superimposition of proteins with PDB IDs 1xxa (red) and 1tig (blue). The left side of the figure represents the structures in the form of Ca atoms. The right side shows superimposed structures. The figure is reproduced from the CLICK webserver- <http://cospi.iiserpune.ac.in/click>

1.6.2 Molecular Dynamics (MD) simulations

The structures of proteins and DNA-protein complexes are a static snapshot of their orientation in an artificial environment. However, within the cells, these molecules are in constant motion, which could be important for their function. These motions can be captured through molecular dynamics (MD) simulations, which solve Newton's equations of motion to derive the putative movements in molecules. Apart from studying molecular motion, MD simulations can also provide some insights into the stability of biomolecules that have bound to each other [139]. In our study, we made use of them to understand the stability of DNA-protein complexes.

The intra- and inter-atomic forces contribute to the stability of biomolecules. These forces have been defined by forcefields such as AMBER [140–142], CHARMM [143,144], OPLS-AA [145,146] and GROMOS [147,148]. All of them employ the same general form for calculating energy that involves computing the bonded interactions and non-bonded interactions. However, they differ from each other in the details of their implementation of these terms.

MD simulations require that all the biomolecular structure components have defined parameters, which will allow the calculation of the energy terms in the forcefields. These parameters are crucial in ensuring the accuracy of the simulations. Once these parameters have been calculated, the biomolecular structure is then minimized so that the simulation starts at the lowest energy state for the structure. This optimum structure is immersed in a solvent (which is water in most cases). The system is neutralized with counter-ions so that the overall charge of the system is zero. To reduce the effects of external factors such as temperature and pressure, steps of equilibration are carried out that will ensure that the system has a constant temperature and pressure. All of these steps can be carried out using multiple simulation packages such as NAMD [149], GROMACS [150,151], GROMOS [152] and AMBER [153] to name a few. There have been many reviews that describe the principles of MD simulations and the recent advances in the field [154–157].

1.6.3. Binding free energy calculations

The binding free energy of association between biomolecules can be estimated through a combination of molecular mechanics energies with either Poisson-Boltzmann (MM/PBSA) [158,159] or generalized Born surface area (MM/GBSA) [160,161]. The binding free energy calculations in this study were performed using NAMD using the MM/GBSA protocol [162]. Using the formula-

$$G_{tot} = H_{MM} + G_{(solv-pol)} + G_{(solv-np)} - T \Delta S_{conf}$$

Where G_{tot} is the total energy of the system, H_{MM} is the sum of bonded and Lennard-Jones energy terms, $G_{solv-pol}$ and $G_{solv-np}$ are the polar and non-polar contributions of solvation energy and T and S_{conf} represent the temperature and conformational entropy terms. For detailed information refer to the cited sources [163–166].

1.6.4 Protein structure modelling

The structures of proteins provide invaluable insights into their function. The experimental determination of three dimensional structure (3D) of proteins is cumbersome

and long process. Within the cells, the proteins fold into their native structure with just the sequence as guidelines. Therefore, the sequence contains all the information needed to generate the 3D structure. Protein modelling tools rely on the sequences along with knowledge-based potentials to generate models that resemble native protein structure conformations. There are multiple modelling softwares such as MODELLER [167,168], ROSETTA [169–171], AlphaFold2 [172] and iTASSER [173,174].

While iTASSER is an *ab-initio* method that does not require structures of homologous proteins to make predictions, MODELLER is a comparative homology-dependent modelling software. ROSETTA on the other hand has both homology-dependent and independent modelling versions. AlphaFold2 is a novel machine learning algorithm that depends solely on the protein sequence to predict its structure. However, it incorporates evolutionary information in the form of multiple sequence alignments (described in the next section).

In this work, we have used models generated by MODELLER and AlphaFold2.

1.6.5. Multiple Sequence Alignments (MSA)

Multiple Sequence Alignments (MSA) are aligned sequences of related proteins. These are valuable in terms of interpreting the evolutionary conservation of an amino acid, which would reflect its importance in protein structure/function [175,176]. The protein homologues are found most often using a BLAST/PSIBLAST algorithm [177,178]. While BLAST performs a single iteration and finds the homologues, PSIBLAST performs multiple iterations. The advantage of using multiple iterations is that it generates a customized profile of the protein after the each iteration, which helps it find distantly related homologues in the next iteration.

The sequences found by BLAST may be formatted into a multiple sequence alignment or there are other methods like Clustal Omega [179], ClustalW [180] or MUSCLE [181] that can generate MSA from the given input sequences. The recent

advances, challenges and limitations of the current methods are described in the cited sources [182–184].

1.6.6 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is a molecular biology technique that allows the amplification of a DNA sequence *in-vitro*. The process involves a target DNA, a heat-stable DNA polymerase (from *Thermus aquaticus*), free deoxyribonucleotides (dNTPs), short DNA sequences complementary to the target DNA sequences (primers) and an appropriate buffer [185–187]. The primers are designed such that they are present at the beginning and end of the target DNA sequence and on opposite strands. There are three steps in a PCR- denaturation, annealing and extension. The denaturation step involves heating the target DNA to high temperatures (90-98°C) to separate the double-stranded DNA into single strands. In the annealing process, the temperature is reduced (50-70°C) and the primers form a duplex with the complementary region on the target DNA. Finally, in the extension process, the DNA polymerase uses free deoxyribonucleotides to extend the primer sequence and generates a copy of the target DNA. This cycle is repeated 25-40 times to generate amplified target sequences. There are many variations of the PCR process, depending on the applications that are described elsewhere [188,189].

1.6.7 Restriction-free (RF) Cloning

Restriction-free (RF) cloning is a method of integrating a target DNA sequence within a circular plasmid without the use of restriction enzymes. The cloning is entirely based on PCR reactions [190]. The design of primers plays a crucial role in this process. The primers must have complementary regions to the plasmid (vector) and the target DNA. In the first PCR reaction, the target DNA is the template and the product is the target DNA with the terminal region complementary to the vector. In the second PCR, the vector acts as the template and the product from the first PCR acts as a primer. The product of the second PCR is the vector with the target DNA of interest.

In order to eliminate the vector without the target DNA, DpnI restriction enzyme is used. The vector is extracted from a bacteria and therefore has hemi-methylated adenines. The PCR product, on the other hand, lacks methylation. DpnI selectively cleaves methylated products and therefore degrades the vector without the target DNA (Figure 1.8).

The DpnI treated product is then transformed and screened to obtain the desired clone. This method was used for cloning Myc, Ets1 and EcoRV in this study.

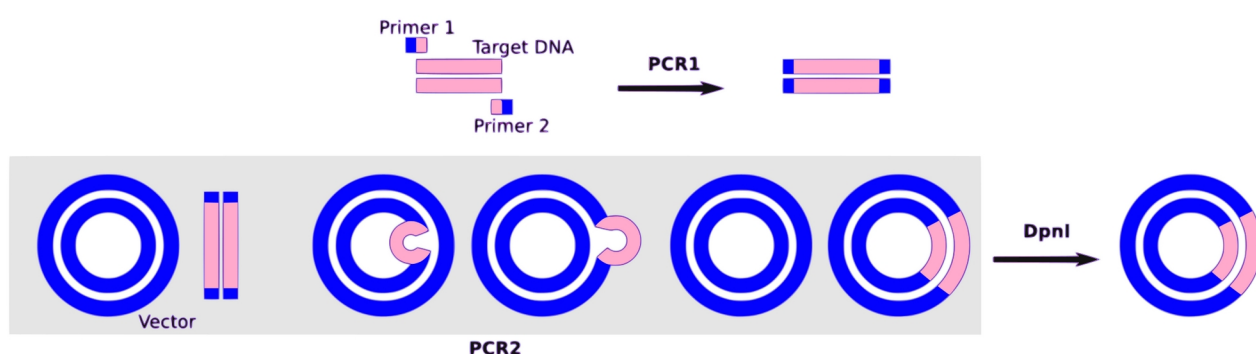


Figure 1.8: Schematic of RF cloning.

The vector is shown in blue and the target DNA in pink.

1.6.8 Electrophoretic Mobility Shift Assay (EMSA)

Electrophoretic Mobility Shift Assay (EMSA) is a semi-quantitative *in-vitro* assay to determine the binding of DNA with proteins [191]. It relies on the molecular weight of the DNA as a readout of binding. If the DNA is unbound, it has a lower molecular weight and can travel faster in an electric field. On the other hand, if the DNA is bound to the protein, the complex is heavier and has a retarded movement in an electric field (Figure 1.9). The DNA is stained with a fluorescent dye and the position of the bands indicates binding. To have an estimate of the binding affinity, the proteins can be added in different concentrations. The formation of the DNA-protein complex can be plotted as a function of the protein concentration to find the apparent binding affinity. There are many reviews that discuss in detail the principles and challenges associated with this technique [192,193].

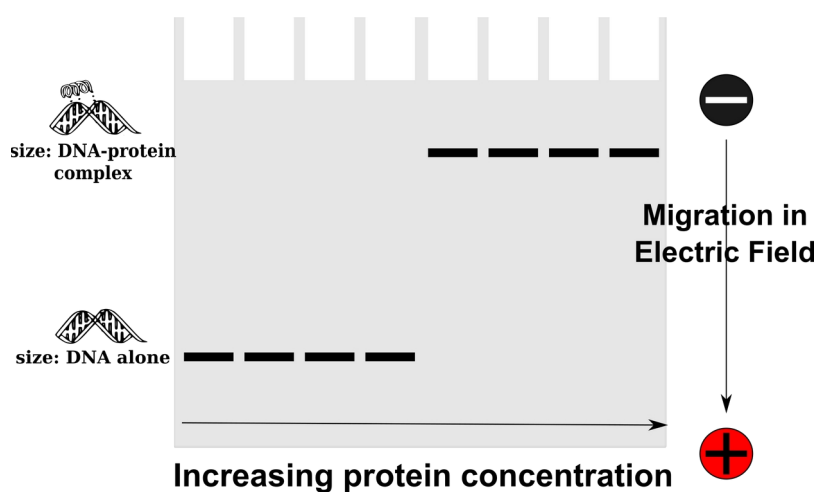


Figure 1.9: Schematic representation of an EMSA experiment.

The black bands indicate the position of the DNA (bottom region is unbound and top region is bound). The direction of movement is from top well to the bottom. The protein concentration is lowest at the first well on the left and increases gradually towards the right.

1.6.9 Microscale Thermophoresis

Microscale Thermophoresis (MST) is an *in-vitro* technique to quantify the binding affinity (k_d) of biomolecules. It is based on thermophoresis, which is the directed movement of molecules in a thermal gradient based on their physical properties like size, charge, hydration shell and conformation [194]. The samples are loaded within a capillary and a temperature gradient is induced by an infrared laser (Figure 1.10 (a)). The movement of molecules is monitored through fluorescence (Figure 1.10 (b)). Therefore the technique requires at least one of the interacting molecules to exhibit either intrinsic fluorescence or be fluorescently tagged. In our study, we have tagged the protein with a fluorescent dye. The fluorescence pattern is different for a bound and an unbound protein (Figure 1.10 (c)). The concentration of one of the interacting molecules is kept constant and serial dilutions are used for the second molecule. This results in a curve, to which we can fit the appropriate binding models depending on the mode of interaction (Figure 1.10 (d)).

The law of mass action is used to calculate the bound fraction of the interacting molecules. The graph of bound fraction can then be used to estimate the K_d as half of the

point at which fluorescence is saturated at the maximum value. This technique has a few advantages over the conventional techniques of binding affinity estimation which include increased sensitivity (fluorescence monitoring) and reduced sample volumes. More details about this experiment can be found elsewhere [195–199].

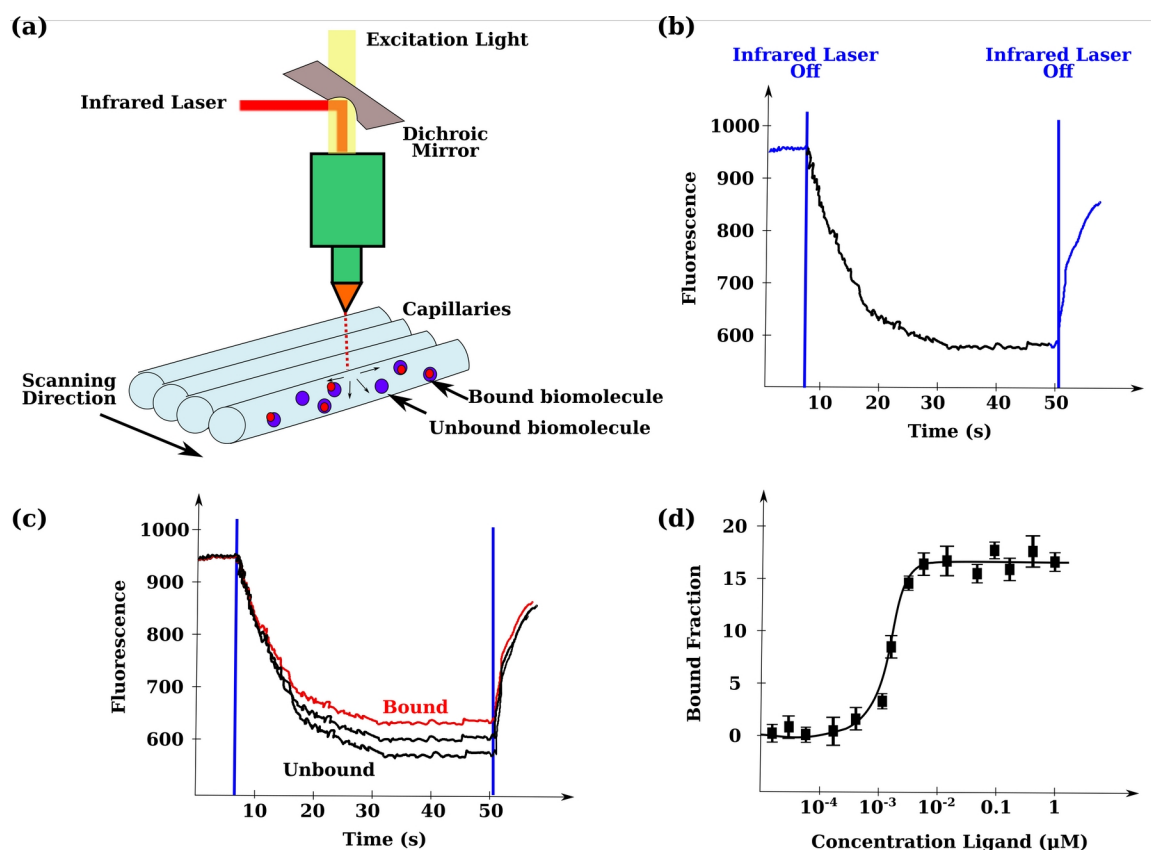


Figure 1.10: Schematic of MST experiment

-depicting experimental setup (a), example readout (b), difference in bound and unbound biomolecules (c) and curve to find K_d (d).

1.6.10. Nano Differential Scanning Fluorimetry (NanoDSF)

NanoDSF is a low volume label-free method that uses the intrinsic fluorescence of tyrosine and tryptophan to monitor the protein conformation [200]. The method relies on the red-shift of fluorescence of the aromatic residues as a result of protein unfolding to calculate melting temperature (T_m). T_m is the temperature at which 50% of the protein is in the unfolded form (see reference for calculation using fluorescence- [201]). In nanoDSF,

the T_m is calculated by monitoring the ratio of fluorescence at 330 nm and 350 nm with increasing temperature [202].

In this study, we have used nanoDSF to analyse the difference in the T_m values when different DNA substrates bind to a DNA-binding protein. The T_m value of a protein is influenced by its immediate environment. The addition of a ligand could increase the number of interactions in the protein, causing the Gibbs free energy to decrease. The decrease, indicative of protein stability results in an increased T_m value in the presence of ligands [203]. The low sample volume and non-requirement of labelling makes it advantageous over the traditional methods of assessing ligand binding. Additional information can be obtained from the following sources- [204,205].

1.7 Chapter summary

This chapter was an introduction to the DNA-protein specificity problem and outlines the approaches used by previous studies to investigate it. We also briefly describe our approach to obtain family-independent nucleotide-amino acid interaction specificity and the techniques used to implement this approach.



CHAPTER 2: Prediction of amino acids for a DNA sequence

OVERVIEW

- Compiling and characterizing a dataset of sequence-specific DNA-protein complex structures
 - Analysis of amino acid composition of DNA-protein complexes
 - Analysis of DNA-protein specificity in molecular interactions
 - π - π stacking
 - Cation- π stacking
 - Hydrogen bonds
 - Comparison of spatial orientation of DNA-protein hydrogen bonds
 - Mononucleotide superimpositions
 - Dinucleotide superimpositions
 - Trinucleotide Superimpositions
 - Scoring scheme for predicting amino acid binding partners of a given DNA
 - Measuring the accuracy of the predictions on training and testing set
 - Chapter Summary
-

The previous chapter illustrated that the analysis of the 3D structures of DNA-protein complexes has led to family-specific recognition preferences and briefly outlined the approach that we have proposed to find family-independent DNA-protein specificity preferences. This chapter deals with the methodology to analyse amino acid-nucleotide specificity preferences. We then used these preferences to predict amino acids that would specifically recognize any given DNA sequence.

2.1 Compilation and characterization of DNA-protein complex structure dataset

As discussed in the previous chapter, the focus of our work was to understand the basis for sequence-specific DNA-protein binding. Therefore, the dataset was curated to include only the DNA-protein complexes where the proteins recognise a specific DNA sequence. We further limited our study to B-form DNA complexes to avoid any possible effects on binding due to the overall shape of the DNA.

The steps followed to construct the dataset include-

- i. We obtained 3467 DNA- protein complex structures from the RCSB Protein Data Bank (PDB) [206] by using the following parameters - 'Molecule Type=complex, Experimental Method=X-RAY, and Chain Type: there is a Protein, and a DNA chain but not any RNA or Hybrid and Resolution is between 0.0 and 3.0 Å.'
- ii. To avoid over-representing certain protein families, we compiled a 40% non-redundant set using h-CD-HIT [207]. We performed three iterations – 1st step of clustering at 90% identity followed by 60% and finally 40%, which resulted in 669 clusters.
- iii. We chose a comparatively permissive distance cut-off and angle range to account for all possible hydrogen-bonded amino acid-nucleotide pairs and experimental errors. The cluster representative was the structure having a chain with the maximum number of DNA-protein hydrogen bonds. Wherever there were multiple chains with maximum hydrogen bonds, the structure with the higher total number of hydrogen bonds (all

chains included) was chosen. If two or more structures had the same number of overall hydrogen bonds, the structure with the better resolution was the representative. From 669 clusters, we removed 81 clusters with no DNA-protein hydrogen bonds, which left us with 588 clusters with 588 representative structures.

- iv. 422 structures remained after screening for double-stranded B-form DNA using Nucleic Acid Database (NDB) [208]. NDB missed filtering some entries that did not belong to the double-stranded B-form DNA category. Further, some of the structures had non-specific DNA-binding activity, such as polymerases. We removed both these types of structures by manual curation using text mining and visualization of structures. This resulted in 285 structures. (A complete list of PDBs that were removed by manual curation is in Appendix 1)
- v. In homo-oligomeric protein assemblies, we retained unique chains. The 285 structures were randomly divided into 235, and 50 PDBs in the training and testing set, respectively.

We characterized the amino acid composition and the molecular interactions in the dataset of DNA-protein complexes.

2.1.1 Analysis of amino acid composition

We computed the amino acid composition of the training and testing dataset. We performed the χ^2 test to check for normality, and it showed a Gaussian distribution at a 95% confidence level. We then compared the amino acid distribution between the training and testing set using Student's t-test, and it did not show any difference at 95% confidence (Appendix 2). The training set was also compared with the dataset of non-DNA-binding proteins to identify amino acid signatures of DNA-binding proteins.

We collated a set of non-DNA-binding proteins from the Uniprot database (2017) [209] by providing keywords with functions unrelated to DNA-binding, such as – cell adhesion, locomotion, behaviour, growth, localization, immune system processes, detoxification, and metabolic processes. To ensure no DNA-binding, the proteins without

any nuclear localization were selected. We also excluded the sequences with putative nuclear localization signals (NLS) using NLSMapper [210] and NucPred [211].

We obtained a 40% redundant set through h-CD-HIT, using the same three-step iteration process that was used to generate the DNA-protein complex dataset. 10-fold cross-validation was done by randomly selecting 235 proteins from the non-DNA-binding and DNA-binding sets to compare the amino acid compositions. We analysed the distribution for each amino acid in the two datasets using the t-test and checked for statistical significance at 95%, 99% and 99.9% confidence intervals (Table 2.1, Figure 2.1 (a)).

Table 2.1: Comparison of amino acid composition in DNA-binding, non-binding proteins and DNA-binding domains.

The amino acids are represented in single-letter codes with their frequency in percentage (%). The p-value is calculated from the mean of the 10-fold cross-validation. The grey rows are not significant at a 95% confidence level.

Amino Acid	Binding protein	Non-binding protein	P- value (Protein)	Binding domain	P-value (Domain)
A	7.08	8.65	3.16E-05	7.64	1.00E-02
C	1.28	1.95	6.12E-04	1.41	2.50E-01
D	4.86	4.52	2.83E-02	3.78	1.47E-12
E	7.42	5.35	1.17E-11	7.16	1.90E-01
F	3.60	4.81	2.11E-08	3.95	5.40E-02
G	5.37	7.31	3.02E-11	4.11	4.77E-11
H	3.03	1.94	5.64E-09	2.43	2.30E-04
I	5.75	6.59	7.61E-04	6.74	2.04E-07
K	8.10	5.50	1.75E-13	8.3	6.70E-02
L	9.79	10.31	4.07E-02	10.8	1.53E-05
M	2.41	3.25	1.19E-07	1.8	3.71E-07
N	4.42	3.87	1.45E-03	3.64	1.47E-06
P	3.95	4.47	3.64E-02	2.03	1.38E-32
Q	4.49	2.77	3.12E-16	4.79	6.70E-02
R	7.54	4.14	2.61E-27	8.93	1.11E-07
S	6.45	6.75	6.07E-02	5.72	3.95E-05
T	4.74	5.54	1.18E-03	4.74	3.90E-01
V	5.55	7.78	1.19E-18	6.52	1.60E-08
W	1.19	1.22	1.90E-01	1.37	3.64E-02
Y	2.99	3.22	3.89E-02	3.74	9.70E-08

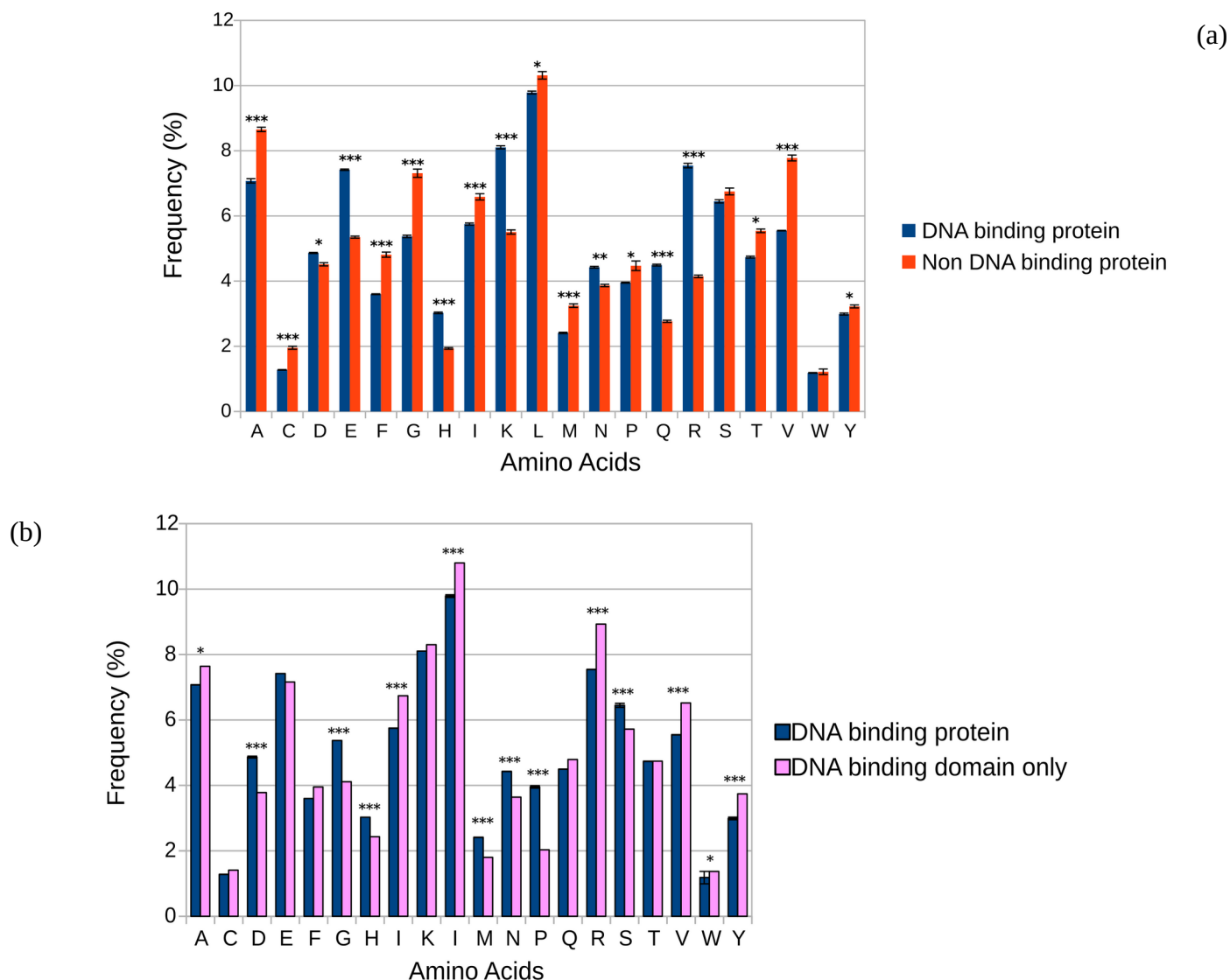


Figure 2.1: Amino acid frequency distribution in protein sequences

comparing DNA-binding (blue) and non-binding proteins (red) in (a) and DNA-binding proteins (blue) and domains only (pink) in (b). The amino acids are represented in single-letter codes on the x-axis. ‘***’, ‘**’ and ‘*’ indicate significance at 99.9%, 99% and 95% confidence levels, respectively. The error bars are standard deviations in the frequency after 10-fold cross-validation.

The t-test showed that all amino acids except serine and tryptophan had statistically significant differences in composition at a 95% confidence level. Glutamate, histidine, lysine, glutamate and arginine were enriched in DNA-binding proteins at a 99.9%

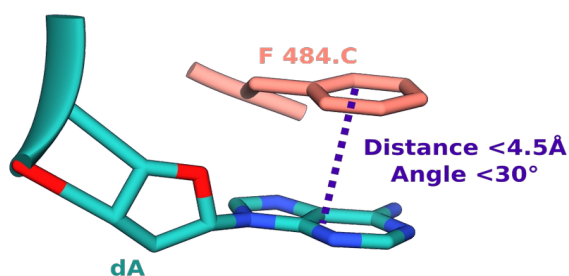
confidence level. Asparagine and aspartate also occurred at a higher frequency in DNA-binding proteins at 99 and 95% confidence, respectively. At 99.9% confidence, we observed that alanine, cysteine, phenylalanine, glycine, isoleucine, methionine and valine had a lower frequency in DNA-binding proteins. This depletion could be because they lack hydrogen bond donors and acceptors in their side chains. Proline, threonine and tyrosine also exhibited a lower abundance in DNA-binding proteins (95% confidence). The trend favoured the enrichment of amino acids with hydrogen bond donor/acceptor atoms in their side chains in DNA-binding proteins (except tyrosine and threonine). Tyrosine and threonine contain hydroxyl groups that could interfere with DNA-binding, but the mechanistic basis is unclear.

We also compared the distribution of amino acids in the protein sequence crystallized with the DNA and the DNA-binding domain alone. The DNA-binding domains were obtained by running RIBFIND [374] on the protein structure and taking all protein clusters with at least one atom within 4 Å of the DNA residues. RIBFIND clusters a protein structure based on rigid bodies detected by secondary structure similarity. The difference in the composition between the entire DNA-binding protein and the DNA-binding domains was evaluated using the t-test. The composition of cysteine, glutamate, phenylalanine, lysine, glutamine and threonine does not change significantly in the DNA-binding domain. However, alanine, isoleucine, leucine, arginine, valine, tryptophan and tyrosine have a higher occurrence in the DNA-binding domain compared to the entire DNA-binding protein. The opposite is true for aspartate, glycine, histidine, methionine, asparagine, proline and serine. The enrichment of arginine and the depletion of aspartate could be due to their affinity and repulsion towards the negatively charged backbone, respectively. The non-polar amino acids are also enriched in the DNA-binding domain. It is possible that the non-polar residues distribute unevenly so the protein can orient itself towards the DNA. Though we could obtain certain inferences about amino acid composition, predicting DNA-binding domains from protein sequences is a challenging problem.

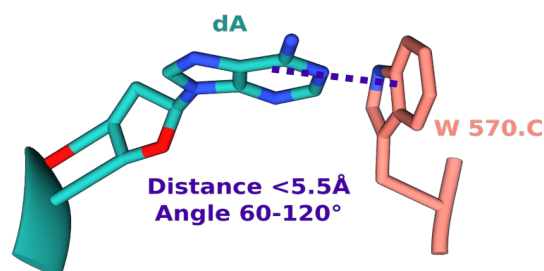
2.1.2 Analysis of molecular interactions

As discussed in chapter 1, nucleotides and amino acids interact in five primary ways- electrostatic interaction [212], hydrogen bonds [213], π - π stacking [61], cation- π interactions [214], and van der Waals interaction [215]. The strength of all of these interactions is inversely proportional to the distance between the participating amino acid-nucleotide pairs. Directionality of interactions (*i.e.*, interactions that rely on relative orientations) is one of the prerequisites for specificity [216]. However, Coulombic, electrostatic, and van der Waals interactions are not directional. Therefore, even though their presence is essential for protein contact with DNA, they cannot confer specificity to the interaction. Hence, the current analysis focused on hydrogen bonds, π - π stacking and cation- π interactions. These molecular interactions are defined by the distance between the interacting atoms and angle (in the case of stacking and hydrogen bonds), which allow their identification in DNA-protein complexes (Figure 2.2).

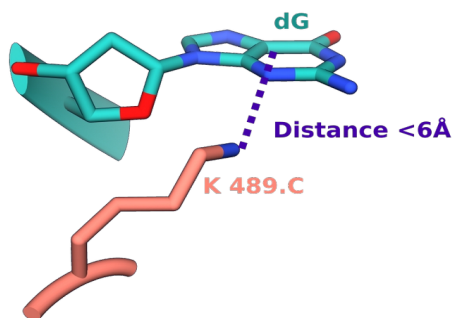
(a) Parallel Stacking



(b) Perpendicular Stacking



(c) Cation- π



(d) Hydrogen Bond

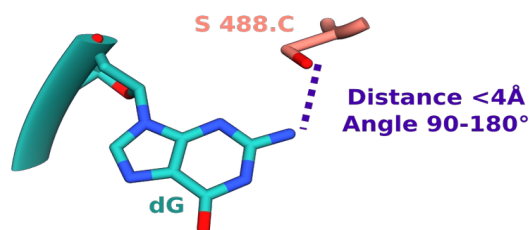


Figure 2.2: Types of nucleotide-amino acid interactions

showing- parallel stacking (a), perpendicular stacking (b), cation- π interaction (c) and hydrogen bond interaction (d). The nucleotides and amino acids are represented in single-

letter codes in cyan and salmon, respectively. The red and blue atoms indicate oxygen and nitrogen atoms, respectively. The purple dotted lines indicate the interaction with the distance and angle constraints for each interaction mentioned. The interactions were obtained from the PDB ID: 5HR4. Chimera was used for visualization of all PDB structures [217].

We calculated the distances using Cell List (https://github.com/neeleshsoni21/Cell_list) and the angles with an in-house script. We normalized the amino acid-nucleotide interactions according to their abundance in the dataset in order to avoid sampling bias.

2.1.2.1 π - π stacking

We identified π - π stacking interactions in the dataset using distance and angle constraints. We calculated the distance between the centroids of aromatic amino acids and nucleosides and the angle between their planes using an in-house python script. Stacking interactions can be of the parallel or perpendicular type depending on the distance and angle constraints (Figure 2.2 (a),(b)). We annotated the interaction with the centroid distance between the aromatic groups <4.5 Å and the angle between the planes $<30^\circ$ as parallel stacking. Interactions with centroid distance <5.5 Å and plane angle in the range 60 - 120° were classified as perpendicular stacking [62,218].

We compared the results for the two kinds of stacking in the DNA-protein training dataset. There were more perpendicular stacking interactions (68.5%) than parallel stacking. Parallel stacking involves inserting amino acid aromatic groups between the DNA base pairs, which is energetically costly due to DNA deformation. This DNA deformation could be another reason for the low abundance of parallel stacking in our dataset because it contains B-form DNA. Adenine had a greater propensity to participate in parallel stacking (74%) compared to perpendicular stacking, whereas guanine (70%), cytosine (56%) and thymine(53%) preferred to interact via perpendicular stacking. However, the number of perpendicular and parallel stacking interactions observed in the dataset was 11 for parallel and 24 for perpendicular stacking interactions. Thus, these interactions may impart stability to the complex and influence specificity, but they are unlikely to be significant

determinants of amino acid-nucleotide specificity. Due to their low numbers, we clubbed parallel and perpendicular stacking interactions into a single category to make inferences about amino acid-nucleotide preferences in stacking. We normalized the observed interactions by the abundance of the nucleotides and amino acids in the dataset.

Cytosine had the highest stacking interactions (46%), followed by adenine (18%), guanine (17%), and thymine (17%) (Figure 2.3 (a)). Even though cytosine had the highest propensity to stack in our dataset, it interacted with all amino acids with a similar frequency. Tyrosine showed interaction with thymine and cytosine, but not with guanine and adenine (Figure 2.3 (b)). Since pyrimidines and tyrosine have six-membered aromatic rings, their stacking could result in better packing, and therefore we observed a higher number of these interactions. The highest number of interactions were seen with tyrosine-thymine. There are only 4 four instances of these in the dataset and each of them have a hydrogen bond with the phosphate backbone accompanying the stacking interaction. The additional hydrogen bond due to the hydroxyl group of tyrosine could be stabilizing this interacting amino acid-nucleotide pair. No observations of tryptophan- guanine stacking could be due to the lower frequency of occurrence of tryptophan.

Adenine-phenylalanine and guanine-histidine pairs showed a higher propensity to interact. Guanine-histidine interaction strength could be enhanced by the formation of both stacking and hydrogen bond interactions. The interaction of phenylalanine with adenine was primarily mediated via parallel stacking. The phenylalanine ring lacks hydrogen bond donors/acceptors. Thus, hydrophobic contact alone could stabilize the interaction, which explains its preference for purines. The number of observed stacking interactions are three and two for adenine and guanine, respectively. Therefore the observed preferences could also be an artefact of lower sampling.

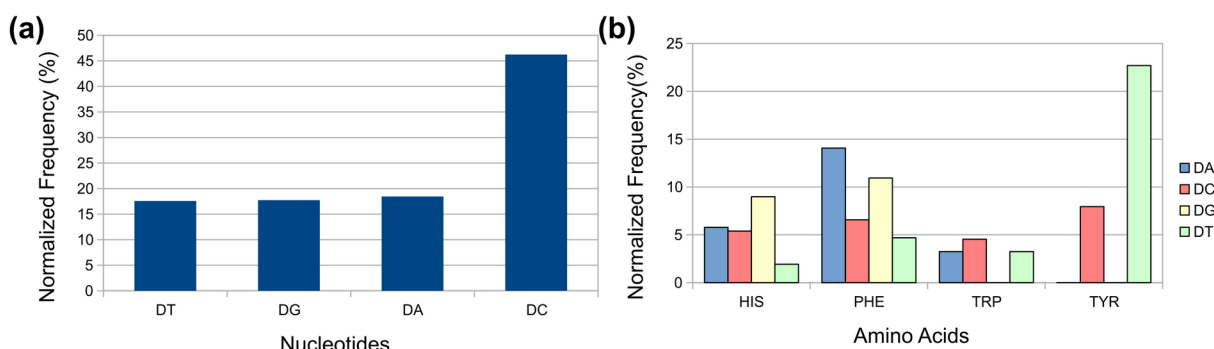


Figure 2.3: Distribution of π - π interactions

-among nucleotides (a) and amino acids (b). The y-axis has the normalized frequencies, and the x-axis has either nucleotides in single-letter codes (a) or amino acids in three-letter representation (b). Red, yellow, purple and green correspond to adenine, cytosine, guanine and thymine, respectively in subplot (b).

2.1.2.2 Cation- π interactions

We identified the cation- π interactions in the dataset by calculating the distance between the centroids of nucleosides and NE and NZ atoms of arginine and lysine, respectively. We characterized the interaction as cation- π interaction if the distance between the positively charged amino acid and the aromatic group of the nucleotide was less than 6 Å [219]. Histidine, in principle, can undergo both π - π and cation- π stacking. However, the charge on histidine is dynamic and pH-dependent. Therefore, for the purpose of this study, we classified its interaction with nucleotides as π - π stacking.

The number of cation- π interactions for arginine (811) is more than twice that of lysine (361). However, the amino acids do not differ in their interaction with the four nucleotides (Figure 2.4 (b)). The propensity of forming cation- π interactions is largest for guanine (50%), followed by thymine (18%), adenine (16%) and cytosine (16%) (Figure 2.4 (a)). The association of arginine with guanine is well documented, and it forms a majority of the cation- π interactions [30,215]. Therefore guanine has the most propensity to form cation- π interactions.

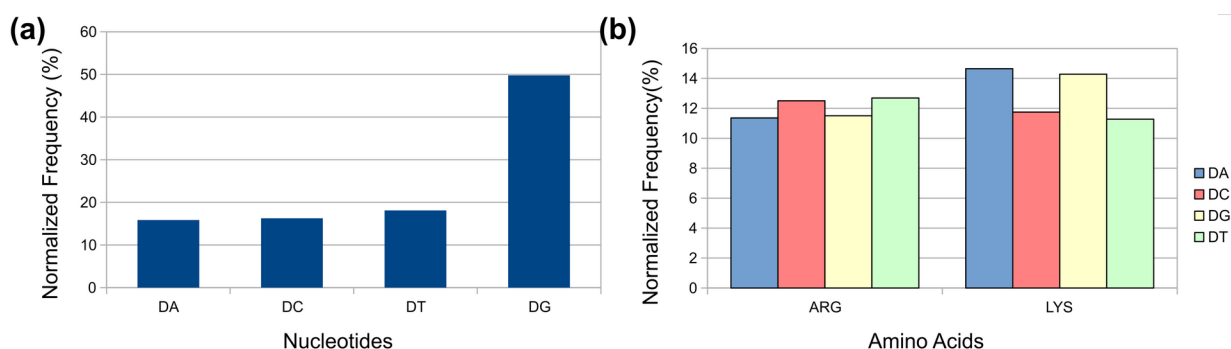


Figure 2.4: Distribution of cation- π interactions

-among nucleotides (a) and amino acids (b) with y and x-axes corresponding to the normalized frequencies and either nucleotides in single-letter codes (a) or amino acids in three-letter representation (b). Red, yellow, purple and green stand for adenine, cytosine, guanine and thymine, respectively in subplot (b).

2.1.2.3 Hydrogen bonds

We identified the hydrogen bonds in the dataset using the distance between hydrogen bond donors and acceptors and their relative orientation. Using Cell List and an in-house python script, we calculated the distance between the hydrogen bond donor/acceptor atoms (List in Appendix 3) and the donor-acceptor-acceptor antecedent angle. A pair of donor-acceptor atoms were annotated as hydrogen bonding if their distance was <4 Å and the angle between the donor-acceptor-acceptor antecedent was $90-180^\circ$ [220]. Since our dataset contains structures up to 3 Å in resolution, we used a higher distance cut-off of 4 Å to ensure we capture all hydrogen bonds that mediate DNA-protein interaction.

The hydrogen bond donors and acceptors are distributed unevenly in the DNA, with more donors and acceptors in the major than in the minor groove. The pattern made by donor-acceptor groups in the major groove can differentiate between AT, TA, GC and CG pairs. However, the AT/TA pair and GC/CG pair are equivalent in the minor groove based on the hydrogen-bonding pattern (Chapter 1: Section 1.2- Figure 1.3).

The major groove is also wider than the minor groove resulting in different physicochemical properties. Therefore, the hydrogen bonding preferences of amino acids in the major and minor grooves were evaluated separately. The main chain of the amino acids also contains a hydrogen bonding donor (NH) and an acceptor (O) atom, but these are common in all amino acids except proline.

The side-chain hydrogen bonding atoms differ among amino acids and have various orientations. Thus, we assessed the hydrogen bond preference of main chain and side chain atoms separately. After calculating the hydrogen bonds in the dataset, we divided the interactions into six categories depending on the region of the DNA and amino acids that interact (Table 2.2). (a) DNA backbone with the amino acid main chain. (b) DNA backbone with amino acid side chain. (c) DNA major groove with the amino acid main chain. (d) DNA major groove with amino acid side chain. (e) DNA minor groove with the amino acid main chain. (f) DNA minor groove with amino acid side chain.

Since we have a small dataset, to make meaningful inferences about hydrogen bond preferences, we needed to employ statistical tests over these data. Hoffman et al. [32] used χ^2 tests to delineate statistically significant and non-significant differences in the interaction patterns. However, one of the caveats of this method is that it is not suitable when the number of observations or the expected value is less than 5. Considering that we have lesser data in the case of some of the nucleotides, we need an alternative to the χ^2 test. Fisher's Exact test is one of the statistical tests that provides relatively more accurate results for scarce data [221].

Fisher's test gives an odds ratio, i.e., the number of observed instances with the number of instances that can occur randomly. An odds ratio >1 implies a favourable interaction, while a value <1 indicates an unfavourable interaction. Each of these odds ratios also has a p-value associated with it. All interactions with a p-value less than 0.05 were considered significant (95% confidence). We performed Fisher's exact test on each of the six nucleotide-amino acid interaction types (Figure 2.5). The odds ratio values are in brackets for all statistically significant interactions.

2.1.2.3.1 DNA-backbone with the main chain of amino acids

Since the DNA backbone is the same irrespective of the identity of the nucleotide, we did not expect its interaction with the main chain of the amino acid to show preferences. However, we observed a statistically significant preference in six cases (Figure 2.5 (a)). Phenylalanine (2.8) and arginine (1.6) main chains showed a positive preference for adenine whereas serine (0.6) showed a negative preference. There were favourable interactions between the glutamate-cytosine (3.3) and histidine-thymine (2.2) pairs with the main chains of these amino acids mediating the interaction. Lysine exhibited an unfavourable interaction with guanine (0.6).

2.1.2.3.2 DNA-backbone with the side chain of amino acids

Amino acid side chains exhibit chemical properties that dictate the likelihood of their interaction with the DNA backbone. The Fisher's Exact test had eight statistically significant interactions (Figure 2.5 (b)). We observed one amino acid that bound favourably to each nucleotide and one that did not. Adenine with asparagine (1.3) had a positive preference, while there was a negative preference with threonine (0.7). The positive preference of cytosine was towards lysine (1.3), whereas tyrosine (0.7) had an unfavourable binding to it. Guanine-histidine (2.1) interaction was favoured, while guanine-cysteine (0.3) was not. Finally, the interaction of thymine with cysteine (3.3) had a positive preference but not with lysine (0.8).

O4' atoms of the DNA backbone had a higher propensity to interact with arginine than other amino acids with ~55% of interaction between them (Figure 2.6). The O4' atom is in the minor groove, and there is a single hydrogen bond acceptor on the nucleotide that faces the minor groove. Due to the proximity of this acceptor to the O4' atom, when arginine interacts with the minor groove nucleotides, it can make a hydrogen bond with the O4' atom leading to the apparent preference of arginine towards the O4' atom.

2.1.2.3.3 DNA major groove atoms with the main chain of amino acids

Due to the unique pattern of donors and acceptors in major and minor grooves (as shown in Chapter 1, Section 1.2, Fig. 3), the propensities of amino acids to interact with the nucleotides would depend on the grooves. The main chain atoms of amino acids show statistically significant preferences for six nucleotide-amino acid interaction pairs in the major groove (Figure 2.5 (c)). These six interactions were composed of two favourable and one unfavourable interaction for both cytosine and guanine. Lysine (4.5) and proline (7.6) had a positive preference for cytosine, while glycine (0.2) had a negative preference. Guanine showed favourable interaction with glycine (5.2) and histidine (2.9), whereas it exhibited an unfavourable interaction with alanine (0).

2.1.2.3.4 DNA major groove atoms with the side chain of amino acids

The interaction of the side chain of amino acids with the atoms of nucleotides exposed to the major groove showed the maximum preferences among all interactions (28) (Figure 2.5 (d)).

Aspartate (196.7) and glutamate (45.9) only had a positive preference for cytosine and a negative or insignificant preference (adenine-glutamate) for all other interactions with nucleotides. Asparagine and glutamine had the same trend for binding with positive preference for adenine (4.0 and 7.9, respectively) and thymine (1.6 and 1.7, respectively) and a negative preference for guanine (0.2 and 0.1, respectively). Serine and threonine also share common preferences with a favourable interaction towards adenine (2.5 and 2.2, respectively) and cytosine (1.3 and 3.8, respectively) and an unfavourable one towards guanine (0.4 and 0.2). The cytosine-serine interaction is not statistically significant. However the odds ratio is greater than 1, indicating its favourable nature albeit with lower confidence. Thus, similar functional groups have the same trend in hydrogen bond preferences.

Histidine had a propensity to interact with only cytosine (2.2) and its interaction with the other three nucleotides was not significant. Tyrosine also has a positive preference for cytosine (2.6) and adenine (3.4) but has a negative preference for guanine (0.3).

Lysine and arginine exhibited a similar preference, where they favoured interacting with guanine alone (3.5 and 5.8, respectively). They had a negative preference for adenine (0.2 and 0.3, respectively) and cytosine (0.03 and 0.02, respectively), and the preference toward thymine was not statistically significant. Despite the similarity in preferences, in some proteins, the replacement of arginine by lysine leads to loss of specific binding [222], whereas in others, it is inconsequential [223]. If both the amino acids have an equal preference for guanine, these differences need to be studied further.

2.1.2.3.5 DNA minor groove atoms with the main chain of amino acids

The pattern of donors and acceptors in the minor groove is different for A/T pairs and G/C pairs, but it cannot differentiate between A/T and T/A or G/C and C/G base pairs). Thus, we expected the minor groove to have preferential hydrogen bond interactions but not as pronounced as the major groove. But when the main chain interactions were analysed by Fisher's Exact test, there were the same number (six) of statistically significant hydrogen-bonded pairs as the major groove (Figure 2.5 (e)).

Arginine main chain had a positive preference for adenine (6.6) and a negative one for guanine (0.0) atoms in the minor groove. Alanine (3.9) and serine (6.0) preferred interaction with cytosine. The interaction of isoleucine (inf) was the most favourable for guanine, and glycine (2.3) interacted most favourably with thymine.

2.1.2.3.6 DNA minor groove atoms with the side chain of amino acids

As we had expected, the minor groove interactions show fewer statistically significant preferences (10) than the major groove (28) (Figure 2.5 (f)).

Both aspartate (inf) and glutamate (inf) show a positive preference for guanine interaction and non-significant for other nucleotides. Asparagine (4.2) and glutamine (17.7) also exhibit a positive preference for guanine, while arginine (0.1) has a negative preference. Adenine did not have a favourable interaction and instead it had a statistically significant unfavourable interaction with glutamine (0.0). Cytosine did not have any statistically significant preferences. Arginine-thymine (3.1) interaction was favourable.

However asparagine (0.3), glutamine (0.04) and tyrosine (0.3) had an unfavourable interaction with thymine.

Table 2.2: Normalized amino acid-nucleotide hydrogen bond distribution in the training set. The amino acids and nucleotides are represented in single-letter codes. Mc, Sc, bb, MJ and MN stand for amino acid main chain, amino acid side chain, DNA backbone, major and minor grooves, respectively.

Amino acids	Mc-bb				Sc- bb				Mc-MJ				Sc-MJ				Mc- MN				Sc- MN			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	0.4	0.5	0.6	0.5	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0
C	0.1	0.1	0.2	0.0	0.2	0.2	0.1	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	0.2	0.2	0.5	0.4	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.2	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0
E	0.1	1.0	0.3	0.6	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.8	2.8	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.7	0.0
F	0.5	0.1	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.3	0.3	0.5	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
H	0.1	0.1	0.1	0.3	0.5	0.4	0.8	0.4	0.0	0.0	0.0	0.0	0.2	0.2	0.6	0.2	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.1
I	0.3	0.4	0.4	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
K	0.5	0.6	0.4	0.5	1.6	1.6	1.6	1.5	0.0	0.2	0.1	0.0	0.2	0.0	3.1	0.4	0.0	0.1	0.1	0.0	0.2	0.1	0.1	0.3
L	0.6	0.6	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
M	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
N	0.3	0.3	0.3	0.3	1.1	0.7	0.9	1.0	0.1	0.1	0.0	0.0	0.9	0.2	0.7	0.4	0.0	0.0	0.1	0.0	0.1	0.1	0.2	0.1
P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Q	0.2	0.2	0.3	0.4	1.0	0.8	1.0	0.9	0.0	0.1	0.0	0.0	1.2	0.2	0.3	0.4	0.0	0.0	0.1	0.0	0.0	0.1	0.3	0.0
R	0.6	0.4	0.4	0.5	1.5	1.3	1.4	1.7	0.0	0.1	0.1	0.0	0.3	0.0	2.7	0.4	0.1	0.0	0.0	0.1	0.2	0.1	0.0	0.4
S	0.3	0.4	0.5	0.5	1.2	1.1	1.4	1.5	0.0	0.1	0.1	0.0	1.0	0.3	1.3	0.4	0.0	0.1	0.0	0.0	0.1	0.1	0.2	0.2
T	0.3	0.3	0.3	0.3	0.8	0.9	1.2	1.3	0.0	0.1	0.0	0.0	0.8	0.6	0.6	0.4	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2
V	0.5	0.3	0.4	0.4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
W	0.1	0.1	0.2	0.0	0.2	0.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Y	0.2	0.1	0.4	0.2	0.8	0.5	0.8	0.8	0.0	0.1	0.0	0.0	0.7	0.3	0.5	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1

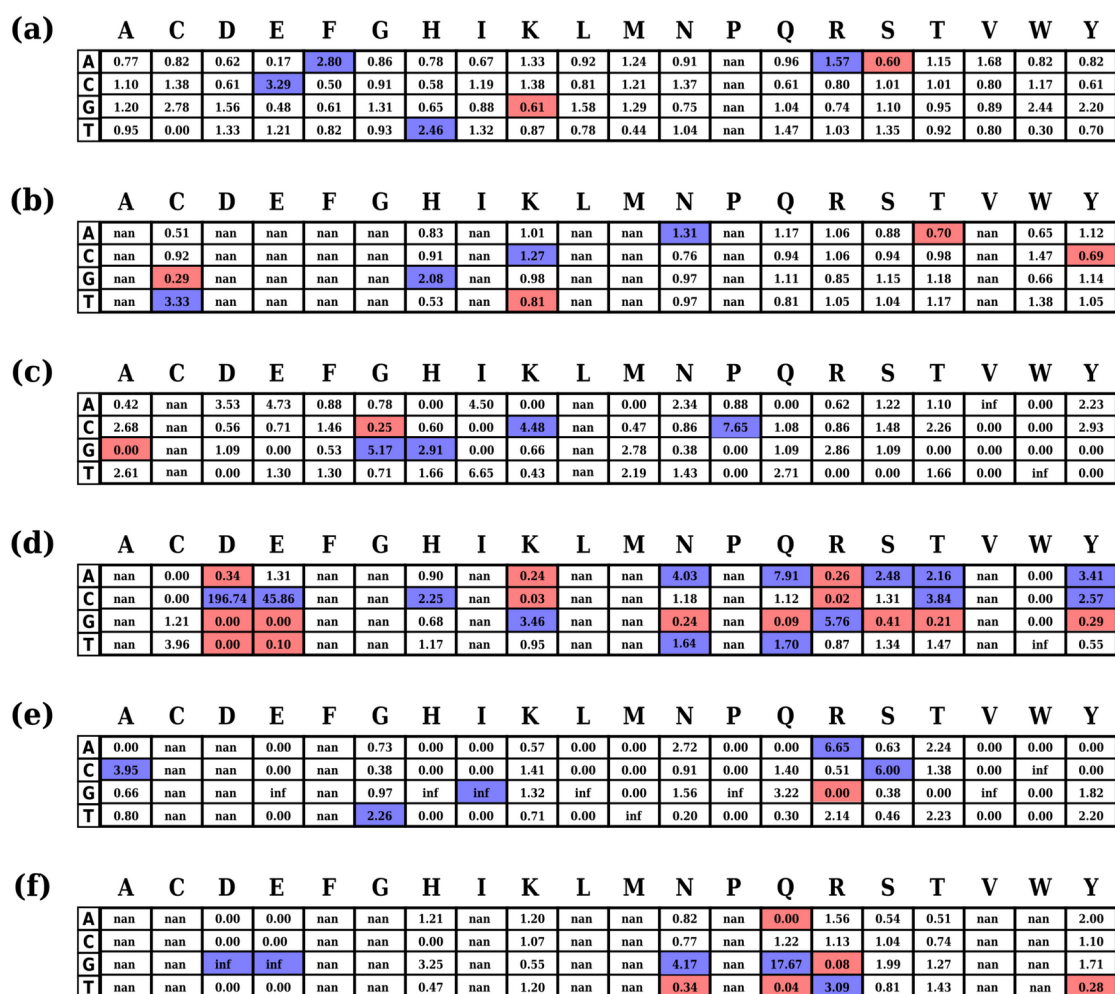


Figure 2.5: Hydrogen bond preferences in the training set.

The nucleotide bases and amino acids are represented as single-letter codes. The values inside the boxes are the odds ratio values from Fisher's Exact test. The interactions not significant at a 95% confidence level are white. The remaining interactions are coloured blue and pink for significantly favourable and unfavourable interactions, respectively. The subplots show the frequency of interaction of atoms of (a) The amino acid main chain with the DNA backbone. (b) The amino acid side chain with the DNA backbone. (c) The amino acid main chain with the atoms in the major groove of DNA. (d) The amino acid side chain with the atoms in the major groove of DNA. (e) The amino acid main chain with the atoms in the minor groove of DNA. (f) The amino acid side chain with the atoms in the minor groove of DNA.

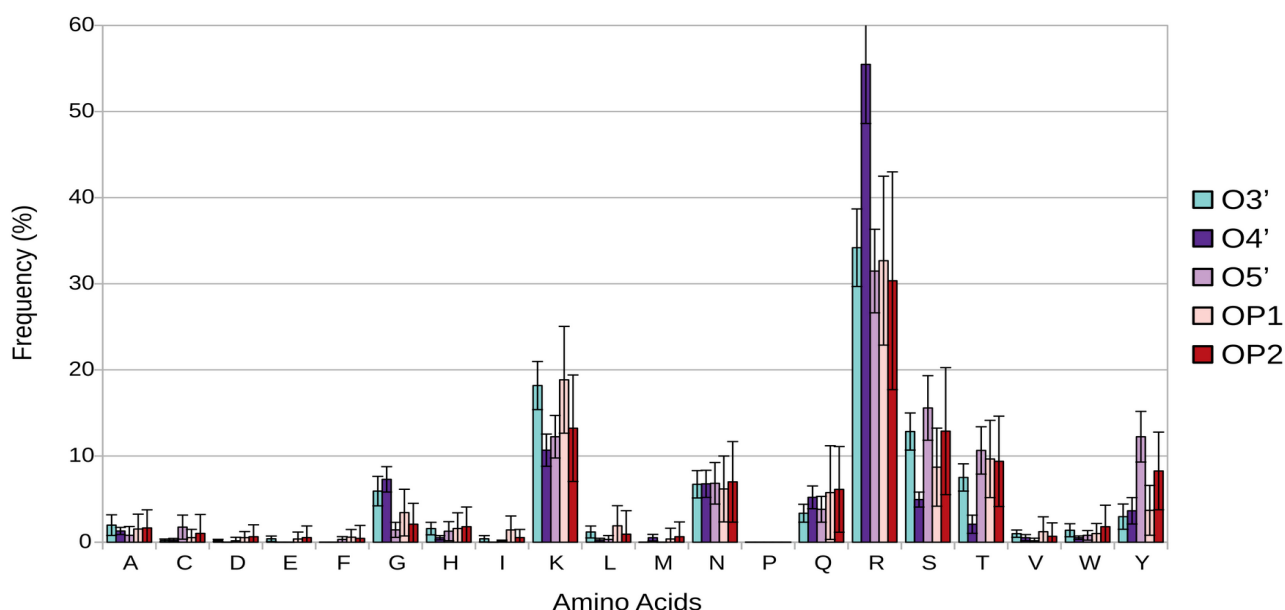


Figure 2.6: Atom-wise frequency of amino acid-DNA backbone hydrogen bonds.

The x-axis has amino acids in single-letter representation. The observed frequency(%) of the 5 DNA backbone atoms -O3' (cyan), O4' (purple), O5' (violet), OP1 (pink), and OP2 (maroon) is on the y-axis. The error bars represent variation across different PDBs.

We also analysed the overall preference of an amino acid-nucleotide pair. We combined the major and minor groove hydrogen bonds to get amino acid-nucleotide interaction frequencies. The phosphate backbone was common in all four nucleotides and would not contribute toward specificity in a generalized scenario. Therefore backbone interactions were not considered further in the analysis.

As observed in section 2.1.2.3.4, the amino acids with similar functional groups have similar preferences in the major groove, which form the majority of all base-specific hydrogen bonds (75%). We could observe statistically significant interactions among these interactions in 28 out of 80 possible interaction pairs. Thus, to increase the confidence in the propensities we grouped amino acids with similar hydrogen bonding functional groups together. Aspartate and glutamate formed the DE class, asparagine and glutamine the NQ class, serine-threonine the ST class, and all main chains were clubbed into the Mc class (Figure 2.7). In the ideal scenario, all these would be allotted individual preferences, but we had to group these amino acids in the paucity of data.

We performed the Fisher's Exact test with these data, with the major and minor grooves combined and the similar functional groups containing amino acids merged (Table 2.3). The preferences in this merged set are similar for all amino acids except the Mc class. While the main chain interactions were not statistically significant before grouping data, after merging, the Mc class shows significant positive (cytosine and thymine) and negative (guanine) preferences. We could marginally improve the statistically significant data from 35% to 45% without the loss in preferential interaction trends.

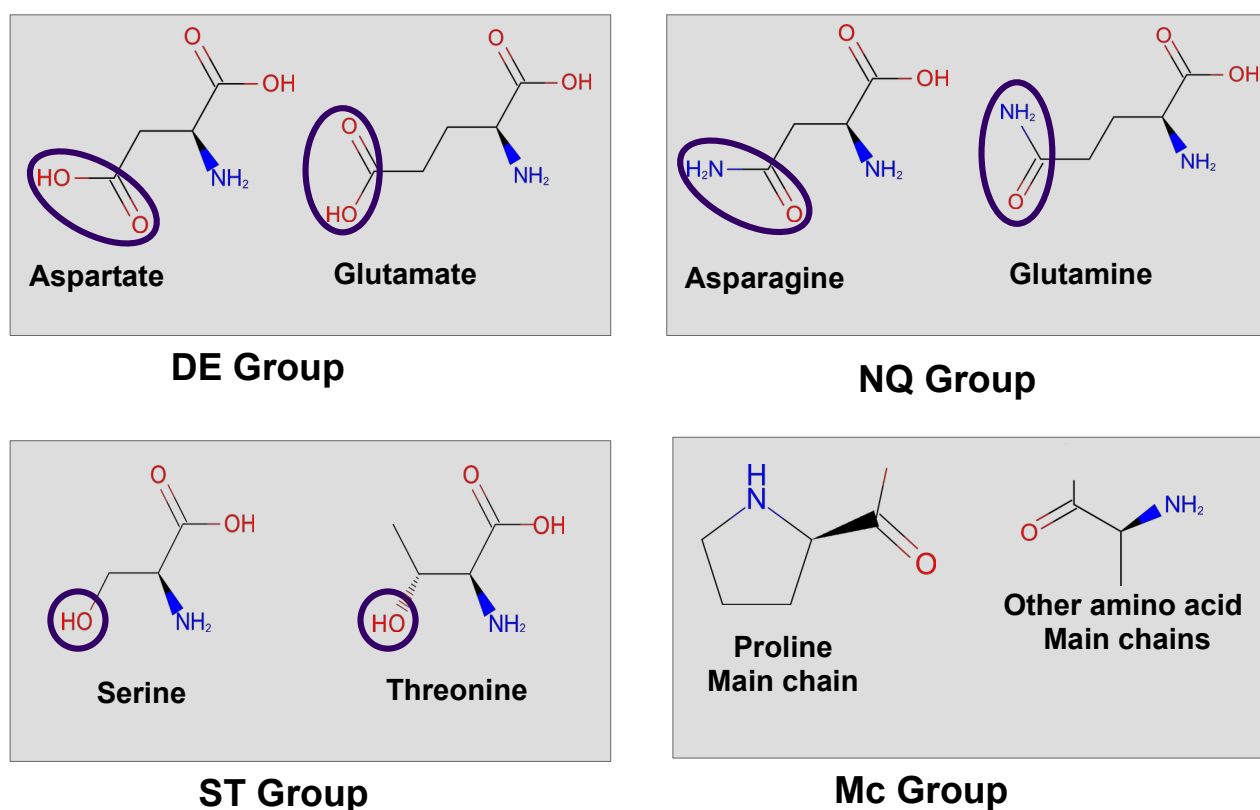


Figure 2.7: Merging amino acids with similar hydrogen bond forming functional groups. Oxygen and nitrogen atoms are in red and blue, respectively. The hydrogen bonding groups are encircled in purple.

Table 2.3: Statistical propensity of nucleotide-amino acid interaction in the training set.

The red and blue cells indicate negative and positive preferences, respectively (significant at 95% confidence). The nucleotides (columns) and amino acids (rows) are represented in single-letter format along with grouped amino acids described in section 2.1.2.3.

	A		C		G		T	
Amino acids	Count	Odds ratio	Count	Odds ratio	Count	Odds ratio	Count	Odds ratio
C	0	0	0	0	2	1.79	1	2.47
DE	21	0.71	124	42.47	8	0.04	1	0.03
H	11	0.91	11	1.44	35	1.01	9	0.78
K	28	0.39	10	0.20	240	2.55	54	0.96
Mc	52	0.87	101	4.05	99	0.36	69	1.41
NQ	180	4.28	48	0.88	133	0.35	70	0.95
R	189	0.42	51	0.12	1066	3.22	284	1.16
ST	63	1.98	33	1.31	79	0.49	41	1.17
W	0	0	0	0	0	0	1	inf
Y	33	3.06	16	1.7	27	0.41	8	0.51

2.2 Comparing hydrogen bonds across DNA-binding protein families

We compared unrelated (non-homologous) structures using CLICK to delineate specificity determinants. CLICK allows topology-independent superimposition with user-defined representative atom selection (refer to Chapter 1, section 1.6.1). Therefore, we aligned non-homologous structures to find atomic-level similarities. The selection of representative atoms introduces flexibility of alignments, facilitating modular superimpositions. We hypothesized that due to the crucial nature of DNA-protein specificity, there would be similarities in hydrogen bond orientations in different DNA-binding protein families. Therefore, we chose hydrogen bond donors and acceptors in nucleotides as representative atoms. The adjacent nucleotides could influence the hydrogen bonding partner of nucleotides. Thus, we performed three kinds of superimpositions-

- Mononucleotide superimpositions: We superimposed all hydrogen-bonded nucleotides in the dataset.
- Dinucleotide superimpositions: We superimposed dinucleotides which were of two types. The first type was composed of the hydrogen-bonded nucleotide with its 3' neighbour (First nucleotide). The second type contained the hydrogen-bonded nucleotide with its 5' nucleotide. The adjacent nucleotide, in this case, may or may not be hydrogen-bonded.
- Trinucleotide superimposition: We superimposed hydrogen-bonded nucleotides with their 3' and 5' adjacent nucleotides. The neighbouring nucleotides on either side may or may not be hydrogen-bonded.

The superimpositions were restricted to the same composition and interaction type. For example, a cytosine major groove atom interacting with an amino acid side chain was superimposed with another cytosine in the dataset that had a side chain interaction via its major groove atom. A successful overlap is where the hydrogen-bonded amino acid donors/acceptors in the aligned substructures have a distance of less than 1.5 Å (Figure 2.8). We calculated the frequency of amino acid-nucleotide interactions in all these successful overlaps.

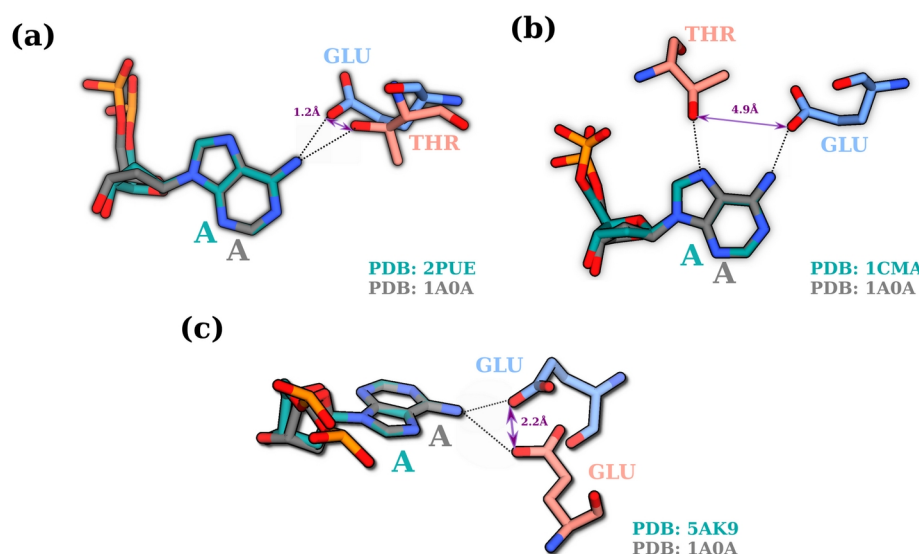


Figure 2.8: Definition of hydrogen bond overlap.

The nucleotide adenine (A) is in grey and cyan, and the amino acids threonine (salmon) and glutamate (blue) are represented in 3 letter codes. The elements O, N and P are

shown in red, blue and yellow colours, respectively. The black dotted lines show the hydrogen bond. Distance between the amino acid hydrogen bond acceptors is in magenta and indicates hydrogen bond overlap in (a) and no hydrogen bond overlap in (b) and (c). The PDB codes for the superimposed structures are given at the bottom right of the subplot.

Since we aimed to find family-independent specificity features, we analysed whether the successful overlaps (indicating similar binding orientations) belonged to the same fold or a different fold. Thus, we assigned CATH IDs to the residues [137,138]. CATH is a protein structure database and annotates folds to regions in the proteins. Wherever a CATH ID was unavailable for assignment, we labelled the fold ambiguous.

A majority (56.7%) of the successful overlaps belonged to a different fold. In comparison, 36.5% belonged to the ambiguous fold, and only 6.8% belonged to the same fold (Figure 2.9 (a)). Since our dataset was non-redundant, many proteins would have different folds. If the hydrogen bond interaction modes were different for each DNA-binding structural class, we would have observed maximum superimposition with the same fold. However, this was not the case, further strengthening our argument on the universality of hydrogen bonding interaction modes.

We further checked if the root mean square deviation (RMSD) between donors and acceptors in the superimposed structures was larger when the two amino acids had different folds (Figure 2.9 (b)). There are more superimpositions with RMSD between hydrogen bond donors and acceptors <0.5 Å for the same fold (35%) as compared to different (25%) or ambiguous fold (24%). The median distance for same, different and ambiguous folds is 0.65 Å, 0.76 Å and 0.75 Å, respectively. One-way ANOVA [224] test at a 95% confidence deems the three distributions different. So there is a better superimposition when the two amino acids belong to the same fold. The difference between the means is 0.1 Å, which is small compared to the resolutions of the structures in our dataset.

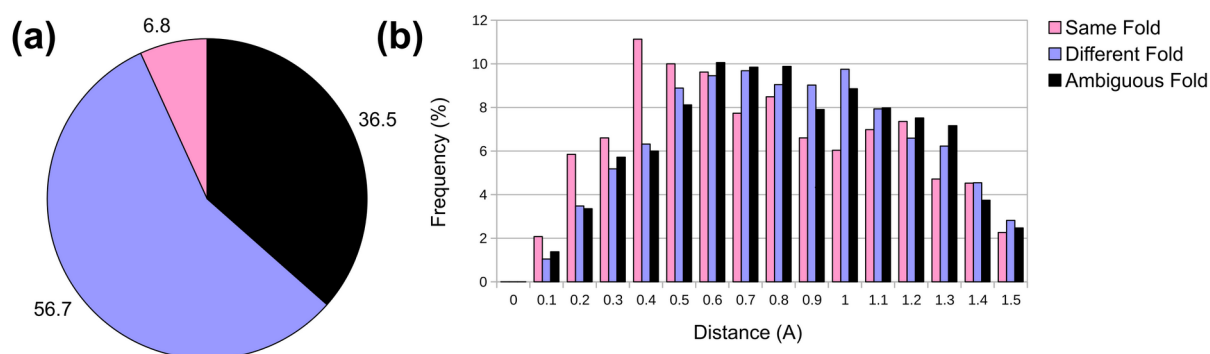


Figure 2.9: Fold-wise analysis of RMSD of superimposed nucleotides.

The same, different and ambiguous fold superimpositions are in pink, blue and black, respectively. (a) Distribution of superimposed folds with their relative frequency (%). (b) Distribution of the frequency (%) of RMSD (Å) in different fold types.

2.2.1 Mononucleotide superimpositions

To obtain information about amino acid preferences, we analysed the hydrogen bonds of the superimposed nucleotides. 22 % of all mononucleotide superimpositions resulted in hydrogen bond overlap. While 22% and 31% of the superimpositions resulted in hydrogen bond overlap in the major and minor grooves, respectively. The hydrogen bond conservation is highest for cytosine, followed by adenine, guanine and thymine (Table 2.4). 99.25% of the backbone alignments did not have hydrogen bond overlap. This supports our argument for not using backbone hydrogen bonds to explain family independent amino acid-nucleotide specificity.

We computed the frequency of superimposition of amino acids (Table 2.5 (raw numbers)). The major groove main chain interactions were most conserved for cytosine in almost all amino acids (exceptions-glutamate, serine and asparagine). The minor groove main chain interactions were most conserved for guanines. The major and minor groove interactions with side chains show variability in conservation depending on the amino acid-nucleotide pair. Therefore, the overall preferences are likely to be influenced most by these interactions.

We calculated the total interactions of a nucleotide-amino acid pair (including major and minor grooves) and analysed the preferences using Fisher's Exact test (Figure 2.10). All the statistically significant nucleotide-amino acid interaction preferences observed before superimposition (Table 2.3) were retained after superimposition as well. The superimposition reinforced the preferences that were too weak to be recognized. After superimposition, the statistically significant preferences increased from 18 to 37 out of 40. Thus, we could obtain statistically significant preferences for most of the nucleotide-amino acid pairs.

	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	0.00	0.42	0.31	0.45	0.24	8.34	0.32	1.94	0.00	2.13
C	0.00	235.07	0.62	0.13	5.45	1.16	0.02	2.22	0.46	4.37
G	1.44	0.00	1.93	2.69	0.29	0.18	5.69	0.40	0.00	0.27
T	3.23	0.00	0.91	0.96	1.82	1.65	1.06	1.26	119.89	0.51

Figure 2.10: Nucleotide-amino acid interaction preferences in the superimposed mononucleotides.

Numbers inside the cells are the values for the odds ratio obtained from Fisher's Exact test. The nucleotide bases and amino acids/groups are shown in single-letter codes. The non-significant interactions at 95% confidence level are white. The remaining interactions are coloured blue and pink for significantly favourable and unfavourable interactions.

Table 2.4: Mononucleotide superimposition frequency.

The nucleotides are shown in single-letter codes and Mc and SC stand for amino acid main chain and side chain, respectively.

Nucleotides	Category	major_Mc	major_SC	minor Mc	minor_SC	Average (%)
A	Superimposed	130	15700	96	3702	
	Total	882	88466	290	8654	
	% Superimposed	14.7	17.7	33.1	42.8	27.1
C	Superimposed	2222	18088	186	1158	
	Total	5128	41518	708	3984	
	% Superimposed	43.3	43.6	26.3	29.1	35.6

G	Superimposed	290	104108	958	984	
	Total	1280	505664	2194	6778	
	% Superimposed	22.7	20.6	43.7	14.5	25.4
T	Superimposed	104	10620	462	6506	
	Total	522	56050	1806	21228	
	% Superimposed	19.9	18.9	25.6	30.6	23.8

Table 2.5: Amino acid-nucleotide mononucleotide superimposition frequency.

The amino acids are in rows, and nucleotides are in columns in single-letter codes. Mc and SC stand for amino acid main chain and side chain, respectively.

	major_Mc				major_Sc				minor_Mc				minor_Sc			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	1	240	0	9	0	0	0	0	0	53	90	15	0	0	0	0
C	0	0	0	0	0	0	140	54	0	0	0	0	0	0	0	0
D	12	91	22	0	231	7074	0	0	0	0	0	0	0	0	41	0
E	16	19	0	1	484	5247	0	0	0	0	2	0	0	0	10	0
F	0	127	12	0	0	0	0	0	0	0	0	0	0	0	0	0
G	15	227	91	16	0	0	0	0	35	24	349	285	0	0	0	0
H	0	95	46	10	142	203	2644	319	0	0	6	0	2	0	4	14
I	5	0	0	7	0	0	0	0	0	0	48	0	0	0	0	0
K	0	307	22	0	699	1	17061	1164	2	18	125	18	611	127	130	1031
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	6	5	1	0	0	0	0	0	0	0	0	0	0	0	0
N	46	301	41	42	3490	1961	3862	1659	20	22	142	19	331	131	277	236
P	2	152	0	0	0	0	0	0	0	0	56	0	0	0	0	0
Q	0	132	5	18	4642	428	1416	1237	0	0	108	4	0	70	170	16
R	3	139	30	0	3685	16	73495	5063	37	0	0	97	2306	673	214	4759
S	26	139	16	0	1492	983	4005	615	1	66	3	18	219	108	61	265
T	4	155	0	0	310	1068	497	393	1	0	0	3	54	8	16	73
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	43	0	3	0	0	0	0	0	0
Y	0	92	0	0	525	1107	988	73	0	0	29	3	179	41	61	112

2.2.2 Dinucleotide superimpositions

We observed preferential hydrogen bond interaction in mononucleotide alignments. We further hypothesized that the neighbouring amino acids could influence the specificity of the interaction. Therefore, we extended this approach to aligning dinucleotides. Depending on the neighbour included in the dinucleotide (3' end or 5' end), there were two alignments: where the first nucleotide is hydrogen-bonded and when the last nucleotide is hydrogen-bonded. With the first nucleotide bonded, 15.5% of the alignments resulted in hydrogen bond overlaps. 16.5% of superimpositions had hydrogen bond overlap with the last nucleotide hydrogen-bonded. The amino acid preferences for some amino acids changed based on the identity of the neighbour. We quantified these preferences using Fisher's Exact test (Figure 2.11).

We observed that DE preferably bound cytosine regardless of the neighbour at the 5' end or 3' end. The following interactions were favourable when the neighbour was only at the 3' end (the neighbour is mentioned in brackets)- lysine-adenine (guanine), arginine-adenine (guanine), ST class-guanine (adenine), Mc class-guanine (cytosine), arginine-guanine (guanine), lysine-guanine (thymine), cysteine-guanine (thymine), ST class-thymine (adenine), Mc class-thymine (guanine) and, ST class-thymine (guanine). Some interactions were more preferred only if the neighbour was at the 5' end (the neighbour is mentioned in brackets) – NQ class-adenine (guanine), ST class-adenine (guanine), Mc class-guanine (adenine), cysteine-guanine (guanine), ST class-guanine (thymine), lysine-thymine (adenine), ST class-thymine (cytosine) and arginine-thymine (guanine).

Guanine-arginine hydrogen bond formed a special case where the interaction was preferred when the adjacent guanine was at the 3' end but not when it was present at the 5' terminus. ST has the most change in preference depending on the neighbour location at the 3' or 5' end. Guanine binding preferences were the most affected by the identity of the adjacent nucleotide. Cytosine binding preferences were unaffected by the identity of the adjacent nucleotide. Guanine as a neighbour changed the hydrogen bonding preference of adenine to the maximum extent.

(a)		3'	C	DE	H	K	Mc	NQ	R	ST	W	Y
A		0.00	0.99	0.31	0.46	0.30	5.30	0.38	2.14	0.00	6.83	
		0.00	1.17	1.91	0.54	0.00	4.94	0.23	4.68	0.00	2.77	
		0.00	0.00	0.00	1.16	0.06	0.62	1.37	0.46	0.00	0.81	
		0.00	0.13	0.12	0.00	0.36	11.78	0.15	5.57	0.00	7.89	
C		0.00	77.65	0.11	0.02	1.33	2.44	0.05	5.63	0.00	10.99	
		0.00	21.85	1.53	0.22	10.93	3.44	0.05	3.38	0.00	7.31	
		0.00	5.97	0.40	0.80	10.87	4.22	0.11	3.00	0.00	2.40	
		0.00	31.01	1.84	0.02	11.55	2.21	0.00	3.92	0.00	10.32	
G		0.00	0.13	0.84	1.14	0.59	0.77	0.79	1.59	0.00	1.85	
		0.00	0.18	2.13	0.86	1.90	0.48	1.71	0.43	0.00	0.00	
		0.00	0.00	2.28	1.61	0.00	0.32	1.25	0.35	0.00	0.24	
		12.22	0.00	0.21	1.08	0.46	0.52	1.54	0.64	0.00	0.08	
T		0.00	0.00	0.89	0.57	9.19	2.12	0.90	1.53	0.00	0.10	
		0.00	0.00	1.65	0.92	0.00	1.24	1.28	0.69	0.00	0.64	
		0.00	0.00	1.64	0.66	2.68	3.74	0.51	2.34	inf	1.46	
		12.45	0.00	0.45	1.01	7.25	1.31	0.74	1.24	0.00	0.24	

(b)	5'	C	DE	H	K	Mc	NQ	R	ST	W	Y
A		0.00	0.00	0.00	0.43	0.40	3.60	1.01	1.85	0.00	1.44
		0.00	7.41	0.00	0.17	0.00	2.97	0.54	5.18	0.00	2.09
		0.00	1.67	0.52	0.13	0.16	6.80	0.62	2.56	0.00	4.06
		0.00	0.00	0.28	0.37	0.11	10.60	0.17	2.15	0.00	11.56
C		0.00	32.15	1.97	0.15	6.81	2.28	0.09	5.98	0.00	2.05
		0.00	14.53	3.61	0.29	35.33	2.04	0.11	1.90	0.00	4.73
		0.00	44.92	0.43	0.13	8.27	2.23	0.00	5.52	0.00	8.03
		0.00	40.44	0.00	0.23	4.86	2.50	0.10	1.52	0.00	16.03
G		0.00	0.00	0.38	1.60	1.64	0.91	0.96	0.23	0.00	0.93
		0.00	0.00	1.30	0.79	0.09	0.73	1.65	0.74	0.00	0.30
		2.82	0.03	2.60	2.89	0.27	0.51	0.80	0.17	0.00	0.05
		0.00	0.02	0.52	0.39	0.06	0.14	1.95	2.66	0.00	1.07
T		0.00	0.00	0.33	1.68	4.70	1.04	0.75	0.51	0.00	0.09
		0.00	0.00	0.00	0.25	2.80	1.67	1.17	2.32	0.00	2.65
		0.00	0.00	0.87	0.97	0.00	1.58	1.47	0.30	inf	0.05
		24.06	0.00	0.34	0.66	4.82	1.83	0.92	1.76	0.00	0.08

Figure 2.11: Nucleotide-amino acid preferences of superimposed dinucleotides.

The nucleotide bases (adenine (red), cytosine (yellow), guanine (purple), thymine (green)) and amino acids/groups are represented in single-letter codes. Numbers inside the cells are the values for the odds ratio obtained through Fisher's test. The interactions not significant at a 95% confidence are white. The remaining interactions are coloured blue and pink for significantly favourable and unfavourable interactions, respectively. The preferences of the cells with slanted lines on them differ when the neighbour is at the 3' or 5' end. (a) The first member of the dinucleotide is hydrogen-bonded, and the neighbour is at the 3' end (Raw data Appendix 4). (b) The second member of the dinucleotide is hydrogen-bonded, and the neighbour at the 5' end (Raw data Appendix 5).

2.2.3 Trinucleotide superimpositions

We observed preferences among mononucleotides and dinucleotides. The dinucleotides revealed different preferences depending on the neighbour at the 3' or 5' end. Therefore, we wanted to study the effect of both 3' and 5' neighbours together. Thus, we performed trinucleotide CLICK superimpositions with a central hydrogen-bonded nucleotide and adjacent nucleotides on either side. The adjacent nucleotides were not required to be hydrogen-bonded. We observed hydrogen bond overlap in 33% of all alignments. 34% of the major groove alignments and 30% of the minor groove alignments had hydrogen bond overlap.

We obtained context-specific preferences by analysing the spatial conservation of hydrogen bond donors and acceptors in aligned hydrogen-bonded trinucleotides. The major and minor groove amino acid counts were merged to analyse preferences and quantified using Fisher's test (Figure 2.12). Cysteine was only hydrogen-bonded favourably with guanine when another guanine was at the 5' end, and thymine was at the 3' end. DE class interaction with guanine was never favoured, while binding to cytosine was almost always favoured regardless of the neighbour. Lysine was only hydrogen-bonded with adenine when thymine was at the 5' end and guanine was at the 3' end. Tyrosine and the NQ class favoured binding guanine only when it had adenine at both ends. Tyrosine favoured binding thymine only when it had cytosine at both ends. ST class-thymine binding was unfavourable only with guanine at the 3' and cytosine at the 5' terminus.

The preference of arginine towards guanine is well known. Therefore, it is interesting that arginine- guanine interactions are unfavourable when the trinucleotides are AGA, GGG and GGT. Negative preference was comparatively stronger with lysine and arginine interaction with cytosine. They did not exhibit a statistically significant interaction or show an unfavourable association irrespective of the 3' and 5' neighbour.

In 199 of 640 nucleotide-amino acid pairs, the preferences were statistically significant in di- (both 5' and 3' neighbours) and trinucleotides. We analysed these 199 interactions and noticed that in 34 cases, the preference was not the same between the

two types of di- and trinucleotides (Table 2.6). In 8 cases, the first and last dinucleotide had the same preference, but trinucleotide had a different preference. In 21 cases, the trinucleotide and last dinucleotide (hydrogen-bonded nucleotide with 5' neighbour) had the same preference, and it was reversed in the first dinucleotide (hydrogen-bonded nucleotide with 3' neighbour). The last 5 cases belonged to the class where the preference was the same for the first nucleotide and trinucleotide but not for the last nucleotide. From these data, we can infer that the 5' neighbour has a more considerable influence on the specificity of the amino acid-nucleotide preference than the 3' neighbour.

In the 165 cases, where the preferences of di- and trinucleotides were the same, DE class interactions were most frequent (35/165). Guanine and cytosine were the hydrogen-bonded central nucleotide in 55 and 54 cases respectively (Appendix 7). We had noticed earlier that cytosine preference toward the DE class is not dependent on the adjacent nucleotide. Arginine and Lysine, which have a preference toward guanine, also have 29 and 27 statistically significant preferences. Both these cases could contribute to similar di- and trinucleotide preferences.

All 64 trinucleotides had statistically significant preferences, among which 5 had only positive preferences, and the remaining had both positive and negative preferences. The blank cells indicating a lack of sufficient data make up 384//640, i.e., 60% of the total possible interactions. Therefore, with an increase in the available structures for DNA-protein complexes, we expect the specificity determinants to be more apparent.

5'	3'	C	DE	H	K	Mc	NQ	R	ST	W	Y
A		0.00	0.00	0.00	0.00	0.00	9.12	0.57	0.36	nan	0.87
		0.00	0.00	0.00	3.34	0.00	1.31	0.09	2.47	nan	28.50
		0.00	0.00	0.00	1.39	0.00	0.00	2.51	0.00	nan	0.00
		0.00	0.00	0.00	0.00	0.00	10.52	0.30	3.33	nan	0.00
		0.00	0.59	0.00	0.31	0.00	1.72	0.36	9.15	nan	4.74
		0.00	120.97	0.00	0.19	0.00	0.00	0.03	0.91	nan	2.94
		0.00	0.45	0.00	0.53	0.00	2.43	1.18	0.35	nan	1.14
		0.00	0.00	0.00	0.00	0.00	24.79	0.00	5.12	nan	4.04
		0.00	0.00	0.00	0.21	0.00	16.75	0.14	0.00	nan	14.42
		0.00	6.99	0.00	0.00	0.00	13.76	0.02	2.90	nan	0.00
		0.00	0.00	0.00	0.28	0.00	0.97	3.03	0.53	nan	0.00
		0.00	0.00	3.18	0.00	0.00	17.99	0.12	2.63	nan	0.00
		0.00	0.00	0.00	0.33	0.00	3.90	0.86	0.64	nan	3.20
		0.00	0.00	1.63	0.00	0.00	21.27	0.04	4.99	nan	0.00
		0.00	0.00	0.00	3.55	0.00	1.00	0.37	1.73	nan	4.51
		0.00	0.00	0.00	0.00	0.00	26.89	0.03	1.05	nan	23.98
C		0.00	87.29	0.00	0.00	0.00	1.72	0.03	0.00	nan	0.00
		0.00	82.26	0.00	0.00	2.04	2.11	0.00	0.51	nan	0.00
		0.00	7.36	0.00	0.00	0.00	3.06	0.00	17.34	nan	0.00
		0.00	63.05	8.91	0.00	0.00	0.00	0.00	2.67	nan	0.00
		0.00	31.11	0.00	0.00	8.15	1.76	0.07	2.08	nan	0.00
		0.00	44.32	6.81	0.00	22.35	0.51	0.00	0.96	nan	0.00
		0.00	27.86	0.00	1.08	13.13	0.39	0.00	0.99	nan	12.37
		0.00	inf	0.00	0.00	0.00	0.00	0.00	0.00	nan	0.00
		0.00	24.87	0.00	0.00	0.00	0.00	0.00	17.01	nan	0.00
		0.00	71.81	0.00	0.14	6.37	0.83	0.00	2.31	nan	0.00
		0.00	11.13	0.00	0.00	17.84	7.39	0.00	2.17	nan	0.00
		0.00	0.00	38.65	0.00	0.00	1.02	0.00	4.33	nan	36.88
		0.00	14.83	0.00	0.05	0.00	5.94	0.02	1.70	nan	22.10
		0.00	28.00	0.00	0.29	0.00	1.02	0.11	0.44	nan	12.44
		0.00	1.47	0.00	0.80	0.00	0.00	0.80	5.79	nan	0.00
		0.00	0.00	0.00	0.00	142.87	9.19	0.00	0.00	nan	0.00
G		0.00	0.00	0.00	1.96	5.78	2.02	0.47	0.92	nan	4.08
		0.00	0.00	1.67	1.22	5.69	0.27	1.59	0.16	nan	0.00
		0.00	0.00	0.11	1.47	0.00	0.29	1.72	0.52	nan	0.00
		0.00	0.00	0.00	1.93	0.36	0.66	1.31	0.00	nan	0.22
		0.00	0.00	0.00	0.88	0.00	1.45	0.90	2.75	nan	0.00
		0.00	0.00	10.37	0.48	0.00	0.47	1.67	0.15	nan	0.00
		0.00	0.00	0.76	1.16	0.00	1.10	1.30	0.61	nan	0.00
		0.00	0.00	1.80	1.26	0.00	0.32	1.69	0.45	nan	0.00
		0.00	0.04	1.37	2.89	0.00	0.31	0.97	0.04	nan	0.17
		0.00	0.00	0.66	1.51	0.81	0.45	1.39	0.70	nan	0.00
		0.00	0.00	11.27	2.61	0.00	0.10	0.70	0.65	nan	0.00
		inf	0.00	0.00	3.77	1.31	0.52	0.81	0.00	nan	0.00
		0.00	0.00	0.00	0.64	0.16	0.04	1.52	4.68	nan	0.00
		0.00	0.00	3.44	0.17	0.00	0.04	6.93	0.00	nan	0.00
		0.00	0.00	0.00	0.44	0.00	0.25	4.02	0.68	nan	0.00
		0.00	0.00	0.00	0.28	0.00	0.19	4.49	1.22	nan	0.00
T		0.00	0.00	0.00	1.86	88.75	0.48	0.30	0.44	nan	0.00
		0.00	0.00	2.81	1.69	0.00	1.60	0.50	2.61	nan	0.00
		0.00	0.00	0.00	1.86	0.00	6.15	0.11	4.34	nan	0.00
		0.00	0.00	0.00	2.01	47.05	5.05	0.11	0.00	nan	0.00
		0.00	0.00	0.00	0.00	0.00	1.53	3.76	0.00	nan	0.00
		0.00	0.00	0.00	0.00	0.00	0.40	4.39	0.00	nan	7.76
		0.00	0.00	0.00	1.11	0.00	2.63	0.50	3.47	nan	0.00
		0.00	0.00	0.00	0.00	8.30	1.84	0.62	6.69	nan	0.00
		0.00	0.00	0.00	0.76	0.00	2.31	1.21	0.35	nan	0.00
		0.00	0.00	3.42	0.58	0.00	0.25	2.86	0.08	nan	0.00
		0.00	0.00	0.00	0.71	0.00	3.03	0.78	1.61	nan	0.00
		0.00	0.00	0.00	2.10	0.00	0.97	0.95	0.52	nan	0.00
		0.00	0.00	0.83	0.72	6.03	2.30	0.63	2.64	nan	0.00
		0.00	0.00	0.00	0.83	0.00	0.00	4.19	0.00	nan	0.00
		0.00	0.00	0.00	0.56	14.27	4.30	0.52	0.82	nan	0.00
		0.00	0.00	1.33	0.26	14.57	7.76	0.14	4.49	nan	0.00

Figure 2.12: Nucleotide-amino acid interaction frequency of superimposed trinucleotides. The nucleotide bases (adenine (red), cytosine (yellow), guanine (purple), thymine (green))

and amino acids/groups are represented as single-letter codes. Numbers inside the cells are the values for the odds ratio obtained through Fisher's test. The interactions not significant at a 95% confidence are white. The remaining interactions are coloured blue and pink for significantly favourable and unfavourable interactions, respectively. The raw data is in Appendix 6.

Table 2.6: Comparison of di- and trinucleotide preferences.

The nucleotides and amino acids/groups are in single-letter codes. First nucleotide means that the first member of the dinucleotide is hydrogen-bonded. Last nucleotide refers to the cases where the second nucleotide of the dinucleotide is hydrogen-bonded. Pink cells indicate cases where the preference of trinucleotides was distinct from dinucleotides. Grey cells represent cases where the last nucleotide and trinucleotide had the same preference. In contrast, blue cells depict the cases where the first nucleotide and trinucleotide had the same preference.

Trinucleotide	Amino acid	Trinucleotide odds ratio	Last Nucleotide odds ratio	First Nucleotide odds ratio
ATC	ST	2.61	0.51	2.32
GGC	R	1.39	0.80	1.65
GGT	H	0.00	2.60	0.52
TGC	ST	0.00	2.66	0.74
TGG	ST	0.68	2.66	0.17
GGT	K	3.77	2.89	0.39
AGG	H	0.11	0.38	2.60
GGT	R	0.81	0.80	1.95
ATC	K	1.69	1.68	0.25
AGT	K	1.93	1.60	0.39
TGA	K	0.64	0.39	1.60
CGG	R	1.30	1.65	0.80
CGT	ST	0.45	0.74	2.66
GTC	ST	0.08	0.30	2.32
GGC	K	1.51	2.89	0.79
ATG	R	0.11	0.75	1.47
GGT	ST	0.00	0.17	2.66
TTA	ST	2.64	1.76	0.51
GGT	C	inf	2.82	0.00
TGA	ST	4.68	2.66	0.23
TGG	H	0.00	0.52	2.60
TGA	Mc	0.16	0.06	1.64
AGT	ST	0.00	0.23	2.66
TTG	Mc	14.27	4.82	0.00

AGC	Mc	5.69	1.64	0.09
TGG	K	0.44	0.39	2.89
CGA	ST	2.75	0.74	0.23
GAG	R	3.03	0.62	0.62
ATG	ST	4.34	0.51	0.30
CGT	K	1.26	0.79	0.39
CAC	NQ	0.00	2.97	2.97
GCA	NQ	0.00	2.23	2.28
AAG	NQ	0.00	3.60	6.80
TAG	K	3.55	0.37	0.13

2.3 Development of the scoring scheme

We determined the frequencies of amino acid-nucleotide interaction for mono-, di- and trinucleotide CLICK superimpositions. We used these data to construct a scoring scheme to predict the amino acids most likely to bind to a given DNA sequence. (equations 1, 2 and 3).

$$S_{xa} = \frac{O_{xa}}{E_{xa}} \quad \text{Eq.1}$$

$$E_{xa} = \frac{F_a F_x \sum O_{xa}}{N} \quad \text{Eq. 2}$$

$$S_{xa}' = \frac{S_{xa}}{\sum S_{xa}} \quad \text{Eq. 3}$$

Where, S_{xa} and O_{xa} correspond to the score and observed frequency of superimposed amino acid a , which interacts with the nucleotide x . E_{xa} is the expected frequency of the interaction by random chance. F_a and F_x are the amino acid and nucleotide frequencies, respectively in the training dataset. N is the number of combinations in which that nucleotide x can occur (1, 4 and 16 in the case of mono-, di- and trinucleotides (e.g., A [A, T, G, C] and [A, T, G, C] A [A, T, G, C])).

The scoring scheme is a variation of the ratio of observed and expected frequency (Eq. 1). We calculated the frequency of amino acids and nucleotides in the dataset to obtain the expected frequency (Eq. 2). If the interaction occurs by chance, this is the

number of times a nucleotide-amino acid pair would interact. A score higher than 1 indicates a favourable interaction. We then normalized the scores to get a range from 0 to 1 for an interaction (Eq. 3).

As the hydrogen bonds in the major and minor groove differ in the amino acid-nucleotide preferences, we incorporated this into the scoring function. The abundance of tri-, di- and mononucleotides differs between the major and minor grooves. This difference in abundance was accounted by f (Eq. 4).

$$f = \frac{O_{xa_{major}}}{O_{xa_{minor}}} \quad \text{Eq. 4}$$

Where, f is the number of times an observation occurs in the major groove Vs. the minor groove, $O_{xa_{major}}$ is the number of observations in the major groove, $O_{xa_{minor}}$ is the number of observations in the minor groove.

The major groove interactions were roughly four times more frequent than minor groove interactions. Thus, the score for the major groove was multiplied by 4 (Eq. 5).

$$S_{xa} = f * 4S_{xa_{major}} + (1-f) * S_{xa_{minor}} \quad \text{Eq. 5}$$

Since we had three matrices corresponding to mono-, di- and trinucleotides, the weight for each was optimized using the grid search method. Within a window size of three, the prediction was made for the central residue by selecting the highest score from the mono-, di- and trinucleotide matrices.

$$Prediction = \max(w_1 * C_{xa}, w_2 * L_{xa}, w_3 * F_{xa}, w_4 * S_{xa}) \quad \text{Eq. 6}$$

Where w_1 , w_2 and w_3 correspond to the weights given to the tri-, di- and mononucleotide matrix, respectively. The trinucleotide matrix provides information about both the neighbours of the hydrogen-bonded nucleotide. Thus, we based these weights on the abundance of data available for trinucleotide. A higher number of data points would increase our confidence in the data. Thus we had a sigmoid function to obtain weight w_1 and distributed the remaining fraction of the weight to w_2 and w_3 (Eq. 7, 8 and 9).

$$w_1 = c + (1 - c) / (1 + e^{p - (q * O_{xa})}) \quad \text{Eq. 7}$$

$$w_2 + w_3 = 1 - w_1 \quad \text{Eq. 8}$$

$$w_2 = d * w_3 \quad \text{Eq. 9}$$

Where c is the y-intercept, p and q are variables that decide the sigmoid slope, O_{xa} is the frequency of occurrence of superimposed trinucleotide, and d is a variable that defines the relationship between w_2 and w_3 .

We performed a grid search in the range of 0 to 1 with an interval of 0.1 for c . The grid search range was between 0 to 10 and 0 to 30, with an interval of 1 for p and q , respectively. The grid search for d was carried out between 0 to 1 with an interval of 0.01. These grid search ranges gave us approximately 3 million combinations.

For each combination of weights, we predicted the amino acid that would specifically bind a given DNA sequence in the training set (Figure 2.13). We had a three-nucleotide sliding window, and a prediction was made for the central nucleotide using mono-, di- and trinucleotide matrices (Eq. 6). The maximum scoring amino acid was predicted, and the prediction was verified using the crystal structure. The combination of weights that resulted in the maximum correct predictions was $p=48$, $q=93$, $d=0.01$ and $c=0$. These resulted in tri-, di- and mononucleotide weights (Appendix 8), which were used to predict the amino acids for the 50 PDBs of the testing set.

The split of 235 and 50 from the 285 non-redundant specific DNA-protein complexes was random. Therefore, we performed this split three times and obtained three sets of weights for the training set. We applied the corresponding weights to the structures not used in the training. If the results are consistent regardless of the structures used for training, it indicates that we have not over-trained or under-trained the weights..

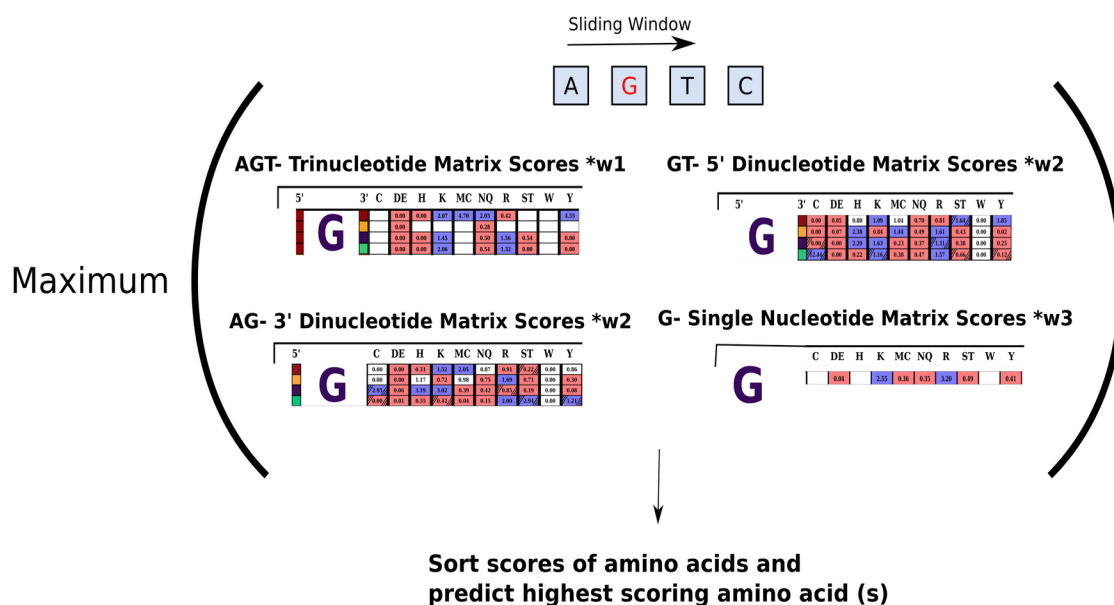


Figure 2.13: Schematic of prediction of amino acids for a given DNA.

In the window size of three, the amino acid for the central nucleotide was predicted (G in red). The scores for AGT from trinucleotide matrix, both AG (Last Nucleotide) and GT (First Nucleotide) dinucleotide scores G mononucleotide score were used for prediction. $w1$, $w2$, and $w3$ are weights optimised for tri-, di- and mononucleotides.

2.4 Accuracy of the predictions on training and testing set

We calculated the DNA-protein hydrogen bonds in the training and testing set. We provided the hydrogen-bonded nucleotides as input to the program and made predictions of the amino acids most likely to bind them. We assessed the prediction as accurate in three ways-

- Strict: If the highest-scoring prediction matched the amino acid in the crystal structure.
- Moderate: If one of the top two highest-scoring predictions matched the amino acid in the crystal structure.
- Lenient: If one of the top three highest-scoring predictions matched the amino acid in the crystal structure.

If a nucleotide had multiple amino acids interacting with it, we classified the prediction as correct if we could predict at least one of the amino acids. Cysteine and

tryptophan did not have any occurrences where they were the only amino acids interacting with a nucleotide.

We obtained 87% and 85% of correct predictions for the training and testing sets, respectively, when the lenient evaluation measure was chosen (Table 2.7). The lowest number of correct predictions were made for histidine (61.7%) and tyrosine (68.1%) in the training set even with the lenient measure. Histidine (50%) and tyrosine (60%) also had the lowest correct predictions in the testing set considering the lenient scheme.

The preference of lysine and arginine towards guanine and thymine was similar. Since arginine interactions are abundant and spatially conserved, lysine was rarely the top-most scoring amino acid. Cysteine and tryptophan did not have sufficient occurrences in the dataset to make accurate predictions. A similar argument can be made for histidine and tyrosine, as evident from the distribution of hydrogen bonds in the training and testing set (Figure 2.14).

The lowest number of hydrogen bonds were observed for cysteine (3 bonds, 0.1%), tryptophan (3 bonds, 0.1%), histidine (68 bonds, 2.7%) and tyrosine (89 bonds, 3%), which also had the lowest number of correct predictions in the training set (0, 0, 57 and 68%, respectively). Histidine (14 bonds, 2.3%) and tyrosine (16 bonds, 2.3%) also had the lowest frequency and lowest correct predictions (50 and 60%, respectively). Pearson's correlation between correct predictions and the percentage of hydrogen bonds was 0.50 and 0.62 for the training and testing set, respectively. This correlation indicates that this method is expected to improve its accuracy with the increased data.

Table 2.7: The correct predictions (%) of amino acids for a given DNA sequence in the training and testing set.

The three assessment ways-strict, moderate and lenient are mentioned for the amino acids and amino acid groups shown in single-letter codes.

	Training			Testing		
Amino Acids	Strict (Top 1)	Moderate (Top 2)	Lenient (Top 3)	Strict (Top 1)	Moderate (Top 2)	Lenient (Top 3)
DE	16.2	72.2	79.6	37.5	92.3	92.3
Mc	51.6	78.3	90.8	43.8	76.9	92.9
H	0.0	24.1	61.7	0.0	0.0	50.0
K	6.5	69.4	89.3	0.0	76.7	89.5
NQ	26.4	63.8	80.4	18.6	53.7	66.7
R	87.4	93.0	93.5	84.6	91.2	91.2
ST	8.3	44.7	72.5	5.6	48.0	75.7
Y	20.0	38.7	68.1	0.0	33.3	60.0
C	0.0	0.0	0.0	0.0	0.0	0.0
W	0.0	0.0	0.0	0.0	0.0	0.0
Overall	54.5	77.3	86.9	53.7	77.3	85.5

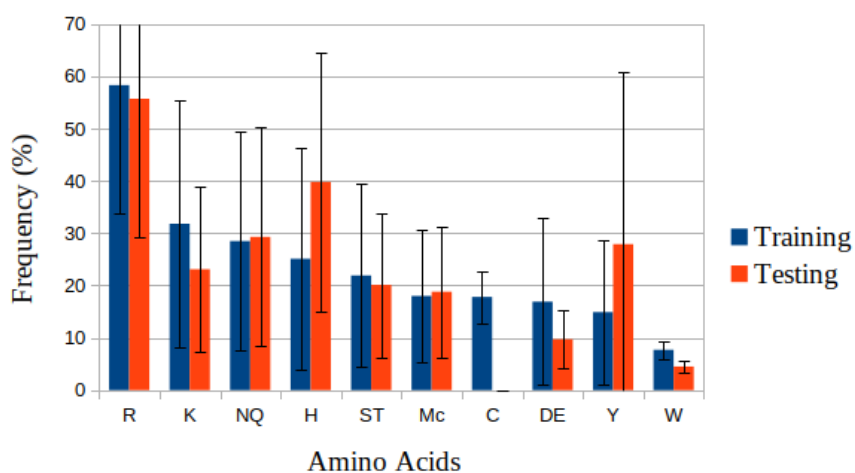


Figure 2.14: Distribution of hydrogen bonds in the training and testing sets.

The testing and training set bonds are blue and orange, respectively. The amino acids/ groups are represented in single-letter codes. The error bars show the deviation across the datasets.

Another reason for the top predictions not matching the amino acid present in the crystal structure could be the probabilistic nature of our predictions. Thus, even though a nucleotide-amino acid pair may form preferential interactions, it may not be present in that particular protein. We performed evolutionary analysis on the 235 proteins in our training set to verify our hypothesis. We used the PSIBLAST module of ncbi-blast-2.7.1+ for 2 iterations with the default e-value of 0.001 [177,178]. We limited the multiple sequence alignment size to 1000. The Multiple Sequence Alignment (MSA) provided a list of amino acids that occur in the binding site in close homologs. The binding site was defined as the hydrogen bonded amino acids in the three-dimensional structure. We correlated the top three predictions for each nucleotide with the amino acids present in the MSA for that protein.

The top 1 prediction was present in the MSA in 74.3% of the cases, while the top 2 and top 3 predictions were present in 86.2 and 90.4% of the cases, respectively. A PDB-wise analysis showed that in 92 PDBs we predicted all hydrogen bond amino acids correctly in the top 1 (Figure 2.15). 137 and 165 PDBs out of 235 had correct predictions for all MSA constituents in top 2 and top 3, respectively.

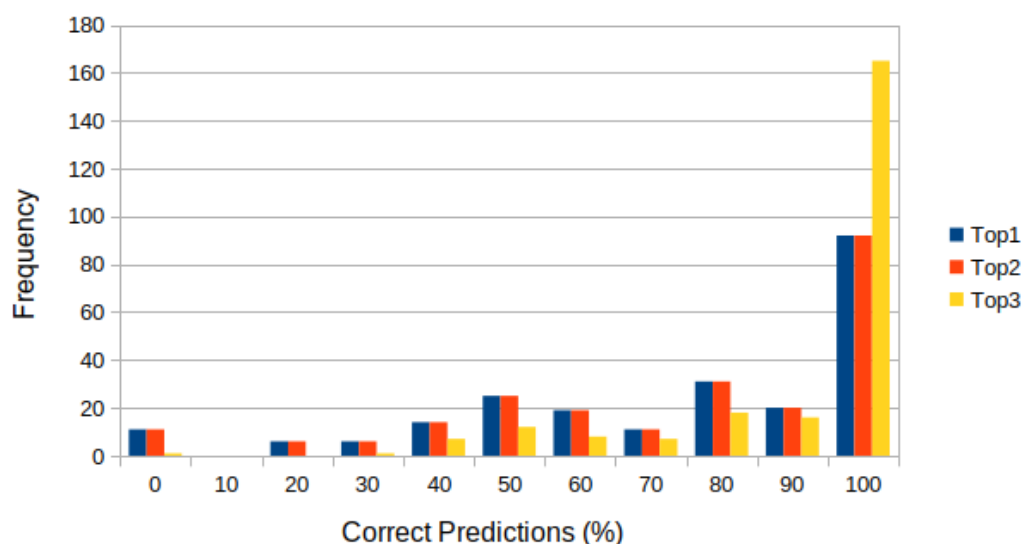


Figure 2.15: Correlating top 3 predictions with the MSA constituents.

The y-axis is the frequency of PDBs with correct predictions (%) on x-axis. Top 1, top 2 and top 3 predictions are coloured blue, orange and yellow respectively.

We can therefore conclude that the apparent low accuracy of our method when only top 1 prediction is considered is partly due to the lower abundance of hydrogen bond data. Since our method is probabilistic, a favourable interaction may not match with the protein that has been crystallized, but can be observed in its close homologs. We validated our method using different *in-silico* (Chapter 3) and *in-vitro* approaches (Chapter 4).

2.5 Chapter summary

This chapter describes the compilation and characterization of the dataset of sequence-specific DNA-protein complexes. The DNA-binding and non-DNA-binding proteins have different amino acid composition profiles. The interactions of the amino acids in DNA-protein complexes are primarily mediated by hydrogen bonds. Amino acid-nucleotide hydrogen bonding partners exhibit statistically significant preferences. We have used superimposition of these hydrogen bonds across the DNA-binding protein families to derive family-independent specificity preferences. We developed a scoring scheme using these preferences to predict the amino acids that would specifically bind a DNA sequence. This approach was validated with a testing set and the MSA of the proteins.

3

CHAPTER 3: *In silico* Validation of Predicting Amino Acids for a DNA Sequence

OVERVIEW

- Validation of our method on the zinc finger proteins
 - With experimentally selected specific zinc finger-DNA complexes
 - Using crystal structures of zinc finger proteins
 - Predicting the MSA from PWMs
 - Prediction of the effect of mutation on DNA-protein binding affinity
 - Validation of our method using Molecular Dynamics Simulations
 - Zif-268
 - Myc
 - Ets1
 - EcoRV
 - Discussions
 - Chapter summary
-

The previous chapter outlined the methodology for the predicting amino acids specifically binding to a given DNA sequence. The current and ensuing chapters deal with the approaches for validating our method, starting with *in-silico* validation in this chapter.

3.1 Validation of our method on the zinc finger proteins

3.1.1 Validation on SELEX selected zinc finger-DNA complexes

The zinc finger proteins have a modular DNA-binding behaviour and units of zinc finger exhibit 1:1 relationship between interacting amino acids and nucleotides (refer to Chapter 1, section 1.4.1). Each unit of the zinc finger proteins recognizes a four base pair sequence, with the first three on the forward strand and the last one on the complementary strand. Many studies have tried to optimise the binding of zinc fingers to DNA triplets that bind on the forward strand. We took the results of one such study by Liu *et al.* [225] that had performed SELEX experiments to determine the best binding and least cross-reactive amino acids to GNN triplets. The GNN triplet is a sequence of trinucleotides, where guanine is the first nucleotide recognised by the zinc finger. We predicted the amino acids binding these GNN triplets and compared our results with the experimentally determined ones (Table 3.1).

As stated in Chapter 2, we combined the amino acids with common hydrogen bonding groups. If the experimentally determined amino acid is the highest scoring amino acid, it is counted in the top 1 prediction. Likewise, the top 2 and top 3 categories contain the experimentally determined amino acid in the top 2 and 3 highest scoring amino acid categories, respectively. Using this prediction scheme, we have 56% correct predictions in top 1, 62% in top 2 and 68% in top 3. We could not identify the serine/threonines and the main chain groups interacting with the nucleotides correctly. This preference may be family-specific and not a generalized method of recognizing nucleotide bases.

Table 3.1: Validation of our method on the Liu et al. zinc finger set.

The amino acids are represented as single letter codes and grouped classes.

	Top 1		Top 2		Top 3	
Amino Acid	Predicted	Not Predicted	Predicted	Not Predicted	Predicted	Not Predicted
DE	0	8	2	6	2	6
H	0	4	0	4	3	1
NQ	7	2	8	1	8	1
R	20	0	20	0	20	0
ST	0	3	0	3	0	3
Mc	0	4	0	4	0	4
% Correct	56.25		62.5		68.75	

3.1.2 Validation on crystal structures of zinc finger proteins

We obtained 33 PDBs from the RCSB PDB that had the zinc finger annotation (1tf6, 4m9v, 2i13, 2drp, 1llm, 6wmi, 1mey, 1p47, 6b0q, 6b0r, 5t00, 5ei9, 5vc9, 3uk3, 6b0o, 5kkq, 5k5l, 6b0p, 5kl4, 5und, 5t0u, 5eh2, 5yeg, 5yef, 1g2f, 5yeh, 1f2i, 5yel, 5v3j, 6ki6, 5v3g, 1g2d, 5yj3). We calculated the hydrogen bonds between the nucleotides and amino acids in these structures using our in-house python script. We then made predictions using the DNA sequences of these structures and checked the accuracy with the hydrogen-bonded partner in the crystal structure (Table 3.2). We have approximately similar accuracy as on the Liu et al. set for top 2 and top 3 predictions. The number of correct predictions in the top 1 category have reduced compared to that in the Liu et al. set.

Table 3.2: Validation of our method on the crystal structures of zinc finger proteins.

The amino acids/groups are represented in single-letter codes.

	Top 1		Top 2		Top 3	
Amino Acid	Predicted	Not Predicted	Predicted	Not Predicted	Predicted	Not Predicted
DE	1	20	7	14	7	14
H	0	11	0	10	9	5
K	0	16	15	6	20	3
NQ	8	15	14	11	18	10
R	73	19	77	19	77	19
ST	4	5	4	5	4	5
Y	0	3	0	0	0	0
Mc	0	1	0	1	0	1
Correct (%)	48.86		63.93		70.31	

3.2 Predicting Multiple Sequence Alignment (MSA) from Position Weight Matrices (PWMs)

The hypothesis behind this analysis was that as a protein sequence diverges, the nucleotide partner that it specifically binds would also change accordingly. This co-evolution can be reconciled by correlating the position weight matrix of the DNA-binding site with the pattern observed in the MSA of the binding protein and its close homologs.

We downloaded 3794 PWMs from TRANSFAC for this validation and restricted the PWM length to less than 10 bases for practical considerations [226]. We further ensured that there was a hydrogen bond between the protein and the nucleotide/(s) in the region specified by the PWM. This filtering resulted in 1925 PWMs. We obtained a 40% non-redundant set using CD-HIT [207,227] and had 452 PWMs in the final dataset (List in Appendix 9).

We generated all nucleotide sequences that satisfy the PWM criteria and predicted the amino acid most likely to bind at each location of the PWM. We took the best prediction for each DNA sequence and compiled all amino acids that could bind at a given DNA location. In order to check if our predicted amino acids matched the ones present in the protein multiple sequence alignment, we carried out PSIBLAST [177,178] for three iterations with an e-value cut-off of 0.001.

We now had a predicted set of amino acids and a set obtained from MSA for each amino acid in the protein that had an interaction with the nucleotide. We checked the ranks of the top 3 predicted amino acids in the MSA. Out of 2193 predictions, the predicted amino acid was present in the MSA in 952 cases (43%). We checked the rank of these 952 predictions and the rank 1 predictions matched with the MSA in 351 or 37% of the cases (Figure 3.1). The lower ranks are also encountered less in the MSA, which is substantiated by the Pearson's correlation coefficient of 0.92 between the rank in the prediction and the frequency in the MSA.

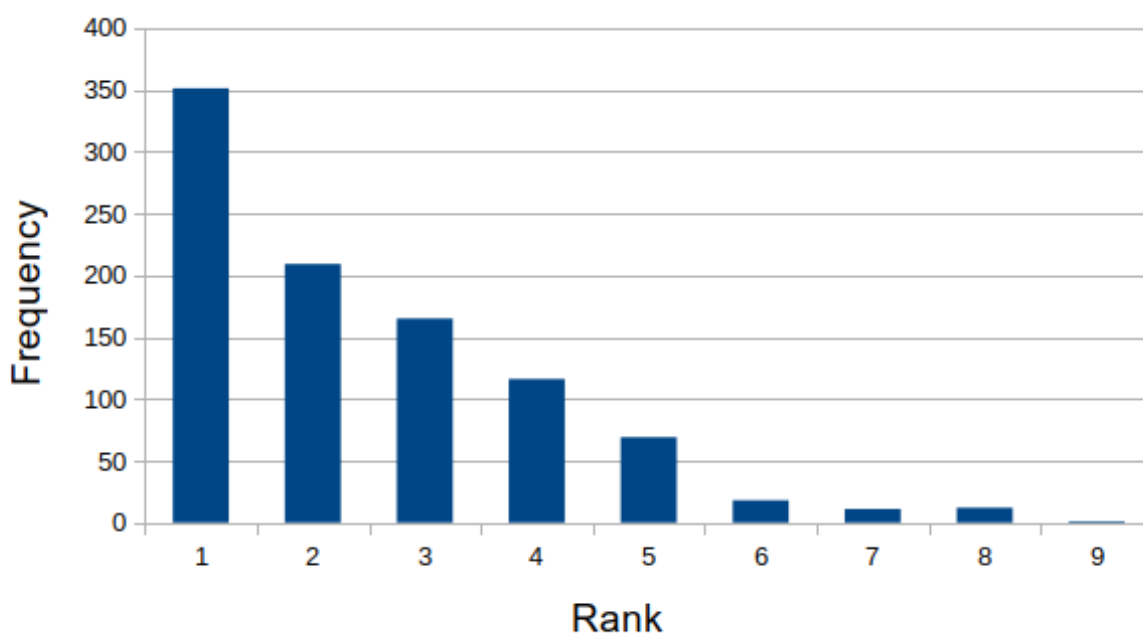


Figure 3.1: Rank distribution of the predicted amino acids that matched with the MSA.

In the 351 cases where the rank 1 of our prediction matched the MSA, 116 were also the highest frequency amino acids in their respective MSA. The rank 1 prediction also matched rank 2 amino acid in the MSA in 74 cases and rank 3 in 50 cases. We analysed

the amino acid composition of these 351 cases and the 116 where the rank 1 of MSA was also rank 1 in our prediction (Figure 3.2). In 33% of the cases, arginine was the top 1 prediction that matched with the MSA at any rank. Arginine was also present in 39% of the cases where the rank 1 prediction also matched the rank 1 MSA.

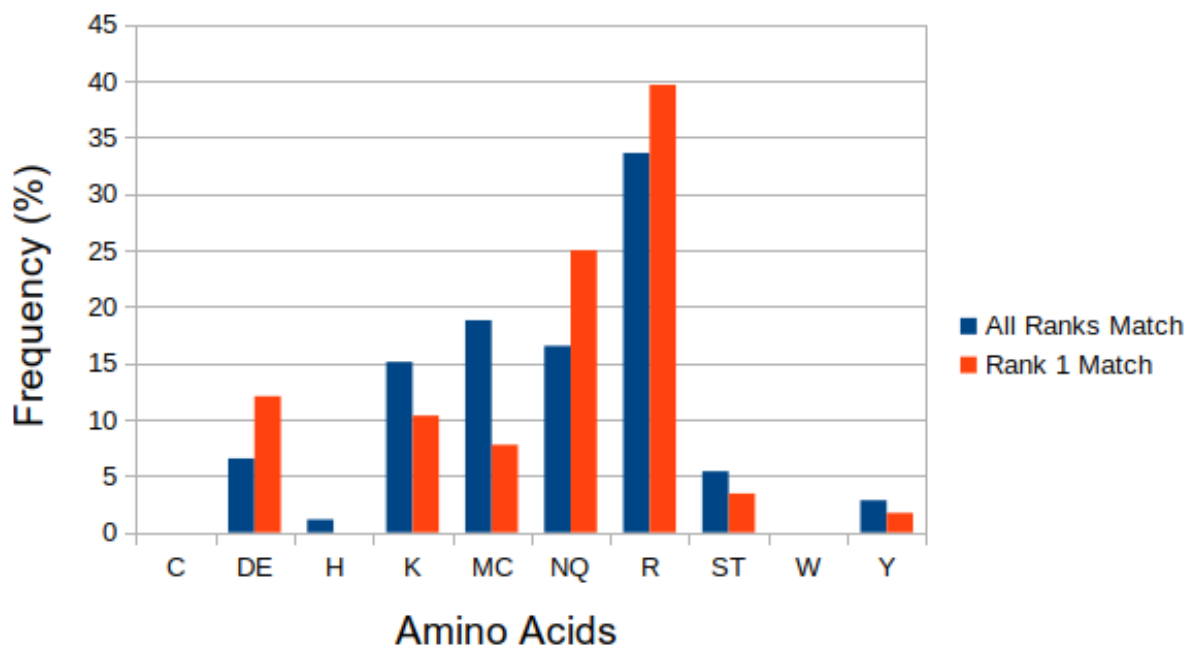


Figure 3.2: Amino acid distribution in the cases where the prediction matched the MSA.

The rank 1 of the prediction matching with all ranks in MSA is shown in blue while the rank 1 of the prediction matching with rank 1 of the MSA is shown in orange.

To quantify the efficacy of the method, we took the three highest frequency amino acids from the MSA and our prediction. We compared the two profiles using Matthews Correlation Coefficient [228] (MCC). The measures for the confusion matrix were defined as follows-

True positive (TP): Present in the prediction and MSA

True negative (TN): Absent in the prediction and MSA

False positive (FP): Present in prediction and not in MSA

False negative (FN): Not present in the prediction and present in MSA

We then calculated the MCC the using the above parameters and the formula-

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TN + FP) * (TP + FN) * (TN + FN)}} \quad \text{Eq. 1}$$

MCC is a balanced statistical score that ranges between -1 and +1, with -1 indicating complete inverse predictions and +1 denoting perfect predictions. The value of MCC is high only if the values for TP and TN are high and that of FN and FP are low.

The MCC value of the MSA prediction using our probabilistic scheme is 0.17. The PDB-wise distribution shows two peaks at 0.1 and 0.6. However, there are many structures for which the MCC values are less than or equal to zero (Figure 3.3).

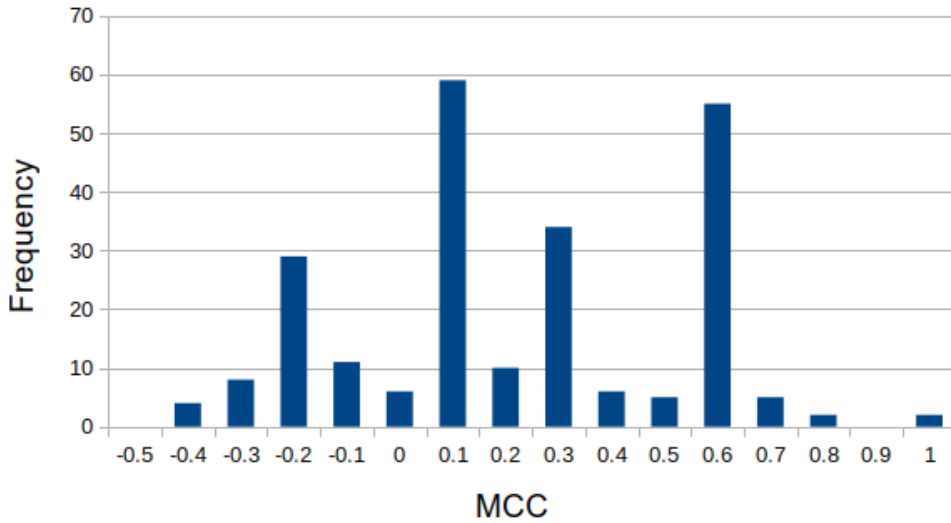


Figure 3.3: Distribution of MCC values for predicting the top 3 amino acids in the MSA

We calculated the rates of false positives, false negatives, true positives and true negatives to derive the reason for the low MCC value using the formulae-

$$\text{True Positive rate (TPR)} = \frac{TP}{TP + FN} \quad \text{Eq. 2}$$

$$\text{True Negative rate (TNR)} = \frac{TN}{TN + FP} \quad \text{Eq. 3}$$

$$\text{False Positive rate (FPR)} = \frac{FP}{TN + FP} \quad \text{Eq. 4}$$

$$\text{False Negative rate (FNR)} = \frac{FN}{TP + FN} \quad \text{Eq. 5}$$

Calculating the MCC using these rates (Eq. 6) is useful when the dataset is imbalanced [229].

$$MCC = \frac{TPR * TNR - FPR * FNR}{\sqrt{(TPR + FPR) * (TNR + FPR) * (TPR + FNR) * (TNR + FNR)}}$$

Using Eq. 6, we obtained an MCC value of 0.19, which is a marginal improvement over the standard MCC formula.

We realized that the true positive rate was low, which led to the overall low MCC (Figure 3.4). This means that while the identification of what is not specific is clear, work still needs to be done on identifying correct pairs of interacting nucleotide-amino acid partners. This could be improved by having more structural data to infer preferences from.

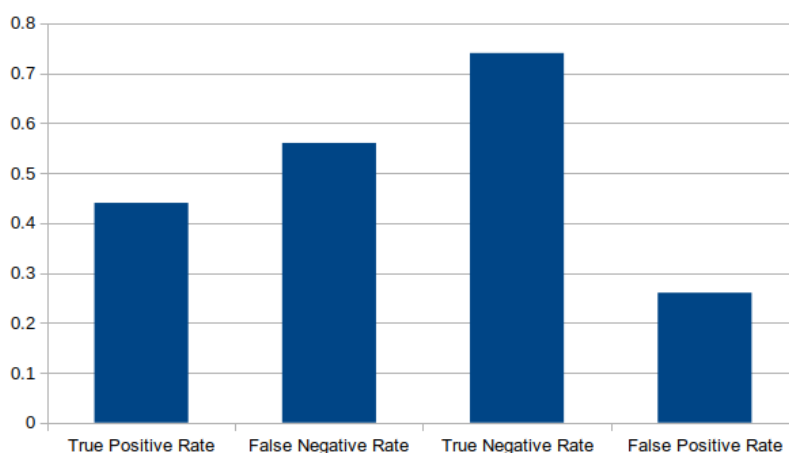


Figure 3.4: Comparison of the contribution of different components toward MCC

3.3 Prediction of the effect of mutation on DNA-protein binding affinity

The study of DNA-protein interaction only considers specificity so far and not affinity between the two moieties. The lack of an adequate dataset for this purpose hinders the binding affinity prediction. The ProNIT resource has multiple binding affinities for a single

DNA protein complex [230]. One of the reasons for multiple values, apart from errors in the entries, is that the binding affinity of DNA-protein complexes is affected by the temperature, pH, ion concentration, presence of co-factors and the experimental technique used to measure the binding affinity. The Mat-Alpha2 protein, for example, has 100 fold difference in its affinity when the ion concentration is changed from 100 to 300 mM [231]. This sensitivity of the binding affinity to experimental conditions limits the dataset we can use to analyse binding affinity changes influenced by specificity changes. Therefore we avoided comparisons of affinity among different proteins. Instead, we took systems where mutations were made on the same DNA-protein complex, either on the protein or the DNA side.

Here, we have considered only the effect of amino acids hydrogen-bonded to DNA. These criteria gave us 83 single-point mutations. 73 of these 83 mutations result in a reduction in the binding affinity. This observation could either be an artefact of lower sampling or could indicate that introducing a mutation in the binding residues has a higher probability of reducing the binding affinity between the protein and DNA. We have also made similar observations with our molecular dynamics simulations (section 3.4).

To predict whether the mutation is likely to increase or decrease the binding affinity of the DNA-protein complex, we employed a normalized trinucleotide odds ratio. The normalization was done so that the sum of its interaction with the amino acids/groups for each trinucleotide equals 1. If the normalized odds ratio for the wild-type trinucleotide-amino acid pair was less than that for the mutant trinucleotide-amino acid pair, we predicted that the binding affinity would increase upon mutation. Similarly, if the normalized odds ratio for the wild-type trinucleotide-amino acid pair was greater than that of the mutant trinucleotide-amino acid pair, we predicted the binding affinity of the DNA protein complex would reduce upon making the mutation.

We compared our prediction with the experimental results. We could correctly predict the resulting change as an increase or decrease in the affinity in 49 cases or 59% of the cases (Table 3.3), which is slightly better than random chance.

However, in 19 of the 34 incorrect predictions, we did not have wild-type or mutant trinucleotide-amino acid pair information. So, when we have trinucleotide information for at least the wild-type or the mutant trinucleotide-amino acid pair, we have an accuracy of 76%. Another point to note is that the assumption is that the hydrogen bond is retained in the mutant protein. For the 15 incorrect predictions with trinucleotide-amino acid information, in 7 cases, the wild-type amino acid has side-chain hydrogen bond donors and acceptors, while the mutant amino acid does not. Along with this, there is a concomitant reduction in binding affinity in these 7 cases. This decrease could either be due to the loss of a hydrogen bond or the mutated amino acid does not have a favourable interaction with the DNA. Therefore, we require more data to judge the method reliably.

We then analysed whether the incorrect predictions were associated with the binding affinity change between the wild-type and mutant proteins. We correlated the predictions with the term $\log(\text{Fold Change})$ defined by the formula-

$$\log(\text{Fold Change}) = \log\left(\frac{K_{d1}}{K_{d2}}\right)$$

where K_{d1} and K_{d2} correspond to the binding affinities of the wild-type and mutant proteins, respectively measured in terms of their dissociation constants.

Table 3.3: Predictions of the effect of mutation on DNA-binding interface.

The correct predictions are highlighted in bold. WT and MT correspond to wild-type and mutant proteins, respectively. The nucleotides and amino acids/groups are depicted in single-letter codes. AA and Tri stand for amino acid and trinucleotide, respectively. $\log(\text{OR})$ is the normalized log odds ratio for that trinucleotide-amino acid interaction. K_d is the experimentally measured dissociation constant used to estimate the binding affinity. $\log(\text{Fold})$ is the log of the ratio of the wild-type binding affinity to the mutant binding affinity.

Sr, No	PDB	WT				MT				log (Fold)
		AA	Tri	log (OR)	K_d	AA	Tri	log (OR)	K_d	
1	1AAY	DE	CAT	0	1.70E-10	Mc	CAT	0	1.00E-10	0.23
2	1AAY	MC	CAT	0	1.00E-10	Mc	GAT	0	2.40E-09	-1.38
3	1AAY	DE	GAT	0	2.00E-08	Mc	GAT	0	2.40E-09	0.92

4	1AAY	R	AGC	0.15	1.70E-10	Mc	AGC	0.54	7.00E-08	-2.61
5	1AAY	R	CGT	0.31	1.70E-10	Mc	CGT	0	1.70E-08	-2
6	1AAY	R	CGT	0.31	1.70E-10	R	CCT	0	2.00E-08	-2.07
7	1AZ0	Mc	CGA	0	1.60E-08	Mc	CAA	0	1.10E-06	-1.84
8	1AZ0	Mc	CGA	0	1.60E-08	Mc	CTA	0	1.40E-06	-1.94
9	1AZ0	Mc	CGA	0	1.60E-08	Mc	CCA	0.19	1.40E-06	-1.94
10	1AZ0	NQ	GAT	0.75	1.60E-08	NQ	GTT	0.21	1.00E-06	-1.8
11	1AZ0	NQ	GAT	0.75	1.60E-08	NQ	GGT	0	1.00E-06	-1.8
12	1AZ0	NQ	GAT	0.75	1.60E-08	NQ	GCT	0.01	2.50E-07	-1.19
13	1BP7	Y	CTA	0	5.60E-07	C	CTA	0	4.70E-08	1.08
14	1BP7	Y	CAA	0.28	4.00E-10	C	CAA	0	2.50E-08	-1.8
15	1BP7	ST	GCA	0.41	4.00E-10	K	GCA	0	5.00E-08	-2.1
16	1BP7	ST	GCA	0.41	4.00E-10	ST	GGA	0.01	5.90E-07	-3.17
17	1BP7	ST	GGA	0.01	5.00E-08	K	GGA	0.5	4.40E-08	0.06
18	1BP7	Y	CAA	0.28	4.00E-10	Y	CTA	0	5.60E-07	-3.15
19	1CW0	R	CTA	0.71	7.40E-10	R	TTA	0.05	7.00E-09	-0.98
20	1CW0	R	CTA	0.71	7.40E-10	R	CTC	0.35	4.10E-08	-1.74
21	1ECR	R	AGT	0.29	1.00E-09	Mc	AGT	0.08	1.30E-07	-2.11
22	1ECR	K	ATG	0.15	1.00E-09	Mc	ATG	0	1.10E-07	-2.04
23	1ECR	NQ	TAG	0.09	1.00E-09	Mc	TAG	0	2.20E-07	-2.34
24	1EWQ	DE	GTG	0	2.20E-09	Mc	GTG	0	2.80E-09	-0.1
25	1J5N	Y	ACA	0	1.00E-08	Mc	ACA	0	4.30E-08	-0.63
26	1J5N	R	ATC	0.05	1.00E-08	Mc	ATC	0	2.80E-08	-0.45
27	1LMB	K	GGT	0	7.00E-12	NQ	GGT	0	5.50E-08	-3.9
28	1LMB	NQ	AGT	0.15	7.00E-12	K	AGT	0.43	1.00E-08	-3.15
29	1LMB	NQ	TAT	0.52	7.00E-12	ST	TAT	0.02	3.90E-09	-2.75
30	1LMB	NQ	TAT	0.52	7.00E-12	Y	TAT	0.46	4.50E-09	-2.81
31	1LMB	NQ	TAT	0.52	7.00E-12	Mc	TAT	0	3.00E-07	-4.63
32	1LMB	ST	TGG	0.13	7.00E-12	Mc	TGG	0	3.00E-07	-4.63
33	1MSE	ST	CTA	0	5.50E-09	Mc	CTA	0	1.80E-08	-0.51
34	1MSE	ST	CTA	0	5.50E-09	Mc	CTA	0	9.00E-09	-0.21
35	1MSE	ST	CTA	0	3.20E-09	Mc	CTA	0	1.80E-08	-0.75
36	1MSE	ST	CTA	0	3.20E-09	Mc	CTA	0	9.30E-09	-0.46
37	1MSE	ST	CTA	0	3.20E-09	Mc	CTA	0	1.30E-09	0.39
38	1MSE	ST	CTA	0	3.20E-09	Mc	CTA	0	1.30E-07	-1.61
39	1MSE	ST	CTA	0	3.20E-09	K	CTA	0	6.10E-08	-1.28
40	1MSE	ST	CTA	0	3.20E-09	DE	CTA	0	1.00E-06	-2.49

41	1MSE	ST	CTA	0	3.20E-09	NQ	CTA	0.29	2.60E-08	-0.91
42	1MSE	ST	CTA	0	3.20E-09	NQ	CTA	0.29	3.60E-08	-1.05
43	1MSE	ST	CTA	0	5.50E-09	ST	CGA	0.46	1.50E-07	-1.44
44	1MSE	ST	CTA	0	3.20E-09	R	CTA	0.71	4.30E-08	-1.13
45	1MSE	ST	CCA	0.05	3.70E-09	Mc	CCA	0.19	1.70E-08	-0.66
46	1MSE	ST	CCA	0.05	3.70E-09	Mc	CCA	0.19	1.20E-08	-0.51
47	1MSE	ST	CCA	0.05	3.70E-09	Mc	CCA	0.19	8.90E-09	-0.38
48	1MSE	ST	CCA	0.05	3.70E-09	Mc	CCA	0.19	7.30E-08	-1.3
49	1MSE	ST	CCA	0.05	3.70E-09	DE	CCA	0.72	1.00E-06	-2.43
50	1MSE	ST	CGA	0.46	2.50E-08	Mc	CGA	0	2.40E-08	0.02
51	1MSE	ST	CGA	0.46	2.50E-08	R	CGA	0.15	2.10E-08	0.08
52	1MSE	K	TGA	0.09	5.50E-09	Mc	TGA	0.02	1.00E-06	-2.26
53	1MSE	K	TGA	0.09	5.50E-09	K	TCA	0	3.70E-07	-1.83
54	1MSE	K	TCA	0	3.70E-07	Mc	TCA	0	3.80E-06	-1.01
55	1MSE	ST	CGA	0.46	2.50E-08	Mc	CGA	0	3.70E-08	-0.17
56	1MSE	ST	CGA	0.46	2.50E-08	NQ	CGA	0.24	5.30E-08	-0.33
57	1MSE	ST	CGA	0.46	2.50E-08	NQ	CGA	0.24	3.90E-08	-0.19
58	1MSE	ST	CGA	0.46	2.50E-08	Mc	CGA	0	3.40E-08	-0.13
59	1MSE	ST	CGA	0.46	2.50E-08	Mc	CGA	0	7.20E-08	-0.46
60	1MSE	ST	CGA	0.46	2.50E-08	K	CGA	0.15	3.60E-08	-0.16
61	1MSE	ST	CGA	0.46	2.50E-08	DE	CGA	0	1.00E-06	-1.6
62	1MSE	ST	CAA	0.54	8.70E-09	Mc	CAA	0	3.30E-08	-0.58
63	1MSE	ST	CAA	0.54	8.70E-09	Mc	CAA	0	2.20E-08	-0.4
64	1MSE	ST	CAA	0.54	8.70E-09	NQ	CAA	0.1	1.50E-08	-0.24
65	1MSE	ST	CAA	0.54	8.70E-09	NQ	CAA	0.1	5.40E-08	-0.79
66	1MSE	ST	CAA	0.54	8.70E-09	Mc	CAA	0	3.70E-08	-0.63
67	1MSE	ST	CAA	0.54	8.70E-09	Mc	CAA	0	1.00E-06	-2.06
68	1MSE	ST	CAA	0.54	8.70E-09	K	CAA	0.02	1.00E-06	-2.06
69	1MSE	ST	CAA	0.54	8.70E-09	R	CAA	0.02	1.00E-06	-2.06
70	1MSE	ST	CAA	0.54	8.70E-09	DE	CAA	0.04	1.00E-06	-2.06
71	1MSE	ST	CCA	0.05	3.70E-09	NQ	CCA	0.04	3.70E-08	-1
72	1MSE	ST	CCA	0.05	3.70E-09	NQ	CCA	0.04	3.90E-08	-1.02
73	1MSE	ST	CCA	0.05	3.70E-09	K	CCA	0	2.20E-08	-0.77
74	1MSE	ST	CCA	0.05	3.70E-09	R	CCA	0	2.20E-08	-0.77
75	1PUF	R	TAC	0	1.80E-07	ST	TAC	0.18	1.10E-07	0.21
76	1PUF	R	GAC	0	1.80E-06	ST	GAC	0.12	4.60E-07	0.59
77	1RUN	DE	CGC	0	1.10E-09	Mc	CGC	0	5.60E-10	0.29

78	1RUN	DE	CAC	0.97	1.40E-11	Mc	CAC	0	1.20E-10	-0.93
79	1RUN	DE	CAC	0.97	1.40E-11	DE	CGC	0	1.10E-09	-1.9
80	1TN9	Y	TAC	0	4.30E-08	Mc	TAC	0	5.87E-07	-1.14
81	1TN9	R	ATT	0	4.30E-08	Mc	ATT	0.87	3.32E-07	-0.89
82	1TN9	K	AGT	0.43	4.30E-08	Mc	AGT	0.08	4.57E-07	-1.03
83	1TN9	R	TAG	0.03	4.30E-08	Mc	TAG	0	9.50E-08	-0.34

The Pearson's Correlation coefficient for the correlation between $\log(\text{fold change})$ and the prediction accuracy is -0.18, while Spearman's Correlation coefficient is -0.17. This suggests that the accuracy of the method is independent of the difference in the binding affinities of the wild-type and mutant proteins. The distributions for the correct and incorrect predictions have a KS test statistic of 0.4 and a p-value of 0.87, indicating that the distributions are similar (Figure 3.5).

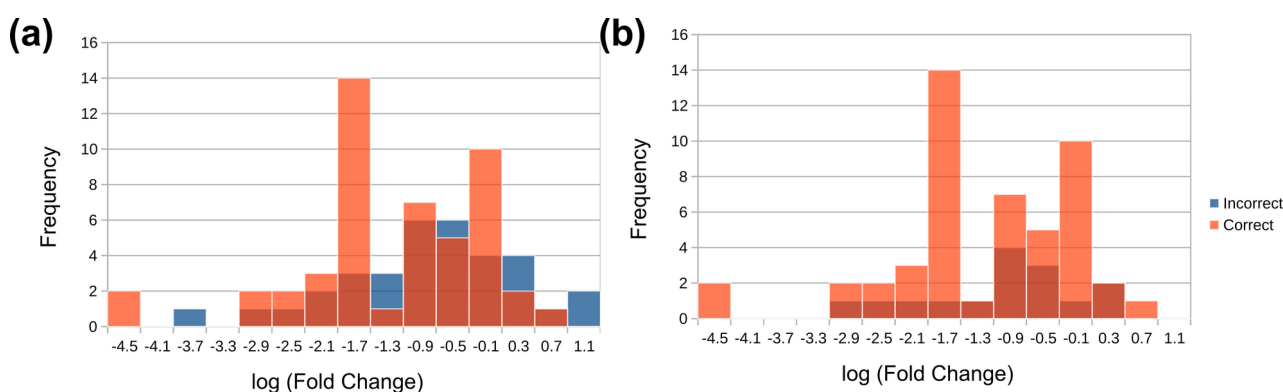


Figure 3.5: Distribution of correct and incorrect predictions of binding affinity change.

(a) Overall distribution for the whole dataset. (b) Distribution after removing the entries that did not have trinucleotide-amino acid data.

The similar distributions indicate that the magnitude of change in the binding affinity is not related to the degree of correctness of our predictions.

3.4 Validation using Molecular Dynamics Simulations

By studying 3D structures of DNA-protein complexes, we have probabilistic preferences that predict the specific binding of DNA and proteins. To validate our method, we used the predicted interaction propensities to change the specificity of a DNA-binding protein from a known consensus sequence to a sequence of our choice (non-consensus).

We predicted the mutation that needed to be made in the protein to make it recognize the query sequence. A model of the mutant protein was generated, and we assessed the binding of the wild-type and mutant proteins with the consensus and non-consensus DNA (Figure 3.6).

The binding assessment was carried out using Molecular Dynamics (MD) simulations discussed in this section and *in-vitro* assays discussed in chapter 4. We selected four protein systems for this analysis- i) Myc, ii) Ets1, iii) EcoRV and iv) Zif-268 as a control. The selected proteins have different architectures, such as α duplicates (Myc), α arrays (Ets1), α/β three-layered sandwiches (EcoRV) and β - β - α zinc fingers (Zif-268). We made a single change in the consensus sequence in the case of the monomer (s1) and two changes in the case of dimers (Myc and EcoRV) since both units bind to half the consensus (Table 3.4, Figure 3.7). For Zif-268 the crystal structures of mutant proteins were already present.

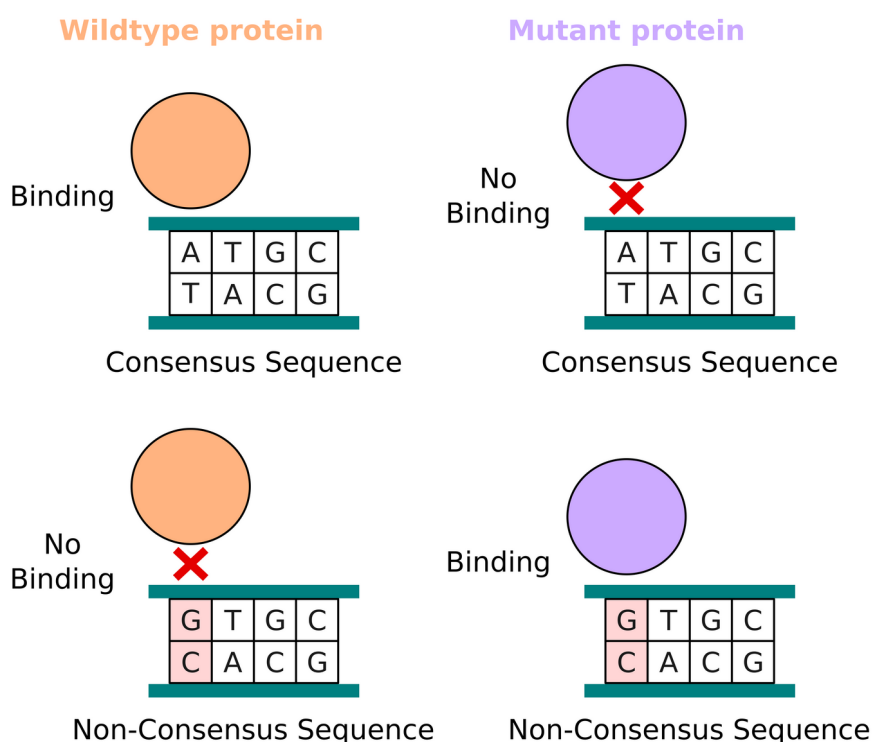


Figure 3.6: Schematic for changing the specificity of a DNA-binding protein.

The consensus sequence binds specifically to the wild-type protein (orange circle). The query sequence binds only the mutated protein (purple circle). The first nucleotide A is changed to G in this example and highlighted with pink.

Table 3.4: Prediction of specificity change in selected DNA-binding proteins.

The nucleotide/(s) changed are highlighted in bold.

Sr. No	Protein	Consensus	Non-Consensus	Predicted protein Mutant	Bonded Nucleotide
1	Myc	CAC GTG	CAT ATG	Arg(20.A)->Gln	CG->TA
2	Ets1	GGA[A/T]	GGTT	Tyr(395.C)->Arg	A->T
3	EcoRV	GATATC	G CTAGC	Ser(183.A)-> Gln	A->C

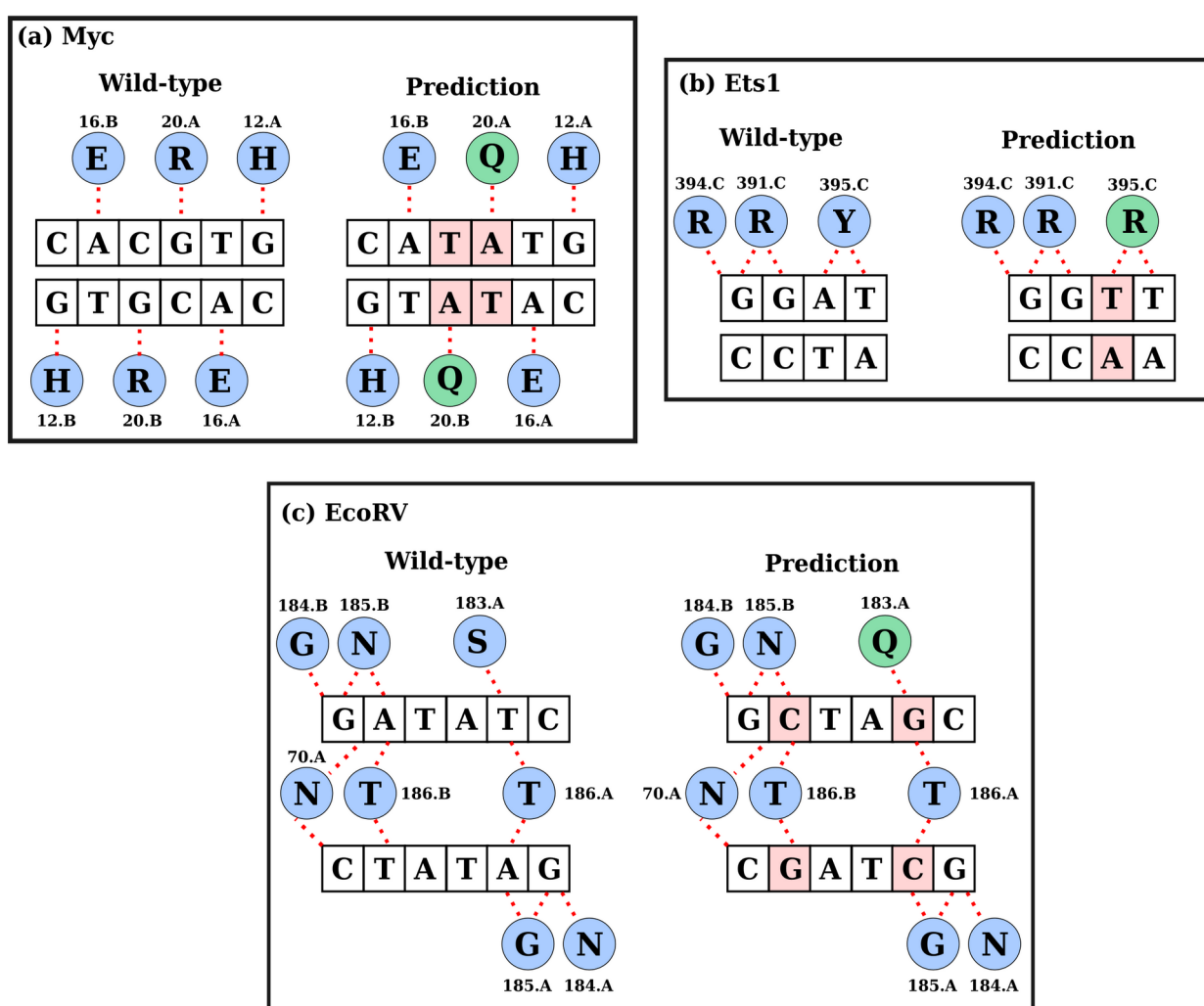


Figure 3.7: Systems for validating the change in specificity of DNA-binding proteins.

The nucleotides are shown in boxes with single-letter codes. Amino acids are represented in single-letter code within circles. The wild-type amino acids are in blue and the mutant are in green. The hydrogen bonds between the nucleotides and amino acids are shown as red dotted lines. The changed nucleotides in the non-consensus sequence are shown pink.

We carried out the MD simulations using NAMD2.14 [149] with the CHARMM [143,144] forcefield for all these systems. We used the CHARMM-GUI webserver to generate the parameter and topology files for the simulation [232,233]. To generate the mutant DNA-protein complex, we introduced mutations in the DNA using the swapna module of chimera [217]. We used CHARMM GUI to introduce protein mutations. We performed simulations for four interaction types- consensus DNA with the wild-type protein, non-consensus DNA with predicted mutant protein, consensus DNA with the predicted mutant protein and the non-consensus DNA with wild-type protein.

We used files from the CHARMM GUI server and generated a neutral charge system with TIP3 water molecules using VMD [234]. We generated a water box to fit the dimensions of the protein using the default settings of the VMD solvate plugin. 10° rotation increments were applied to the protein to minimize the water molecules in the system. Therefore the box size and shape differed in the different protein systems. The system was neutralized with Na⁺ and Cl⁻ counter ions using the ionize plugin (Refer to Appendix 10 for the system sizes for different DNA protein complexes).

We minimized this neutral system for 100 steps using the default conjugate gradient technique, with each step corresponding to 2 fs (femtoseconds) using NAMD2.14 and CHARMM36 m forcefield. The minimized system was equilibrated for 50000 steps or 100 ps. The equilibration process involved thermal equilibration to 300 K and pressure equilibration to 1 atm pressure. We used periodic boundary conditions and Particle Mesh Ewald (PME) full electrostatics with the Lennard Jones cut-off distance as 12 Å for the NPT ensemble. The production run of 50 ns was performed for in triplicates for all simulations.

The RMSD and RMSF of the complexes were calculated using the formulae-

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (x_i - x_i^{ref})^2}{N}} \quad \text{Eq. 6}$$

$$RMSF = \sqrt{\frac{\sum_{i=1}^t (x_i - x_i^{ref})^2}{t}} \quad \text{Eq. 7}$$

Where, x_i and x_i^{ref} are the coordinates of the MD snapshot at a particular time and the reference structure, respectively. N and t correspond to the total number of residues and the time of the simulation respectively. Therefore, RMSD is the deviation across time and calculated at each time point, and RMSF is the deviation in each residue for the course of the simulation.

To calculate the binding free energy (refer to Chapter 1, section 1.6.3), we followed implicit solvation protocol [162]. We solvated the simulated protein alone, DNA alone and the complex in an implicit solvent. We obtained the total energy of all three of these systems at every time point. The binding free energy was calculated as-

$$\Delta G_{bind} = G_{total}(\text{protein+DNA}) - G_{total}(\text{protein}) - G_{total}(\text{DNA}) \quad \text{Eq. 8}$$

We took the last 2 ns of the trajectory for comparison of binding energy across different complexes. The binding energy values for the different systems were compared to the wild-type crystal structure value using the one-way ANOVA test to check for statistical significance [224]. The RMSD, RMSF, distance between the DNA and proteins and the DNA-protein hydrogen bonds were also compared between the DNA-protein complexes using ANOVA. The p-values reported in the following sections are with respect to the respective wild-type DNA-protein crystal structures for all systems.

3.4.1 Zif-268

Zif-268 is a mouse transcription factor associated with early development. It belongs to the zinc finger protein family, which has been extensively studied and there exists a specificity code for it. The zinc finger protein consists of an α and β sheet, co-ordinated by a Zn^{+2} ion [235,236].

Each finger in a zinc finger protein recognizes a four base-pair sequence. Multiple such fingers together recognize the entire recognition sequence. Since our objective in the validation process was to compare the binding of mutant and wild-type protein to different substrates, we used this system as a control. The crystal structures of different Zif-268 variants have been solved with the DNA that they specifically recognize [237]. Comparing

MD results of binding between these proteins would provide a baseline for judging our predictions. The wild-type protein has a peptide RDER, the amino acids in which correspond to positions -1, +2, +3 and +6 and recognizes the GCGT site. The first mutant (MT1) has a peptide DSNR that specifically recognizes the GACC site. While the second mutant (MT2) has the most affinity toward GCAC site when the peptide sequence is QGSR. To simulate the scenario where we generate models of the mutant protein and check the binding using simulations, we also generated the models of the proteins MT1 and MT2. To assess the effect of errors introduced by modelling, we also modelled the wild-type structure from MT2 protein.

The analyses of the MD simulation results involved plotting the average RMSD (calculated from the average structure), RMSF, the distance between the centroids of the protein and DNA, total DNA-protein hydrogen bond frequency, nucleotide-specific hydrogen bond frequency and the binding free energy of the complexes. The average values for these are available in Table 3.5 for all DNA-protein complexes and detailed run-wise analysis is in Appendix sections 11-18. The average RMSF and RMSD values would have similar trends. Therefore, we have reported the values for average RMSD in Appendix 11.

We observed that MT1 (2.9 Å) and MT2 (3 Å) crystal structures had a higher RMSD compared to the wild-type Zif-268 protein (1.7 Å) (p-value <0.00001 for both, Figure 3.8 (a)). However, both MT1 and MT2 had lower RMSD (1.2 Å for both) when the models were used instead (p-value <0.00001 for both, Figure 3.9 (a)). The lowest RMSD was observed for Zif-268 MT1 protein -wild-type DNA complex (1.15 +/- 0.3 Å, Table 3.5).

The RMSF values of the wild-type Zif-268 protein with the wild-type, MT1 and MT2 DNA were comparable with the range 1.4 to 1.7 Å (p-values 0.98 and 0.57, respectively, Figure 3.8 (b)). The MT1 and MT2 DNA-protein complexes had higher RMSF values compared to the wild-type crystal structure complex with an average RMSF of 2.3 to 2.6 Å (p-values 0.02 and 0.0007). The MT1 protein in complex with wild-type DNA also differed from the wild-type complex, but with the wild-type DNA, it has a lower RMSF (1.1 Å) (p-value 0.004). The MT1 and MT2 models also had a significantly lower RMSF than the wild-

type DNA-protein complex (1.2 Å) (p-values 0.0002 and 0.0001, respectively Figure 3.9(b)). The wild-type model and crystal structure however had similar RMSF profiles (~1.4 Å, p-value 0.18). The RMSD values of the wild-type protein with the MT1 and MT2 substrates is lower (1.3-1.4 Å) than with its consensus substrate (1.7 Å) (p-values 0.00002 and 0.01, respectively).

The distance between the DNA and protein may indicate if the complex is getting dissociated. However, since we had a very short simulation time, we did not expect to see any such behaviour. Yet, the MT1 protein with the wild-type DNA shows a 30% increase in the distance between the protein and DNA by the end of 50 ns (Figure 3.8 (c)). The smallest distance between the DNA and protein is with the MT2 crystal structure complex with its cognate DNA (3.2 Å). The crystal structures of MT1 and MT2 proteins with their respective DNA has a similar centroid distance compared to the wild-type crystal structure, which are ~1.1 to 1.2 Å (p-values 0.2 and 0.6). However, the models of wild-type, MT1 and MT2 proteins have a higher distance between the DNA and protein compared to the wild-type protein (distance range: 3.5 -3.8 Å, p-values <0.00001 for all, Figure 3.9 (c)). The wild-type protein in complex with MT1 DNA has a lower distance (2.5 Å) while in complex with MT2 DNA has a higher distance (3.6 Å) between the two macromolecules when compared to wild-type DNA-protein complex (3.1 Å) (p-values <0.00001 for both).

Hydrogen bonds are one of the interactions that stabilize DNA-protein binding. We looked at all DNA-protein hydrogen bonds (Figures 3.8 and 3.9 (d)) and the hydrogen bonds between the nucleotide base and amino acids (Figures 3.8 and 3.9 (e)). Our hypothesis was that while other interactions enhance binding between DNA and protein in a general manner, specificity is provided by the bonds with the nucleotide bases. Over the course of the simulation, we observed a reduction of ~30% in total and ~23% in specific hydrogen bonds in all the DNA-protein complexes.

However, the wild-type protein complex with MT1 DNA seemed to regain the lost specific bonds towards the end of the simulation. It also had the largest number of total hydrogen bonds (47 bonds, p-value <0.00001). The largest number of specific bonds are formed between the model of MT2 protein and its preferred DNA (17 bonds, p-value

<0.00001). MT1 model and crystal structure have higher number of total (47 and 40, respectively) and specific hydrogen bonds (16 and 14, respectively) compared to the wild-type with 37 total bonds and 13 specific bonds (p-values for all <0.00001). The wild-type protein in complex with MT1 (47 bonds) and MT2 DNA (44 bonds) have a higher number of total bonds than the wild-type DNA-protein complex (37). A similar trend is seen for nucleotide-specific bonds of wild-type protein with MT1 and MT2 DNA.

The binding energy evaluation showed that the best binding energy was with the MT2 DNA-protein complex (-96 kcal/mol, p-value <0.00001, Figure 3.8 (f)). The model of MT2 DNA also had a better binding energy (-78.9 kcal/mol) than WT (-66.5 kcal/mol) and MT1 models (-66.1 kcal/mol) (p-value <0.00001, Figure 3.9 (f)). The wild-type protein had the best binding energy with the MT1 substrate (-83.5 kcal/mol) while the MT1 protein had the best binding energy for the wild-type substrate (-70.7 kcal/mol) (p-value <0.00001 for both). This result directly contradicts the experimental study, where they performed competitive assays using the wild-type DNA and the mutant proteins had a better binding to the reported sequence [237].

The wild-type DNA protein model had a better binding energy (-66.5 kcal/mol) compared to the crystal structure (-62.4 kcal/mol) (p-value <0.00001). While the MT1 model had a better energy (-61.1 kcal/mol), the MT2 model had a worse energy compared to their respective crystal structures (-61.9 for MT1 and -96.6 kcal/mol for MT2) (p-value <0.00001 for both). However, the trend between the three proteins is maintained where MT2 has the best binding followed by the wild-type and MT1. In our experiment, we used the wild-type crystal structure and models of the mutant. The results from the Zif-268 mutants show that they have a significantly increased (MT2) or reduced (MT2) binding energy depending on the protein. Thus, modelling errors could affect the estimate of the energy units.

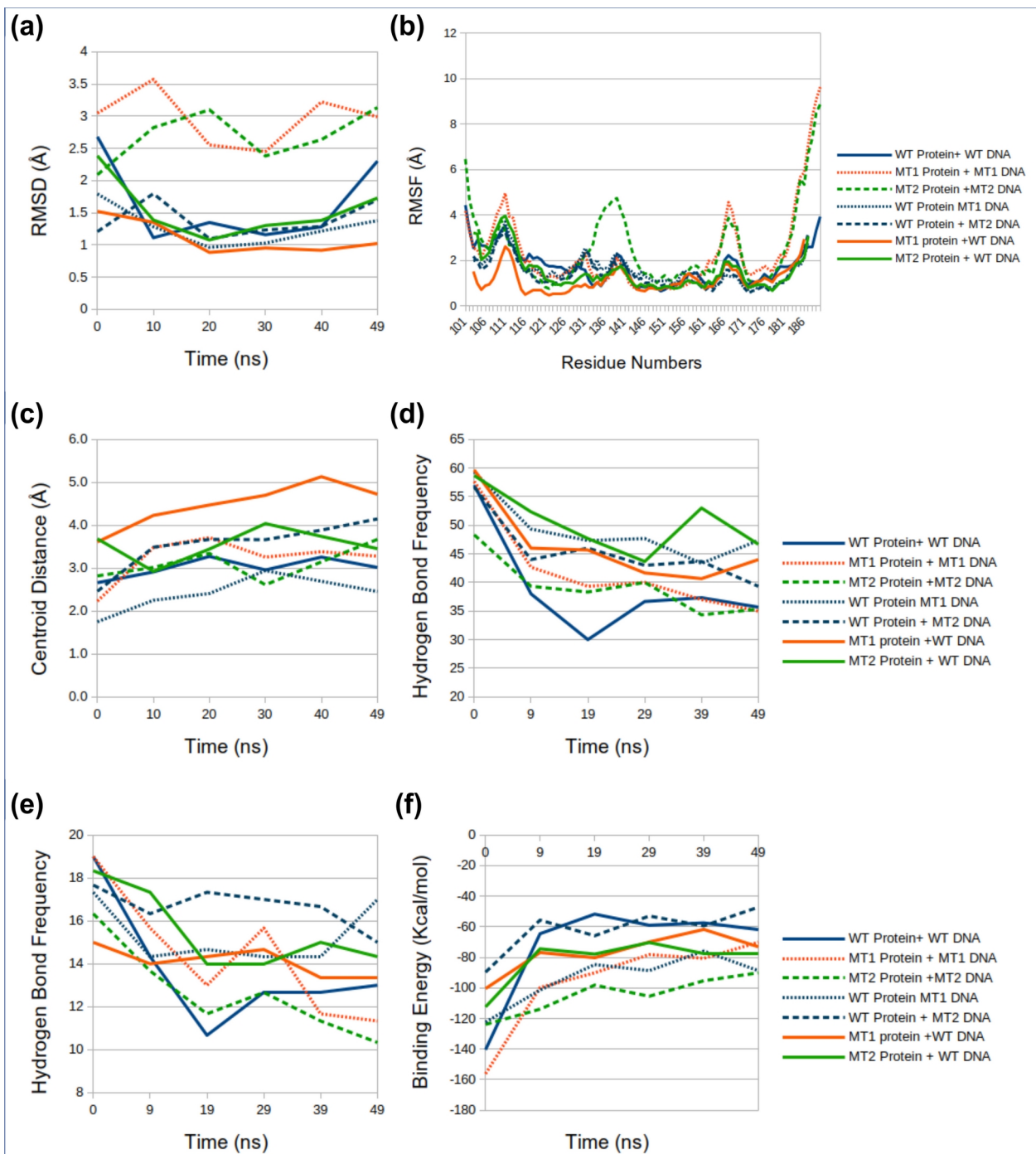


Figure 3.8: MD results of Zif-268 wild-type and mutant proteins

-depicting RMSD (a), RMSF (b), Centroid distance between DNA and protein (c), Total hydrogen bonds (d), specific hydrogen bonds with nucleotide base (e) and Binding energy (f). The wild-type (WT), MT1 and MT2 proteins are shown in blue, orange and green respectively. The WT, MT1 and MT2 DNA are depicted in solid, dotted and dashed lines respectively.

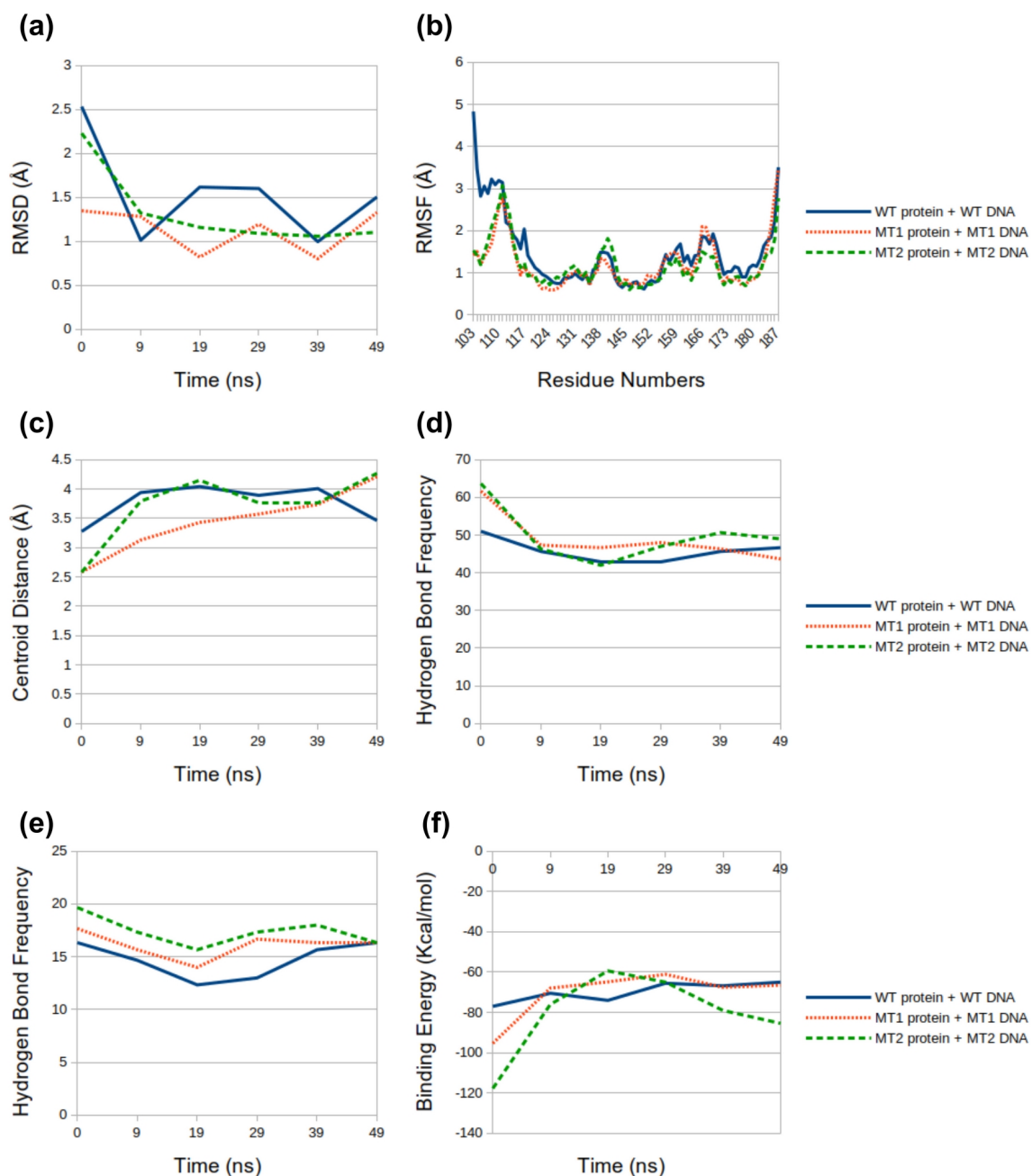


Figure 3.9: MD results of Zif-268 wild-type and mutant protein models

-depicting RMSD (a), RMSF (b), Centroid distance between DNA and protein (c), Total hydrogen bonds (d), specific hydrogen bonds with nucleotide base (e) and Binding energy (f). The wild-type (WT), MT1 and MT2 protein models are shown in blue, orange and green, respectively. The WT, MT1 and MT2 DNA are depicted in solid, dotted and dashed lines, respectively.

3.4.2 Myc

The Myc protein family consists of c-Myc, l-Myc and n-Myc. These are transcription factors that are encoded by regulatory genes and proto-oncogenes [238]. They are present in different tissues at different stages of the life cycle. c-Myc dysregulation is most frequently associated with cancer [239]. Due to its relevance in cancer, it has been extensively studied and therefore we chose it as one of the proteins to validate our method [240,241].

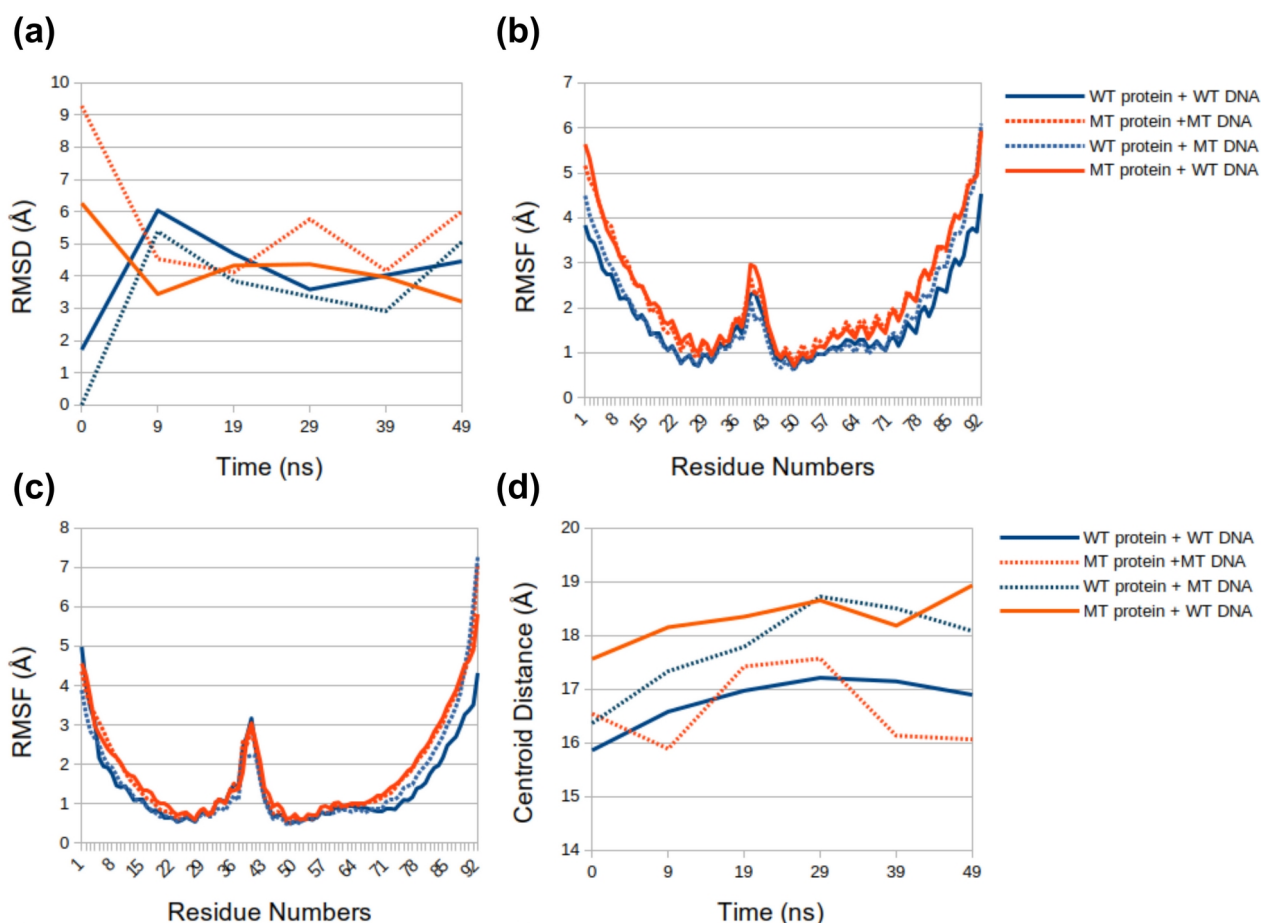
The Myc protein family has a basic-helix-loop-helix-leucine zipper structure. There is a variant of the Myc protein called Omo-Myc, which inhibits the tumorigenic activity of Myc while still promoting normal cell proliferation. This variant is dominant-negative, which means that it overrides the oncogenic protein's activity and holds therapeutic potential. We have tried to study the specificity of the Omo-Myc protein (PDB ID 5I50) [242]. The consensus sequence of Myc is the E-box sequence- CACGTG [243,244]. Since Myc binds as a dimer, each unit recognizes the 3 nucleotide half-site. Therefore, to change the third nucleotide, we also needed to change the fourth nucleotide. We mutated the central CG to TA to change cytosine at position 3 to thymine. For recognizing this new sequence (non-consensus) specifically, we predicted the arginine to glutamine mutation in the protein.

We analysed the results of this prediction by plotting the average RMSD (calculated from the average structure), RMSF, the distance between the centroids of the protein and DNA, total DNA-protein hydrogen bond frequency, hydrogen bond frequency between the nucleotide base and amino acids and the binding free energy of the complexes.

The mutant protein with wild-type DNA has a lower RMSD (3.8 Å) compared to the wild-type DNA-protein complex (4.5 Å) (p-value 0.04). The RMSF of the wild-type DNA-protein complex is significantly lower (1.5 Å) than that of the mutant protein in complex with non-consensus (1.96 Å) and consensus DNA (1.99 Å) (p-values 0.03 and 0.01, respectively, Figures 3.10 (a),(b) and (c)). The lowest RMSD is observed for Myc mutant protein in complex with wild-type DNA (3.86 Å).

The distances between the mutant protein with the wild-type DNA (17.9 Å) and the wild-type protein-non-consensus DNA complexes (18.5 Å) are higher than the wild-type DNA-protein complex (16.8 Å) (Figure 3.10 (d)). The wild-type protein-consensus DNA and the mutant protein-non-consensus DNA complexes have an average distance of ~1.6 Å between the DNA and protein (p-value 0.95).

The wild-type DNA-protein complex had the maximum total hydrogen bonds (43 bonds, Figure 3.10 (e)). When we considered only the hydrogen bonds with the nucleotide base, the wild-type protein had 7 and 6 nucleotide-specific bonds with the non-consensus and consensus DNA, respectively (p-value 0.0002, Figure 3.10 (f)).



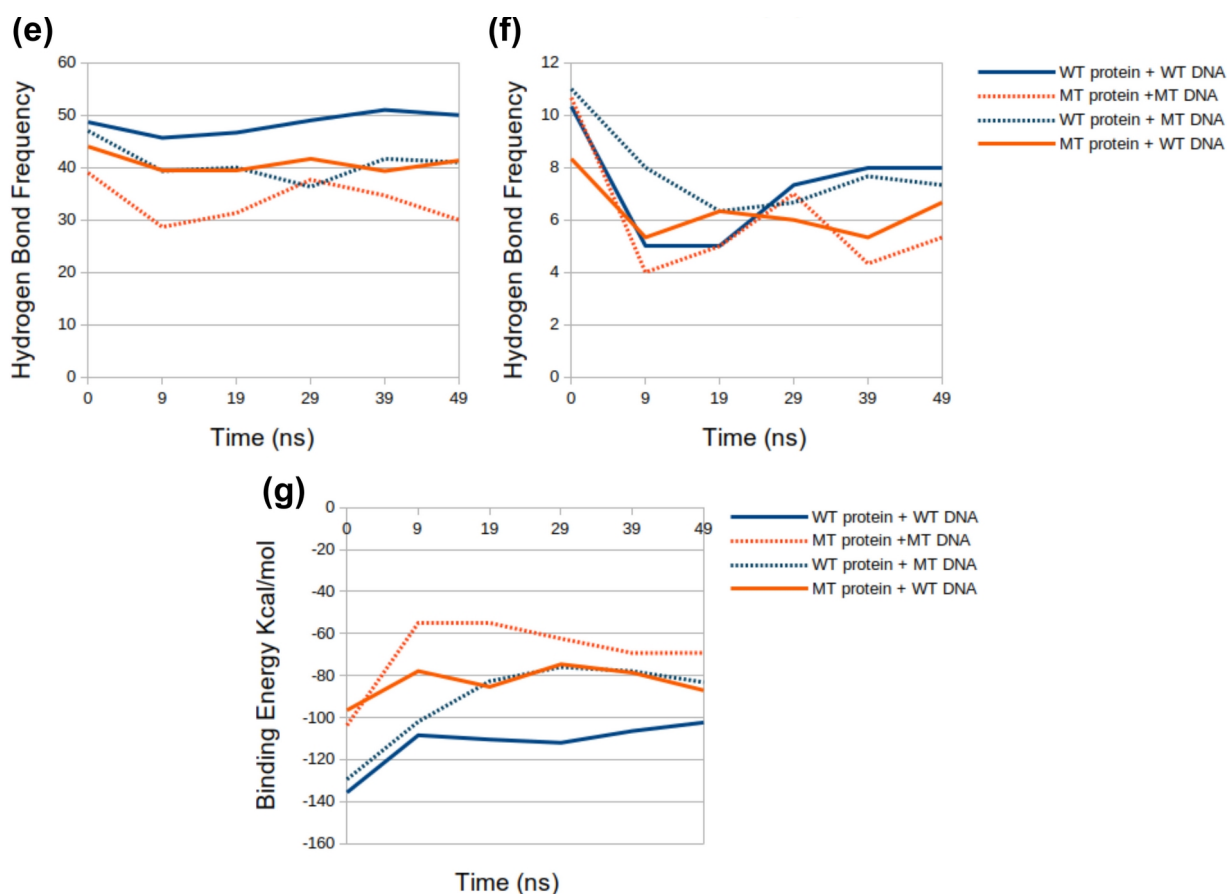


Figure 3.10: MD results of Myc wild-type and mutant proteins

-depicting RMSD (a), RMSF (chain A) (b), RMSF (chain B) (c), Centroid distance between DNA and protein (d), Total hydrogen bonds (e), Total hydrogen bonds with nucleotide base (f) and Binding energy (g). The wild-type (WT) and mutant (MT) proteins are shown in blue and orange, respectively. The WT and MT DNA are depicted in solid and dotted lines, respectively.

The binding energy of the wild-type DNA-protein complex was -107.6 kcal/mol, which was better than all other complexes (p-value <0.00001, Figure 3.10 (g)). The mutant DNA-protein complex had a binding energy of -65.9 kcal/mol, which was the largest among all complexes. The binding energy of the wild-type protein-mutant DNA and mutant-protein-wild-type DNA complexes was comparable with -82 to -83 kcal/mol (p-value 0.38). Among the wild-type and mutant DNA substrates, the Myc mutant protein had a better binding energy toward the wild-type DNA.

3.4.3 Ets1

c-Ets1 is a transcription factor that regulates the expression of cytokines and chemokines encoded by the Ets1 gene. The amino acid sequence of this protein is highly conserved across species indicating its crucial role [245]. The Ets1 protein family is defined by its characteristic DNA-binding domain, a winged helix-turn-helix motif. Due to its function in immune modulation, it is also a proto-oncogene [246].

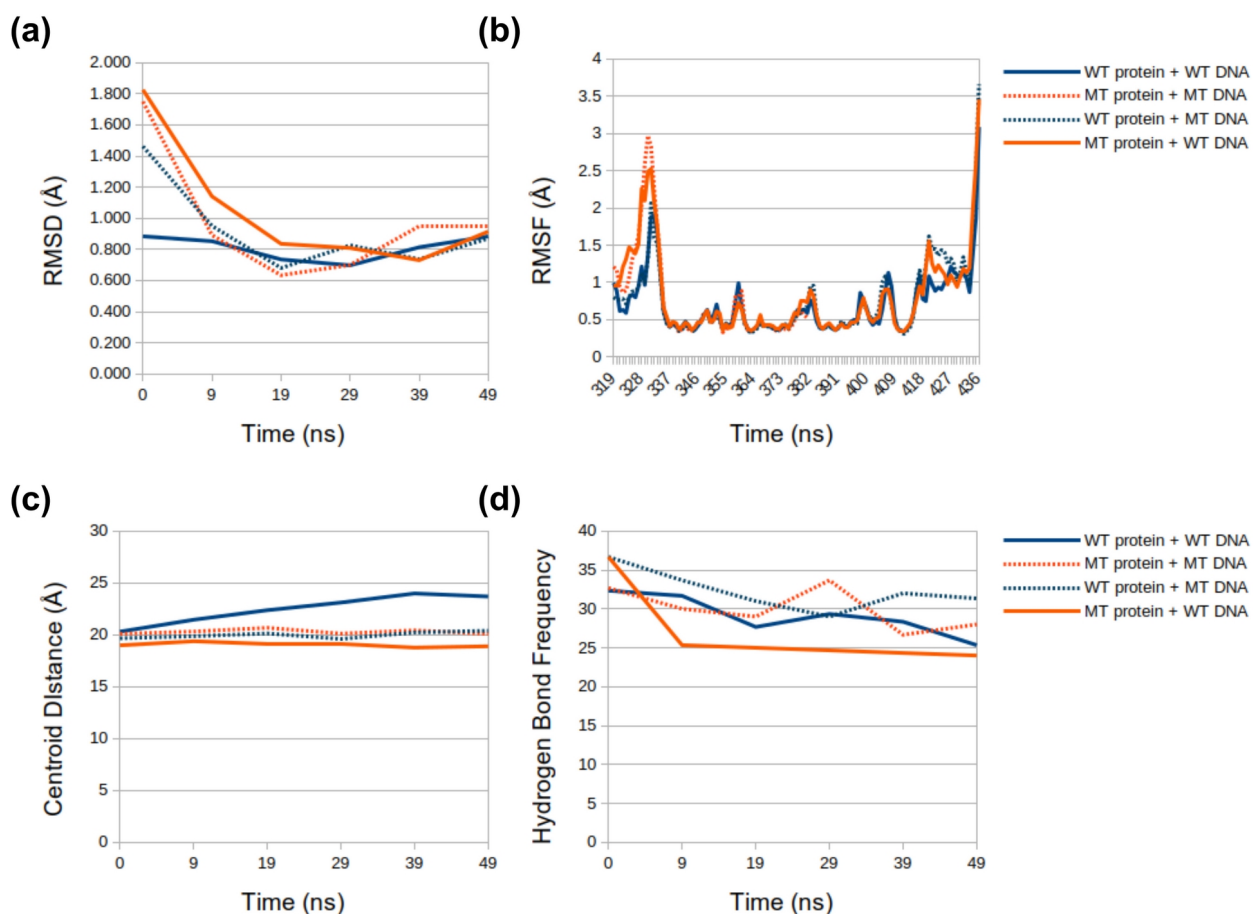
The winged helix-turn-helix domain consists of three α -helices and four β -strands. It recognizes the consensus sequence GGA[A/T] [247]. While Ets1 regulates gene expression individually, other transcription factor complexes also assist Ets1. We took the structure of Ets1 from PDB ID: 3WTY, where Ets1, Runx1, and CBF β together bind the TCR α enhancer [248]. Ets1 binds as a monomer and binds to the GGAT DNA sequence in this structure. We changed the adenine to thymine, so the non-consensus sequence is GGTT. As per our specificity preferences, the tyrosine, which was interacting with adenine, should be mutated to arginine for specific binding to the non-consensus sequence.

To analyse the relative specificity of the predicted DNA-protein complex, we simulated it along with the complexes of wild-type DNA-protein, wild-type protein-non-consensus sequence and mutant protein-wild-type DNA. We plotted the average RMSD (calculated from the average structure), RMSF, the distance between the centroids of the protein and DNA, total DNA-protein hydrogen bond frequency, hydrogen bond frequency between the nucleotide base and amino acids and the binding free energy of the complexes.

The average RMSD of the wild-type protein with the consensus (0.75 Å) and non-consensus DNA (0.84 Å) is similar (p-value 0.13) but differs from mutant protein with consensus (0.93 Å) and non-consensus DNA (0.87 Å) (p-values 0.001 and 0.03, respectively) (Figure 3.11 (a)). However, the RMSF values are similar for all four systems in the range 0.7 to 0.8 Å (Figure 3.11 (b)).

The centroid distance between DNA and protein 22.5 Å for the wild-type DNA-protein complex, significantly larger than other complexes (p-value <0.00001). The shorted DNA-protein centroid distance is 19.1 Å between the mutant protein and consensus DNA

(Figure 3.11 (c)). The overall hydrogen bonds have a similar frequency for wild-type and mutant DNA-protein complexes with ~30 bonds (p value 0.63). However, the Ets1 wild-type protein has ~32 hydrogen bonds with the non-consensus substrate, which is the largest among all complexes (p-value 0.0001) (Figure 3.11 (d)). The wild-type protein with consensus (7.5) and non-consensus DNA (7.7) had the highest and similar number of nucleotide-specific hydrogen bonds (p-value 0.39). Despite this, the binding energy of the wild-type protein is maximum with the non-consensus DNA at 03.7Kcal/mol (p-value <0.00001, Figure 3.11 (e)). The mutant protein has a better binding energy with the predicted non-consensus DNA (-66.2 kcal/mol) as compared to the wild-type DNA (62.7 kcal/mol) (p-value <0.00001).



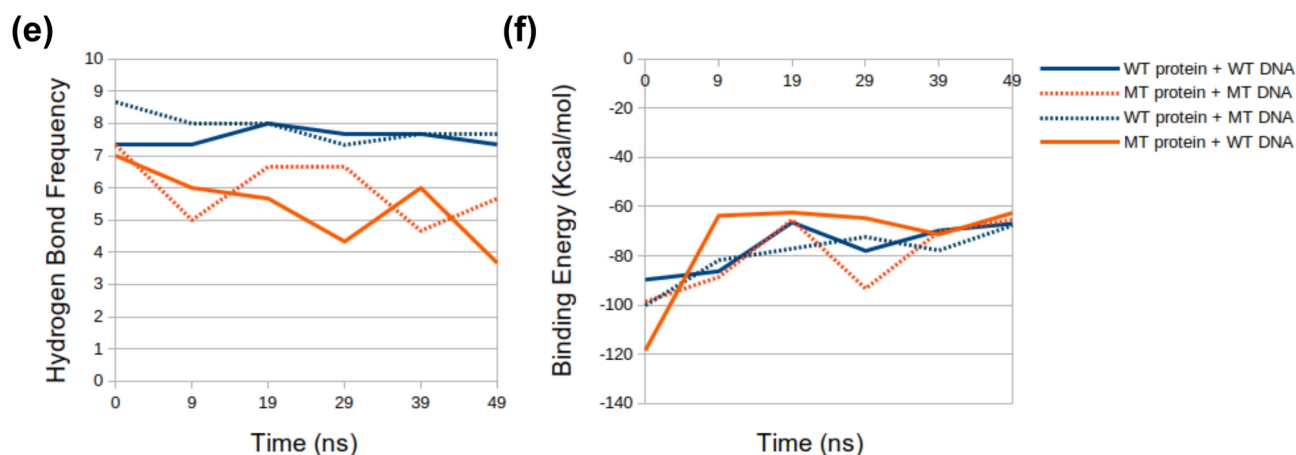


Figure 3.11: MD results of *Ets1* wild-type and mutant proteins

-depicting RMSD (a), RMSF (b), Centroid distance between DNA and protein (c), Total hydrogen bonds (d), Total hydrogen bonds with nucleotide base (e) and Binding energy (f). The wild-type (WT) and mutant (MT) proteins are shown in blue and orange, respectively. The WT and MT DNA are depicted in solid and dotted lines, respectively.

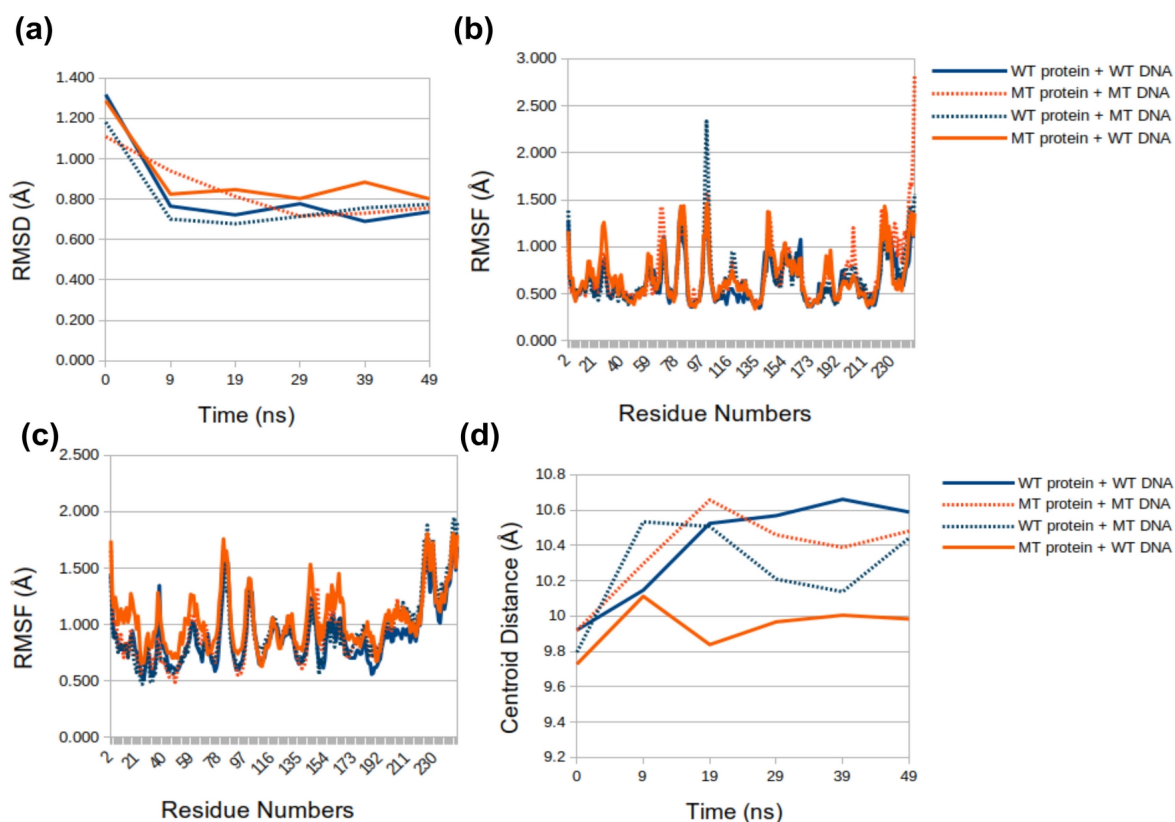
3.4.4 EcoRV

EcoRV is a restriction enzyme from *Escherichia coli* that is produced by the bacteria to protect itself against foreign DNA. The enzyme recognizes the GATATC sequence and makes a blunt cut at the central TA (GAT/ATC). The bacteria protect the self genome from cleavage by methylating the first adenine in the recognition sequence. The methylated adenine-containing consensus sequence is not cleaved by EcoRV [249,250].

EcoRV has a mixed α/β structure that recognizes the 6-nucleotide cognate sequence as a dimer in the presence of Mg^{+2} ions [251]. The enzyme-bound DNA conformation is bent, and it is hypothesized that the substrate recognition is dependent on direct and indirect readout [252]. Since restriction enzymes have more stringent sequence recognition than transcription factors, we chose this system for validation. We changed the central A to C, and since the recognition site is a palindrome, the non-consensus sequence generated was GCTAGC. The thymine was involved in a hydrogen bond interaction with a serine (residue number 183). Our method predicted that a serine to glutamine mutation

would allow recognition of GCTAGC, while minimizing the interaction with the consensus sequence.

To validate the prediction, we performed MD simulations of the wild-type DNA-protein, the mutant protein-non-consensus DNA, wild-type protein-non-consensus DNA and mutant protein-wild-type DNA complexes. We analysed the results using some measure such as the average RMSD (calculated from the average structure), RMSF, the distance between the centroids of the protein and DNA, total DNA-protein hydrogen bond frequency, hydrogen bond frequency between the nucleotide base and amino acids and the binding free energy of the complexes. The RMSD, RMSF and the DNA-protein centroid distance of the wild-type DNA-protein complex was different from the mutant protein-wild-type DNA complex (p-values 0.001, <0.00001 and <0.00001, respectively, Figure 3.12).



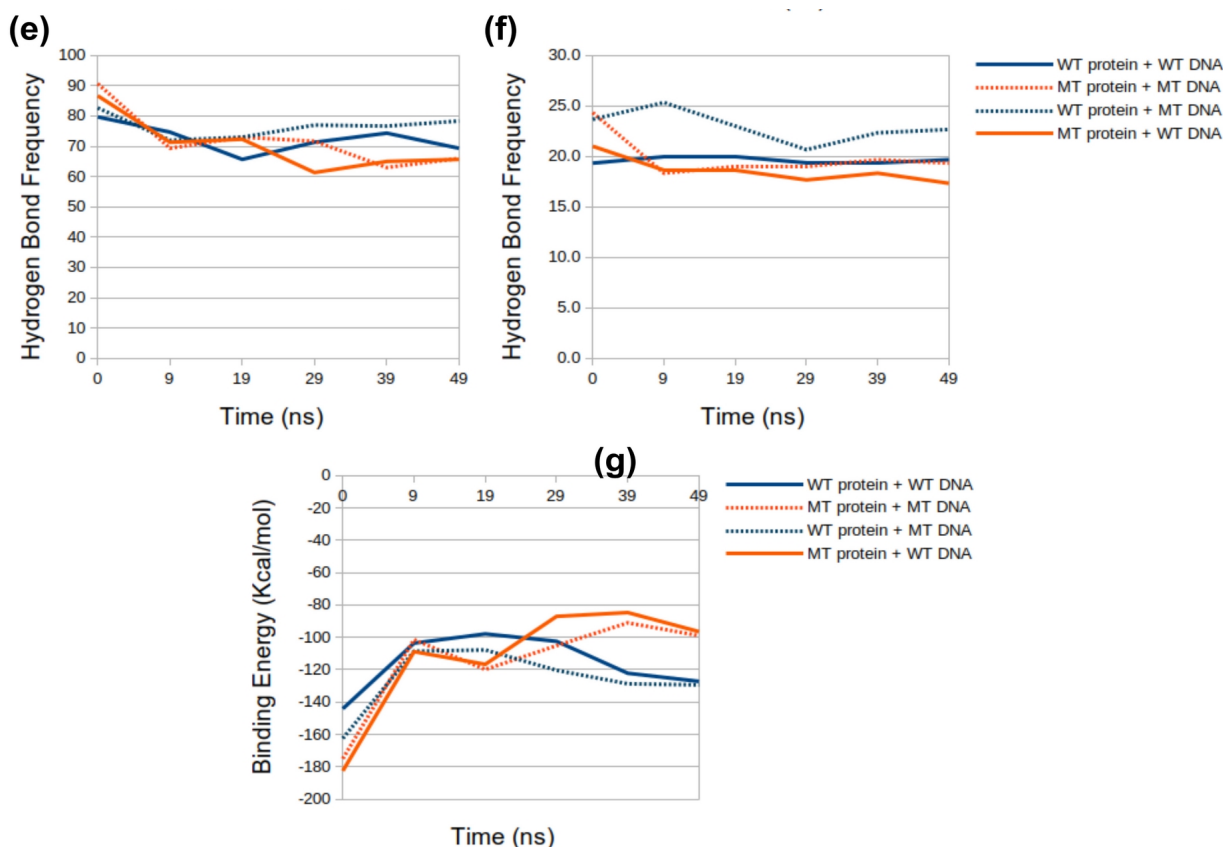


Figure 3.12: MD results of EcoRV wild-type and mutant proteins

-depicting RMSD (a), RMSF (chain A) (b), RMSF (chain B) (c), Centroid distance between DNA and protein (d), Total hydrogen bonds (e), Total hydrogen bonds with nucleotide base (f) and Binding energy (g). The wild-type (WT) and mutant (MT) proteins are shown in blue and orange, respectively. The WT and MT DNA are in solid and dotted lines, respectively.

None of the other complexes showed a difference in RMSD or RMSF with the averages in the range of 0.7 to 0.8 Å. The EcoRV mutant protein had 10 Å between DNA and protein molecules, which was significantly lower than the other systems (p-value <0.00001). The total and nucleotide-specific hydrogen bonds were the highest in the complex of wild-type protein with the non-consensus DNA with 74 and 22 bonds, respectively (p-values 0.0047 and <0.00001, respectively). However, the binding energy was comparable between the wild-type protein interaction with mutant DNA with binding energies, -122.5 and -123.7 kcal/mol, respectively. (p-value 0.24). The EcoRV mutant protein also had a similar binding energy profile with the wild-type consensus and the predicted non-consensus DNA with -103.4 and -10.4 kcal/mol (p-value 0.37).

3.4.5 Summary of validation with MD simulations

We performed MD simulations for Zif-268, Myc, Ets1 and EcoRV to validate the procedure of changing the specificity of a DNA-binding protein. We compared the RMSD, distance between the centroids, overall and specific hydrogen bonds and the binding energy to analyse the relative binding of the wild-type and mutant proteins to different DNA substrates (Table 3.5, Appendix 11-15).

Zif-268 MT2 protein-DNA complex had the best binding among all Zif-268 complexes at -96.6 kcal/mol. However, it did not have the best values for total and nucleotide-specific hydrogen bonds. Therefore, the contribution from non-hydrogen bond interactions may be responsible for deciding the binding energy. The wild-type Zif-268 protein had better binding to MT1 DNA (-83.5 kcal/mol) than to its consensus (-62.4 kcal/mol) (p-value <0.00001). The binding energy of the MT1 protein was greater for the wild-type DNA (-70.7 kcal/mol) than the experimentally determined substrate (-619 kcal/mol) (p-value <0.00001).

Since the Zif-268 mutants were crystallized and we had modelled the mutations for Myc, Ets1 and EcoRV, we modelled the Zif-268 mutants as a control. We compared the binding energy of models of the Zif-268 wild-type, MT1 and MT2 DNA-protein complexes with each other and to the corresponding crystal structures. The modelled wild-type and MT2 complexes had a statistically different binding energy profile compared to their crystal structure counterparts (p-value <0.00001). While the binding energy improved for the wild-type after modelling it from -62.4 to -66.5 kcal/mol, it became worse for MT2 from -96.6 to -78.9 kcal/mol. However, MT1 model did not have a difference in binding energy compared to its crystal structure with average values of -61.0 and -61.9, respectively (p-value 0.48) or the crystal structure of the wild-type protein (-62.4) (p-value 0.06). Thus the simulation performed by modelling of the mutation may not correspond to the actual specificity as the model accuracy may influence the outcome.

The wild-type Myc DNA-protein complex was the best binding system for this protein with an energy of -107.6 kcal/mol. The total number of hydrogen bonds between the DNA and protein were also highest for Myc wild-type DNA-protein complex (47). The

Myc mutant protein had a better binding energy with the wild-type DNA (-83.5 kcal/mol) than the predicted non-consensus DNA (-65.9 kcal/mol) (p-value <0.00001).

Ets1 wild-type protein had a better binding energy with the non-consensus (-73.7 kcal/mol) than its cognate DNA (-69.2 kcal/mol). The overall and nucleotide-specific hydrogen bonds are also maximum for this complex (32 and 8, respectively). The Ets1 mutant protein with the non-consensus DNA (-66.2 kcal/mol) has a higher binding energy compared to the wild-type protein with this DNA (-62.7Kcal/mol). However, the wild-type protein also shows better binding to the predicted non-consensus substrate and hence confounds the results.

The binding energy of the EcoRV wild-type protein was similar towards the consensus (-122.5 kcal/mol) and the non-consensus (-123.7 kcal/mol). Similarly, there was no distinction between the binding energy of the mutant protein with the two substrates.

The initial assumption was that the wild-type protein should have a better affinity to its consensus sequence than any other sequence. However, Zif-268, Ets1 and EcoRV violate this assumption. Zif-268 wild-type protein interaction with the wild-type and the MT1 DNA had similar binding energy values (p-value 0.66). The binding energy for Ets1 wild-type protein was better with the mutant substrate (p-value <0.00001) and for EcoRV it was comparable between the wild-type DNA and non-consensus substrate (p-value 0.24).

We analysed the correlation between the different criteria used to compare the DNA-protein complexes towards the binding energy. The Pearson's Correlation coefficient was the highest between the total number of hydrogen bonds (0.75) and the binding energy. This is expected since the binding energy has a bonded interaction term for its calculation. However, the correlation between the nucleotide-specific hydrogen bonds and the binding energy is only 0.39 (about half of that with the total bonds). Since the specific interactions distinguish between the different DNA-protein complexes, this is the term that we should be monitoring to assess whether the prediction was successful. However, in Ets1, Myc and EcoRV the highest specific interactions are observed between the wild-type protein with the mutant DNA substrate. In addition, we also checked the retention of the hydrogen bonds present initially in the crystal structure (Appendix 18).

Table 3.5: MD analysis of wild-type and mutant proteins with different DNA substrates. The average RMSD (Å), Centroid distance between DNA-protein (Å), the total DNA-protein hydrogen bonds, the nucleotide-specific hydrogen bonds and the binding energy (kcal/mol) are reported. WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The values in bold represent the best value for that protein (lowest RMSD, centroid distance and binding energy; highest total hydrogen bonds and nucleotide-specific hydrogen bonds). The p-value was obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal structure. The cells coloured green do not have a statistically significant p-value. For detailed triplicate data and p-values refer to Appendices 10-14.

Sr. No	Protein	DNA	RMSD	Centroid	Total H-bonds	Specific H-bonds	Binding Energy
1	Zif-268 WT	WT	1.71 +/- 0.7	3.06 +/- 0.7	37.39 +/- 11.5	13.21 +/- 3.7	-62.4 +/- 10.9
2	Zif-268 MT1	MT1	2.87 +/- 0.9	3.21 +/- 0.8	40.39 +/- 10.7	14.36 +/- 2.6	-61.9 +/- 19.5
3	Zif-268 MT2	MT2	3.01 +/- 1.1	3.16 +/- 1.0	38.33 +/- 10.3	11.91 +/- 2.5	-96.6 +/- 26.4
4	Zif-268 WT #	WT	1.55 +/- 0.6	3.85 +/- 0.7	45.81 +/- 5.3	14.47 +/- 3.5	-66.5 +/- 10.8
5	Zif-268 MT1 #	MT1	1.22 +/- 0.4	3.46 +/- 0.8	46.94 +/- 6.0	15.89 +/- 2.6	-61.1 +/- 14.9
6	Zif-268 MT2 #	MT2	1.16 +/- 0.3	3.70 +/- 0.7	45.44 +/- 6.6	16.65 +/- 2.7	-78.9 +/- 18.6
7	Zif-268 WT	MT1	1.31 +/- 0.6	2.55 +/- 0.8	47.09 +/- 5.1	14.23 +/- 1.9	-83.5 +/- 13.0
8	Zif-268 WT	MT2	1.44 +/- 0.6	3.60 +/- 0.9	43.90 +/- 5.7	16.62 +/- 1.8	-47.6 +/- 8.8
9	Zif-268 MT1	WT	1.15 +/- 0.3	4.57 +/- 0.5	43.80 +/- 5.4	14.65 +/- 2.0	-70.7 +/- 16.9
10	Zif-268 MT2	WT	1.50 +/- 0.6	3.55 +/- 0.6	46.89 +/- 7.8	14.92 +/- 2.7	-77.9 +/- 23.7
11	Myc WT	WT	4.50 +/- 1.4	16.86 +/- 1.6	46.66 +/- 5.4	6.45 +/- 2.2	-107.6 +/- 18.8
12	Myc MT	MT	4.97 +/- 1.7	16.81 +/- 1.3	32.43 +/- 4.5	5.22 +/- 1.8	-65.9 +/- 12.0
13	Myc WT	MT	4.28 +/- 2.2	17.91 +/- 1.1	39.72 +/- 5.2	7.42 +/- 2.3	-83.3 +/- 13.0
14	Myc MT	WT	3.86 +/- 1.6	18.48 +/- 1.0	39.75 +/- 6.2	5.89 +/- 3.3	-82.3 +/- 25.8
15	Ets1 WT	WT	0.76 +/- 0.2	22.47 +/- 1.6	29.59 +/- 3.9	7.55 +/- 3.8	-69.2 +/- 9.2
16	Ets1 MT	MT	0.87 +/- 0.3	20.37 +/- 0.2	29.79 +/- 4.5	6.11 +/- 4.5	-66.2 +/- 11.9
17	Ets1 WT	MT	0.85 +/- 0.3	20.02 +/- 0.9	31.66 +/- 3.6	7.78 +/- 3.6	-73.7 +/- 16.4
18	Ets1 MT	WT	0.93 +/- 0.3	19.11 +/- 1.1	25.85 +/- 4.6	5.31 +/- 4.6	-62.7 +/- 11.8
19	EcoRV WT	WT	0.75 +/- 0.2	10.45 +/- 0.7	70.06 +/- 7.2	19.17 +/- 1.8	-122.5 +/- 22.4
20	EcoRV MT	MT	0.81 +/- 0.2	10.47 +/- 0.5	70.93 +/- 7.9	19.48 +/- 2.8	-103.4 +/- 20.4
21	EcoRV WT	MT	0.79 +/- 0.2	10.40 +/- 0.5	74.37 +/- 10.6	21.7 +/- 4.8	-123.7 +/- 22.3
22	EcoRV MT	WT	0.85 +/- 0.2	10.04 +/- 0.4	68.45 +/- 8.8	19.02 +/- 2.3	-104.4 +/- 24.5

The Zif-268 MT2 model had the highest retention of the original bonds from the crystal structure with the bonds being retained in 43.5 snapshots on an average out of 50. Myc wild-type protein had the highest retention of bonds with the non-consensus DNA (26.2/50). However, in the other two cases, i.e., Ets1 and EcoRV, the wild-type proteins with their cognate sequences had the highest retention at 43.8 and 36.0, respectively. The molecular dynamics system maintains the overall energy of the system relatively constant. The original bonds can be replaced by water molecules without any energetic loss and therefore, looking at hydrogen bond retention as a proxy for specificity may not work.

Thus, to conclude whether the predicted mutations result in specificity change according to MD simulations, we analysed the relative binding energy. However, the predicted mutants had a higher binding energy than the wild-type in all cases. Therefore, we compared the binding energy of the mutant protein towards the predicted and wild-type substrate. The Myc mutant protein does not show better binding to the predicted substrate as compared to the wild-type substrate. Ets1 mutant shows a better binding towards the predicted substrate compared to the wild-type DNA and EcoRV exhibits similar binding towards both substrates.

3.5 Discussions

The current chapter discussed the *in-silico* validation techniques to demonstrate the applicability of the DNA-protein recognition preferences such as- the prediction of amino acids involved in zinc finger protein recognition, the prediction of the MSA of a protein from the PWM of its binding site, changing the specificity of a DNA-binding protein and predicting the effect of mutations of DNA-binding domain.

We validated the predictions from our method using the zinc finger code. We used two zinc finger datasets for the validation. The first set was optimized using SELEX experiments to bind the intended target sequences. The second set was obtained through the PDB and had the zinc finger protein annotation. We can predict the experimentally optimized sequences comparatively more accurately than the 3D-structures from the PDB. We have noted in Chapter 2 that the nucleotides have interactions with all amino acids

with various degrees. The analysis of multiple sequence alignments also revealed the presence of the predicted amino acid in the close homologs of the DNA-binding protein. The crystal structure has one of the variants that can specifically bind and the prediction is not necessarily incorrect if the predicted and actual amino acid in the protein do not match. This hypothesis is further supported by better predictions on the experimentally determined specific DNA-protein complexes.

The second validation was done on the prediction of the MSA of a protein from the PWM of its recognition sequence. The differences in the DNA consensus site may have their origin in the changed amino acid partners in their recognizing proteins. We predicted the amino acids for the PWMs and compared them with the amino acids obtained through evolutionary analysis (PSIBLAST). The MCC of our method was 0.17, which is not very high. We observed that the low true positive rate was primarily responsible for the low MCC, while the true negative rate was high. This implies that correct amino acid identification is a problem, but what is not specific is very clear. To increase the correct predictions, we need more structural data, which will allow us to explore machine learning techniques.

Another *in-silico* validation was to predict the effect of a mutation in the DNA-binding protein when the amino acid was involved in hydrogen bonding with DNA. We had an accuracy of 59% in predicting whether the mutation would result in an increase or a decrease in the binding affinity. However, this accuracy increases to 76% if we exclude the nucleotides for which we have no observations in our dataset. Thus, with the increase in availability of data in the future, we can expect the accuracy of this method to increase.

The ProNIT dataset contains Myb protein (PDB ID: 1MSE), which was mutated to multiple amino acids and the binding affinity for all of them with DNA was determined by isothermal calorimetry [253]. There are 26 unique mutations of amino acid-trinucleotide pair in the protein. However, since we grouped amino acids into classes based on their functional groups, some of the mutations share the same amino acid-trinucleotide pairs. We compared the binding affinity change when the wild-type and mutated pair was the same in the dataset. With the same mutation, the binding affinities ranged between a 100-

fold reduction to twice that of the wild-type. For the same kind of mutation, this variation in the binding affinity makes it difficult to generate a generalized method for quantifying the change in affinity. However, with more data, rather than just predicting the increase or decrease, we could attempt to quantify the change in affinity in the future.

The final validation involved predicting the mutation in DNA-binding proteins to bind a sequence of interest. Changing the specificity of a DNA-binding protein has been attempted in previous studies [254–256]. Most of these methods rely on the randomization of amino acids and screening of the complexes. The rational design of a DNA-binding protein to recognize a sequence of interest has been a difficult problem to solve due to the absence of a universal recognition code. Thus, we employed our probability-based method to understand if our method could solve this problem. To evaluate whether the predicted protein binds to its intended target specifically, we employed MD simulations.

As an experimental control, we used zinc finger wild-type protein and its variants, which have been experimentally optimized to recognize a target sequence. We noticed that the use of models of these proteins instead of the crystal structure had higher binding energies. Also, the wild-type protein bound to the non-cognate DNA with a better binding energy than to the cognate sequence. A similar observation was made with one of the variants, which had a better binding energy towards the wild-type cognate site and not the experimentally determined site. Since these control experiments did not give the expected results, it decreases the confidence in the validation of the predictions using MD simulations.

The binding energy showed that the predicted protein mutant bound with a better energy with the intended target for Ets1. In the case of EcoRV mutant, the binding energies were comparable between the intended target and the wild-type site. However, Myc showed that the mutant protein retains the original specificity towards the wild-type consensus site. We also noticed that the binding energy of the predicted mutant proteins was lower than their wild-type counterparts. It has been noted that the engineering of specificity can lead to reduced affinity [257]. We have also noticed this in the dataset that

we curated from ProNIT. Therefore, we validated these predicted mutations using *in-vitro* techniques (discussed in Chapter 4).

3.6 Acknowledgements

Atreyi Chakraborty helped with some of the molecular dynamics simulations for Ets1 and Zinc Fingers. Saswati helped with EcoRV wild-type simulations.

3.7 Chapter summary

In this chapter, we have discussed the four *in-silico* validation approaches for the prediction of amino acids specific for a DNA sequence.

The validation using zinc finger proteins showed better results in top1 (56%) than our training and testing set as well as the zinc finger protein crystal structures.

The prediction of MSA of a protein from its recognition site showed that our ranks are correlated with the occurrence of the amino acid in the MSA of the protein constructed from sequences of close homologs.

The prediction of the effect of mutation on the binding affinity could successfully predict the increase or decrease in binding affinity in 76% of the cases where the trinucleotide data was available.

Finally, we validated our predictions by changing the DNA-binding specificity of proteins and analysing them using Molecular Dynamics (MD) simulations. However, the control system of Zif-268 showed that the wild-type protein had a better binding energy towards the non-cognate DNA than its consensus sequence.

Moreover, the modelling of the mutations increases the binding energy of the DNA-protein complex in two cases and increases it in one case.

The results for the predicted mutations showed that Myc mutant had better binding to the wild-type consensus sequence. Ets1 mutant had a better binding towards the non-consensus DNA that it was predicted to bind. There was no significant difference in binding of the EcoRV mutant or wild-type proteins to bind either the cognate or non-cognate sites.

4

CHAPTER 4: *In-vitro* Validation of Predicting Amino Acids for a DNA Sequence

OVERVIEW

- Changing the specificity of DNA-binding protein
 - Methods for molecular cloning, expression and purification
 - Molecular cloning
 - Expression and purification
 - *In-vitro* Binding Assays for Ets1 and Myc
 - Electrophoretic Mobility Shift Assay (EMSA)
 - Microscale Thermophoresis (MST)
 - Nano Differential Scanning Fluorimetry (NanoDSF)
 - Comparison of binding using different *in-vitro* techniques
 - Discussions
 - Chapter summary
-

The previous chapter dealt with the *in-silico* validation of the method for predicting amino acids specific to a given DNA sequence. In this chapter, we describe the results for *in-vitro* validation.

4.1 Changing the specificity of DNA-binding protein

We have a probabilistic measure of specific binding between nucleotides and amino acids by analysing DNA-protein complex structures. In chapter 3, one of the validations was to change the specificity of a DNA-binding protein from a known consensus sequence to a sequence of our choice. We had evaluated the predictions using molecular dynamics simulations, and it did not reveal any clear outcomes about the accuracy of our method. Therefore, we attempted to verify the mutations *in-vitro*.

We had chosen the same systems from chapter 3 for validation- EcoRV, Ets1 and Myc. We made predictions for mutations to be made on the protein for it to recognize a sequence of interest as described in chapter 3, Table 3.3. An additional double mutant for Myc was also generated (Table 4.1).

Table 4.1: The experimental systems selected for changing their DNA-binding specificity

Sr. No	Protein	Wild-type consensus	Non-consensus	Predicted Mutation in protein
1	Ets1	GGA[A/T]	GGTT	Tyr (395.C)-> Arg
2	Myc	CACGTG	CATATG	Arg (20.A)-> Gln
3	Myc	CACGTG	ATCGAT	His (12.A)→Gln +Glu(16.A)->Tyr
4	EcoRV	GATATC	GGTACC	Ser (183.A)-> Gln

4.2 Methods for molecular cloning, expression and purification

The wild-type genes for these proteins were synthesized by Genscript. The EcoRV enzyme cleaves the unmethylated GATATC sites [249,251,252]. Therefore, we required the EcoRV methyltransferase to express this enzyme without killing the host bacteria. We obtained the plasmid containing the EcoRV methyltransferase gene from Prof. Dr. Albert Jeltsch (University of Stuttgart).

4.2.1 Cloning

We followed the restriction-free (RF) strategy (described in the introduction to techniques, section 7) for cloning. This technique involved polymerase chain reactions (PCR) to obtain the required gene.

4.2.1.1 PCR protocol

For all steps of the PCR, we used autoclaved distilled water (D/W), 10x pfu polymerase buffer (Table 4.2), Roche deoxynucleotides (dNTPs) and polymerase purified by Dr. Saikrishnan's lab. The components for making the buffer were from Sigma and available in the common biology lab supplies. We followed a general PCR protocol and made changes to the extension time as per the size of the product to be amplified (Tables 4.3 and 4.4). The list of primers used in this study are mentioned in Table 4.5.

Table 4.2: Composition of 10X pfu polymerase buffer (pH 8.3 at 25°C)

Component	Concentration
Tris.HCl	100 mM
MgCl ₂	20 mM
KCl	500 mM

Table 4.3: PCR reaction setup for cloning

Component	Concentration	Volume
Pfu buffer	10X	5 µL
dNTPs	10mM of each dATP, dCTP, dGTP, dTTP	2 µL
Forward Primer	10 µM	2 µL
Reverse Primer	10 µM	2 µL
Template	Up to 50ng for 50µL reaction	variable
Polymerase	-	0.8 µL
Autoclaved D/W	-	Make up volume to 50 µL

Table 4.4: *The PCR protocol used for cloning*

Repetition	Step	Temperature (°C)	Time
1	Initial denaturation	95	2 minutes
35	Denature	95	30 seconds
	Anneal	55	30 seconds
	Extend	72	1Kb per min
1	Final Extension	72	2x(1Kb per min)

4.2.1.2 RF cloning

We used the RF strategy to clone the mutants for EcoRV, Ets1 and Myc. The genes received from Genscript were in pET22 (+) plasmid and had a pelB signal sequence for ease of purification. However, this pelB sequence was not cleaved in the expression strains (BL21A1 and BL21DE3). Therefore we cloned the wild-type and mutant genes into the pHis17 plasmid from Dr. Saikrishnan's lab. For allowing expression of the gene, pHis17 has T7 promoter and T7 terminator regions. The T7 forward primer is the promoter sequence near the transcription start site recognized by T7 RNA polymerase [258]. The T7 reverse primer is the sequence present after the gene and forms a stable stem-loop structure, which acts as a termination signal for the T7 RNA polymerase [259].

To change the plasmid backbone of the wild-type genes, we needed to perform two steps of PCR. In the first step, we designed the forward and reverse primers such that part of it had complementarity with the gene, and part of it was complementary to pHis17. We used the Genscript-provided plasmid with our gene of interest as the template. So the product generated by PCR had the gene with ends that were complementary to the vector plasmid. We purified this DNA fragment using Qiagen PCR purification kits to remove unused primers, dNTPs and polymerase. We replaced the gene in the pHis17 plasmid with our gene of interest in the second step of PCR. We used the pHis17 plasmid with the McrBC gene as the template. The product from the first PCR acted as the forward and reverse primers for the second step. Thus, during the PCR, only the remaining backbone of the pHis17 plasmid would get added to the first PCR product.

To introduce a mutation as well, we performed two steps of PCR. In the first step, the forward primer was complementary to the region where we wanted to mutate, except for the codon where we wanted to make the change. We used the wild-type gene in the pHis17 plasmid as the template and T7 reverse for this PCR. Thus, the resultant PCR product had part of the gene, the desired mutation, and part of the pHis17 plasmid. The product from the first PCR acted as the forward and reverse primers for the second step. We used the wild-type gene-containing plasmid as the template. Therefore, the second step extended the first PCR product to have the whole gene and the plasmid and mutation of interest.

For EcoRV, we needed to transform two genes into bacteria- restriction enzyme and the methyltransferase. Since only one plasmid is sufficient for ampicillin resistance, we would not know whether both genes have got transformed. Therefore, we cloned the EcoRV methyltransferase in a plasmid with ampicillin resistance and the EcoRV restriction enzyme in a plasmid with kanamycin resistance. Therefore, during selection, we can use both antibiotics to obtain the bacteria that have the methyltransferase and the restriction enzyme.

Table 4.5: *The sequences of primers used for cloning.*

F and R indicate forward and reverse, respectively.

Sr. No	Primer Name	Length	Primer Sequence
1	T7F	20	TAATACGACTCACTATAGGG
2	T7R	19	GCTAGTTATTGCTCAGCGG
3	pHis EcoRV Gene F	40	CTTTAAGAAGGAGATATACATATGATGAGTCTTCGTTCTG
4	pHis EcoRV Gene R	45	GATGATGATGATGATGGGATCCTTATTTTCTTCCTCGGTATATCC
5	EcoRV S183Q F	37	GGAGATTTGGCAGGACAAGGAAACACAACAAATATAG
6	Myc R20Q F	24	GTCAGCGTCAGAACGAACTGAAAC
7	pHis EcoRV Methyltransferase F	45	CTTTAAGAAGGAGATATACATATGAAAGATAAAGTTTTTGTTCCG
8	pHis EcoRV Methyltransferase R	46	GATGATGATGATGATGGGATCCTCACTCTTCAATTACCAAATATC
9	pHis Ets1 Gene F	49	CTTTAAGAAGGAGATATACATATGAGCCTGCAACGTGTGCCGAGCTATG

10	pHis Ets1 Gene R	28	CCGGACGCGGATGAACTCGAGCACCACC
11	pHis Myc Gene F	45	CTTTAAGAAGGAGATATACATATGAAGCACCACCACC ACCACCAC
12	pHis Myc Gene R	35	GATGATGATGATGATGGGATCCTTACGCGCAGCTG
13	Ets1 Y395R F	29	GCCGTGGTCTGCGTCGTTACTACGATAAG

4.2.1.3 Transformation protocol

We needed to ensure that only the desired PCR product gets inserted into the bacteria and none of the other templates used during PCR. Therefore, we added 0.5 µL of DpnI enzyme to 9.5 µL of the final PCR product and incubated it for 3 hours at 37°C. This enzyme cleaves the DNA, which has methylated adenine in the GATC sites. Since we have obtained the PCR templates from bacteria, they will have methylated adenine. However, our PCR products would not have any methylation and escape degradation.

We transformed these DpnI digested products into bacterial cells by electroporation. We added the DpnI digested product to 80 µL of NEB Turbo electro-competent cells. The DNA was incubated with the cells for 10 minutes on ice to allow surface adsorption. We then added this mixture to an electro-cuvette (Gene Pulser®/MicroPulser, Biorad) and diluted it using 100 µL chilled st. MiliQ. The cuvette was wiped thoroughly with a soft tissue to remove moisture from the ice. A pulse of 2500 volts was supplied for 4 ms to the electro-cuvette using the Biorad electroporator. We added 200 mL of 2X Luria Bertani, Miller (LB) broth (HiMedia) to the electro-cuvette and incubated it for 30 minutes at 37°C.

The plasmid has an ampicillin resistance gene. Therefore, we spread the LB broth with the transformed cells on a Petri-plate containing LB agar with ampicillin. After incubation at 37°C for 12-16 hours, we selected a few colonies from the plate for sequencing. We inoculated individual colonies into LB broth with 100 µg/mL ampicillin. After 16 hours of incubation at 37°C, the culture was centrifuged at 4500 rpm at 4°C for 15 minutes. The plasmid was extracted from the cells using the Qiagen mini-prep kit. The size of the plasmids was verified and we chose some for sequencing.

4.2.2 Expression and purification

4.2.2.1 Myc

We used BL21AI cells to express Myc wild-type, R20Q mutant and Myc QY double mutant proteins. We followed the same protocol for all three proteins. The transformed cells were incubated in LB +ampicillin plate overnight at 37°C. A few colonies from the plate were transferred to LB broth with ampicillin (100 µg/ml) at 37°C for 3 hours for primary culture. 0.1% of the primary culture was inoculated into 2 L media with ampicillin (100 µg/ml) and grown at 37°C till OD₆₀₀ (Optical density at a wavelength of 600 nm) was 0.8. The culture was then induced with 0.1% arabinose and incubated at 14°C for 18 hours. We centrifuged the culture at 4500 rpm and, after discarding the LB, redissolved the pellet in lysis buffer (Table 4.6).

We sonicated the cells with 1 second on- 3 seconds off cycles for 3 mins. We repeated the sonication with the same settings after 10 minutes of incubation on ice. We used ultracentrifugation at 37000 rpm for 50 mins at 4°C to separate the soluble and insoluble proteins. The supernatant was loaded on a Ni-NTA column (HisTrap HP, 5 mL, GE Healthcare) and eluted with a gradient of B250 and B1000 buffers (Table 4.7). 10 mL protein fractions were collected at 0, 10, 20, 30, 50 and 100% B1000. We denatured the protein fractions in 2x Sodium Dodecyl Sulfate (SDS) dye (Table 4.8) at 95°C for 10 minutes and centrifuged it at 14000 rpm for 10 minutes. We loaded the supernatant on 18% SDS-polyacrylamide gel (Table 4.9)

We dialysed the purest fraction in SnakeSkin dialysis tubing (3.5K MWCO 22mm, ThermoFischer) using 2 L dialysis buffer (Table 4.10) for 2 hours at 4°C (Figure 4.1 (a), Lanes 4,5 and 4.1 (b), Lanes 5,6 and Lane 7 in 4.1 (c)). We concentrated the dialysed protein to 1.2 mg/mL in centricons (Vivaspin Turbo 15, 10,000 MWCO PES, Sartorius), flash-freezed and stored at -80°C. The Myc wild-type protein is a strong leucine zipper protein and does not dissociate into individual units even after SDS denaturation. Therefore, the wild-type protein does not correspond to the expected protein size.

Table 4.6: Composition of Myc lysis buffer

Component	Concentration
Tris.Hcl (pH 8.0)	20 mM
NaCl	250 mM
Ethylenediamine tetra-acetic acid (EDTA)	1 mM

Table 4.7: Composition of Ni-NTA buffers for Myc

Component	B250	B1000
Tris.Hcl (pH 8.0)	20 mM	20 mM
NaCl	250 mM	250 mM
Imidazole	250 mM	1 M

Table 4.8: Composition of 2x SDS dye

Component	Concentration
Tris.Hcl (pH 6.8)	125 mM
SDS	4 %
Bromophenol blue	0.02 %
Glycerol	20 %
Dithiothreitol (DTT)	200 mM

Table 4.9: Composition of 18% SDS-polyacrylamide gel

Component	Stacking Gel (2 mL)	Resolving Gel (5 mL)
1.5M Tris.Hcl (pH 8.8)	-	1.25 mL
1.5 M Tris.Hcl (pH 6.8)	250 µL	-
30 % Acrylamide-bisacrylamide (AB) mix	330 µL	3 mL
10% SDS	20 µL	50 µL
10% Ammonium persulfate (APS)	20 µL	50 µL
N, N, N', N'- Tetramethylethylenediamine (TEMED)	4 µL	5 µL
D/W	1.4 mL	645 µL

Table 4.10: Composition of Myc dialysis buffer

Component	Concentration
Tris.Hcl (pH 8.0)	20 mM
NaCl	500 mM
DTT	1 mM

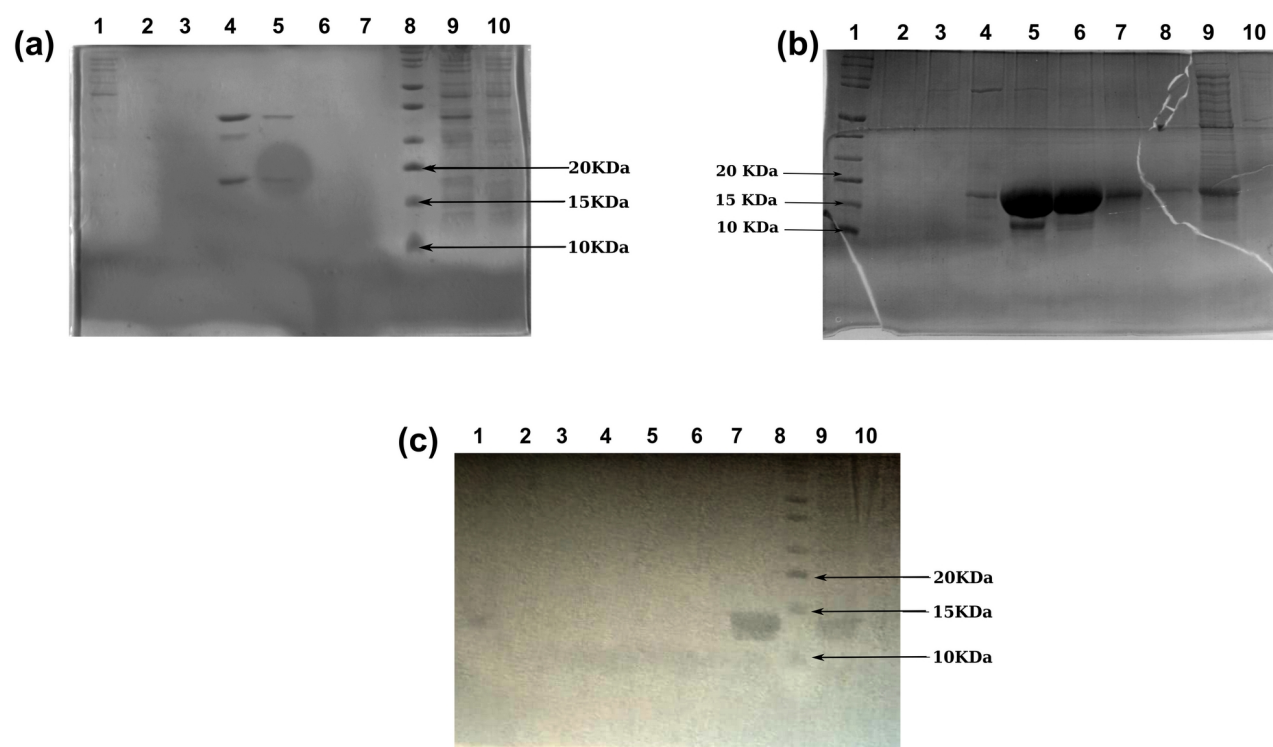


Figure 4.1: SDS-PAGE gel after HisTrap column elution for Myc -wild-type (a), R20Q mutant (b) and Myc QY double mutant (c). Lane 9 and 10 are the total protein loaded on the column and the unbound protein, respectively in subplots (a), (b) and (c). Lane 8 corresponds to the protein ladder in figures (a) and (c), while lane 1 is the ladder in subplot (b). Other lanes contain eluted fractions. The protein monomer is 14 KDa in size and is expected near the 15 KDa band.

4.2.2.2 Ets1

The Ets1 wild-type and Y395R mutants were both expressed and purified using the same protocol. The plasmid with the gene was transformed into BL21AI cells. The transformed cells were incubated in LB with ampicillin plate overnight at 37°C. A patch of colony was picked up and transferred to LB broth with ampicillin (100 µg/ml) at 37°C for 3 hours for primary culture. 0.1% of the primary culture was inoculated into 2 L media ampicillin (100µg/ml) and grown at 37°C till OD₆₀₀ was 0.8. The culture was then induced with 0.1% arabinose and incubated at 37°C for 3 hours. The culture was pelleted down at 4500 rpm and the pellet was redissolved in lysis buffer (Table 4.11).

It was then sonicated with 1 second on 3 second off cycles for 3 mins. This was repeated after 10 minutes of incubation on ice. To separate the soluble and insoluble proteins, ultracentrifugation was done at 37000 rpm for 50 mins at 4 °C. The supernatant was loaded on Ni-NTA column (HisTrap HP, 5 mL, GE Healthcare) and eluted with a gradient of B250 and B1000 (Table 4.12). 10mL protein fractions were collected at 0, 10, 20, 30, 50 and 100% B1000. They were loaded on 18% SDS-polyacrylamide gel and the fraction with purest protein (Lane 6, for both Figure 4.2(a) and 4.2(b)) was collected in SnakeSkin dialysis tubing (3.5K MWCO 22mm, ThermoFischer). It was dialysed in 2L dialysis buffer (Table 4.13) for 16 hours at 4°C. The dialysed protein was concentrated to 2 mg/mL in centricons (Vivaspin Turbo 15, 10,000 MWCO PES, Sartorius). It was then flash-frozen and stored at -80°C.

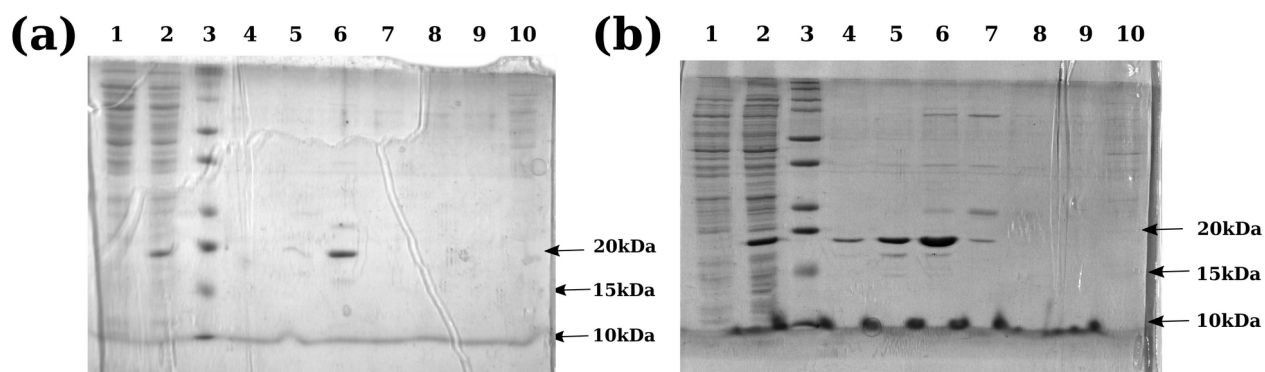


Figure 4.2: SDS PAGE gel after HisTrap column elution of Ets1 -wild-type (a) and Ets1 Y395R mutant (b) proteins. Lane 2 is the total protein loaded on the column, Lane 1 is the unbound fraction and Lane 3 is the protein ladder in both figures. Other lanes contain eluted fractions. The protein is 19 KDa in size and is expected near the 20 KDa band.

Table 4.11: Composition of Ets1 lysis buffer

Component	Concentration
Potassium phosphate buffer (pH 7.4)	50 mM
KCl	500 mM
MgCl ₂	10 mM
Imidazole	20 mM
Glycerol	10%

Table 4.12: Composition of Ni-NTA buffer for Ets1 buffer

Component	B250	B1000
Potassium phosphate buffer (pH 7.4)	50 mM	50 mM
KCl	500 mM	500 mM
Imidazole	250 mM	1 M

Table 4.13: Composition of Ets1 dialysis buffer

Component	Concentration
Sodium phosphate buffer (pH 7.4)	50 mM
KCl	100 mM
DTT	1 mM

4.2.2.3 EcoRV

As described earlier, EcoRV is a restriction enzyme that cleaves unmethylated GATATC sites. The host bacterium is also susceptible to cleavage if the EcoRV methyltransferase does not methylate this site. We cloned the two genes into plasmids with different selection markers (ampicillin and kanamycin). Therefore, we used culture media containing ampicillin and kanamycin antibiotics to select bacteria with both genes using-

- Co-transformation: We added equal concentration (200 ng) of both the plasmids and transformed them into BL21AI cells. However, we could not obtain colonies on an LB plate with ampicillin and kanamycin antibiotics.
- Transformation of methyltransferase containing bacteria: Since transformation of both plasmids together was not successful, we transformed methyltransferase enzyme in BL21AI cells. We then made these bacteria competent (Figure 4.4) using Acid Salt Buffer (ASB, Table 14). After that, we transformed the restriction enzyme on these BL21AI cells containing the methyltransferase gene.

Table 4.14: Composition of ASB

Component	Concentration
MnCl ₂	100 mM
CaCl ₂	100 mM
Sodium Acetate	40 mM
Adjust pH to 5.5 using acetic acid and filter with 0.22 µM filter into autoclaved bottle	

We tried expression at 30°C (5 hours), 37°C (3 hours) and 14°C (16 hours). However, the OD₆₀₀ of the culture after incubation reduced from 0.8 to ~0.3, indicating cell lysis. There was no reduction in O.D. if we did not induce the culture. We also tried different cell types apart from BL21AI, such as BL21DE3, C43, and Rosetta. However, we did not observe protein expression in any of them. The SDS-PAGE gel did not show a difference in protein levels in uninduced and induced cultures with either BL21AI cells or BL21DE3 cells (Figure 4.3). Instead, the protein levels seemed to be higher in the uninduced culture. Around the expected size of the enzyme, there was no prominent band

in the induced culture. Therefore, we could not proceed with the purification of the enzyme and characterization of the mutant.

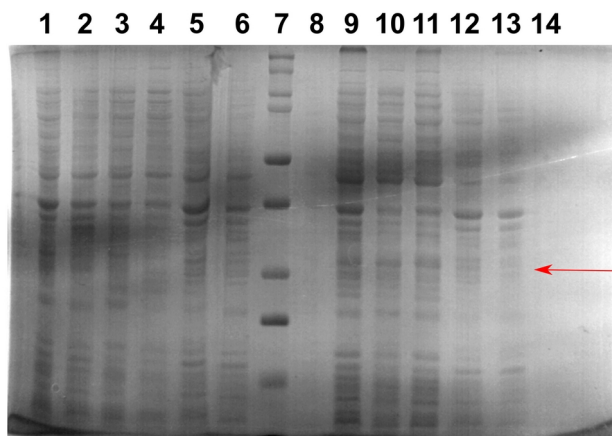
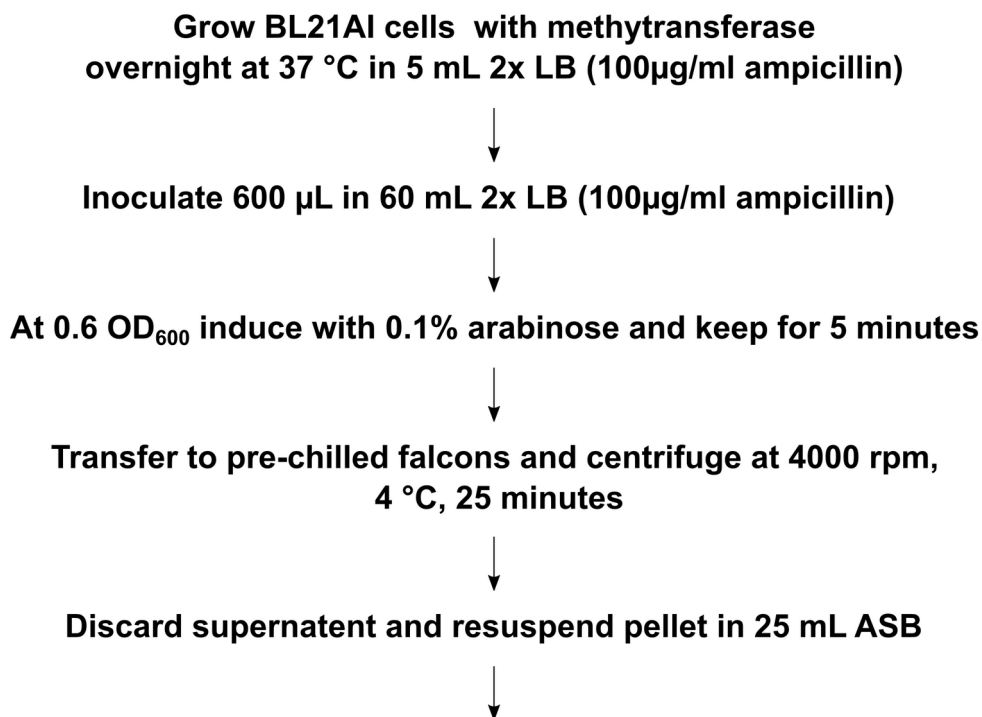


Figure 4.3: Expression check gel for *EcoRV*

-using BL21AI (lanes 1-6) and BL21DE3 cells (lanes 9-14). The different fractions are- uninduced lysate (lanes 1 and 9), induced lysate (lanes 2 and 10), uninduced supernatant (lanes 3 and 11), induced supernatant (lanes 4 and 12), uninduced pellet (lanes 5 and 13) and induced pellet (lanes 6 and 14). The ladder is shown in lane 7 and the expected enzyme size is around the red arrow.



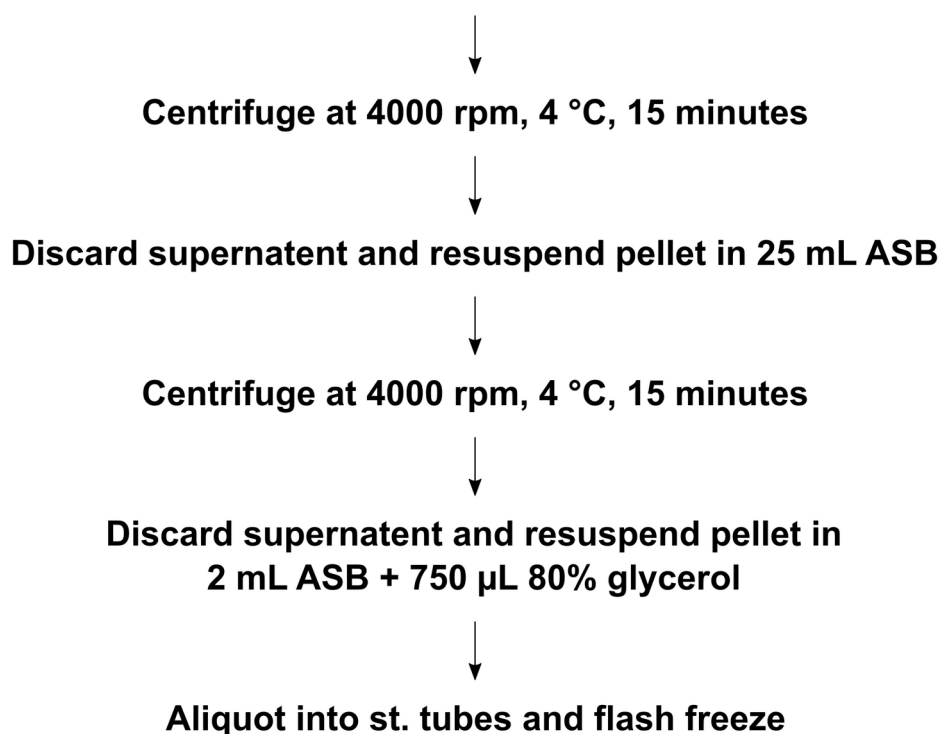


Figure 4.4: *Protocol for competent cell preparation containing methyltransferase gene*

4.3 Binding Assays for assessing specificity of Ets1 and Myc

We wanted to assess the binding of the mutant and wild-type proteins with different DNA. Therefore, we employed three in-vitro techniques for this- Electrophoretic Mobility Shift Assay (EMSA), Microscale Thermophoresis (MST) and Nano Differential Scanning Fluorimetry (Nano-DSF). We procured lyophilized single-stranded (ss)-DNA sequences from Sigma and Eurofins (Table 4.15) and used autoclaved D/W to obtain 100 µM DNA concentration. We annealed these sequences to form double-stranded (ds) DNA (Table 4.16) in the binding buffers (Tables 4.17 and 4.18). The PCR steps gradually reduced from 99°C to 25°C, each step lasting 1 minute. The annealed DNA concentration is half of that in the ss-DNA form. Therefore, 20 µM ss-DNA resulted in 10 µM ds-DNA.

Table 4.15: DNA substrates for binding assays

Protein	Substrate	Forward	Reverse	Length (bases)
Ets1	Consensus	ACCGGATGTGGCTTC	GAAGCCACATCCGGT	15
	Non-consensus	ACCGGTTGTGGCTTC	GAAGCCACAACCGGT	15
	Non-specific	GCATGCATGCATGCA	TGCATGCATGCATGC	15
Myc	Consensus	CACCCGGTCACGTGGCCTAC AC	GTGTAGGCCACGTGACCGG GTG	22
	Non-consensus (Single Mutant)	CACCCGGTCATATGGCCTACA C	GTGTAGGCCATATGACCGGG TG	22
	Non-consensus (Double Mutant)	CACCCGGTATCGATGCCTACA C	GTGTAGGCATCGATACCGGG TG	22
	Non-specific	GGTCCGAAGATTCTAGCATAA T	ATTATGCTAGAATCTTCGGAC C	22

Table 4.16: DNA annealing protocol

Component	Final Concentration	Step	Temperature	Time
Forward primer	20 µM	1	99°C	1 minute
Reverse primer	20 µM	2	98°C	1 minute
Binding buffer	1x	:	:	:
Water	Make up volume	77	25°C	1 minute

Table 4.17: Composition of Myc binding buffer (1x)

Component	Concentration
Tris.Hcl (pH 8.0)	10 mM
NaCl	500 mM
MgCl ₂	5 mM
EDTA	1 mM
DTT	1 mM

Table 4.18: Ets1 binding buffer (1x)

Component	Concentration
Phosphate buffer (pH 7.4)	50 mM
KCl	100 mM
MgCl ₂	2 mM
DTT	1 mM

4.3.1 Electrophoretic Mobility Shift Assay (EMSA)

EMSA is a semi-quantitative assay to assess the binding of a protein with DNA [192,193]. It involves adding increasing concentration of the protein to a constant DNA concentration. The DNA migration is the readout where unbound DNA migrates faster in an electric field, compared to protein-bound DNA. The binding assay was carried out by incubating 500 nM ds-DNA with different protein concentrations at 12°C for 20 min in the binding buffers (Tables 4.17 and 4.18 for Myc and Ets1, respectively). The reaction was then loaded on 8% native polyacrylamide gel and run at 4°C, 150 V in Tris-Borate EDTA (TBE) buffer (Table 4.19, Table 4.20) using the BioRad mini-Protean system.

Table 4.19: Composition of TBE buffer (1x)

Component	Concentration
Tris.HCl (pH 8.0)	130 mM
Boric acid	45 mM
EDTA	2.5 mM

Table 4.20: Composition of 8% polyacrylamide gel (8 mL)

Component	Concentration
D/W	4.19 mL
30% AB mix	2.1 mL
5x TBE	1.6 mL
10% APS	100 µL
TEMED	10 µL

The gel was stained with Ethidium Bromide (EtBr) and the DNA intensity was quantified by ImageJ software [260]. The unbound DNA intensity was used to find the apparent binding affinity using the formula-

$$K_d = \left(\frac{C + [P]_{max}}{y} \right) - [P] \quad \text{Eq. 1}$$

where, K_d is the apparent binding affinity in terms of the dissociation constant, C is the intensity of unbound DNA with no protein added, $[P]_{max}$ is the maximum protein concentration, y is the intensity of unbound-DNA when protein concentration $[P]$ is used.

4.3.1.1 EMSA results for Myc

The native PAGE showed binding of Myc wild-type protein with all the substrates used (Figure 4.5). However, the apparent K_d was lowest for the consensus sequence (Table 4.21).

In the case of the Myc R20Q single mutant, the DNA-shift occurs at the same concentration for the wild-type consensus sequence and the predicted substrate for this mutation (Figure 4.6). The binding affinity values calculated by Eq. 1 are also similar. However, compared to the wild-type protein, the affinity of the Myc R20Q mutant is lower for both the wild-type consensus DNA substrate and the random substrate. The proteins now bind at similar binding affinities indicating a loss of specificity.

The Myc QY double mutant also shows DNA-shift at the same concentration for the wild-type consensus sequence and the predicted substrate for this mutation (Figure 4.7). However, as opposed to the single mutant, there is an improvement in binding to the predicted substrate over the wild-type protein. Comparing the binding affinity of the Myc QY double mutant protein with that of the wild-type protein towards the predicted substrate shows a difference of 2 nM (standard deviation 1.6 nM). Binding to a random substrate does not change much from 4.0 +/- 2.1 nM in the mutant protein to 4.7 +/- 1.3 nM in the Myc double mutant.

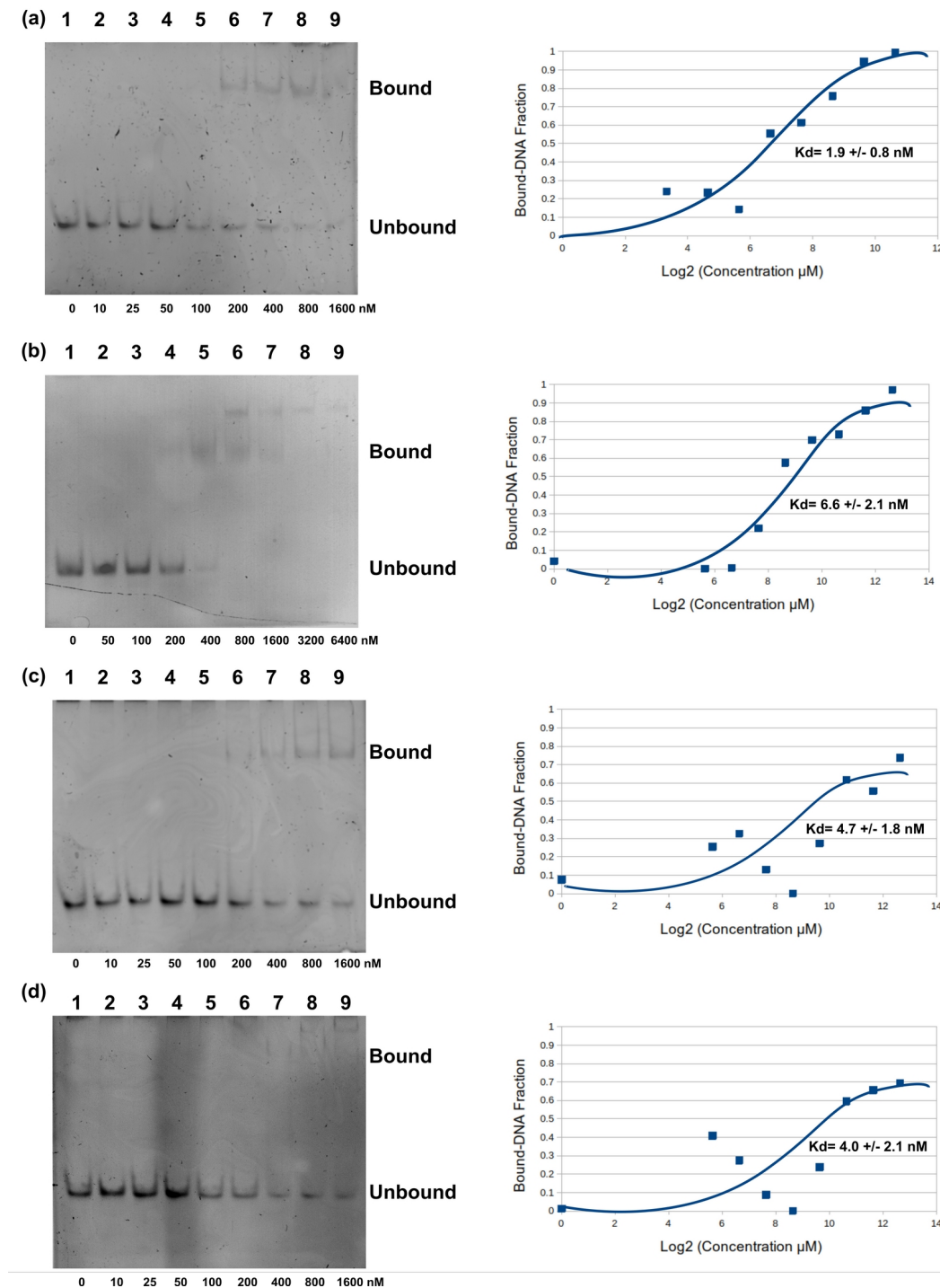
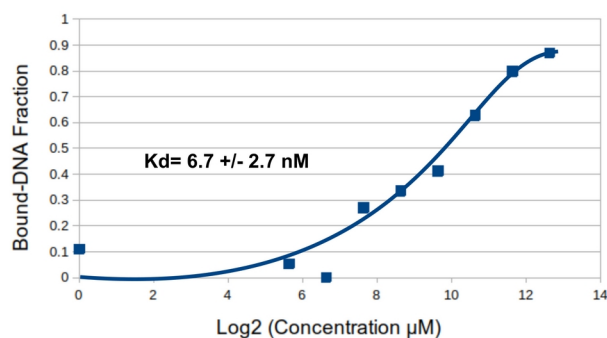
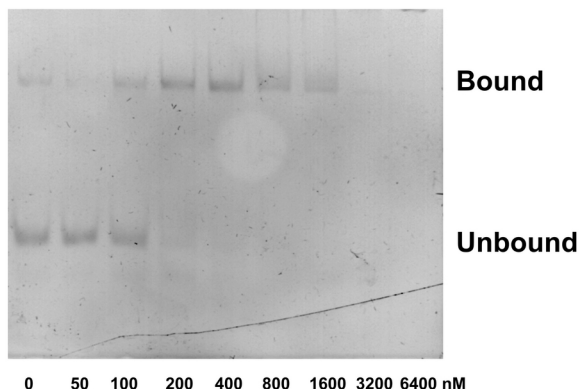


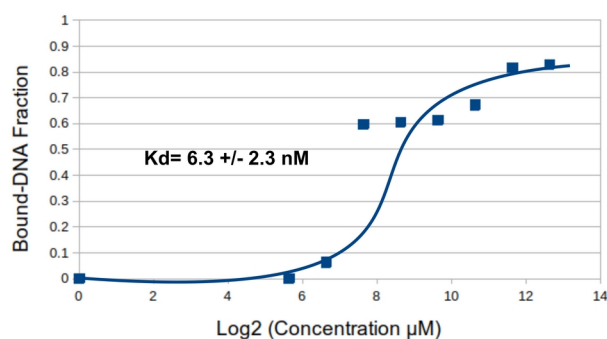
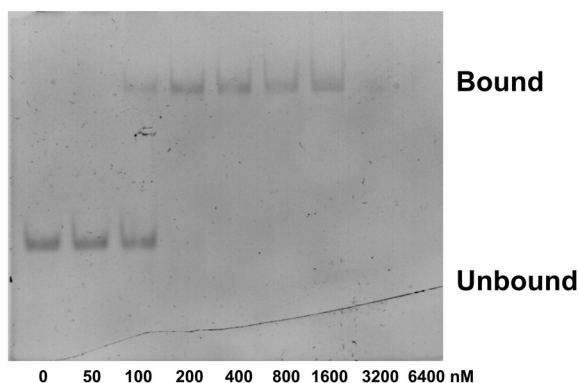
Figure 4.5: EMSA of Myc wild-type protein

-with wild-type consensus DNA (a), non-consensus single mutant (b), non-consensus double mutant (c) and random DNA substrates (d). The protein concentrations are mentioned below each lane. The fraction of bound DNA (x-axis) as a function of $\log_2$ of protein concentration is plotted on each sub-figure to the right.

(a) 1 2 3 4 5 6 7 8 9



(b) 1 2 3 4 5 6 7 8 9



(c) 1 2 3 4 5 6 7 8 9

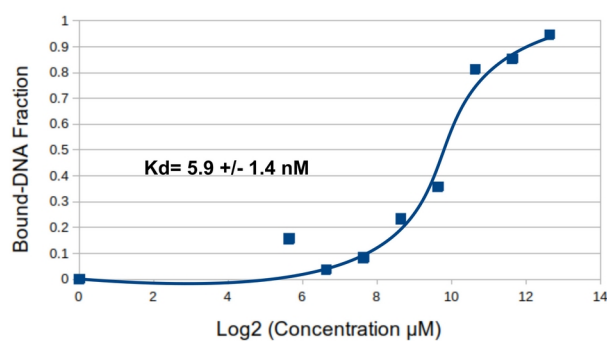
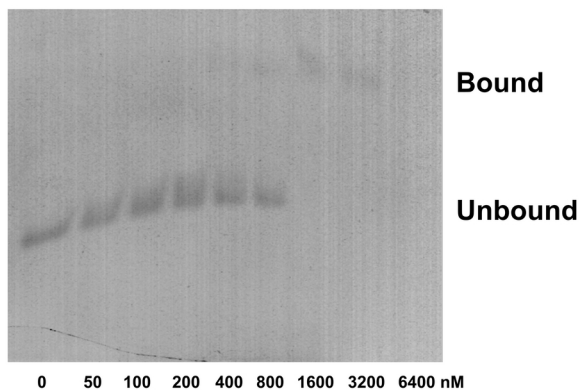
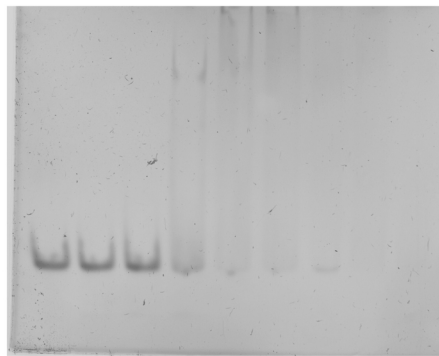


Figure 4.6: EMSA of Myc R20Q single mutant protein

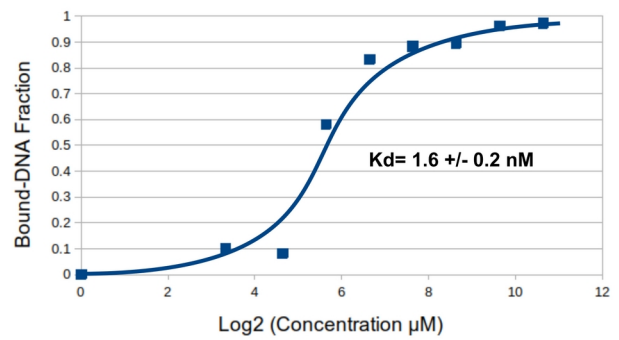
-with wild-type consensus DNA (a), non-consensus single mutant (b) and random DNA substrates (c). The protein concentrations are mentioned below each lane. The fraction of bound DNA (x-axis) as a function of $\log_2$ of protein concentration is plotted on each sub-figure to the right.

(a) 1 2 3 4 5 6 7 8 9



0 10 25 50 100 200 400 800 1600 nM

Bound
Unbound

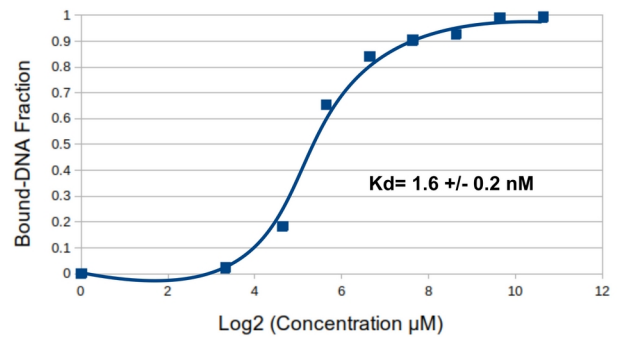


(b) 1 2 3 4 5 6 7 8 9

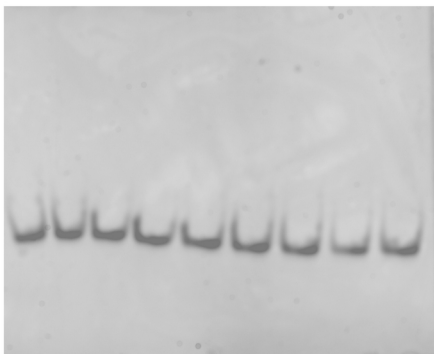


0 10 25 50 100 200 400 800 1600 nM

Bound
Unbound



(c) 1 2 3 4 5 6 7 8 9



0 10 25 50 100 200 400 800 1600 nM

Bound
Unbound

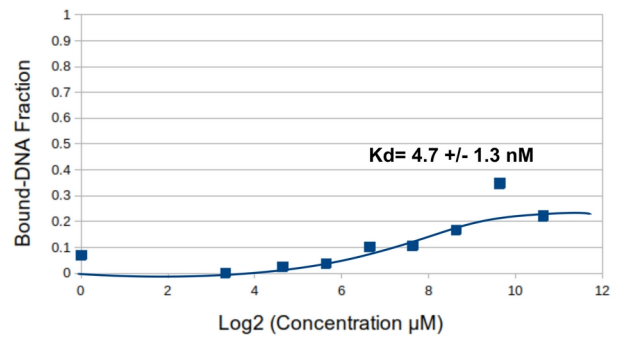


Figure 4.7: EMSA of Myc QY double mutant protein

-with wild-type consensus DNA (a), non-consensus double mutant (b) and random DNA substrates (c). The protein concentrations are mentioned below each lane. The fraction of bound DNA (x-axis) as a function of $\log_2$ of protein concentration is plotted on each sub-figure to the right.

Table 4.21: EMSA results of Myc wild-type and mutant proteins with different substrates.

The values in bold indicate better binding affinity between the mutant and wild-type proteins. The predicted substrate is the non-consensus single mutant substrate for Myc R20Q and non-consensus double mutant substrate for Myc QY protein.

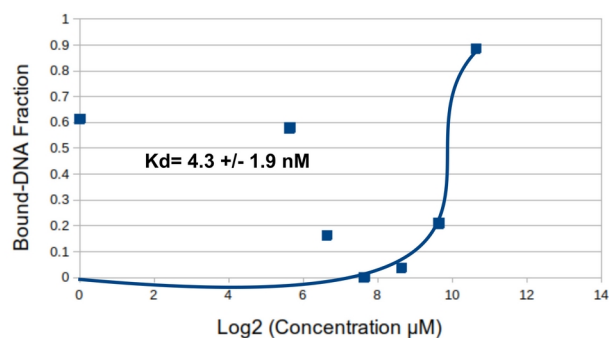
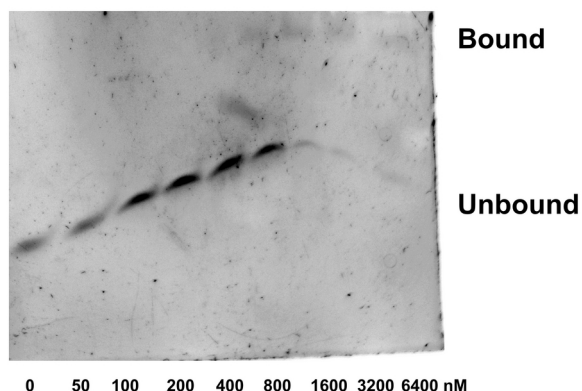
Sr. No	Protein	Substrate	Apparent K _d (nM)	
			Mutant protein	Wild-type protein
1	Myc R20Q single mutant	Wild-type Consensus	6.7 +/- 2.7	1.9 +/- 0.8
		Predicted Substrate	6.3 +/- 2.3	6.6 +/- 2.1
		Random Substrate	5.9 +/- 1.4	4.0 +/- 2.1
2	Myc QY double mutant	Wild-type Consensus	1.6 +/- 0.2	1.9 +/- 0.8
		Predicted Substrate	1.6 +/- 0.2	3.6 +/- 1.6
		Random Substrate	4.7 +/- 1.3	4.0 +/- 2.1

4.3.1.2 EMSA results for Ets1

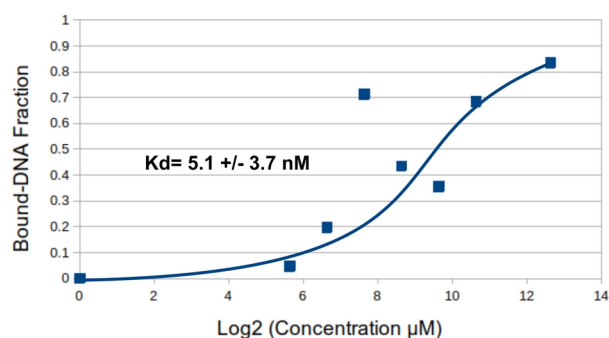
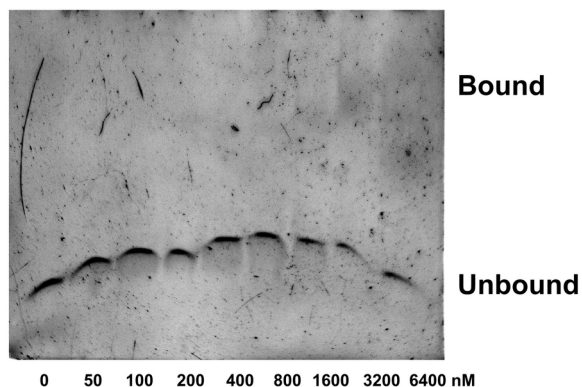
The EMSA for Ets1 wild-type protein showed comparatively more efficient binding to the wild-type consensus sequence (4.3 nM, Figure 4.8). The binding towards the non-consensus substrate predicted to be specific towards the mutant protein was weaker (5.1 nM), yet it was better than the random DNA substrate (44 nM).

The Ets1 Y395R mutant exhibited comparatively better binding towards the wild-type consensus sequence (5.8nM, Figure 4.9). Overall, the wild-type protein had better binding than mutant protein for all substrates, except the random substrate (Table 4.22). However, the Ets1 protein showed a lower binding affinity than the Myc protein.

(a) 1 2 3 4 5 6 7 8 9



(b) 1 2 3 4 5 6 7 8 9



(c) 1 2 3 4 5 6 7 8 9

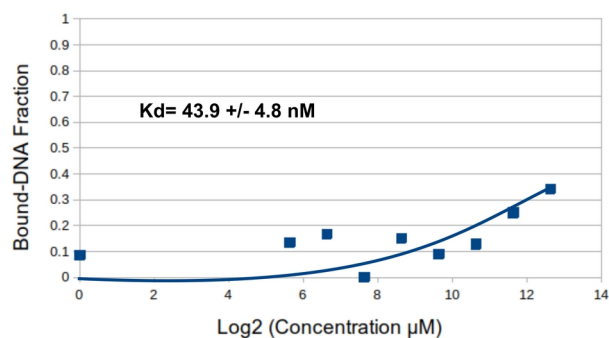
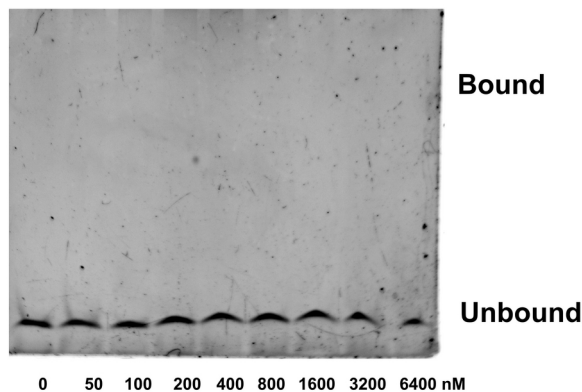
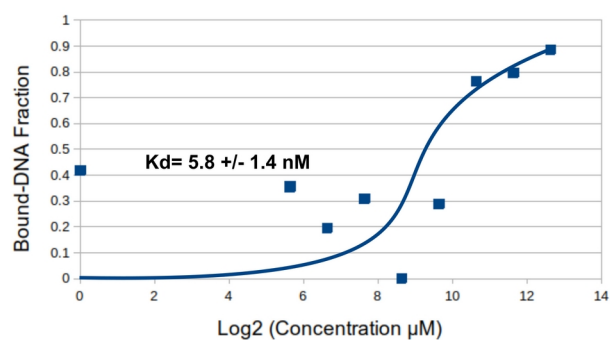
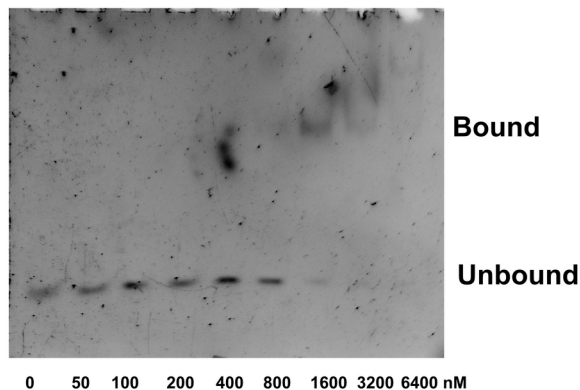


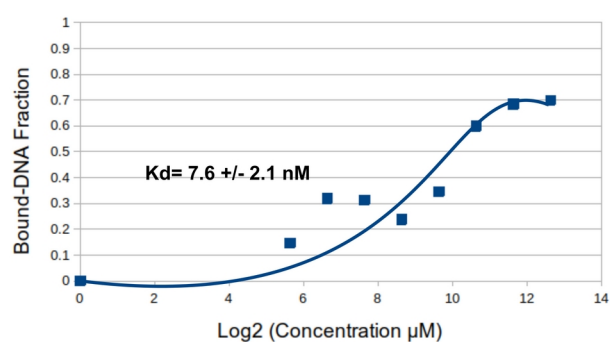
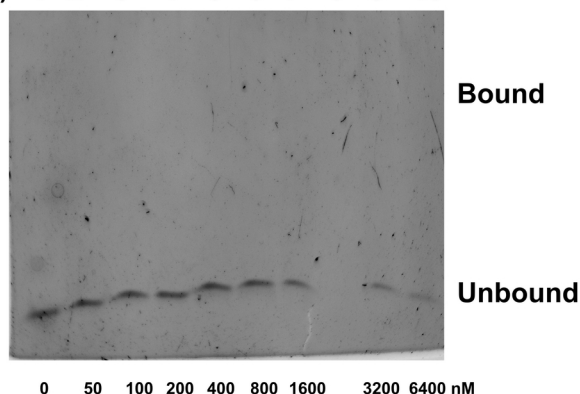
Figure 4.8: EMSA of Ets1 wild-type protein

-with wild-type consensus DNA (a), non-consensus mutant (b) and random DNA substrates (c). The protein concentrations are mentioned below each lane. The fraction of bound DNA (x-axis) as a function of $\log_2$ of protein concentration is plotted on each sub-figure to the right.

(a) 1 2 3 4 5 6 7 8 9



(b) 1 2 3 4 5 6 7 8 9 10



(c) 1 2 3 4 5 6 7 8 9

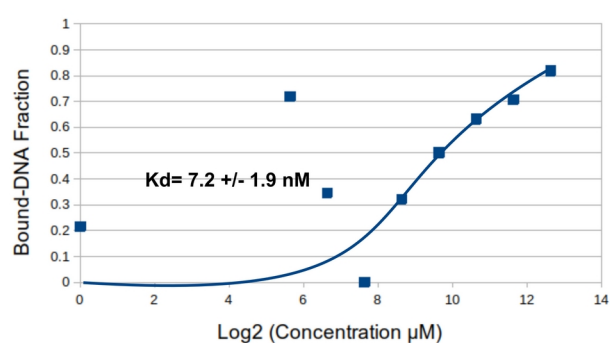
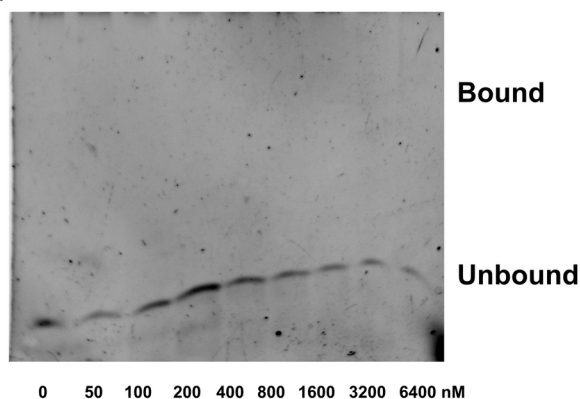


Figure 4.9: EMSA of Ets1 Y395R mutant protein

-with wild-type consensus DNA (a), non-consensus mutant (b) and random DNA substrates (c). The protein concentrations are mentioned below each lane. The fraction of bound DNA (x-axis) as a function of $\log_2$ of protein concentration is plotted on each sub-figure to the right. Note: Lane 8 in Figure (b) is empty.

Table 4.22: EMSA results of Ets1 proteins with different substrates.

The better binding affinity values for each DNA substrate are highlighted in bold.

Sr. No	Protein	Substrate	Apparent K_d (nM)	
			Mutant protein	Wild-type protein
1	Y395R single mutant	Wild-type Consensus	5.8 +/- 1.4	4.3 +/- 1.9
		Predicted Substrate	8.3 +/- 1.8	5.1 +/- 3.7
		Random Substrate	7.2 +/-1.9	43.9 +/- 4.8

4.3.2 Microscale Thermophoresis (MST)

MST is an *in-vitro* fluorescence-based technique to quantify the binding of biomolecules (described in the introduction to techniques, section 9) [196,199]. Generally, the ligand is fluorescently labelled, which is DNA in our case. However, we had purchased DNA substrates without a label. Therefore, we labelled the proteins using an amine-labelling dye (NHS) kit (NanoTemper Technologies). The dye has a reactive NHS-ester group that forms covalent bonds with primary amines (lysine residues in proteins).

To label the proteins, we made a 100 μ L solution of 10 μ M concentration. In the case of Myc, the protein was purified in Tris buffer and had DTT in the dialysis buffer, while Ets1 did not have Tris but had DTT. These two components interfere with labelling. Therefore, we used a buffer exchange column to remove them. We first added 200 μ L of 150mM HEPES (pH 7.0) and centrifuged the column for 1 minute at 1500 g. We then discarded the flowthrough and repeated this twice. After this, we added the protein and centrifuged it for 2 minutes at 1500 g. The flowthrough with our buffer-exchanged protein was collected in a 1.5 mL tube. The dye was in powder form, and it was resuspended in 30 μ L of 100% dimethyl sulfoxide (DMSO). This generated a dye with a concentration of 470 μ M. In order to ensure that all the proteins get labelled, we add five times more dye than protein. Since our protein concentration was 10 μ M, we added 10 μ L of 470 μ M dye (50 μ M). The mixture was kept in the dark for 30 minutes.

The excess dye had to be removed from the mixture. Therefore, we used a second column (also from the kit). We equilibrated the column by adding 2 mL 150 mM HEPES buffer (pH 7.0) thrice. The column has an outlet that is closed with a stopper. The buffer

was allowed to move out through gravity. After this, we added the labelled protein and 300 μL of 150 mM HEPES buffer. The first few drops were ignored. The remaining was collected in 1.5 mL tubes.

The highest annealed DNA concentration for Myc was 40 μM , and for Ets1, it was 20 μM . We made 1:1 serial dilutions of DNA for all proteins. The total reaction volume was 10 μL , so we added 5 μL each of the DNA and protein. The DNA concentration ranged from 6×10^{-10} to 1×10^{-5} M for Ets1, whereas for Myc, it was 6×10^{-10} to 2×10^{-5} M. We used blue and red labelling dyes with excitation wavelengths 495 and 521 nm, respectively. The experiment was run on Monolith NT.115 with 100% excitation power and 40% MST power. The accompanying software from NanoTemper Technologies (MO.Affinity Analysis v3.0.3) was used to measure the binding affinities (Table 4.23). The software calculates the binding affinity using FNorm function-

$$F_{norm} = \frac{F_1}{F_0} \quad \text{Eq, 2}$$

where F_0 and F_1 correspond to the fluorescence values before and after switching on the laser, respectively.

While fitting the K_d model, if the highest concentration shows an upward trend, then the FNorm values are plotted. However, if the curve indicates saturation, then the maximum fluorescence value is assumed to correspond to the bound state. The FNorm values are normalized to plot the fraction of DNA-bound.

The binding reaction was carried out in phosphate buffer for Myc R20Q since labelling was performed in that buffer. Both the wild-type and Myc R20Q mutant proteins exhibit better binding to the random DNA compared to the consensus and predicted substrates. However, for both proteins, the binding to the two substrates is similar. The difference in binding of Myc R20Q to consensus and predicted DNA is 1.4 μM , while the standard deviations are 4.4 and 6.8 μM (Table 4.23, Figure 4.11). Similarly, the wild-type protein binding to the two substrates has a binding affinity difference of 0.8 μM , with standard

deviations of 0.8 and 3 μM . (Figure 4.10). This implies that the wild-type protein does not differentiate between these two substrates and neither does the Myc R20Q mutant.

The Myc QY double mutant protein aggregated in a solution without Tris. Therefore, after labelling, the binding reaction was carried out in Tris buffer. The reactions with Myc wild-type protein were also carried out in Tris for comparison. The Myc QY double mutant shows better binding towards the predicted substrate (non-consensus double mutant) compared to the wild-type consensus DNA substrate. The binding of Myc QY double mutant to the random DNA substrate could not be detected by MST (Figure 4.12). In the case of the Myc wild-type protein with the Tris buffer, the binding affinity is better than that in the phosphate buffer, but in both cases, in the duplicate experiment binding was not seen. However, the trend remains the same. The predicted non-consensus double mutant substrate binds better than the wild-type consensus sequence. The random sequence has a ~ 3 -fold higher K_d value than the non-consensus double mutant substrate.

The Ets1 proteins showed aggregation after labelling with the NHS dye. We tried changing different compositions of buffer (phosphate buffer with Tween, Tris buffer). Finally, we re-labelled the proteins and immediately performed the experiment. While the 20-second time for the experiment worked for Ets1 wild-type protein, we used 10-second experiment for the Ets1 Y395R mutant to avoid aggregation.

Neither Ets1 wild-type nor Y395R mutant binds the random DNA. The wild-type protein binding to its consensus shows a large variation of 0.4 μM compared to its binding affinity i.e., 0.3 μM . However, the predicted mutant binds to the non-consensus DNA with twice the affinity (Figures -4.13, 4.14).

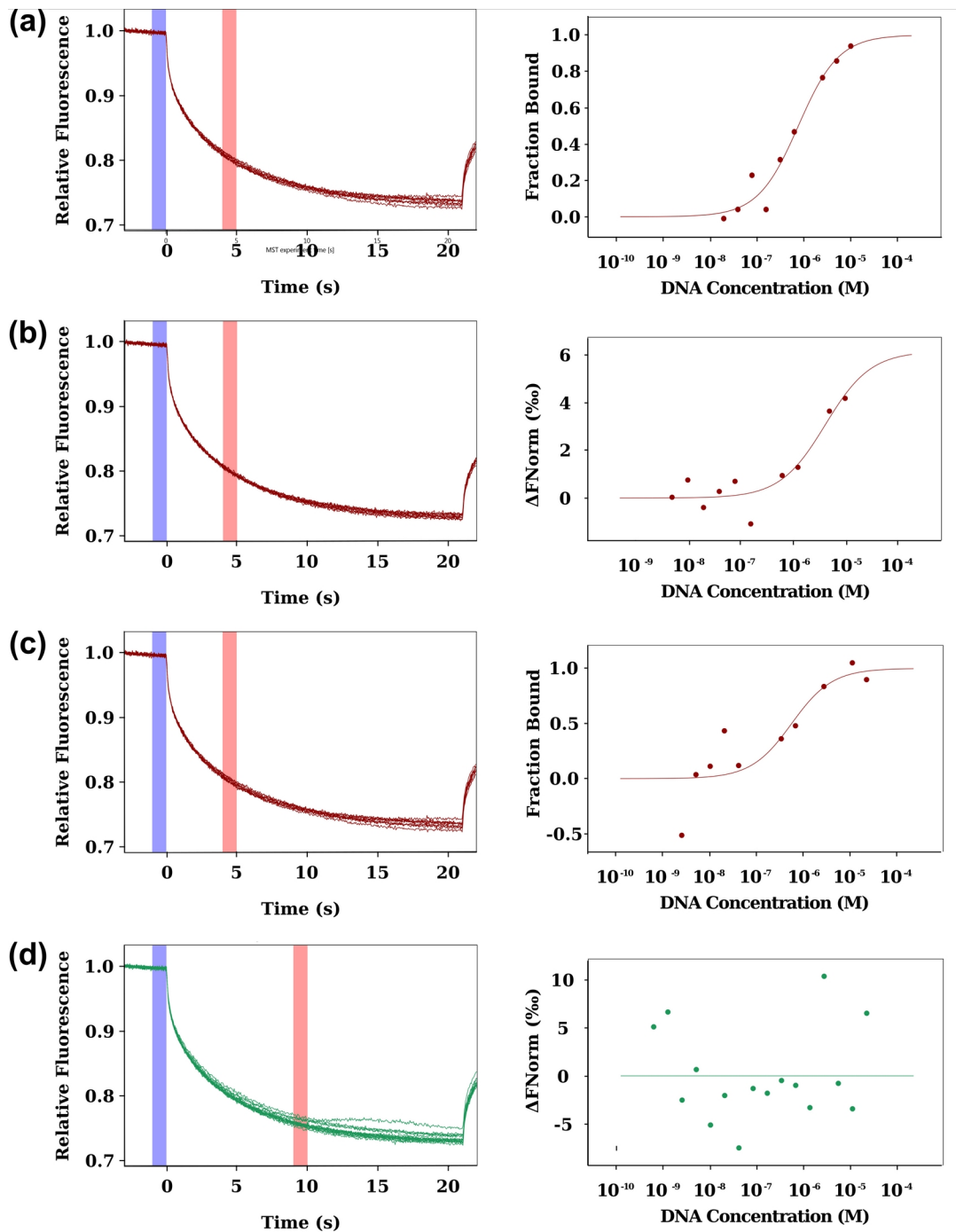


Figure 4.10: MST data for Myc wild-type protein

-interacting with wild-type consensus (a), non-consensus single mutant (b), non-consensus double mutant (c), and random DNA substrates. The left side of each sub-figure shows the relative fluorescence values during an MST experiment. The binding affinity calculation involved the region between the blue and red bars. The right side shows

the curve used to calculate the binding affinities with the DNA concentration on the x-axis and the fraction bound or F_{Norm} on the y-axis. The data that could not fit a binding curve is depicted in green (d).

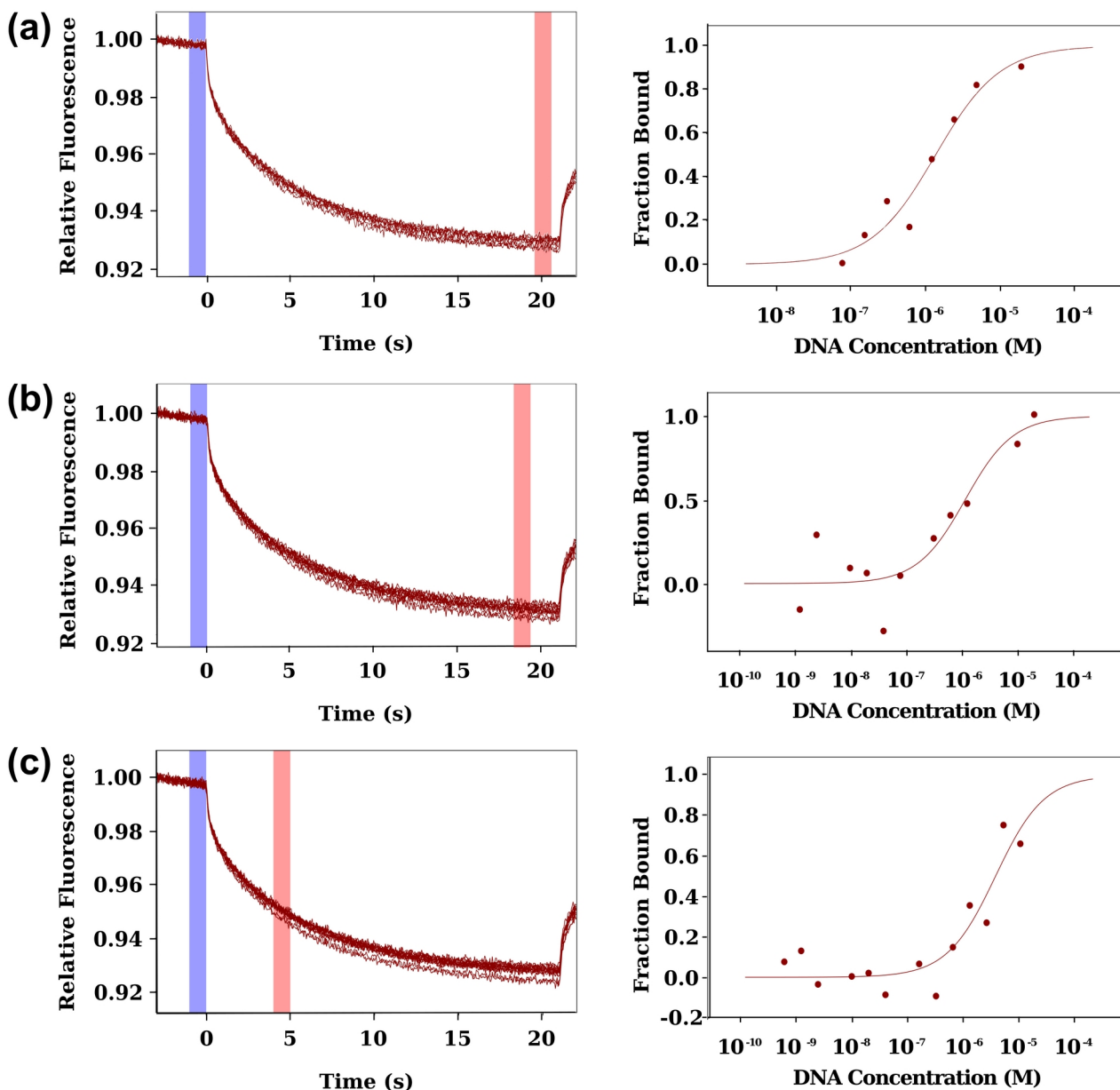


Figure 4.11: MST data for Myc R20Q single mutant protein -interacting with wild-type consensus (a), non-consensus single mutant (b) and random DNA substrates (c). The left side of each sub-figure shows the relative fluorescence values during an MST experiment. The binding affinity calculation involved the region between the

blue and red bars. The right side shows the curve used to calculate the binding affinities with the DNA concentration on the x-axis and the fraction bound or FNorm on the y-axis.

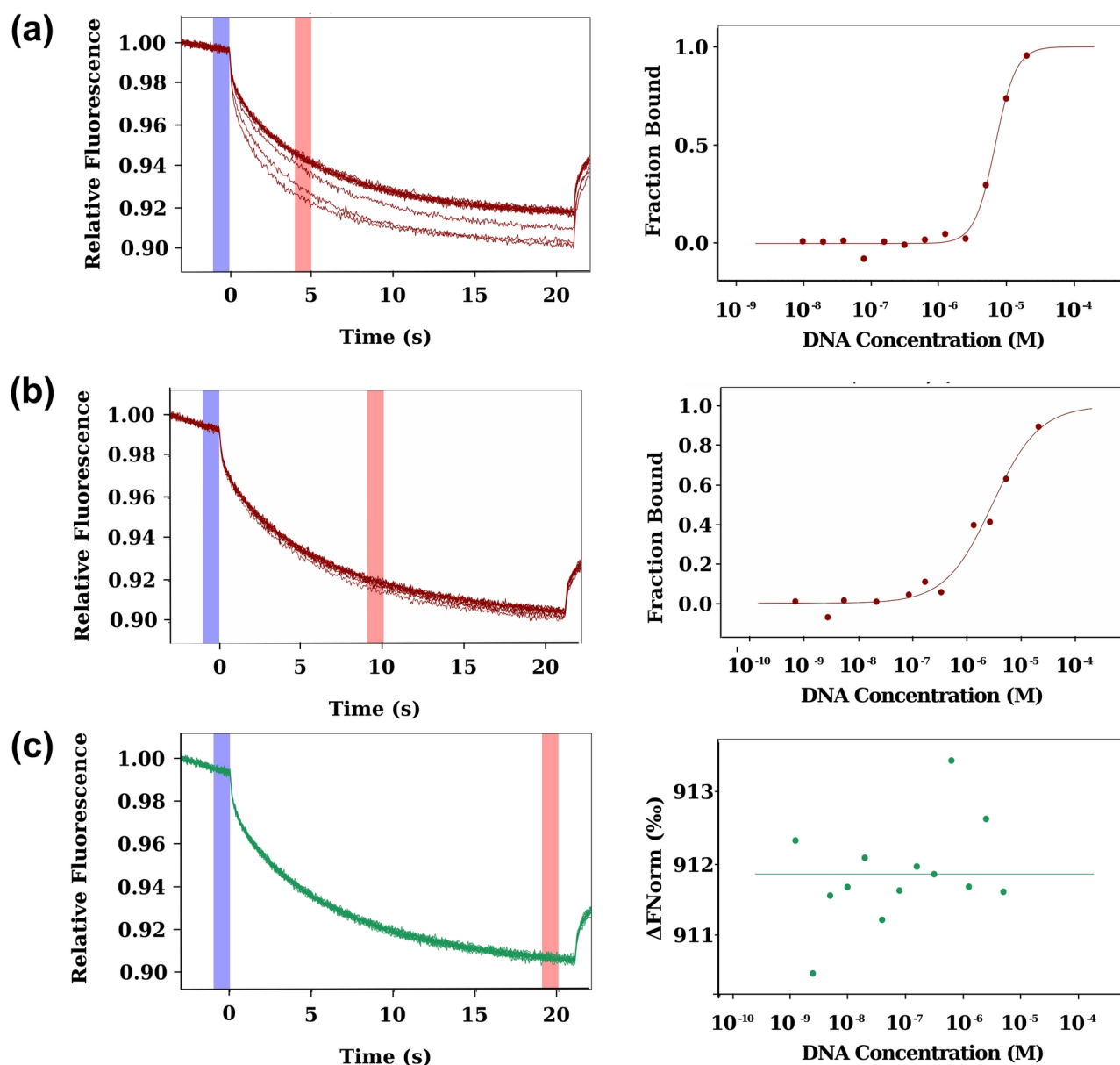


Figure 4.12: MST data for Myc QY double mutant protein -interacting with wild-type consensus (a), non-consensus single mutant (b) and random DNA substrates (c). The left side of each sub-figure shows the relative fluorescence values during an MST experiment. The binding affinity calculation involved the region between the blue and red bars. The right side shows the curve used to calculate the binding affinities

with the DNA concentration on the x-axis and the fraction bound or FNorm on the y-axis. The data that could not fit a binding curve is depicted in green (c).

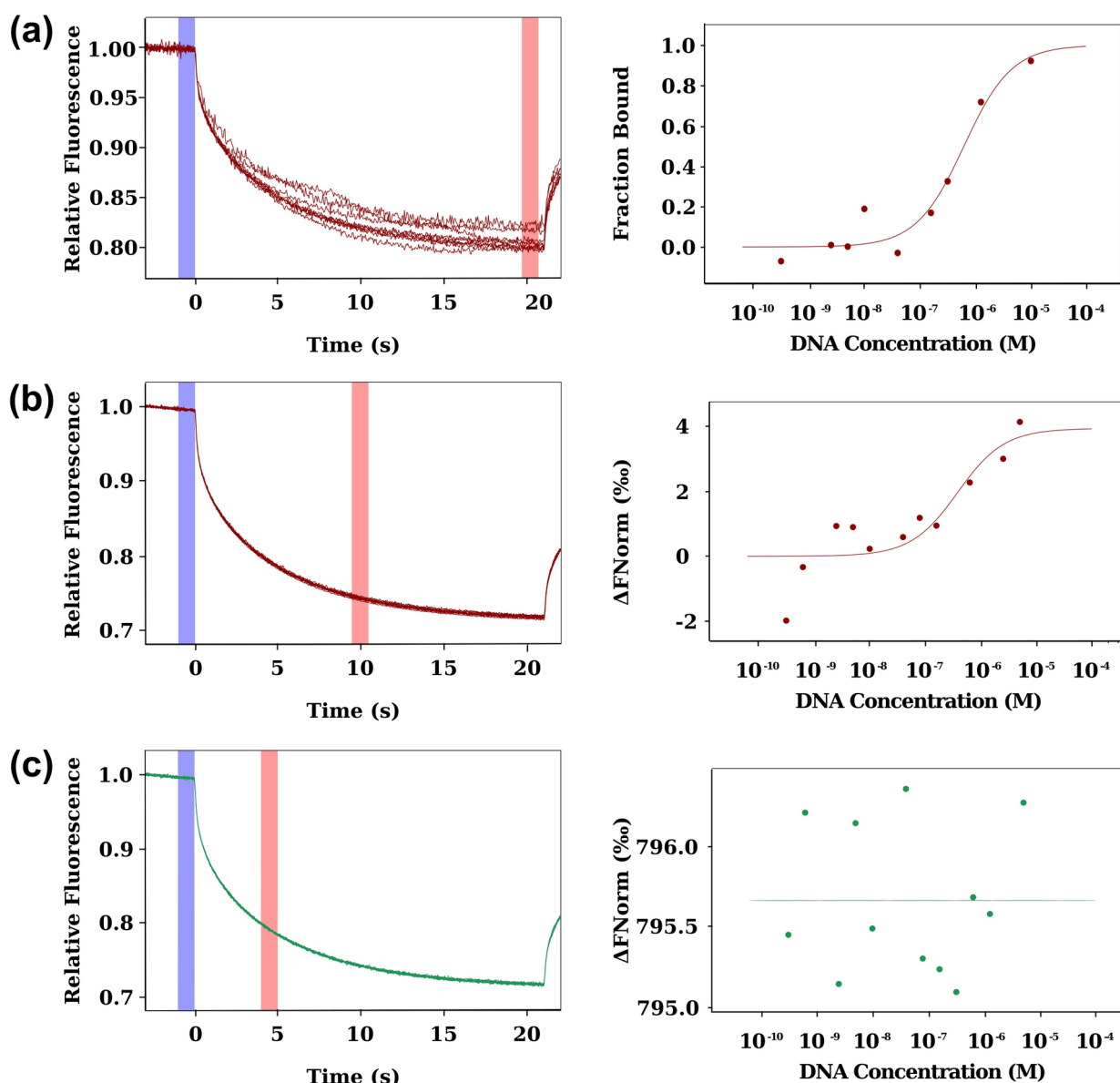


Figure 4.13: MST data for Ets1 wild-type protein

-interacting with wild-type consensus (a), non-consensus single mutant (b) and random DNA substrates (c). The left side of each sub-figure shows the relative fluorescence values during an MST experiment. The binding affinity calculation involved the region between the blue and red bars. The right side shows the curve used to calculate the binding affinities

with the DNA concentration on the x-axis and the fraction bound or F_{Norm} on the y-axis. The data that could not fit a binding curve is depicted in green (c).

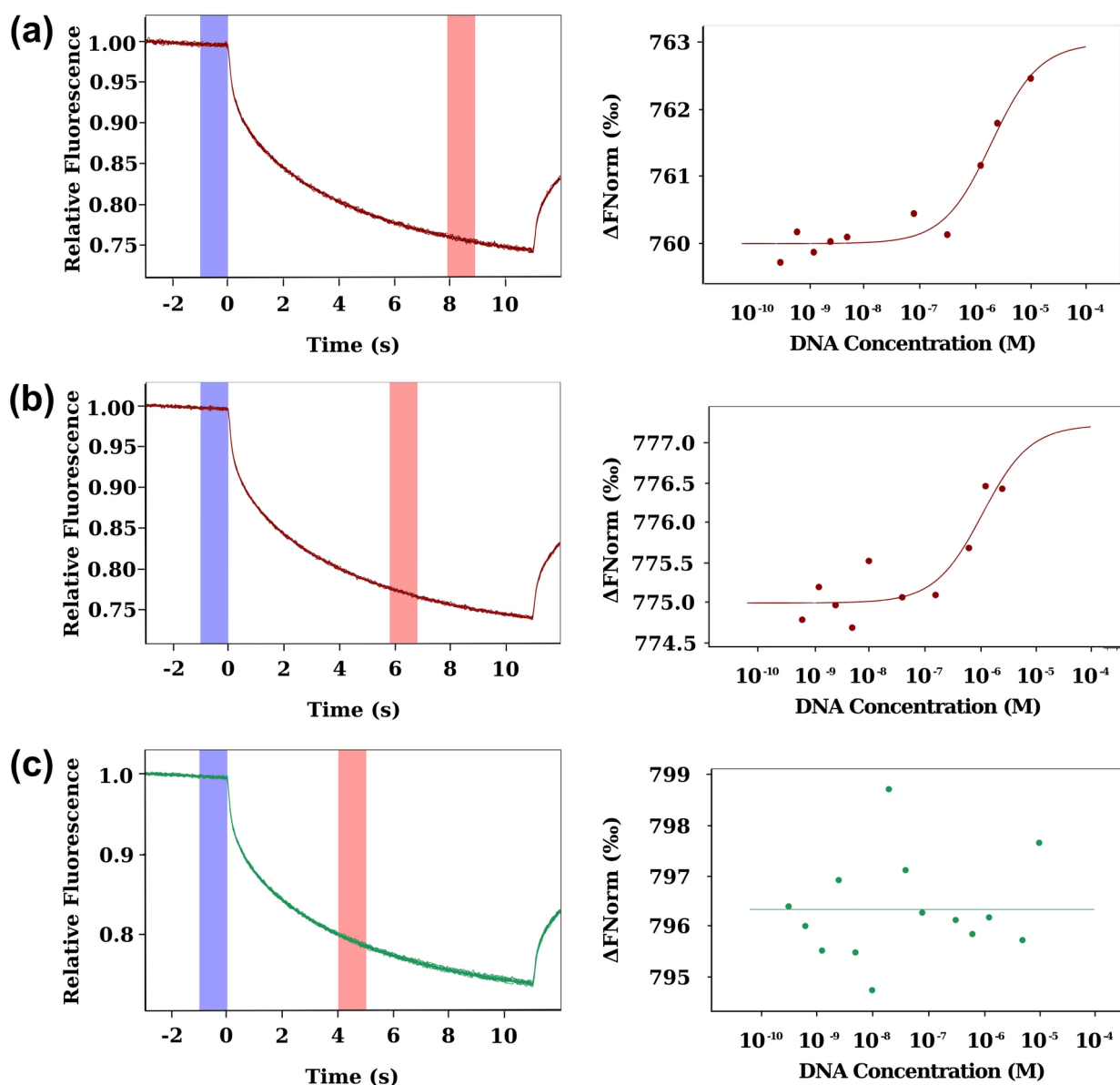


Figure 4.14: MST data for *Ets1* Y395R protein

-interacting with wild-type consensus (a), non-consensus single mutant (b) and random DNA substrates (c). The left side of each sub-figure shows the relative fluorescence values during an MST experiment. The binding affinity calculation involved the region between the blue and red bars. The right side shows the curve used to calculate the binding affinities with the DNA concentration on the x-axis and the fraction bound or F_{Norm} on the y-axis. The data that could not fit a binding curve is depicted in green (c).

Table 4.23: Binding affinity values from the MST experiment.

The better binder between the wild-type and mutant proteins for each substrate is highlighted in bold. 'No fit' indicates reactions where the data does not fit a binding curve and therefore the K_d cannot be determined. * indicates no binding was detected in one of the duplicates. # indicates data unavailable for the experiment

Sr. No	Protein	Substrate	K_d (μ M)	
			Mutant protein	Wild-type protein
1	Myc R20Q single mutant	Wild-type Consensus	4.5 +/- 4.4	5.0 +/- 0.8
		Predicted	5.9 +/- 6.8	4.2 +/- 3.0
		Random	4.0 +/- 0.6	1.6 +/- 0.9
2	Myc QY double mutant	Wild-type Consensus	7.1 +/- 0.4	0.4 +/- 0.4
		Predicted	3.6 +/- 1.3	0.2 +/- 0.3
		Random	No fit	1.8*
3	Ets1 Y395R	Wild-type Consensus	1.1 +/- 1.1	0.31 +/- 0.4
		Predicted	0.7 +/- 0.5	0.36 +/- 0.01
		Random	No fit	No fit

4.3.3 Nano Differential Scanning Fluorimetry (nanoDSF)

NanoDSF is an *in-vitro* technique to estimate the stability of proteins (refer to Chapter 1, section 1.6.10) [202,204]. In our case, we used it to check the protective effect of ligands on the unfolding of proteins.

We ran the experiment on Prometheus NT.48 and varied the temperature between 20 to 95 °C with a heating ramp of 1°C/minute. The intrinsic fluorescence of tryptophan and tyrosine in the presence and absence of different DNA substrates indicated the protein conformation. The reaction mixture consisted of 10 μ L each of 10 μ M DNA and 10 μ M protein (in a protein only reaction, we used 20 μ L of protein instead). We loaded 10 μ L of this into a capillary and measured the fluorescence intensity at 350 and 330 nm. Using the ratio of these intensities and the first derivative of this ratio, the software outputs a T_m value (peak of the curve).

A higher T_m value compared to the T_m value of the protein alone indicates the stabilizing effect of the DNA. In the case of the Myc wild-type protein, the highest T_m value is with the

wild-type consensus sequence. The Myc R20Q single mutant has the best T_m for the protein alone, while the Myc QY double mutant has the best T_m for the non-consensus sequence. Rather than having a protective effect, the T_m value seems to be lower for Myc R20Q. The measurement of T_m is through the absorbance of aromatic residues. The occlusion of aromatic groups could be one of the reasons for this apparent reduction in T_m upon DNA-binding.

Ets1 wild-type protein shows a higher T_m for the non-consensus substrate. However, the difference between the wild-type and non-consensus substrates is 0.4 °C and within experimental error limits. However, the Ets1 Y395R mutant has a better T_m with the wild-type substrate than other DNA substrates (Table 4.24, Figures 4.15 and 4.16).

Table 4.24: T_m values measured by nanoDSF.

The highest T_m value for each protein is highlighted in bold.

Sr. No	Protein	Substrate	T_m (°C)		ΔT_m (°C)	
			Mutant	Wild-type	Mutant	Wild-type
1	Myc R20Q single mutant	-	61.5	48.0	-	-
		Wild-type Consensus	60.2	51.0	-1.3	3.0
		Predicted Substrate	60.3	49.1	-1.2	1.1
		Random Substrate	61.1	50.7	-0.4	2.7
2	Myc QY double mutant	-	43.5	48.0	-	-
		Wild-type Consensus	46.7	51.0	3.2	3.0
		Predicted Substrate	47.1	48.6	3.6	0.6
		Random Substrate	46.1	50.7	2.6	2.7
3	Ets1 Y395R	-	52.9	53.2	-	-
		Wild-type Consensus	55.6	54.2	2.7	1.0
		Predicted Substrate	54.9	54.6	2.0	1.4
		Random Substrate	55.0	53.8	2.1	0.6

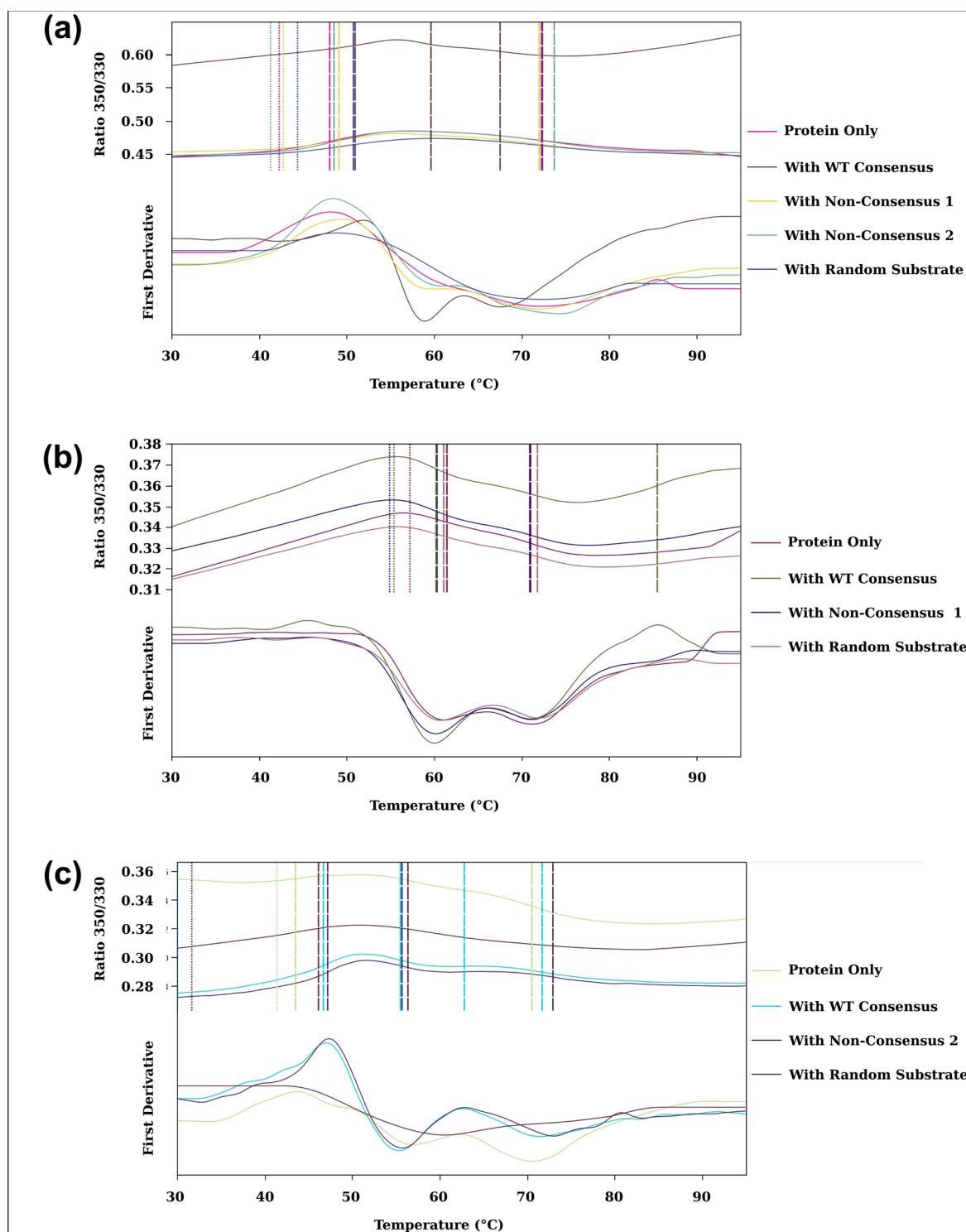


Figure 4.15: NanoDSF data for Myc

-wild-type (a), Myc R20Q single mutant (b) and Myc QY double mutant (c) proteins. The top section of each figure shows the fluorescence intensity ratio at 350 and 330 nm

wavelengths. The bottom section shows the first derivative of this ratio. The non-consensus substrate 1 corresponds to the non-consensus single mutant substrate for Myc single mutant, and the non-consensus substrate 2 is the non-consensus double mutant substrate for Myc double mutant. The colours in the legend correspond to either protein alone or with the different substrates.

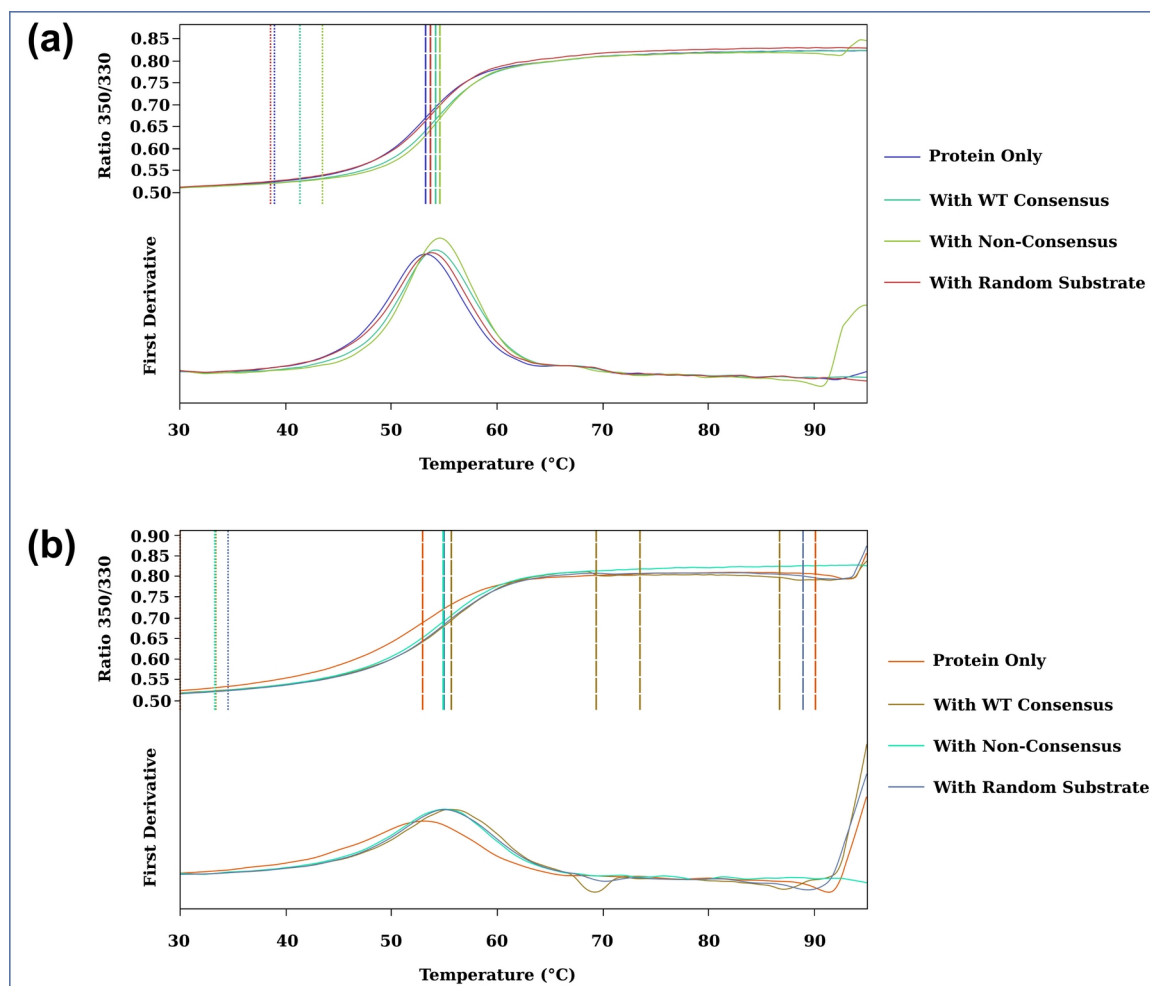


Figure 4.16: NanoDSF data for *Ets1*

-wild-type (a) and *Ets1* Y395R mutant (b) proteins. The top section of each figure shows the fluorescence intensity ratio at 350 and 330 nm wavelengths. The bottom section shows the first derivative of this ratio. The colours in the legend correspond to either protein alone or with the different substrates.

We calculated the ΔT_m for each protein-DNA substrate by subtracting the T_m value of the protein alone from that of the complex. The ΔT_m can indicate the effect of that DNA sequence on the wild-type and mutant proteins. For Myc QY mutant, the stabilizing effect of the wild-type consensus and random DNA is similar to the wild-type and mutant proteins (ΔT_m is similar). However, the ΔT_m for Myc QY double mutant with the predicted non-consensus is six times that of the wild-type protein. We did not observe a difference of this magnitude in any other case.

4.4 Comparison of binding with different *in-vitro* techniques

We have analysed the binding between protein and DNA using three different techniques. EMSA was semi-quantitative, MST was quantitative, and nanoDSF was an indirect assay. Finding the same observation using different methods increases confidence in the results. Therefore, we compared the binding of mutant and wild-type proteins towards each substrate (Table 4.25).

- Myc R20Q protein

Myc R20Q single mutant protein exhibited anomalous behaviour in nanoDSF where the DNA-binding reduces the T_m value of the protein. As discussed previously, the T_m is detected using intrinsic fluorescence, which could be occluded due to ligand binding. Therefore, we only compared the EMSA and MST results for this mutant. The interaction of the mutant protein with the wild-type consensus DNA substrate is similar to the wild-type protein in the MST experiment (K_d ratio 1.1). In the EMSA experiment, the values of the apparent K_d for the mutant and wild-type proteins differ by 4.8 nM, which is larger than the standard deviations (2.7 nM). Therefore, it is difficult to reliably conclude the actual preference of the Myc R20Q protein towards the two substrates compared to the wild-type protein.

Myc R20Q protein interaction with the predicted non-consensus substrate is better than that of the wild-type protein in EMSA but not in MST experiments. In EMSA, the difference between apparent K_d of wild-type and mutant proteins with the

substrate is 0.3 nM and within the standard deviation (~ 2 nM). Similarly, there is a large standard deviation of 6.8 μ M associated with the binding of Myc R20Q to the predicted substrate.

The binding of random substrate to Myc R20Q mutant and wild-type proteins is better than the cognate or predicted DNA. The random DNA sequence is "GGTCCGAAGATTCTAGCATAAT" which could putatively contain another signature of Myc binding regions apart from the E-box. Both the Myc R20Q mutant and Myc wild-type proteins exhibit similar trends of binding. This implies that one mutation is perhaps insufficient to change the specificity of a transcription factor such as Myc.

- Myc QY double mutant protein

Myc QY double mutant protein interaction with different substrates showed the same trend with EMSA and nanoDSF but not with MST. The wild-type and the Myc QY double mutant proteins had a similar affinity towards the wild-type consensus DNA substrate (difference 0.3 nM, standard deviation 0.8) in EMSA. However, the wild-type protein exhibited ~ 10 -fold better affinity towards the wild-type consensus DNA substrate than Myc QY double-mutant protein in MST. In nanoDSF, the difference between ΔT_m of the wild-type and Myc QY double mutant protein with the wild-type consensus DNA substrate is 0.2 $^{\circ}$ C (higher for Myc QY double mutant). Therefore, both EMSA and nanoDSF showed that the Myc QY double mutant protein was a better binder to the wild-type consensus DNA sequence, but the difference in affinity is small. However, there was a 10-fold better binding by the wild-type protein in the MST experiment, indicating that the wild-type protein was a better binder to the wild-type consensus sequence than Myc QY double mutant.

The binding of Myc QY double mutant protein with the predicted non-consensus double mutant DNA substrate was better than that of the wild-type proteins as measured by EMSA and nanoDSF. MST showed that the wild-type protein bound to this substrate at a five-fold better affinity. However, the affinity of the wild-type protein was better than the Myc QY double mutant protein in all three substrates. The ΔT_m of the Myc QY double mutant protein with the non-consensus double

mutant DNA substrate was six times that of the wild-type protein. Moreover, in EMSA, the binding affinity for the interaction of Myc QY double mutant protein with the non-consensus double mutant DNA substrate was twice that of the wild-type protein. Thus, we can conclude that the binding of the Myc QY double mutant protein is more favourable towards the predicted non-consensus double mutant DNA substrate than the wild-type protein.

Myc QY double mutant showed similar binding towards the random DNA substrate as the wild-type protein with EMSA and nanoDSF. The ΔT_m of the Myc QY double mutant and wild-type protein differed by 0.1 °C (higher for wild-type protein). Also, the apparent K_d from EMSA differed by 0.7 nM (standard deviation 2.1 nM). Thus, there was no difference in binding the random substrate when these two methods were considered. However, MST showed wild-type protein binding to the random substrate (1.8 μ M) but not Myc QY double mutant. Therefore, since there was a significant difference in the binding in MST while negligible in others, we would consider the MST results. The wild-type protein has a better affinity for the random sequence than Myc QY double mutant.

- Ets1 Y395R mutant protein

The Ets1 wild-type protein had a better affinity toward the wild-type consensus sequence than the Ets1 Y395R mutant in EMSA and MST experiments. In nanoDSF, the Y395R mutant had a better ΔT_m than the wild-type protein when binding to the wild-type consensus DNA substrate (~3 fold). With EMSA, even though the binding of the wild-type protein to the wild-type consensus was better (with the difference of 1.3 nM), it was in the standard deviation range (1.9 nM). Since nanoDSF is an indirect measure of assessing binding and MST is a more sensitive instrument, we concluded that the wild-type protein had a better affinity toward the wild-type consensus sequence.

With the non-consensus substrate, the wild-type Ets1 protein had a better affinity compared to the Y395R mutant in both EMSA and MST experiments. However, in nanoDSF, the ΔT_m was higher for the Ets1 Y395R mutant binding the non-

consensus DNA substrate (0.6 °C). In EMSA, the difference between the affinities of Ets1 wild-type and Y395R mutant proteins towards the non-consensus DNA substrate was within the standard deviation range. The MST experiment showed three-fold better binding of wild-type protein with non-consensus DNA substrate than Ets1 Y395R mutant. Since the difference in ΔT_m was not very high, we speculated that the binding of Ets1 wild-type protein to the non-consensus DNA substrate was better than Ets1 Y395R.

There was no binding of Ets1 wild-type and Ets1 Y395R mutant to the random DNA substrate in an MST experiment. However, in both nanoDSF and EMSA, the Ets1 Y395R mutant protein had a better ΔT_m and apparent K_d , respectively, for binding to the random substrate. Thus, the Ets1 Y395R mutant protein had a comparatively higher propensity of binding a random DNA substrate.

So far, we have discussed the comparative analysis of the wild-type and mutant proteins with different substrates. However, we also need to analyse the preferred substrate for each protein in the different experiments. The Myc and Ets1 wild-type proteins preferred binding their respective consensus sites in EMSA, not in MST. In nanoDSF, Myc wild-type was stabilized the most by the wild-type consensus. However, Ets1 had a comparable stabilization with the wild-type consensus and predicted non-consensus DNA substrates.

The Myc R20Q single mutant preferred binding to the predicted non-consensus single mutant and the consensus substrate in both MST and EMSA. The Myc QY double mutant preferred the predicted non-consensus double mutant DNA substrate in all three in-vitro experiments. However, in the Ets1 Y395 mutant, results for EMSA and nanoDSF showed that it bound preferably to the wild-type consensus sequence rather than the predicted sequence.

Thus, if we only consider protein binding to the predicted substrate, we have been successful in changing the specificity of one protein, In one case, the change was not quantified as different due to high standard deviation of the experiment and potentially less difference in binding specificity.

Table 4.25: Comparison of results from different in-vitro techniques.

The predicted substrate is the non-consensus single mutant DNA substrate and non-consensus double mutant DNA substrate for Myc R20Q single and Myc QY double mutant proteins, respectively. The values indicate the ratio of wild-type K_d or ΔT_m to that of the mutant protein. Pink cells indicate that the results from this method are not to be considered (see text). '#' indicates that there was no binding in the wild-type. '§' indicates no binding in the mutant. The bold values reflect better binding in mutant compared to the wild-type protein.

Sr No.	Protein	Substrate	EMSA	MST	NanoDSF
1	Myc R20Q	WT Consensus	0.28	1.1	
2		Predicted	1.05	0.7	
3		Random	0.68	0.4	
4	Myc QY	WT Consensus	1.19	0.06	0.94
5		Predicted	2.25	0.05	0.17
6		Random	0.85	§	1.04
7	Ets1 Y395R	WT Consensus	0.74	0.3	0.37
8		Predicted	0.61	0.5	0.7
9		Random	6.10	#§	0.28

4.5 Discussions

This chapter has discussed the *in-vitro* experiments performed to validate the specificity preferences described in Chapter 2. We selected two transcription factors- Ets1 and Myc and a restriction enzyme -EcoRV for the validation. We changed the nucleotide sequence of the preferred binding site of these proteins and used our algorithm to predict the amino acids that must be mutated in the proteins to bind the new non-consensus DNA.

We tested these predictions using in-vitro DNA-binding assays. The in-vitro assays required proteins to be purified. However, due to standardization problems, we could not purify the EcoRV restriction enzyme. One of the strategies to employ in the future is to have the methyltransferase on the same plasmid upstream of the restriction enzyme. This strategy would ensure that the methyltransferase gets transcribed first and then the

restriction enzyme, protecting the host bacterium from cleavage. We cloned and purified the wild-type and mutant proteins for Ets1 and Myc.

The assays performed were EMSA, MST and nanoDSF. Overall, the EMSA reactions do not show a significant difference in binding among the different substrates and proteins. Considering the standard deviations, we can confidently conclude relative binding affinity for five pairs of protein-DNA complexes. The Myc wild-type protein prefers the wild-type consensus sequence over the predicted non-consensus single mutant DNA substrate. The Myc wild-type protein also has a better binding to the wild-type consensus protein than the Myc R20Q single mutant protein. The predicted non-consensus double mutant DNA substrate has a better binding to the Myc QY double mutant protein as compared to a random DNA. The binding affinity between the Myc QY double mutant and the predicted non-consensus double mutant DNA substrate is better than that of the wild-type protein for this substrate. The Ets1 Y395R also shows a higher non-specific binding activity than the Ets1 wild-type protein. Out of 18 possible comparisons, only five are significantly different. This shows that the technique is not very sensitive in detecting differences in binding.

When we analysed the results from different experiments, MST results do not match with EMSA in any of the cases where the mutant protein has a better binding (Table 4.25). Another observation was that the wild-type protein bound to the predicted non-consensus DNA substrate with a better affinity than the wild-type consensus sequence. This has not been observed in EMSA. Therefore, if this is not an artefact of fluorescence labelling it could be because both these proteins are transcription factors. Regulation of gene expression necessitates specific binding and dissociation at the appropriate signals. A strong association may hinder the recognition process, and therefore these proteins have a comparatively lower binding affinity towards the wild-type consensus sequence.

NanoDSF relies on the redshift of aromatic residues upon unfolding to measure the T_m . Myc lacks tryptophan residues and entirely relies on the tyrosine residues to provide intrinsic fluorescence. We have observed that for the Myc R20Q mutant, upon addition of the DNA substrate, there is a reduction of T_m . This reduced T_m could either arise because

the protein is destabilized upon complexation with DNA or the occlusion of tyrosine residues by the binding mode. We believe that it is more likely for the tyrosine residues to be occluded. Thus, we did not consider the nanoDSF results for Myc R20Q protein. Apart from Myc QY double mutant interaction with the random DNA substrate, all other substrates show an increase in ΔT_m values with the mutant proteins compared to the wild-type proteins. It could be interpreted as the mutants having a better affinity than the wild-type protein towards these substrates.

We had predicted that mutation from arginine to glutamine at residue number 20 would make the protein recognize the CATATG site. However, in both EMSA and MST, the binding of the mutant and wild-type proteins to the consensus and non-consensus DNA is within standard deviation. This is in line with our method since we have a probabilistic scoring scheme. We have a list of favourable interactions and all of them are preferred to a degree. We believe that due to the broad specificity of Myc, one mutation was not sufficient to change the specificity of the enzyme. The result of Myc double mutant showing the expected change lends some credence to this hypothesis. It has lower binding affinities than the wild-type protein. However, we have not attempted to quantify the difference in the binding affinities when a nucleotide-amino acid pair is replaced by another. Therefore, comparing the wild-type protein affinities with the mutant protein affinities would not validate our method. Instead, a more appropriate validation would be whether the predicted sequence is a better binder than other sequences. We have achieved this validation for Myc QY double mutant protein. Thus, we could validate one of the three predictions using in-vitro binding assays with confidence and cannot comment upon the second one since the difference may be too small to be quantified.

4.6 Acknowledgements

The experimental techniques were taught primarily by Sujata Sharma and Vinayak Tumuluri. Atreyi Chakraborty contributed to part of the *in-vitro* experiments including Myc and MST experiments.

4.7 Chapter summary

We discussed the in-vitro validation of the prediction of the amino acid specific for a DNA sequence in this chapter. We predicted the amino acid/(s) that should be mutated to recognize a DNA sequence of interest. We cloned and purified the predicted and wild-type proteins for Ets1 and Myc. Using these purified proteins, we performed three binding assays, EMSA, MST and nanoDSF. Some of the results were not consistent across experiments. Also the mutants had a reduced binding affinity compared to the wild-type counterparts. Thus, we only compared the affinity of the mutant protein with the predicted and wild-type substrates. We observed that the Myc double mutant showed better binding toward the predicted non-consensus DNA than the original consensus in multiple *in-vitro* assays. Myc single mutant and the wild-type protein showed similar binding to both substrates. Hence, we were successful in validating one out of three predictions experimentally.

5

CHAPTER 5: Predicting DNA-Protein Interface for a Protein Structure

OVERVIEW

- Introduction to the problem
 - Method for the predicting the DNA-binding interface for a given protein structure
 - Assessment of the robustness of JEDII in predicting the DNA-protein interface
 - On the dataset curated by us- PDNA-285
 - In the absence of close homologs
 - With the apo conformation of DNA-binding proteins
 - With protein models as inputs
 - Comparison of JEDII with other methods
 - Predicting amino acids binding DNA
 - Comparison with other methods on benchmark datasets
 - Comparison of other methods on PDNA-285
 - Predicting DNA sequence binding a protein
 - Validation on ChIP-Seq dataset
 - Comparison of our method with DBD2BS
 - Comparison of our method with Farrel and Guo
 - Case study of cro repressor
 - Discussions
 - Chapter summary
-

So far in the thesis we have described the prediction of specific amino acids that would bind to a given DNA sequence. Since, our goal was to infer specific association preferences between DNA and proteins, we also attempted to solve the analogous problem of the prediction DNA sequence specific for a given protein, discussed in this chapter.

The prediction of the DNA-protein interface for a given protein structure is generally treated as two sub-problems- predicting the DNA-binding residues on the protein and the prediction of nucleotides specifically binding to that interface. In this present study we have integrated both these problems and proposed a novel predictor that can predict the DNA-binding residues and the DNA sequence they would bind. We have not attempted to classify proteins as DNA-binders as there are many methods that are available for this that have achieved a good accuracy. Among the best methods is StackDPPred [261], which has an accuracy of 90% and MCC of 0.8.

5.1 Introduction to the problem

Many crucial cellular processes depend on the interaction of protein and DNA [2,262-274]. The three-dimensional (3D) structures of the protein-DNA complexes resulting from these interactions provide mechanistic details of their function [50,275–279]. These structural data can provide information at the level of individual residues. This information from structures could help identify/rationalize the residues at the DNA-protein interface whose disruption could lead to diseased conditions [280–282].

The experimental determination of DNA-protein structures is time-consuming, and the structures of all protein families are not represented. Even among the protein families with DNA-protein complex data, the sampling of the constituent members is not uniform in the Protein Data Bank (PDB) [283]. There are also experimental methods that provide indirect information about the DNA-protein binding interface [3,250,284-293]. However, the slow progress by experimental methods in generating DNA-protein interface data initiated the development of a plethora of computational methods. These methods can provide high throughput results in a relatively short time.

The prediction of the DNA-protein interface has been treated as two separate problems. One set of methods solely predicts the DNA-binding residues, while a second set looks at predicting the DNA sequence recognized by a protein.

Among the methods that predict DNA-binding residues, there are three classes based on the features they use to make the prediction- sequence-based, structure-based and a combination of sequence and structural features.

5.1.1 Sequence-based DNA-binding residue prediction

Sequence-based methods rely only on the protein sequence to predict the binding residues. The sequence-based algorithms employ various methods for predicting DNA-binding residues such as Support Vector Machines (SVMs) [294–301], neural networks [118,302–306], random forest classification[307,308], ensemble learning [309], deep learning [310] and other combinations of machine learning techniques [311–319]. Since the binding of DNA to their protein counterparts is a three dimensional event, the accuracy of the sequence-based methods is lower compared to when structural data is incorporated in the predictions.

5.1.2 Structure-based DNA-binding residue prediction

Structure-based methods exclusively employ features from structural data to predict DNA-binding residues such as electrostatics, geometry and surface curvature [320–324]. While they perform better than methods using sequence alone, they are still not the best at predicting DNA-binding residues.

5.1.3 Sequence and Structure-based DNA-binding residue prediction

The final class of methods uses sequence and structural information for DNA-binding residue prediction. Since they incorporate all available data for a given protein, they are relatively more accurate than sequence or structure-based methods. The methods that use both sequential and structural features are more diverse in the techniques that they use for predicting DNA-binding residues. The techniques include

SVM [300,325–327], ensemble learning [328], neural networks [329–331] docking [332], clustering [333–335], light gradient boosting [336], and a combination of machine learning and template-based techniques [337]. This class incorporates the most useful data from sequences as well as structures and therefore GraphBind is currently the best method for the prediction of DNA-binding residues.

The above-mentioned methods have successfully predicted the DNA-binding residues to various degrees. However, predicting DNA-protein interface not only involves predicting amino acid interacting partners, but also necessitates predicting the DNA sequence for a protein. Most of the computational methods for the prediction of DNA sequence deal with transcription factor binding sites [338,339,348,340–347]. These methods require a DNA sequence as an input and search for the transcription factor binding sites within this region. However, there are also methods that predict the DNA sequence for all DNA-binding proteins with DNA sequence alone or DNA and protein sequences as input [349]. Some prediction tools also use structural data as input [350–353]. However, no single method identifies both DNA and protein binding regions.

Our proposed method does not treat the identification of DNA-protein interface identification as separate problems and predicts the binding residues on protein as well as DNA. The method employs topology independent structural superimpositions to derive the DNA-binding interface as described in the following sections.

5.2 Methodology

In this study, we developed a method to predict the DNA-binding residues along with the DNA-binding pose from a given protein structure. We hypothesized that the arrangement of hydrogen bond donor and acceptor atoms is key to binding between DNA and protein molecules. Thus, we used a topology independent structural alignment software, CLICK [40] to align the donor and acceptors in a query protein against a database of X-ray structures of DNA-protein complexes.

5.2.1 Compiling datasets used in this study

The datasets for the prediction of the DNA-binding interface (algorithm in section 5.2.2)-

5.2.1.1 Query Dataset

The query dataset consists of the proteins from 285 DNA-protein complexes defined in chapter 2 (PDNA-285).

5.2.1.2 Template Dataset

The template database consisted of only X-ray crystal structures having a resolution better than 3 Å and composed only of DNA and protein chains that resulted in 3909 structures (May 2020). Wherever rotational data was present, we performed symmetry operations and used the biological unit of these structures. We obtained a 40% NR template database with 700 sequences using these 3909 PDBs as input to h-CD-HIT to assess the performance of our method in the absence of close homologs in the template dataset.

5.2.1.3 Benchmark Datasets

We validated the prediction of amino acids binding DNA using datasets from previous studies- PDNA-62 [118], P224 [327], PDNA-129 [330], HOLO-APO 82 [334], where the numbers in the dataset name indicate their size.

For the validation of DNA-sequence prediction, apart from the PDNA-285 dataset, we have used three datasets-

1. A set of 7 proteins used by DBD2BS in their study [354].
2. Position Weight Matrices (PWMs) generated by ChIP-seq data from JASPAR for comparison with the predicted DNA sequence [355]. JASPAR had 331 entries for PWMs, of which 72 had a three-dimensional (3-D) protein structure. These proteins did not possess a 3-D structure in complex with the DNA and formed the validation set for DNA sequence prediction.
3. 27 Transcription factor chains from Farrel and Guo study [352]: Their dataset had 27 protein chains, which bound as monomers to their binding site. We obtained the PWMs for

each of these proteins from JASPAR. We selected the nucleotide with highest frequency at each position to generate reference DNA sequences against which we evaluated our method.

5.2.1.4 Validation using models

All previously described validations used crystal structures extracted from the PDB. We wanted to evaluate the efficacy of our prediction method when using models of proteins instead of their experimentally determined structures. To this end, we constructed models of the proteins in the PDNA-129 dataset separately using MODELLER [356] and AlphaFold2 [172].

For generating models with MODELLER, we used the protein with the best identity with the query excluding the crystal structure as the template for modelling. We obtained 122 models as seven sequences did not find alternate templates. We analysed the TM-score of these 122 models using TM-align to ascertain their similarity to the experimental structures [133]. 97 of the 122 models had a TM-score >0.5 , where a higher TM-score indicates greater structural similarity.

We also modelled the sequences from PDNA-129 using AlphaFold2 (using either the AlphaFold2 database or modelling with the google colab notebook). 127 of these 129 sequences have a TM-score >0.5 .

5.2.2 Prediction Algorithm

The prediction algorithm relies on the similarity of the hydrogen bond pattern at the interface of DNA-protein interactions. We used hydrogen bond donor and acceptor atoms in protein residues to guide structural alignments using CLICK [40,41]. The hydroxyl group in serine, threonine and tyrosine can act as both donor and acceptor. As a practical consideration, we treated this group as a donor in this study, since it acts as a donor in 70% of cases in the PDNA-285 dataset. We used default CLICK settings to align the query structure with all structures in the template database (except for the same PDB structure). The default settings use solvent accessibility and secondary structure parameters for alignment.

The DNA coordinates of the template were transferred to the superimposed query to build a model. To remove unsatisfactory models, we used the criteria of clashes and contacts. A *clash* was defined as a distance less than 1.5 Å between the C, C^α atom of protein, and the DNA atoms. We discarded all models with more than one clash.

The O and N atoms of the main chain are not explicitly considered in the checking of clashes as these atoms are hydrogen bond acceptors and donors, respectively. Our method involves the superimposition of the query with a template and the subsequent transfer of coordinates of the DNA to form the DNA-protein complex. This complex is not energy minimized or processed in any other way after superimposition. The clash criteria is the only means of eliminating possible bad models. This is also why the clash criteria is somewhat permissive (1.5 Å threshold, instead of a larger value). In addition, we have also made the hydrogen bond criteria permissive as we are allowing all donor-acceptor distances below 4 Å. Potentially these could include clashes (below 1.5 Å) - but we observed that this occurs in ~1% of the N and O atoms in the data (Appendix 19). With these permissive hydrogen bond and clash criteria, we believe we can identify all plausible hydrogen-bonded pairs.

A contact between DNA and the protein consisted of a distance less than 3.5 Å between the hydrogen bond donor/acceptor atoms of protein and DNA. We removed the models with less than three contacts between DNA and protein.

To select the best model, we define a term called the query overlap (%) -

$$QueryOverlap(\%) = 100 * \left(\frac{n}{t}\right) \quad \text{Eq.1}$$

where n is the number of donors and acceptors in the query protein superimposed with the template and t is the total number of donors and acceptors in the query protein. We predicted the model with the highest query overlap as the putative binding pose. The DNA-binding residues were the amino acid residues with any atoms within a distance of 4 Å from any DNA atom. We predicted the DNA sequence of the template as the putative nucleotide sequence that the protein would bind. We labelled this protocol of predicting the DNA-binding interface as Juxtaposition Enabled DNA-binding Interface Identifier (JEDII) (Figure 5.1). We demonstrated the efficacy of JEDII using PDNA-285 dataset and validated it using the different benchmark datasets.

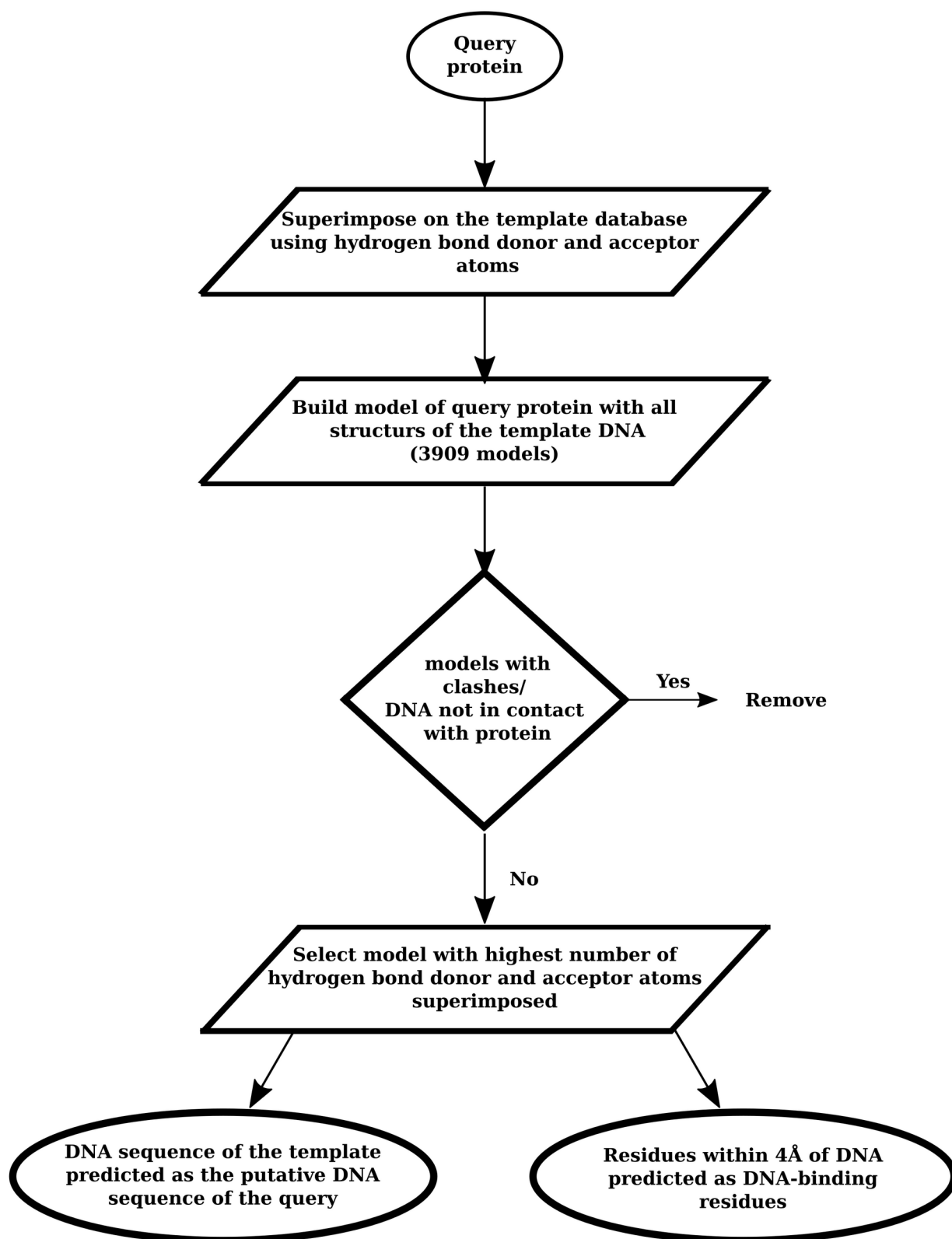


Figure 5.1: Flowchart of the steps in the JEDII protocol

5.2.3 Comparison of JEDII with GraphBind and HDock on PDNA-285 dataset

We compared the performance of JEDII with GraphBind [330] and HDock [357,358] on the PDNA-285 dataset. We used the GraphBind standalone version to make predictions on proteins from the PDNA-285 dataset. Out of 285 PDBs in our dataset, 107 are a part of the GraphBind training set. Therefore, we predicted the DNA-binding residues for the remaining 178 proteins. We parallelized GraphBind jobs using GNU Parallel [359]. For predictions using HDock, we docked the DNA structure from PDB 6EL8 on each of the proteins in the dataset and predicted the residues within 4 Å of the DNA in the top model as the DNA-binding residues.

5.2.4 Evaluation Metrics

The number of DNA-binding residues is lower than the non-DNA-binding residues. Moreover, the structural data that can be used to derive the properties of DNA-binding residues is limited. Therefore, this limits the methods to make correct predictions only in the cases where the input protein resembles the binding modes that the method has been trained for. One of the ways of reducing the over-fitting is by using a balanced measure for training the method. Since the number of non-DNA binding residues is larger, the true positives will be greater and hence the accuracy will in turn have a higher value. An extreme case is that even when all the residues are predicted as non-DNA-binding, the accuracy will still have a high value. Thus, using a balanced measure such as MCC [228] (defined in chapter 3), F1 or balanced accuracy can help in better evaluation of the method.

Thus the predicted results were evaluated using MCC, Accuracy (Acc), Specificity (SP), Sensitivity (SN), Balanced Accuracy (Bal. Acc), Precision (P) and F-measure (F1), which are defined below-

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TN + FP) * (TP + FN) * (TN + FN)}} \quad \text{Eq. 2}$$

Where,

TP = true positive (predicted binding residue correctly),

TN= true negative (predicted non-binding residue correctly),

FP= false positive (predicted a non-binding residue as a binding residue),

FN= false negative (predicted a binding residue as non-binding).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Eq. 3}$$

$$SN = \frac{TP}{TP+FN} \quad \text{Eq. 4}$$

$$SP = \frac{TN}{TN+FP} \quad \text{Eq. 5}$$

$$Bal. Acc = \frac{SP+SN}{2} \quad \text{Eq. 6}$$

$$P = \frac{TP}{TP+FP} \quad \text{Eq. 7}$$

$$F1 = 2 * \left(\frac{P * SN}{P + SN} \right) \quad \text{Eq. 8}$$

While MCC ranges from -1 to +1, all other measures range between 0 and 1. However, we have reported the values in % for SP, SN, P, Acc and Bal. Acc.

In the JEDII protocol, we selected the template DNA sequence as the sequence of the DNA likely to bind the input protein. To evaluate the prediction, we aligned the predicted sequence and the actual sequence using the malign module of MODELLER. We then calculated the similarity score as -

$$\frac{m_{aln}}{L_{DNA}} \quad \text{Eq. 9}$$

where m_{aln} is the number of matched nucleotides and L_{DNA} is the length of the actual DNA sequence.

5.3 Assessment of the robustness of JEDII in predicting DNA-protein interface

5.3.1 Assessment of the performance of JEDII on PDNA-285

We predicted the DNA-binding interface for the proteins in the PDNA-285 dataset. We have segregated the prediction of the DNA-protein interface into the predictions of amino acid residues binding DNA and the prediction of DNA sequence that would bind the protein. In each section, we have described the results for predicting amino acid residues binding DNA, followed by the DNA sequence prediction.

From the predicted DNA-protein model for each query protein, we annotated all the amino acid residues within 4 Å of the DNA as the DNA-binding residues. Using this criterion, JEDII achieved an overall accuracy of 94% and an MCC of 0.70 on this dataset (Table 5.1). The MCC for the dataset shows a bimodal distribution with one maxima at 0 and another at 0.9 (Figure 5.2 (a)).

We evaluated two factors contributing to the performance of the method- query overlap and sequence identity between the template and the query. We plotted these parameters in the PDNA-285 dataset against the MCC (Figure 5.2 (b) and 5.2 (c)). The average query overlap was $49.1 \pm 32.5\%$, while the average sequence identity between template-query was $59.1 \pm 34.6\%$. The query overlap exhibits a stronger positive correlation with the MCC (Spearman's correlation coefficient= 0.84) than the sequence identity between the template and the query (Spearman's correlation coefficient=0.60).

Table 5.1: Performance of JEDII on PDNA-285 dataset with redundant and 40% non-redundant template database.

The terms Acc, Bal. Acc, MCC, SN, P and F1 stand for Accuracy, Balanced Accuracy, Matthew's Correlation coefficient, Sensitivity, Precision and F1 score, respectively.

Method	Dataset	Acc (%)	Bal. Acc (%)	MCC	SN (%)	SP (%)	P (%)	F1
JEDII	Redundant	94	81	0.70	65	98	84	0.73
JEDII	NR-40	90	62	0.45	38	97	66	0.48

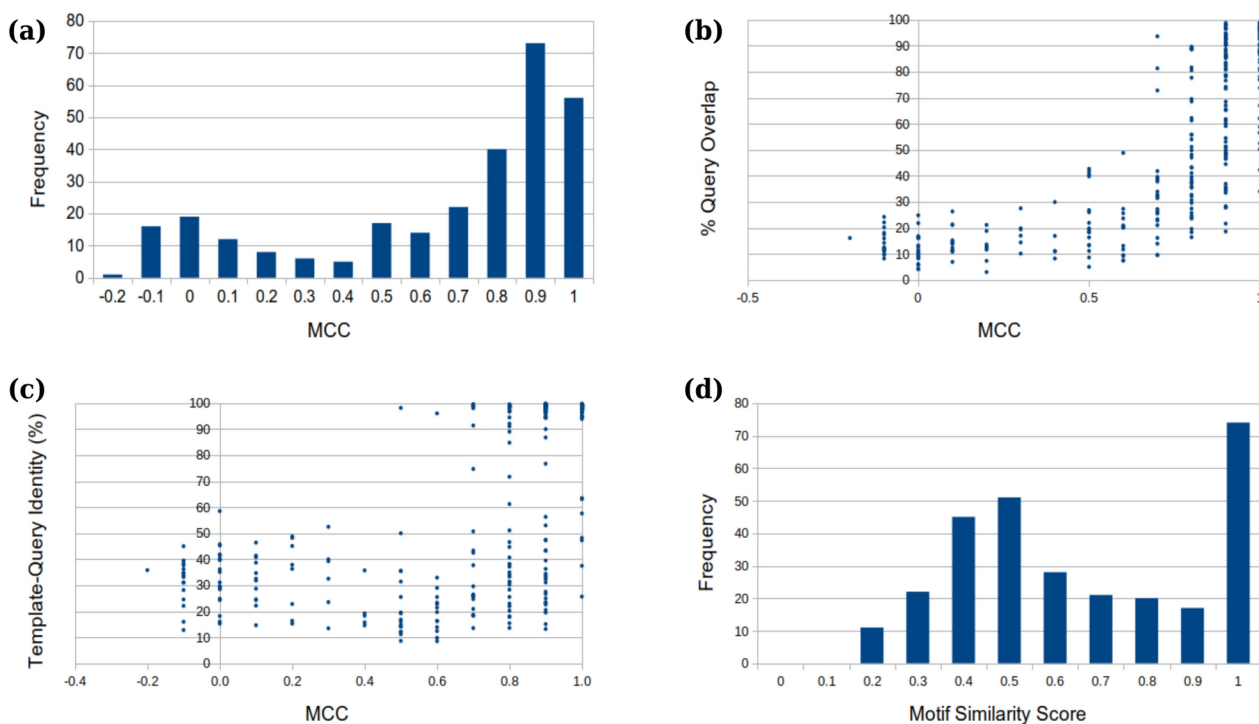


Figure 5.2: Performance of JEDII on the PDNA-285 dataset.

Subplots (a), (b) and (c) show MCC distribution, variation in MCC with query overlap (%) and average sequence identity between template-query (%), respectively for the prediction of DNA-binding residues. (d) is the distribution of similarity score for the DNA-sequence prediction.

87.5% of the perfectly-identified binding sites (49 out of 56) have a template identity >90%. Among the remaining seven proteins with a DNA-binding residue prediction of 1, the lowest sequence identity between the template and query was 25.6% (PDB ID: 1R8D). Moreover, in 14 PDBs with $MCC \geq 0.4$, the template selected had a different CATH [137,138] fold from the query.

JEDII had an average similarity score of 0.62 for predicting the DNA sequence on the PDNA-285 dataset (Figure 5.2 (d)). For 62 out of 289 unique proteins, we could identify the DNA sequence with a similarity score of 1.

We then assessed the robustness of our method when the query protein has $\leq 40\%$ identity with the template, when unbound protein conformation is given as an input and when models are given as inputs.

5.3.2 Accuracy in the presence of close sequence homologs

We have established in section 5.3.1 that the accuracy of predictions is correlated with the query overlap (0.84) and the sequence identity (0.60) between template-query. However, sequential similarity also implies structural similarity, resulting in a better overlap of the query protein. Therefore, we wanted to assess the method's robustness in the absence of close sequential homologs.

We used 700 protein-DNA complexes from the NR-40 database as templates to make predictions for proteins in the PDNA-285 dataset. The accuracy of the method for DNA-binding residue prediction reduces to 90.0%, with an MCC of 0.45 (Table 5.1). We can still identify DNA-binding residues of 4 query proteins with an MCC of 1. The correlation between the MCC and the template-query sequence identity is -0.05, while that between MCC and % query overlap is 0.67. Therefore, in the absence of close homologs, the identification of DNA-binding residues relies on the structural similarity of the DNA-binding interface.

Without close sequential homologs in the template dataset, the prediction of the DNA sequence has an average similarity score of 0.40. Though there has been a reduction in the accuracy, JEDI could identify the DNA sequence for one of the query proteins (PDB ID: 5GZB) with a similarity score of 1.

5.3.3 Accuracy using the unbound protein conformation

We employed the HOLO-APO 82 dataset to estimate the efficiency of our method when provided with the unbound conformation of DNA. this dataset consists of 82 pairs of proteins in both apo (DNA-unbound) and holo (DNA-bound) conformation. The accuracy of the method for DNA-binding residue prediction is 95% and 93% for holo and apo

conformation, respectively. The MCC of JEDII on the holo conformation is 0.67, while for the apo conformation, it is 0.55 (Table 5.5). In only 17 out of 82 cases, JEDII identified the corresponding holo-structure as the template. The average RMSD between apo and holo-structures was 1.2 Å irrespective of whether JEDII identified the holo-structure as a template or not.

We studied the effect of the degree of conformational change on the efficacy of the method using correlations. We correlated the RMSD between the apo and holo-structures with the MCC of amino acid residues binding DNA. The Pearson's correlation coefficient for these two parameters was -0.08, indicating that the MCC is independent of the degree of conformational change. The better MCC for holo-structures could be due to their similarity with the structures of DNA-protein complexes in the template database.

JEDII had an average similarity score of 0.66 for predicting the DNA sequence on both the apo and holo datasets. Out of 82 sequences, the crystal structure DNA sequence was precisely identified (similarity score = 1) in 23 and 30 cases in holo and apo datasets, respectively.

5.3.4 Accuracy using protein models as inputs

We assessed the ability of our method to identify DNA-binding interface using models generated by MODELLER and AlphaFold2 from the protein sequences in the PDNA-129 dataset. The MCC values for the prediction of amino acids binding DNA on models built by MODELLER are better than that for AlphaFold2 with 0.32 and 0.25, respectively.

To rationalize this result, we used TM-align to superimpose the models with the respective experimental structures in the PDNA-129-dataset. The average TM-score for AlphaFold2 models is higher (0.88) than that of MODELLER (0.74). 97 models built by MODELLER and 127 models from AlphaFold2 have TM-scores >0.5. Both these factors indicate that the models from AlphaFold2 are better. However, on these models with TM-scores >0.5, JEDII has MCCs of 0.37 and 0.25 for models built by MODELLER and AlphaFold2, respectively which contradicts the model quality assessment.

Therefore, we checked the DNA-protein interface regions. After superimposition, both AlphaFold2 and MODELLER have 63 models where the DNA-protein interface shows $>5\text{\AA}$ distance between the residues. Neglecting these 63 models, the average RMSD of hydrogen bond donor and acceptor atoms is $3.1\text{\AA}$ for MODELLER, and $2.6\text{\AA}$ for AlphaFold2 built models. Among the 63 models, 34 have unaligned interface regions when either MODELLER or AlphaFold are used to generate models. When we superimposed the models generated by MODELLER and AlphaFold2 with each other, from the 122 models, 22 have a TM-score <0.5 . If we only consider the 97 MODELLER models with TM-score >0.5 with respect to the experimental structure, there is no MODELLER model with TM-score <0.5 with respect to AlphaFold2 models.

We realized that we had 70 AlphaFold structures from the AlphaFold2 database. Many of these structures were entire proteins and not just the DNA-binding domains (which we had modelled with MODELLER). Therefore, we removed the extra regions from these models and computed the new MCC of JEDII, which was 0.33. Therefore, if only DNA-binding domains are given as inputs, JEDII has an equivalent performance on both MODELLER and AlphaFold2 models (with AlphaFold2 having a marginally better performance).

In the case of DNA sequence prediction with models as inputs, JEDII has a similarity score of 0.41 on MODELLER models and 0.40 on AlphaFold2 models. Perfectly identified DNA sequences (similarity score =1) are 6 and 7 for MODELLER and AlphaFold2 models, respectively.

5.4 Comparison of JEDII with other methods

We wanted to assess the accuracy of our method in comparison to other existing algorithms. However, no software performs both functions of predicting amino acids binding DNA and DNA-sequence prediction. Therefore we compared JEDII with the different classes of methods separately.

5.4.1 Predicting amino acid residues binding DNA

5.4.1.1 Comparison with other methods on benchmark datasets

We employed PDNA-62, P224 and PDNA-129 datasets that have been used in previous studies to analyse the performance of JEDII on DNA-binding residue prediction. We also compared the accuracy of JEDII with other methods on the APO-HOLO82 dataset and on models built from sequences derived from PDNA-129 dataset.

We have compared JEDII to sequence-based methods (ADASYN, BindN, BindN-RF, BindN+, CNNsite, DBS-Pred, DBS-PSSM, DNAPred, DRNAPred, EL_PSSM-RT, SVMnuc, TargetDNA, TargetS), structure-based methods (COACH-D) and methods that use a combination of both sequence and structural features (DNABind, DP-Bind, GraphBind, NucBind, PDNAsite, PDRLGB, PreDNA, RBSCORE). We obtained the values for all evaluation metrics from the respective papers for all methods except GraphBind. In the case of GraphBind, the program was run on models from PDNA-129 and on the 178 proteins from PDNA-285 dataset.

On both the PDNA-62 and P224, we perform better than all other methods in every assessment measure except sensitivity (Table 5.2 and Table 5.3). On the PDNA-129 dataset, we perform better than or at par with most methods, except GraphBind. GraphBind performs better than JEDII in all measures except precision (Table 5.4). The PDNA-129 dataset contains 27 entries that have been solved using electron microscopy. Since our template database only has DNA-protein complexes with resolution <4 Å, we have a worse performance on this dataset compared to GraphBind.

JEDII performs better on the HOLO-APO 82 dataset as compared to other methods. (Table 5.5). While on the holo dataset, JEDII outperforms other methods in all evaluation parameters, in the apo dataset, it has lower F1 and sensitivity compared to JET2DNA.

We compared the prediction of DNA-binding residues from protein models with GraphBind. We used the models built by MODELLER and AlphaFold2 for evaluation. On both, the models built by MODELLER and AlphaFold2, JEDII has a better accuracy, specificity and precision but GraphBind has better MCC and F1 values (Table 5.6). This trend holds even when we only consider the models that have TM-score >0.5.

Table 5.2: Comparison of JEDII with existing algorithms on PDNA-62 dataset.

The terms Acc, MCC, SN and SP stand for Accuracy, Matthew's Correlation coefficient, Sensitivity and Specificity, respectively.

Values for BindN [294], BindN-RF [307], BindN+ [295], DBS-Pred [118], DBS-PSSM [302], EL_PSSM-RT [309], DNABind [337], DP-Bind [312], PDNAsite [328], PDRLGB [336] and PreDNA [327] were taken from their respective papers.

Methods	Acc (%)	MCC	SN (%)	SP (%)
Sequence-based				
BindN	70	-	69	70
BindN-RF	78	-	78	78
BindN+	79	0.44	77	79
DBS-Pred	79	-	40	82
DBS-PSSM	66	-	68	66
EL_PSSM-RT	81	0.51	85	80
Sequence+ Structure-based				
DNABind	-	0.57	82	-
DP-Bind	78	0.49	79	77
PDNAsite	85	0.58	86	85
PDRLGB	81	0.52	86	81
PreDNA	83	0.50	80	84
Structure-based				
JEDII	91	0.64	72	94

Table 5.3: Comparison of JEDII with existing algorithms on P224 dataset.

The terms Acc, MCC, SN, P and F1 stand for Accuracy, Matthew's Correlation coefficient, Sensitivity and Specificity, respectively. The values for ADASYN [305], CNNsite [306], EL_PSSM-RT [309], PDNAsite [328], PDRLGB [336], PreDNA [327] and RBSCORE [335] were obtained from their respective papers.

Method	Acc (%)	MCC	SN (%)	SP (%)
Sequence-based				
ADASYN	83	0.48	77	84
CNNsite	84	0.40	77	84
EL_PSSM-RT	78	0.34	80	78
Sequence + Structure-based				
PDNAsite	82	0.40	83	82
PDRLGB	80	0.38	83	80
PreDNA	82	0.35	76	82
RBSCORE	85	0.40	63	88
Structure-based				
JEDII	94	0.66	68	97

Table 5.4: Comparison of JEDII with existing algorithms on PDNA-129 dataset.

The terms MCC, SN, P and F1 stand for Matthew's Correlation coefficient, Sensitivity, Precision and F1 score, respectively.

Values for DNAPred [296], SVMnuc [300], TargetDNA [301], TargetS [319], DNABind [337], GraphBind [330], NucBind [300] and COACH-D [320] were taken from their respective papers.

Method	MCC	SN (%)	P (%)	F1
Sequence-based				
DNAPred	0.33	40	35	0.37
SVMnuc	0.30	-	37	0.34
TargetDNA	0.29	42	28	0.34
TargetS	0.26	24	37	0.29
Sequence+ Structure-based				
DNABind	0.41	60	35	0.44
GraphBind	0.50	68	42	0.52
NucBind	0.31	-	37	0.35
Structure-based				

COACH-D	0.30	-	36	0.34
JEDII	0.41	42	47	0.45

Table 5.5: Comparison of performance of JEDII on HOLO-APO 82 dataset with other methods.

The terms Acc, MCC, SN, SP, P and F1 stand for Accuracy, Matthew's Correlation coefficient, Sensitivity, Specificity, Precision and F1 score, respectively. *iJET2 has divided the holo set to two parts (74 and 8 proteins) based on the stoichiometry. For comparison we have used the average value of the evaluation metrics. We have considered the values of their predictions that were obtained from a consensus of 8 runs out of 10. Values for iJET2 [334], DRNAPred [314], DISPLAR [329] and MultiVORFFIP [332] were taken from their respective papers.

Method	Dataset	Acc (%)	MCC	SN (%)	SP (%)	P (%)	F1
	Holo	Sequence + Structure-based					
iJET2 *		84	0.50	60	91	61	0.59
		Structure-based					
JEDII		95	0.67	67	98	74	0.70
	Apo	Sequence-based					
DRNAPred		85	0.21	23	95	40	0.24
		Sequence + Structure-based					
DISPLAR		87	0.39	41	94	54	0.43
IJET2 *		88	0.52	63	92	58	0.59
MultiVORFFIP		87	0.45	44	95	64	0.5
		Structure-based					
JEDII		93	0.54	51	97	65	0.57

Table 5.6: Comparison of the accuracy of JEDII on models built by MODELLER and AlphaFold2.

The terms Acc, MCC, SN, P, SP and F1 stand for Accuracy, Matthew's Correlation coefficient, Sensitivity, Precision, Specificity and F1 score, respectively.

indicates models with TM-scores >0.5. The values for GraphBind were obtained by running the standalone version [330].

Method	Dataset	Modelling Software	Acc (%)	MCC	SN (%)	P (%)	SP (%)	F1
GraphBind	PDNA-122	MODELLER	89	0.39	58	34	92	0.43
JEDII			92	0.32	31	40	96	0.36
GraphBind	PDNA-97 [#]		90	0.46	64	40	93	0.49
JEDII			92	0.37	38	44	96	0.41
GraphBind	PDNA-129	AlphaFold2	91	0.26	42	22	93	0.29
JEDII			94	0.25	26	30	97	0.28
GraphBind	PDNA-127 [#]		91	0.27	43	23	93	0.30
JEDII			94	0.25	27	30	97	0.28

5.4.1.2 Comparison with other methods on PDNA-285 dataset

For comparison of our accuracy, with other methods on the PDNA-285 dataset, we used GraphBind and HDock. JEDII performs better than GraphBind and HDock on the PDNA-285 dataset in all measures except sensitivity, where GraphBind is better (Table 5.7). Overall, except for PDNA-129, JEDII has a better performance than other sequence-based, structure-based and methods using both sequence and structure data. We consistently have a better specificity than other methods, however sensitivity of the method is one of the limitations of the method.

Table 5.7: Comparison of JEDII with GraphBind and HDock on PDNA-285 dataset.

* As mentioned in the methods, predictions were made for 190 proteins not in the GraphBind training set. The terms Acc, MCC, SN, SP, P and F1 stand for Accuracy, Matthew's Correlation coefficient, Sensitivity, Specificity, Precision and F1 score, respectively.

Method	Dataset	Acc (%)	MCC	SN (%)	SP (%)	P (%)	F1
GraphBind	PDNA-285*	92	0.69	86	93	63	0.73
HDock	PDNA-285	91	0.55	49	97	72	0.58
JEDII	PDNA-285	94	0.70	65	98	84	0.73

5.4.2 Validation of DNA sequence prediction

5.4.2.1 Validation using ChIP-seq data

ChIP-seq experiments generate DNA-binding sequence data for proteins, that is generally stored as PWMs in databases such as JASPAR. We took 72 proteins from JASPAR that did not a structure in complex with DNA, but had a structure of the protein to assess the accuracy of JEDII in DNA sequence prediction. To derive the DNA-binding site, we expanded the PWMs to individual sequences that satisfy the PWM criteria. When we compared the DNA sequence predicted by JEDII with the sequences from PWM, we obtained an average similarity score was 0.81 (Figure 5.3).

This is the highest similarity score that we have obtained for JEDII among all inputs. One of the reasons for this could be that the ChIP-seq data is not very high resolution. This means that the number of positions where all four nucleotides have been shown to interact with the protein is high. Since we look for all variants of the PWM and verify the prediction, this could be artificially boosting the accuracy.

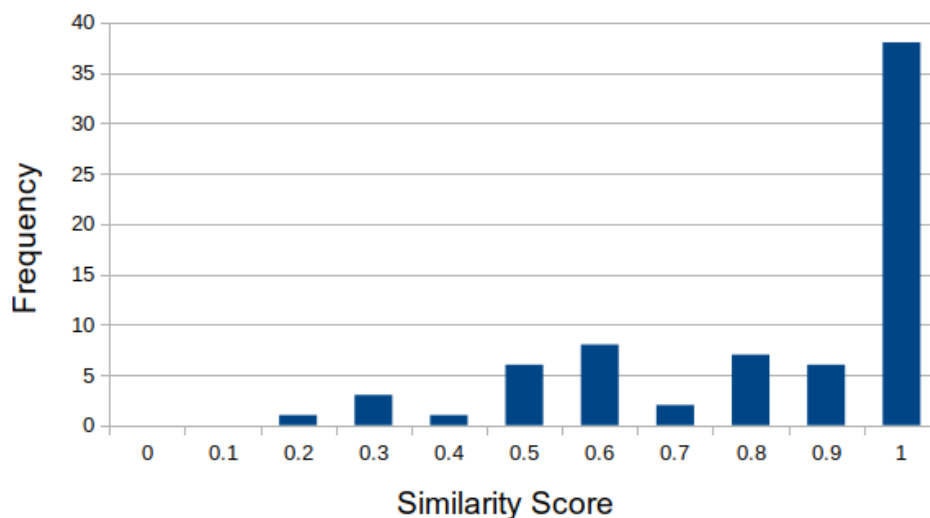


Figure 5.3: *Distribution of motif similarity score in ChIP-seq dataset*

5.4.2.2 Comparison with DBD2BS

For comparison of DNA-sequence prediction we used DBD2BS, which generates a PWM for a given protein structure based on knowledge-based potentials [353]. The 7 proteins that they have used are in their unbound conformation with an average RMSD of 1.1 Å. We have an average similarity score of 0.51 on these dataset. Since JEDII outputs a single DNA sequence and DBD2BS outputs a PWM, we could not compare the two methods directly. Therefore, we used their definition and compared the number of times the highest occurring nucleotide in the PWM is correctly predicted (Table 5.8). JEDII has a better performance than DBD2BS in 4 cases. In 1 case the number of correctly matched positions is the same for both methods, while in 2 of the 7 cases, DBD2BS performs better than JEDII.

Table 5.8: Comparison of JEDII with DBD2BS.

The values are the number of nucleotide positions where the predicted nucleotide also had the highest frequency of occurrence in JEDII and DBD2BS.

Sr. No	PDB ID	JEDII	DBD2BS
1	1NFI	5/10	5/10
2	1MI7	9/19	5/19
3	2A63	13/17	7/17
4	1M6U	3/12	6/12
5	1GV2	4/11	5/11
6	1MH3	7/18	6/18
7	2GZW	9/10	5/10

5.4.2.3 Comparison with study by Farrel and Guo

Though DBD2BS is the only algorithm that predicts the binding site for a given protein, there are other algorithms, which are specific for transcription factors. One such algorithm was developed by Farrel and Guo [352]. They also predict PWMs from a protein structure, however instead of a template-based algorithm, they use a fragment-based method to generate a complex of a transcription factor with DNA. The sequence of DNA is then generated using permutations and combinations and evaluating each sequence with an energy function to output a PWM. In their study, they have evaluated the number of positions in the PWM that their method has correctly predicted. We used this measure to compare the method with JEDII on the 27 protein chains they have used in their study. JEDII achieved an average similarity score of 0.62, while the Farrel and Guo method had a similarity score of 0.53. We performed better than Farrel and Guo in 18 out of 27 cases. In two cases, both methods had the same similarity score and in 7 cases Farrel and Guo's study had a better similarity score than JEDII (Figure 5.4).

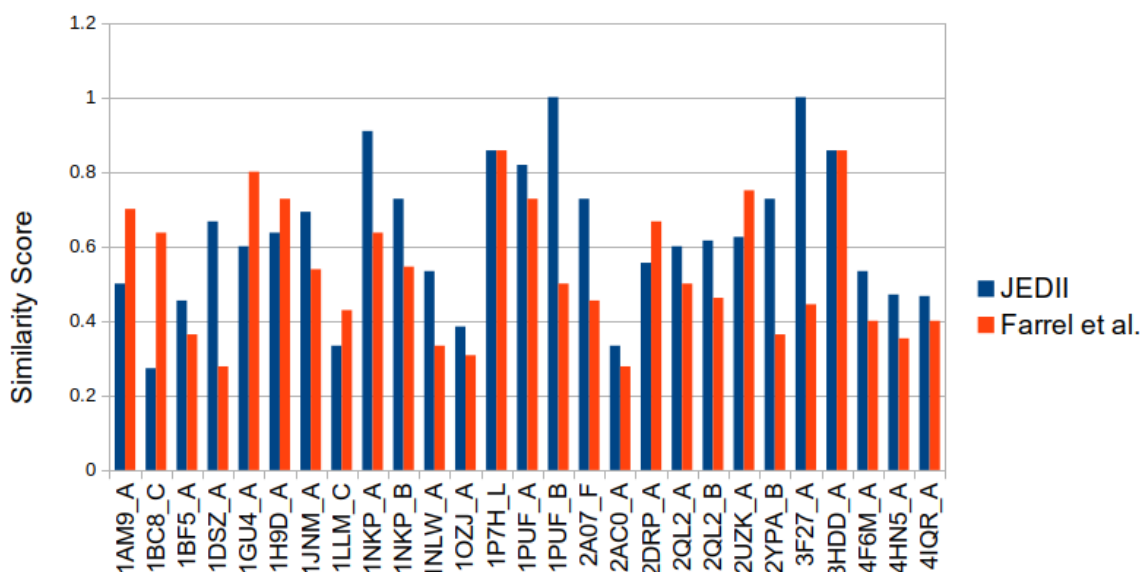


Figure 5.4: Comparison of JEDII with Farrel and Guo's study.

The y-axis shows the similarity score for each chain mentioned on the x-axis. The similarity scores for JEDII and Farrel and Guo's study are depicted in blue and orange, respectively.

5.5 Case Study of Cro Repressor

We predicted the DNA-binding pose and the DNA sequence for cro repressor of bacteriophage lambda (PDB ID: 6CRO). The template selected by JEDII was cl protein of bacteriophage lambda (PDB ID: 1RIO), which had a sequence identity of 31% with the cro repressor. The two proteins have different CATH IDs [137,138]. The cro repressor has 3.30.240.10 while the cl protein has 1.10.260.40 and 1.10.10.10. The superimposition also occurs between atoms in an α helix (cl) and those in a β sheet (cro repressor) (Figure 5.5). The predicted and the actual binding pose are very similar with the angle between the helical axis as 18.22° and the RMSD as $0.78\text{\AA}$. The MCC for the DNA-binding residue prediction is 0.9 and when compared with the PWM, the predicted DNA site has 14/16 sites matched (87.5% similarity).

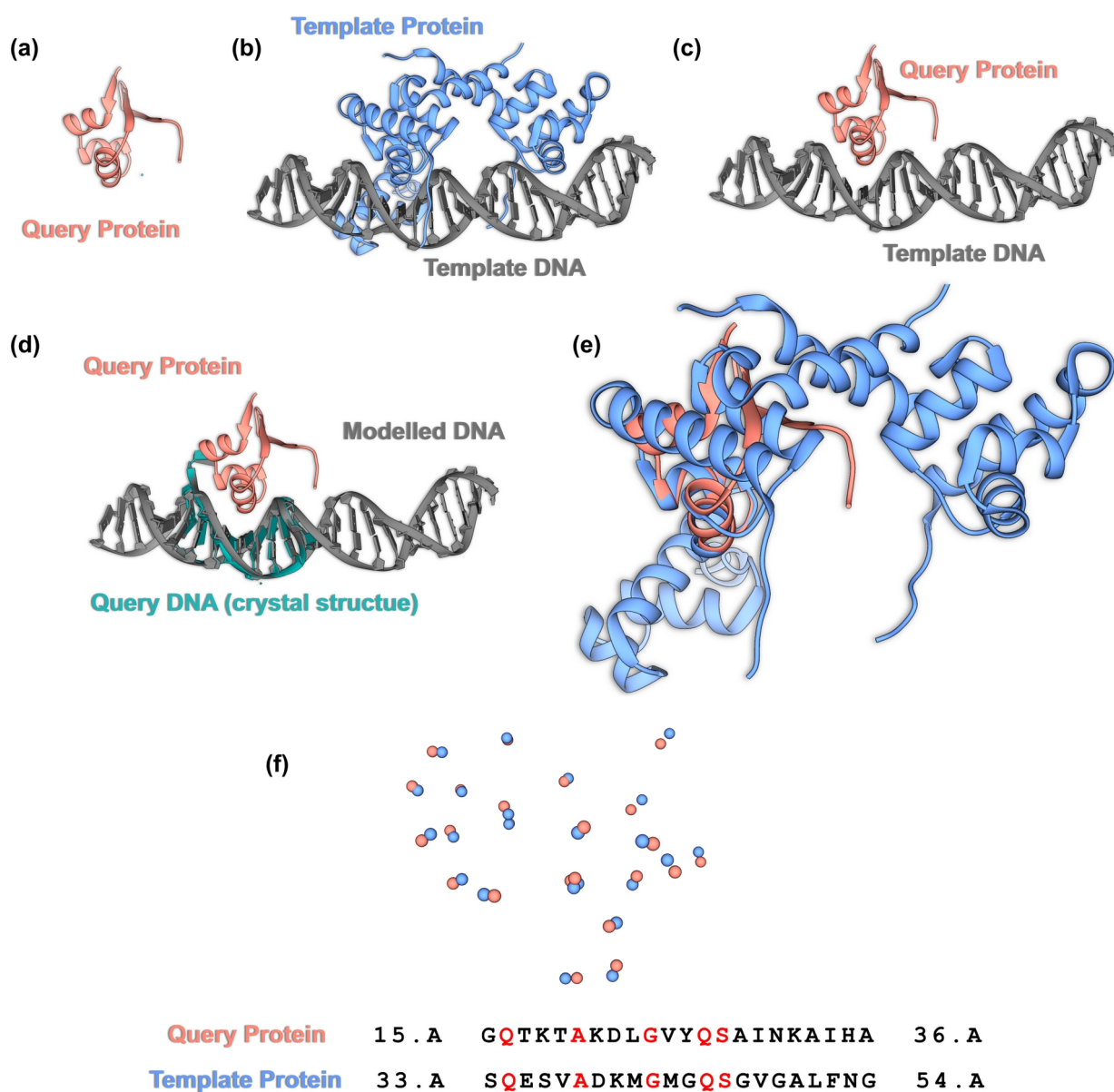


Figure 5.5: The prediction of binding pose of cro repressor.

The query and template proteins are in salmon and blue colour, respectively. The query and template DNA are depicted in green and grey, respectively. The panels show the structure of the query protein (a), template DNA-protein complex (b), predicted query protein-DNA model (c), superimposition of query DNA from crystal structure and predicted DNA (d) superimposition of query and template proteins and superimposition of the donor and acceptor atoms along with the DNA sequence of the superimposed region (f).

5.6 Discussions

This study proposes an algorithm to predict the DNA sequence that would specifically bind a given protein structure. The method involves the superimposition of hydrogen bond donors and acceptors in the query protein against a dataset of known DNA-protein complexes. Hydrogen bonds are the most significant contributors to DNA-protein specificity [15,30,31]. Therefore, we hypothesized the conservation of the orientation of hydrogen bond donors and acceptors in different DNA-binding proteins. Hence, the algorithm relies on the maximum match between the hydrogen bond donors and acceptors of the query and the template for template selection. We know that the protein region at the interface is complementary to the DNA. Therefore, if the query protein has matched well with the template protein, the DNA sequence of the template would also match with that of the query. Thus, we predict the DNA sequence of the template as the putative DNA sequence of the query.

We curated the PDNA-285 dataset to obtain only the protein-DNA complexes that exhibit sequence-specific binding. We used this dataset to test our hypothesis that the orientation of hydrogen bond-donor acceptor atoms shows conservation in specific DNA-protein complexes. The performance of JEDII is highest on the PDNA-285 dataset, lending credence to our hypothesis.

The MCC distribution of JEDII on the PDNA-285 dataset has a bimodal shape with peaks near 0 and 0.9. The sensitivity value is reflective of this distribution pattern. If we identify the optimal template, JEDII can predict the interface very accurately, while the binding site is completely missed in the other cases. We could identify the DNA sequences with an average similarity score of 0.62 on the PDNA-285 dataset. While we could identify the DNA-binding residues in 56 proteins with an MCC of 1, we obtained a similarity score of 1 for the DNA sequence prediction in 62 proteins. This implies that even though the DNA-binding pose may differ slightly, affecting the prediction of DNA-binding residues, our method can identify the DNA-protein interface correctly.

We analysed the contribution of the query identity and the average sequence identity between the template and query towards the MCC. Spearman's correlation coefficient was high for both parameters. However, the correlation with the query overlap was comparatively higher. Thus, JEDII tends to perform better in the presence of homologous structures bound to DNA. However, the performance is more dependent on the accuracy of the alignment between the query and the template proteins. To understand the contributions of the sequence identity between template-query, we employed the 40% NR template database to make predictions on the proteins from PDNA-285 dataset.

With the 40% NR template dataset, there was a reduction in MCC from 0.70 to 0.45. However, there were still 4 query proteins with an MCC of 1. The correlation between the MCC and the template-query sequence identity reduced to -0.05 from 0.60 with the 40% NR template dataset. However, the correlation between the MCC and the query identity remained high at 0.67. Therefore, in the absence of close homologs, the identification of DNA-binding residues relies on the structural similarity of the DNA-binding interface. There was a reduction in the efficacy of identification of the DNA sequence with the 40% NR template dataset. Since JEDII is a template-based method, a reduction in efficacy in the absence of close homologs was an expected outcome.

However, unlike other methods, there are examples where JEDII has a good performance without close sequential templates. One such example is that of the cro repressor. The selected template has a sequence identity of 31%, and the query and template have different CATH folds. Despite these differences, JEDII predicts the DNA-binding residues, DNA-binding pose and the predicted DNA sequence of the cro repressor protein with high accuracy. However, making uniformly good predictions in the absence of close homologs, needs improvements on the current algorithm. We plan to incorporate nucleotide changes based on the nucleotide-amino acid specificity rather than just relying on the geometric similarity of the interface.

Another evaluation of robustness was to check the performance of JEDII on unbound DNA-binding protein conformations. Previous studies have noted a conformational change in both DNA and proteins when they interact to form a complex [360,361]. Most of the

studies utilizing structural information to predict DNA-binding residues look at the conformation of the protein in the DNA-bound state. However, the aim of developing these methods is for the DNA-binding residue predictions of proteins that do not have a DNA-bound conformation.

Thus, it would influence the accuracy of the methods in the cases where conformations differ in the bound and unbound states. We used the APO-HOLO 82 dataset to analyse the performance of JEDII. While the accuracy does not change significantly with either apo or holo-structures as inputs, the MCC values are lower for the apo structure (0.55) compared to the holo-structure (0.67). Despite the conformational change, JEDII can identify 17/82 or 20% of the holo-structures as the templates for their respective apo structures, which points to its robustness.

The DNA sequence prediction is also unaffected with apo or holo-structures as inputs, with both giving a similarity score of 0.66. The identification of DNA sequence may have an advantage with the apo structure. JEDII identified the exact DNA sequence in 23 out of 82 cases using the holo-structure as query, while this number increased to 30 with the apo structure as input. The overall similarity score is the same for apo and holo, indicating that the selected templates for the holo structure have more near-native identifications than the apo protein.

We speculate that the template database does not have many structures similar to apo conformations. Thus if JEDII identifies the appropriate template, the DNA sequence matches perfectly. Otherwise, the similarity score is low. When we compare the performance of JEDII with other methods on these dataset, we outperform the existing methods on both datasets.

We can also obtain the unbound conformation of a DNA-binding protein through protein models. Especially with the development of AlphaFold2, there has been improvement in obtaining near-native conformations of proteins [172]. Therefore, we tested the robustness of JEDII on models. While JEDII has a worse performance on both MODELLER and AlphaFold2 models than experimentally determined structures as inputs.

When the DNA-binding domain alone is given as an input, the performance is slightly better with AlphaFold2 models.

This could be because AlphaFold2 models have more models with TM-scores >0.5 and also better average TM-score. Moreover, the RMSD between hydrogen bond donor and acceptor atoms, which are used for superimposition, is also lower for AlphaFold2 models. The AlphaFold models from the database contain the entire protein and not just the DNA-binding domain. Since our method looks for template similarity, AlphaFold2 models have a poorer MCC when entire protein structures are directly given as an input instead of just the DNA-binding domain.

GraphBind has a better performance with respect to MCC, sensitivity and F1 score when compared to JEDII. This holds true for experimentally determined structures of the PDNA-129 dataset and models from these dataset. The PDNA-129 dataset contains 27 proteins whose structures have been determined by electron microscopy and the average resolution of these structures is 4.4 Å. Since JEDII relies on finding appropriate templates using hydrogen bond-donor acceptor superimposition, the comparatively poor resolution may hinder the prediction accuracy. The template dataset, which is curated from X-ray crystal structures may also not have structures that resemble the proteins solved by electron microscopy. Therefore, future improvements to JEDII could involve incorporation of evolutionary information to overcome the lack of structural data.

While JEDII is a valuable tool, we wanted to compare its performance with the existing methods. Since no single method can predict both the DNA-binding residues and the DNA sequence, we compared the two kinds of predictions separately. For the DNA-binding residue prediction comparison, we used PDNA-62, P224 and PDNA-129 benchmark datasets. We perform better than other methods on the PDNA-62 and P224 datasets but not on the PDNA-129 dataset, on which GraphBind is better. We suspect GraphBind's better performance has to do with their incorporation of evolutionary data. ~80% of its MCC contribution of GraphBind is from evolutionary information alone. As discussed before, the 27 electron microscopy structures also contribute to JEDII lagging

behind GraphBind and incorporating evolutionary information is likely to improve our efficacy.

Apart from using benchmark datasets, we compared the performance of JEDII with GraphBind and HDock on the PDNA-285 dataset. GraphBind is currently the best method for predicting DNA-binding residues and incorporates both structural and sequential features for making predictions. While HDock is a protein-nucleic acid docking program that predicts the binding pose of the DNA. We perform better than GraphBind and HDock in all measures except the sensitivity.

To validate the DNA sequence prediction, we used data from ChIP-seq experiments, since it is one of the most common methods to identify the DNA sequence binding a protein. We used data from the PWMs generated by ChIP-seq experiments and compared the DNA sequence predicted by JEDII with all the sequences that would satisfy a given PWM. In this case, we have a motif similarity of 0.81.

One of the reasons for the high similarity score values in ChIP-seq experiments could be the presence of ambiguous positions in the PWM. By ambiguous positions, we mean that all four nucleotides occur at that position (Ns) in the consensus sequence. While it is possible that at a particular position no single nucleotide is favoured, the experimental conditions also influence the binding of proteins to DNA. Therefore, less stringent conditions could enhance non-specific binding resulting in a PWM with more Ns. Thus, it is crucial to rationalize the data obtained from ChIP-Seq experiments with known DNA-protein specificity patterns. Considering the time and resources involved in performing ChIP experiments, JEDII can be used to obtain a preliminary hypothesis of the putative binding site.

For comparing the DNA sequence prediction by JEDII with other computational methods, there were only two candidates- DBD2BS and a study by Farrel and Guo. While DBD2BS is a method that can be used for any DNA-binding protein, the method by Farrel and Guo is specific for transcription factors. DBD2BS aligns the query protein to templates and uses knowledge-based potentials to predict a PWM. Since JEDII can predict a single DNA sequence, thus for comparison, we used their measure of a correct prediction. They

consider a prediction as correct if the predicted nucleotide is also the nucleotide with the highest frequency at that position. JEDII performs better than DBD2BS in 4 out of 7 test cases and equivalent in 1.

The method by Farrel and Guo is a fragment-based method for predicting the PWM bound by transcription factors. Since a direct comparison was not possible, we compared JEDII with their method using similarity scores. For each position in the PWM, we consider it as a correct prediction if that nucleotide has the highest frequency for that position. Farrel and Guo also predicted if they could correctly identify a position in the PWM, but with the PCC method [362]. We compared the similarity scores for both methods and performed better than Farrel and Guo's study in 18 out of 27 proteins. Both methods have the same similarity score in two cases, and JEDII has worse performance than Farrel and Guo in 7 cases. As mentioned previously, incorporating nucleotide-amino acid interaction probabilities could improve the performance of JEDII for DNA sequence prediction.

In the different datasets, a common limitation of JEDII is its comparatively lower sensitivity. The lower value of sensitivity is due to its inability to make a correct prediction without an appropriate template. With the increasing structural data, incorporation of evolutionary information and protein-nucleotide specificity information, JEDII can be improved in the future. Overall, despite being a solely structure-based method, JEDII outperforms not only other structure and sequence-based methods, but most methods that employ a combination of structure and sequence. The unique feature of this method is that it can predict the DNA sequence for a given protein regardless of its structural or functional class. It integrates the problems of prediction of DNA-binding residues and the prediction of DNA sequences into the common umbrella of DNA-protein interface prediction.

5.7 Chapter summary

In this chapter we discussed the development of an algorithm that integrates the prediction of the DNA-binding residues, DNA-binding pose and the DNA sequence that a protein structure would bind. Our method, JEDII, is a template-dependent method that relies on the hydrogen bond orientation similarity across DNA-binding protein families to

find the most appropriate template. This method can predict the DNA-binding residues and the sequence in the absence of close homologs, with the unbound conformation of DNA-binding proteins and with the protein models as inputs. The method outperforms the existing algorithms and to our knowledge is the only one that predicts the DNA-binding residues and the DNA sequence from a protein structure.



CHAPTER 6: Conclusions and Future Directions

OVERVIEW

- Summary of this study
 - Analysis of DNA-protein specificity in non-B-form DNA
 - Predicting binding affinities of the DNA-protein complexes
 - Engineering specific DNA-protein complexes
 - Concluding remarks
-

6.1 Summary of this study

The molecular details of what dictates specific interaction between DNA and proteins have not been elucidated for DNA-binding proteins. While there seems to be a recognition code in protein families such as the zinc finger proteins and the TAL type III effector proteins, no such code is known for the other sequence specific DNA-binders. Owing to the necessity of specific DNA-protein binding in crucial biological functions, we believe that there is a universally conserved molecular signature behind this process. This molecular signature, in our view, is largely determined by hydrogen bonds due to their directional nature and their abundance in DNA-protein complexes. Previous studies focused on hydrogen bonds in conjunction with other interactions [32,72,118–121]. But these efforts only focussed on interactions of a single nucleotide or a particular family of DNA-binding proteins. Statistical analysis of DNA-protein crystal structures has led to knowledge about amino acid nucleotide preferences but not to a generalized recognition code [31–34].

The lack of a general code, according to us was a consequence of not comparing the DNA-binding preferences across structural folds. However, in our study we have studied nucleotide-amino acid interaction preferences across DNA-binding protein families. We superimposed structures using hydrogen bond donors/acceptors, which were independent of the protein conformation, and therefore we learned molecular details of recognition. Moreover, we adopted a holistic approach where we have used these hydrogen bond donor/acceptor superimpositions to solve two problems:

- i. Prediction of amino acids specifically binding a given DNA sequence-based
- ii. Prediction of the DNA sequence that would be recognized by a protein structure

The first 4 chapters discussed our approach and results for the first problem, while Chapter 5 dealt with the second problem.

To study the first problem of predicting the amino acid specific for a given protein, we compiled a 40% non-redundant set of 285 sequence-specific DNA-protein complex structures. This set was randomly divided into 235 structures in the training set, from which we gleaned the nucleotide-amino acid specificity preferences. Based on these preferences

we predicted the amino acids that would bind the DNA sequences in the remaining 50 structures of the testing set. We analysed the molecular interactions in the training set and observed that hydrogen bonds formed the majority of interactions with the DNA (except van der Waals).

Thus, we focussed on explaining the specificity preferences using hydrogen bond propensities. However, we did not observe sufficient number of interactions to make statistically significant inferences of the binding preferences for many nucleotide-amino acid pairs. Thus, we clubbed the amino acids with similar hydrogen bonding groups together. This resulted in 45% of the nucleotide-amino acid pairs with statistically significant favourable or unfavourable preferences. In order to compare the spatial orientation of the hydrogen bonds, we superimposed the hydrogen-bonded nucleotides. Upon superimposition of two nucleotides, we considered it as a hydrogen bond overlap if the hydrogen-bonded amino acid partners of these nucleotides were at a distance <1.5 Å.

The superimposition of the hydrogen bonds enhanced the confidence levels in the propensities and we could obtain nucleotide-amino acid interaction preferences in 92.5% of the cases. We also superimposed the nucleotides adjacent to the hydrogen-bonded one and observed that the nucleotide-amino acid specificity was influenced by the neighbouring nucleotides. We found that the neighbour at the 5' end had a greater influence on specificity than the neighbour at the 3' terminus. We compiled these data from the different superimpositions into a probability-based scoring scheme. We analysed this prediction scheme using the testing set and it gave comparable results as the training set, which points to the robustness of the method. Since our method is probabilistic we do not have a binary classification of binding and non-binding amino acids for a DNA sequence. This is also supported by the multiple sequence alignments where the predicted amino acid features in the close homologs, if not in the crystal structure of that protein. We validated the method using *in-silico* and *in-vitro* techniques.

We performed the *in-silico* validation using four approaches. In the first approach we used zinc finger proteins that had a known recognition code and predicted the amino acids likely to bind the DNA sequence. The prediction accuracy for the highest scoring

amino acid was similar to the training and testing set when the zinc finger proteins with SELEX optimised sequences were chosen for comparison. We could predict 56, 62 and 69% of the amino acids in the optimized sequences. The predictions were incorrect for serine/threonine and main chains. While our prediction scheme ranks the amino acids likely to be specific based on the frequency of occurrence, we cannot quantify the difference in specificity between the ranks. It is possible that for certain trinucleotides, multiple amino acids are favoured. In the cases where there are many favourable amino acids, the choice of the amino acid could be family-specific. However, due to the generalized nature of our predictions, we cannot predict them accurately.

The second *in-silico* approach was predicting the multiple sequence alignment (MSA) of a protein from the position weight matrix (PWM) of its recognition site. For each sequence that satisfied the PWM, we predicted the most likely amino acid partners. We then compared this list of predicted amino acids with those present in close homologs of the protein. The amino acids predicted by our method were present in the MSA in 43% of the cases. The occurrence in the MSA was also rank-dependent, with rank 1 occurring with 37% frequency. This finding was substantiated by the Pearson's correlation coefficient of 0.92 between the rank in the prediction and the frequency in the MSA.

We also quantified the efficacy of the method in terms of MCC, which was 0.17. Even though there are no other methods that can predict the MSA from the DNA sequence, this MCC value needs improvement. The reason for the low MCC was the low true positive rate. This means that while the identification of what is not specific is clear, identifying correct pairs of interacting nucleotide-amino acid partners is still a challenge. This could again be due to the fact that our method is probabilistic and when there are multiple favourable interactions, identification of the one in that structure poses a problem. We have analysed the effect of nucleotide neighbourhood on specificity. There could be a similar neighbourhood influence at the protein interface. Studying the protein neighbourhood could provide insights on which amino acids would be preferred when many of them can interact favourably with a nucleotide.

The third *in-silico* validation of our method was predicting the effect of mutations on the DNA-protein interface. We made predictions based on the trinucleotide odds ratio. If we only consider the mutations for which we have trinucleotide data, we can make 49 out of 64 predictions correctly. Among the remaining 15, 7 mutations are made to amino acids lacking side chain hydrogen bond donors/acceptors and there is a reduction in binding affinity of the mutant. By definition, we have considered them as hydrogen bonds formed by the main chain atoms. However, if the hydrogen bond is not formed, we could be incorrectly penalizing our method. In the absence of structures for these mutations, however, we can only speculate about the true reason for these incorrect predictions.

Finally, we validated our predictions by changing the DNA-binding specificity of proteins and analysing them using Molecular Dynamics (MD) simulations. We changed the original recognition sequence of proteins and predicted the amino acid changes that would bind the new non-consensus sequence. We used two transcription factors- Myc and Ets1 and one restriction enzyme- EcoRV for this validation.

As a control system, we used Zif-268, for which the mutants had a crystal structure with their preferred DNA sequence. Since the predicted amino acids would be incorporated in the crystal structure by modelling, we also modelled the zinc finger wild-type and mutant proteins as another control. However, the control experiments did not give expected results in the simulations. The use of models of the Zif-268 proteins instead of the crystal structure had higher binding energies. Also, the wild-type protein bound to the non-cognate DNA with a better binding energy than to the cognate sequence. Thus, MD simulations may not be an appropriate way of assessing the predictions. The results from our prediction showed that the Ets1 mutant exhibited the specificity predicted by us, while Myc did not. EcoRV has similar affinity towards both the original recognition site and the non-consensus DNA for the wild-type as well as the mutant protein. To understand the change in specificity, we employed *in-vitro* assays on these systems.

Out of the three systems that we simulated- Myc, Ets1 and EcoRV, we could obtain the wild-type and mutant proteins for Myc and Ets1. We also predicted a double mutant for Myc. However, the *in-vitro* results for Ets1 and Myc were reversed. Ets1 Y395R mutant

exhibited better binding towards the wild-type consensus sequence and not the non-consensus. Myc double mutant showed better binding towards the predicted non-consensus DNA than the original consensus in multiple *in-vitro* assays. While the Myc single mutant did not have quantifiable increase or decrease in affinity for the predicted substrate. Thus, we were successful in changing the specificity of one protein out of three predictions experimentally.

These validations increased our confidence in the use of hydrogen bond donor/acceptor atoms in studying specificity. Therefore, we used the same approach to tackle the problem of DNA sequence prediction from a protein structure. The DNA sequence prediction requires knowledge of the amino acids that would interact with it. While there are many tools that either predict the DNA-binding residues or the DNA sequence, to the best of our knowledge, none of them predict both. We developed a method that integrates the DNA-binding residue prediction with the DNA sequence prediction. We adopted a template-dependent superimposition methodology called Juxtaposition Enabled DNA-binding Interface Identifier (JEDII).

JEDII uses hydrogen bond donors and acceptors to find the most appropriate template and the DNA from the template is used to build a model for the query-DNA complex. The amino acids interacting with DNA are predicted as DNA-binding amino acids and the DNA sequence of the template is predicted as the putative DNA sequence that the protein would bind. The rationale was that the protein region at the interface is complementary to the DNA. Therefore, if the query protein matched with the template protein, the DNA sequence of the template would also match with that of the query. Our method had an MCC of 0.70 for DNA-binding residue prediction and a similarity score of 0.62 (maximum value 1) on the 285 proteins that we had compiled.

We assessed the robustness of our method by checking its accuracy in the absence of close homologs, with the unbound conformation of the protein as input and protein models as inputs. While there was some loss in the MCC with these validations, JEDII still performed better than or at par with other methods. The similarity score was less affected by these measures of robustness. Since it is a template-based method, JEDII tends to

perform better in the presence of homologous structures bound to DNA. However, the accuracy is more dependent on the efficiency of alignment between the query and the template proteins.

We compared our method against existing algorithms on four independent benchmark sets for DNA-binding residue prediction and two sets for DNA sequence prediction. We also validated the DNA sequence prediction using PWMs from ChIP-seq data and had a similarity score of 0.81. Considering the time and effort involved in such an experiment, JEDII can be used to obtain preliminary information about the DNA-binding protein. The unique feature of this method is that it can predict the DNA sequence for a given protein regardless of its structural or functional class. It integrates the problems of prediction of DNA-binding residues and the prediction of DNA sequences into the common umbrella of DNA-protein interface prediction.

Therefore, with the development of a methodology to predict the amino acid for a given DNA and to predict the DNA sequence binding a protein we fulfilled our two objectives. Among many of the extensions of this project, future work can be directly based on-

- Analysis of DNA-protein specificity in non-B-form DNA
- Predicting binding affinities of DNA-protein complexes
- Engineering specific DNA-protein complexes

6.2 Analysis of DNA-protein specificity in non-B-form DNA

The thesis was based on understanding the specificity of the DNA-protein complexes. However, we had only focussed on the classical B-form DNA. The primary limitation was the availability of data. The B-form DNA is the most abundant form of DNA inside cells and therefore, it had a larger volume of DNA-protein complex structural data available. However, we attempted to do a preliminary analysis for other DNA-protein complexes, which can be pursued in future with the availability of more information. We analysed double-stranded A-form, double-stranded Z-form and single-stranded DNA.

The A-form of DNA acts as a defence mechanism to protect the DNA under dehydration [363,364]. Single-stranded DNA is an intermediate conformation in processes such as replication [80,81], repair [82–84] and recombination [85–87]. Z-form DNA is characterized by the left-handed helix with a zig-zag orientation of its phosphate groups [96]. The details of these alternate DNA forms are described in Chapter 1, section 1.3.2. To study the specificity of this conformation we obtained 3D structures from NDB [208].

There were 210, 137 and 13 protein-DNA complexes annotated as single-stranded, double-stranded A-form and Z-form DNA respectively. However, none of the Z-form DNA have hydrogen bonds with the nitrogenous base and amino acids. The Z-DNA is hypothesized to have a structural recognition of its zig-zag pattern by the DNA-binding proteins [115]. Therefore, we did not proceed with the Z-form DNA.

Among the 137 A-DNA complexes, 86 were DNA/RNA hybrid complexes, 48 were double-stranded DNA complexes and 3 were RNA-protein complexes. Thus, we divided the A-form conformation to two sets double-stranded A-form DNA and DNA/RNA hybrid structures.

We obtained a 40% NR set from each set separately using CD-HIT [227]. This resulted in 68 structures for single-stranded DNA and 11 structures in each for double stranded A-DNA and DNA/RNA hybrid.

We calculated the hydrogen bonds between the DNA and protein in the these sets using our in-house python script and parameters described in Chapter 2, section 2.1.2. We considered the hydrogen bonds made by the protein with only the DNA strand in the hybrid structures. The single-stranded, double-stranded A-form and DNA/RNA hybrid DNA-protein complex sets had 1066, 230 and 145 hydrogen bonds, respectively between the nucleoside and amino acid.

We analysed the preferences of these interactions by Fisher's Exact Test described in Chapter 2. Among the 40 interaction pairs of nucleotides and amino acids, the single-stranded DNA-protein complex has statistically significant interactions in 6 cases at 95% confidence. The double-stranded A-form and DNA/RNA hybrid complexes have 11 and 4 statistically significant interaction pairs, respectively. Despite having more hydrogen bonds,

the single-stranded DNA has fewer significant preferences than double-stranded A-form DNA. This could imply that the interactions of single-stranded DNA with their cognate proteins are less specific in nature. We compared the statistically significant preferences in the alternate DNA conformations with the double-stranded B-form DNA-protein complexes (Figure 6.1).

(a)	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	nan	1.39	1.28	0.66	1.00	0.68	1.11	0.66	2.13	1.62
C	nan	3.16	0.81	0.29	1.15	1.12	0.97	0.65	0.00	1.44
G	nan	1.10	0.76	2.38	0.88	0.72	1.21	0.81	0.25	0.63
T	nan	0.33	1.25	0.92	1.00	1.24	0.87	1.59	3.75	0.89

(b)	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	0.00	0.00	0.00	0.00	1.11	3.78	0.17	6.14	0.00	0.00
C	0.00	9.10	7.33	3.68	1.54	1.82	0.14	0.77	0.00	3.52
G	0.40	0.34	0.40	0.19	0.59	0.30	6.37	0.65	0.82	0.82
T	18.36	0.00	0.00	3.67	1.45	0.81	0.80	1.18	8.12	0.00

(c)	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	nan	0.00	nan	0.00	0.00	0.87	0.60	8.62	nan	3.33
C	nan	inf	nan	inf	0.00	0.39	0.19	0.00	nan	2.61
G	nan	0.00	nan	0.00	0.63	1.07	3.05	1.33	nan	0.46
T	nan	0.00	nan	0.00	10.75	2.21	1.09	0.00	nan	0.00

(d)	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	0.00	0.71	0.91	0.39	0.86	4.26	0.42	1.97	0.00	3.05
C	0.00	42.53	1.44	0.20	4.05	0.86	0.12	1.31	0.00	1.70
G	0.89	0.04	1.01	2.52	0.36	0.35	3.22	0.49	0.00	0.41
T	4.93	0.03	0.77	0.96	1.40	0.95	1.15	1.17	inf	0.51

Figure 6.1: Comparison of binding preferences in different DNA conformations - single-stranded(a), double-stranded A-form (b), DNA/RNA hybrid (c) and double-stranded B-form (d) DNA-protein complexes. The values inside the cells are the odds ratio values obtained through Fisher's Exact test. The white cells are not significant at 95% confidence.

Blue and pink cells indicate significant interactions that have a positive and a negative preference, respectively.

The single-stranded DNA had four preferences that matched with the B-DNA out of six. In the remaining two preferences, the interacting pairs were not significant in B-DNA. The double-stranded A-form DNA and the DNA/RNA hybrid structures did not have any difference in preferences in the four pairs that are significant in the hybrid structure set. However, compared to the double-stranded B-form DNA, there is a difference in the lysine preferences. Lysine-cytosine interaction is favoured in the double-stranded A-form and the DNA/RNA duplex structures. But in the double-stranded B-form DNA, lysine-cytosine interaction has a negative preference. Whereas the opposite is true for lysine-guanine interaction. The lysine-guanine interaction is not favoured in the double-stranded A-form DNA, but is favourable for double-stranded B-form DNA.

Considering that we have based our inference on only 22 structures, we cannot claim there is a difference in preference with confidence. This needs to be investigated in the future when more structural data of A-form DNA-protein complex become available. It is well known that certain DNA sequences are more prone to deformation [365–367]. If they adopt an A-like conformation locally, this could mean that their recognition is mediated by different preferences and therefore a universal code of recognition has not been discovered yet.

6.3 Predicting binding affinities of the DNA-protein complexes

In the current study, we have analysed the role of hydrogen bonds in imparting specificity to DNA-protein complexes. In Chapter 3, section 3.4, using these specificity cues we predicted the affect of a point mutation in the DNA-protein complex towards the binding affinity. However, we only predicted whether it would increase or decrease the affinity compared to the wild-type DNA-protein complex. As discussed in Chapter 3, the experimental conditions such as pH, temperature and ion concentration can affect the binding affinity by orders of magnitude [231]. Also, for the same wild-type to mutation

change there was up to a 100-fold change in the binding affinity. These factors make absolute binding affinity prediction a difficult problem to solve.

A possible approach that can be used in future to tackle this problem would be to include other nucleotide-amino acid interactions in the prediction. As stated in Chapter 1, section 1.3, there are other interactions that also contribute to the stability of the interactions. Therefore a mutation may not only change the hydrogen bond interaction, but also change electrostatics and hydrophobicity of the interface locally. Thus, for an estimate of the binding affinity we need to incorporate all the interactions in the complex.

There are computational methods that predict the binding affinity [368–370]. However, we believe that the use of trinucleotide preferences with the other interactions could improve the binding affinity predictions. Moreover, combined with JEDII, we could employ the method to derive binding affinity of unbound proteins to different DNA substrates.

6.4 Engineering specific DNA-protein complexes

In Chapters 3 and 4 we attempted to validate the predicted specificity preferences by changing the specificity of DNA-binding proteins. However, we changed one or two amino acids at most for this validation. In order to generate a protein to specifically bind a DNA sequence we can follow a reverse engineering process.

In a typical design protocol, the backbone is first generated and the sequences that best fit this backbone are selected [371,372]. Our proposed method aims to do the reverse. We have structures of DNA-protein complexes with their inter-molecular interactions. Depending on the DNA sequence for which we want to design a protein, we can select appropriate templates of the DNA-protein complexes. We have shown that the hydrogen-bonded amino acid groups show conservation across different DNA-binding protein families. Thus, we can take the hydrogen-bonded atoms from the template proteins. We can thread these atoms through different backbones. The binding affinity calculations described in section 6.2 can help us decide the best protein binder for the

DNA sequence of interest. A second approach would be to use Rosetta suite [373] and select the best sequence according to the lowest RMSD with the desired interface.

6.5 Concluding remarks

In this thesis we have studied the specificity between DNA-protein interactions primarily using hydrogen bonds. We have demonstrated multiple instances where the different proteins exhibit the same hydrogen bond pattern. We have developed a methodology to predict the amino acids that will bind to a given DNA sequence specifically. We have also developed a technique to predict the DNA-binding residues, the DNA-binding pose and the DNA sequence that a protein structure would bind.

We have validated our methods with diverse *in-silico* and *in-vitro* approaches. Our findings provide previously unknown insights into the specificity determining factors in DNA-protein complexes. These may be applied in various spheres, some of which we have shown in our thesis and proposed others that may be pursued as a direct extension from the results of this thesis. The amino acid prediction from the DNA sequence may be employed to study the evolution of transcription factors. The current methodology can predict the effect of the mutation on DNA-protein interface and therefore could help rationalize diseases that occur as a result of this. An extension of this would be to predict the binding affinity of the DNA-protein complexes.

We have shown that the DNA sequence prediction can be used to obtain binding site information for a protein and has a good correlation with the site obtained through Chromatin Immunoprecipitation sequencing (ChIP-seq). In the future, using the hydrogen bond donor/acceptor derived interface, engineering proteins with required specificity could be achieved without regard for conforming to a specific fold. A similar approach may be adopted to analyse the specificity in non-B form DNA and RNA-protein complexes or any interface that relies on hydrogen bonds for recognition of the binding partner. The applications and implications of this study are immense and the results from our analysis would be useful additions to the scientific community. A common limitation that we have encountered throughout our study was the paucity of data. In the future, with that increase in structural data, we believe that these hydrogen bond preferences could pave the way for a molecular recognition code.

Appendix

1. List of PDBs removed by manual curation to retain only B-form sequence-specific DNA-protein complex structures

1A35, 1CEZ, 1CKT, 1EWN, 1EWQ, 1FN7, 1K2, 1L3S, 1M5R, 1MTL, 1OE5, 1QUM, 1RXW, 1T3N, 1TEZ, 1WVL, 1X9N, 1XPX, 1Y6F, 1YFJ, 2AOQ, 2AQ4, 2DNJ, 2FVQ, 2H7G, 2IHN, 2IS2, 2OFI, 2P0J, 2P6R, 2PFO, 2R9L, 2RGR, 2WWY, 2XRZ, 2ZKD, 3AAF, 3CVY, 3CW7, 3DSD, 3ECP, 3F2B, 3FBD, 3G0R, 3GQC, 3GYH, 3H25, 3I0W, 3IAY, 3JR5, 3JSM, 3JY1, 3K5L, 3KHH, 3KNT, 3KXT, 3M4A, 3NGI, 3NH2, 3O1M, 3ODA, 3ODE, 3PT6, 3Q0D, 3QMC, 3RZL, 3S6I, 3SLP, 3SPD, 3TED, 3TWM, 3VH0, 3VK7, 3VXV, 3ZVN, 4B23, 4C2U, 4CEI, 4CEJ, 4DKJ, 4ER8, 4FLW, 4FZX, 4GCL, 4GLE, 4GLX, 4HT4, 4HTU, 4IRC, 4J9Q, 4LLN, 4LVJ, 4NM6, 4O3M, 4O6A, 4P0Q, 4P4P, 4PAR, 4PTF, 4PZI, 4Q0Z, 4RI8, 4RKG, 4SKN, 4TMU, 4UBB, 4XQK, 4XVK, 4YPH, 4Z3A, 4ZYD, 5BPD, 5CDM, 5CHZ, 5CO8, 5D4D, 5DAC, 5DS5, 5G34, 5GKG, 5HP4, 5HRB, 5I44, 5J5P, 5L2X, 5NW9, 5T5C, 5T9J, 5UM9, 5V0E, 5VHV, 5VZD, 5W2A, 5XFQ, 6AS7, 6CNQ, 6EMY

2. Amino acid (three-letter representation) composition (%) in training and testing sets

Amino Acids	Training Set	Testing Set
ALA	7.45	7.09
CYS	1.20	1.34
ASP	5.14	5.15
GLU	7.42	7.10
PHE	3.84	4.13
GLY	5.67	5.99
HIS	2.36	2.23
ILE	6.25	6.61
LYS	7.27	7.33
LEU	10.45	9.51
MET	1.89	2.09
ASN	4.30	4.39
PRO	3.87	4.19
GLN	4.29	4.05
ARG	6.93	7.02
SER	6.01	6.08
THR	4.91	4.59

VAL	6.24	6.33
TRP	1.15	1.45
TYR	3.35	3.33

3. Hydrogen bond donor-acceptor list used in this study

Amino Acid	Acceptor	Acceptor Antecedent	Donor
ALA	O	C	N
ARG	O	C	NH2
			NH1
			NE
			N
ASN	OD1	CG	ND2
	O	C	N
ASP	OD2	CG	N
	OD1	CG	
	O	C	
CYS	O	C	SG
			N
DA	N3	C2	N6
	N3	C4	
	N7	C5	
	N7	C8	
	N9	C4	
	N9	C8	
	O3'	C3'	
	O4'	C4'	
	O5'	C5'	
	OP1	P	
	OP2	P	
DC	N1	C6	N4
	N1	C2	
	O2	C2	

	O3'	C3'	
	O4'	C4'	
	O5'	C5'	
	OP1	P	
	OP2	P	
DG	N3	C2	N2
	N3	C4	
	N7	C5	
	N7	C8	
	O6	C6	
	N9	C4	
	N9	C8	
	O3'	C3'	
	O4'	C4'	
	O5'	C5'	
	OP1	P	
	OP2	P	
DT	N1	C6	N3
	N1	C2	
	O2	C2	
	O4	C4	
	O3'	C3'	
	O4'	C4'	
	O5'	C5'	
	OP1	P	
	OP2	P	
GLN	OE1	CD	NE2
	O	C	N
GLU	OE2	CD	N
	OE1	CD	
	O	C	
GLY	O	C	N

HIS	O	C	ND1
	NE2	CD2	NE2
	NE2	CE1	N
	ND1	CG	
	ND1	CE1	
ILE	O	C	N
LEU	O	C	N
LYS	O	C	NZ
			N
MET	O	CA	N
PHE	O	C	N
PRO	O	C	
SER	O	C	OG
	OG	CB	N
THR	O	C	OG1
	OG1	CB	
TRP	O	C	NE1
			N
TYR	O	C	OH
	OH	CZ	N
VAL	O	C	N

4. Superimposition frequency when the first nucleotide of the dinucleotide is hydrogen-bonded.

The amino acids/groups (columns) and nucleotides (rows) are shown in single-letter representation.

H-bonded Nucleotide	3' Neighbour	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	A	0	6	2	86	2	113	149	89	0	83
	C	0	3	5	42	0	47	43	69	0	17
	G	0	0	0	398	1	41	706	53	0	29
	T	0	1	1	0	3	234	89	219	0	113
C	A	0	196	1	6	12	84	37	245	0	157

	C	0	58	6	28	37	52	17	81	0	57
	G	0	13	1	53	24	39	21	47	0	14
	T	0	34	3	1	17	15	0	38	0	32
G	A	0	7	43	1347	33	177	1900	522	0	197
	C	0	7	68	820	65	82	1890	130	0	0
	G	0	0	152	2821	0	159	4123	302	0	69
	T	41	0	23	2212	48	225	4178	455	0	22
T	A	0	0	7	129	59	67	314	84	0	2
	C	0	0	16	234	0	53	472	53	0	16
	G	0	0	6	68	10	52	107	57	1	13
	T	10	0	3	165	42	37	238	59	0	4

5. Superimposition frequency when the second nucleotide of the dinucleotide is hydrogen-bonded.

The amino acids/groups (columns) and nucleotides (rows) are shown in single-letter representation.

5' Neighbour	H-bonded Nucleotide	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	A	0	0	0	122	4	114	389	121	0	30
C		0	25	0	26	0	49	134	129	0	21
G		0	8	4	25	1	118	185	99	0	48
T		0	0	3	92	1	214	102	119	0	145
A	C	0	37	5	10	13	17	13	62	0	9
C		0	43	20	42	97	35	39	61	0	43
G		0	103	3	23	39	44	0	154	0	76
T		0	82	0	32	20	39	33	47	0	101
A	G	0	0	20	1234	61	140	1538	80	0	81
C		0	0	61	794	4	113	1891	223	0	28
G		15	4	262	4476	45	293	4543	241	0	19
T		0	2	70	1251	8	69	4645	1228	0	215
A	T	0	0	4	314	40	38	326	38	0	2
C		0	0	0	11	4	9	62	22	0	8

G		0	0	17	364	0	90	733	38	4	2
T		11	0	5	200	48	75	432	135	0	2

6. Superimposition frequency of trinucleotides

5' Neighbour	H-bonded Nucleotide	3' Neighbour	C	DE	H	K	Mc	NQ	R	ST	W	Y
A	A	A	0	0	0	0	0	48	47	2	0	1
		C	0	0	0	3	0	1	1	1	0	2
		G	0	0	0	12	0	0	48	0	0	0
		T	0	0	0	0	0	95	60	29	0	0
C		A	0	1	0	2	0	6	14	13	0	2
		C	0	50	0	2	0	0	3	3	0	2
		G	0	3	0	13	0	31	98	3	0	2
		T	0	0	0	0	0	16	0	5	0	1
G		A	0	0	0	1	0	18	5	0	0	4
		C	0	10	0	0	0	25	1	6	0	0
		G	0	0	0	8	0	16	139	5	0	
		T	0	0	2	0	0	25	6	5	0	0
T		A	0	0	0	8	0	41	81	5	0	5
		C	0	0	1	0	0	25	2	8	0	0
		G	0	0	0	47	0	12	46	11	0	6
		T	0	0	0	0	0	175	13	14	0	4 2
A	C	A	0	30	0	0	0	6	2	0	0	0
		C	0	54	0	0	1	13	0	2	0	0
		G	0	1	0	0	0	1	0	2	0	0
		T	0	22	4	0	0	0	0	4	0	0
C		A	0	32	0	0	3	9	6	6	0	0
		C	0	25	4	0	5	2	0	2	0	0
		G	0	40	0	12	6	3	0	4	0	9
		T	0	20	0	0	0	0	0	0	0	0
G		A	0	34	0	0	0	0	0	32	0	0
		C	0	88	0	3	5	10	0	14	0	0
		G	0	6	0	0	2	8	0	2	0	0

		T	0	0	4	0	0	1	0	2	0	3
T		A	0	116	0	3	0	123	9	29	0	50
		C	0	22	0	2	0	4	6	1	0	5
		G	0	1	0	2	0	0	9	4	0	0
		T	0	0	0	0	2	2	0	0	0	0
A	G	A	0	0	0	71	10	49	120	14	0	12
		C	0	0	3	19	4	3	76	1	0	0
		G	0	0	1	103	0	16	365	15	0	0
		T	0	0	0	97	1	26	259	0	0	1
C		A	0	0	0	6	0	6	26	6	0	0
		C	0	0	31	18	0	11	164	2	0	0
		G	0	0	7	90	0	56	353	18	0	0
		T	0	0	16	99	0	19	392	14	0	0
G		A	0	1	11	157	0	16	289	1	0	1
		C	0	0	8	143	4	33	469	27	0	0
		G	0	0	82	215	0	8	377	26	0	0
		T	2	0	0	402	10	62	626	0	0	0
T		A	0	0	0	85	1	4	566	148	0	0
		C	0	0	26	15	0	2	453	0	0	0
		G	0	0	0	68	0	25	757	34	0	0
		T	0	0	0	82	0	35	1356	101	0	0
A	T	A	0	0	0	10	14	2	13	1	0	0
		C	0	0	5	25	0	16	48	14	0	0
		G	0	0	0	5	0	8	3	4	0	0
		T	0	0	0	9	8	12	5	0	0	0
C		A	0	0	0	0	0	2	12	0	0	0
		C	0	0	0	0	0	1	21	0	0	2
		G	0	0	0	3	0	4	8	3	0	0
		T	0	0	0	0	1	3	9	5	0	0
G		A	0	0	0	6	0	10	33	1	0	0
		C	0	0	12	21	0	6	180	1	0	0
		G	0	0	0	12	0	26	59	9	0	0
		T	0	0	0	37	0	13	82	4	0	0

T		A	0	0	3	24	8	41	105	27	0	0
		C	0	0	0	6	0	0	40	0	0	0
		G	0	0	0	2	2	7	10	1	0	0
		T	0	0	1	2	4	20	8	9	0	0

7. Statistically significant preferences common between Di- and Trinucleotides

Trinucleotide	Amino acid	Trinucleotide odds ratio	Last Nucleotide odds ratio	First Nucleotide odds ratio
GAG	K	0.28	0.13	0.13
GAT	NQ	17.99	6.80	10.60
GGG	ST	0.65	0.17	0.17
CAG	NQ	2.43	2.97	6.80
GGA	ST	0.04	0.17	0.23
CAT	NQ	24.79	2.97	10.60
TGT	R	4.49	1.95	1.95
AAT	NQ	10.52	3.60	10.60
CCA	K	0.00	0.29	0.15
AGA	Mc	5.78	1.64	1.64
GAC	K	0.00	0.13	0.17
GTT	DE	0.00	0.00	0.00
GCA	ST	17.01	5.52	5.98
CCA	DE	31.11	14.53	32.15
TAT	DE	0.00	0.00	0.00
TCA	NQ	5.94	2.50	2.28
GAC	ST	2.90	2.56	5.18
CCC	Mc	22.35	35.33	35.33
TGG	R	4.02	1.95	0.80
ACG	R	0.00	0.09	0.00
TAA	DE	0.00	0.00	0.00
GTG	DE	0.00	0.00	0.00
ACC	DE	82.26	32.15	14.53
TAT	R	0.03	0.17	0.17
CCG	R	0.00	0.11	0.00
TAC	ST	4.99	2.15	5.18
GGA	K	2.89	2.89	1.60
ACT	R	0.00	0.09	0.10

GGT	NQ	0.52	0.51	0.14
TCC	DE	28.00	40.44	14.53
TTG	NQ	4.30	1.83	1.58
TGT	NQ	0.19	0.14	0.14
TAG	R	0.37	0.17	0.62
TTA	DE	0.00	0.00	0.00
AAA	K	0.00	0.43	0.43
AGA	DE	0.00	0.00	0.00
TAA	K	0.33	0.37	0.43
GCC	ST	2.31	5.52	1.90
CGT	NQ	0.32	0.73	0.14
CAT	K	0.00	0.17	0.37
TCA	Y	22.10	16.03	2.05
GCG	R	0.00	0.00	0.00
TGG	DE	0.00	0.02	0.03
ACA	DE	87.29	32.15	32.15
CGT	R	1.69	1.65	1.95
GCA	DE	24.87	44.92	32.15
CGC	K	0.48	0.79	0.79
TGC	R	6.93	1.95	1.65
TGC	NQ	0.04	0.14	0.73
GGG	NQ	0.10	0.51	0.51
GCG	Mc	17.84	8.27	8.27
TCA	DE	14.83	40.44	32.15
CAC	R	0.03	0.54	0.54
TGA	H	0.00	0.52	0.38
ATA	R	0.30	0.75	0.75
GCT	Y	36.88	8.03	16.03
CCA	R	0.07	0.11	0.09
TCT	Mc	142.87	4.86	4.86
AAA	DE	0.00	0.00	0.00
CGG	DE	0.00	0.00	0.03
TGT	Mc	0.00	0.06	0.06
TAA	NQ	3.90	10.60	3.60
TTT	Mc	14.57	4.82	4.82
TCA	ST	1.70	1.52	5.98
GCG	DE	11.13	44.92	44.92

TGA	DE	0.00	0.02	0.00
TTA	Mc	6.03	4.82	4.70
TCT	R	0.00	0.10	0.10
CCC	R	0.00	0.11	0.11
TAC	K	0.00	0.37	0.17
CCA	Mc	8.15	35.33	6.81
ACA	K	0.00	0.15	0.15
TGT	DE	0.00	0.02	0.02
ACC	K	0.00	0.15	0.29
GGC	DE	0.00	0.03	0.00
TAC	NQ	21.27	10.60	2.97
CAC	DE	120.97	7.41	7.41
CCC	H	6.81	3.61	3.61
TCC	R	0.11	0.10	0.11
AGT	H	0.00	0.38	0.52
ACA	R	0.03	0.09	0.09
AGC	ST	0.16	0.23	0.74
TAC	R	0.04	0.17	0.54
GTG	NQ	3.03	1.58	1.58
AGC	DE	0.00	0.00	0.00
ACC	NQ	2.11	2.28	2.04
CAT	R	0.00	0.54	0.17
GGC	Y	0.00	0.05	0.30
CCG	DE	27.86	14.53	44.92
GGG	Y	0.00	0.05	0.05
GAT	K	0.00	0.13	0.37
TAG	Y	4.51	11.56	4.06
CCG	Y	12.37	4.73	8.03
GCC	K	0.14	0.13	0.29
GGG	DE	0.00	0.03	0.03
TGT	H	0.00	0.52	0.52
GCA	K	0.00	0.13	0.15
GAT	R	0.12	0.62	0.17
GCC	R	0.00	0.00	0.11
CGC	ST	0.15	0.74	0.74
TGC	DE	0.00	0.02	0.00
ACC	R	0.00	0.09	0.11

CTC	K	0.00	0.25	0.25
ACT	K	0.00	0.15	0.23
TGG	NQ	0.25	0.14	0.51
CGG	Y	0.00	0.30	0.05
TGT	K	0.28	0.39	0.39
GGA	DE	0.04	0.03	0.00
ACG	ST	17.34	5.98	5.52
GGG	K	2.61	2.89	2.89
AAT	K	0.00	0.43	0.37
GAA	NQ	16.75	6.80	3.60
CGG	ST	0.61	0.74	0.17
AAT	ST	3.33	1.85	2.15
TGG	Mc	0.00	0.06	0.27
TAT	NQ	26.89	10.60	10.60
TCG	ST	5.79	1.52	5.52
GAC	NQ	13.76	6.80	2.97
TGC	K	0.17	0.39	0.79
CAT	ST	5.12	5.18	2.15
GCA	R	0.00	0.00	0.09
CGT	Mc	0.00	0.09	0.06
GAC	R	0.02	0.62	0.54
TCC	Y	12.44	16.03	4.73
TAT	K	0.00	0.37	0.37
TAT	H	0.00	0.28	0.28
ATT	Mc	47.05	4.70	4.82
CAG	K	0.53	0.17	0.13
TCA	R	0.02	0.10	0.09
GCG	NQ	7.39	2.23	2.23
GGG	H	11.27	2.60	2.60
CCC	DE	44.32	14.53	14.53
CCT	R	0.00	0.11	0.10
AGG	K	1.47	1.60	2.89
CGC	R	1.67	1.65	1.65
GCT	R	0.00	0.00	0.10
AAT	DE	0.00	0.00	0.00
CTT	ST	6.69	2.32	1.76
TCA	K	0.05	0.23	0.15

GGC	NQ	0.45	0.51	0.73
CGC	DE	0.00	0.00	0.00
AGG	DE	0.00	0.00	0.03
TTT	NQ	7.76	1.83	1.83
CCT	DE	inf	14.53	40.44
ATA	Mc	88.75	4.70	4.70
CCC	K	0.00	0.29	0.29
GGG	Mc	0.00	0.27	0.27
GCC	DE	71.81	44.92	14.53
TTT	ST	4.49	1.76	1.76
ACT	DE	63.05	32.15	40.44
AGG	ST	0.52	0.23	0.17
CGT	DE	0.00	0.00	0.02
AGT	DE	0.00	0.00	0.02
GGT	DE	0.00	0.03	0.02
AAA	NQ	9.12	3.60	3.60
TAT	Y	23.98	11.56	11.56
CGC	NQ	0.47	0.73	0.73
CAA	ST	9.15	5.18	1.85
CCG	Mc	13.13	35.33	8.27
GCC	Mc	6.37	8.27	35.33
CAC	K	0.19	0.17	0.17
GGG	R	0.70	0.80	0.80
AGA	K	1.96	1.60	1.60
AGA	H	0.00	0.38	0.38
CTC	Y	7.76	2.65	2.65

8. Weights optimized by grid-search for the prediction of an amino acid specific for a given DNA sequence.

The nucleotides are in single-letter representation.

Trinucleotide	Trinucleotide Weight	Dinucleotide weight	Single Nucleotide Weight
AAA	1	0	0
AAC	0	0.01	0.99
AAG	1	0	0
AAT	1	0	0

ACA	0	0.01	0.99
ACC	1	0	0
ACG	0	0.01	0.99
ACT	0	0.01	0.99
AGA	1	0	0
AGC	1	0	0
AGG	1	0	0
AGT	1	0	0
ATA	0	0.01	0.99
ATC	1	0	0
ATG	0	0.01	0.99
ATT	0	0.01	0.99
CAA	1	0	0
CAC	1	0	0
CAG	1	0	0
CAT	0	0.01	0.99
CCA	0.98	0	0.02
CCC	1	0	0
CCG	1	0	0
CCT	0	0.01	0.99
CGA	0	0.01	0.99
CGC	1	0	0
CGG	1	0	0
CGT	1	0	0
CTA	0	0.01	0.99
CTC	0	0.01	0.99
CTG	0	0.01	0.99
CTT	0	0.01	0.99
GAA	0	0.01	0.99
GAC	0.18	0.01	0.81
GAG	1	0	0
GAT	1	0	0
GCA	1	0	0
GCC	1	0	0
GCG	0	0.01	0.99

GCT	0	0.01	0.99
GGA	1	0	0
GGC	1	0	0
GGG	1	0	0
GGT	1	0	0
GTA	1	0	0
GTC	1	0	0
GTG	1	0	0
GTT	1	0	0
TAA	1	0	0
TAC	0	0.01	0.99
TAG	1	0	0
TAT	1	0	0
TCA	1	0	0
TCC	1	0	0
TCG	0	0.01	0.99
TCT	0	0.01	0.99
TGA	1	0	0
TGC	1	0	0
TGG	1	0	0
TGT	1	0	0
TTA	1	0	0
TTC	0.03	0.01	0.96
TTG	0	0.01	0.99
TTT	0.18	0.01	0.81

9. List of chains selected for MSA prediction from PWM

1A0A_B, 1ZET_A, 3D6Z_A, 4H5Q_B, 5GNJ_I, 6IFM_B, 1A3Q_A, 1ZM5_A, 3DPG_B, 4HC9_A, 5GZB_A, 6J9C_D, 1AM9_A, 1ZME_D, 3ERE_D, 4H DU_A, 5H1C_B, 6JBX_A, 1B69_A, 1ZS4_A, 3GOX_A, 4HF1_A, 5H3R_A, 6JE4_K, 1B94_B, 1ZX4_A, 3GZ6_A, 4HP3_C, 5HF7_A, 6JRP_D, 1BF5_A, 1ZZI_A, 3H0D_A, 4HQE_A, 5HLG_E, 6JYW_A, 1BVO_A, 2A3V_C, 3HQF_A, 4HT4_A, 5HP4_A, 6L8E_C, 1CF7_A, 2A6O_B, 3HQG_A, 4I2O_B, 5HR4_C, 6M0W_A, 1CF7_B, 2AOR_A, 3I0X_A, 4IHS_B, 5I2D_R, 6M3D_C, 1CKT_A, 2AS5_N, 3IAY_A, 4IX7_B, 5I44_I, 6MRJ_D, 1D02_A, 2AYB_B, 3IGM_A, 4IZZ_A, 5J2Y_A, 6NSM_B, 1D3U_A, 2BAM_B, 3IMB_B, 4J19_B, 5JUB_B, 6NY1_Y, 1D5Y_A, 2BNZ_D, 3JCA_E, 4J1J_A, 5K1Y_F, 6O6P_B, 1DC1_A, 2BOP_A, 3JR5_A, 4JBM_A, 5K5J_A, 6O8E_A, 1DFM_A, 2C9L_Z, 3JRH_B, 4JQD_E, 5L6L_D, 6OD4_B, 1DH3_C, 2D5V_B, 3JSP_B, 4K8X_A, 5LD2_D, 6ON0_B, 1DMU_A, 2DNJ_A, 3JTG_A,

4KDP_B, 5LEJ_B, 6OPM_B, 1DP7_P, 2DPD_B, 3JXD_L, 4L18_E, 5M0Q_A, 6P0S_E, 1EA4_K, 2DRP_D,
 3K57_A, 4L62_M, 5MHK_A, 6P1A_A, 1F4S_P, 2ERG_A, 3KDE_C, 4LMG_C, 5MPF_B, 6P7E_B, 1FIU_B,
 2EVJ_A, 3KET_A, 4LUP_A, 5N9F_A, 6PBD_A, 1FOK_A, 2F03_C, 3KQL_B, 4LVL_A, 5NPK_D, 6PIJ_G,
 1G38_D, 2FIO_A, 3L1P_A, 4MDE_B, 5O63_B, 6PQU_A, 1GCC_A, 2FL3_A, 3LSP_A, 4MTE_D, 5O6D_B,
 6PSR_L, 1GD2_G, 2FO1_A, 3M4A_A, 4MZR_B, 5SWW_C, 6Q2B_B, 1H6F_B, 2G1P_A, 3M7K_A,
 4N0O_E, 5T7X_B, 6QEC_A, 1H88_B, 2GB7_A, 3MKW_B, 4NHJ_A, 5TH3_B, 6QFD_A, 1HBX_A,
 2GLO_A, 3MZH_A, 4NID_A, 5TKZ_B, 6QIB_A, 1IAW_B, 2H27_A, 3N6S_A, 4NQA_H, 5TRD_B, 6RYL_E,
 1IXY_B, 2H3A_B, 3NIC_E, 4ON0_B, 5U07_D, 6S48_A, 1J46_A, 2HAP_D, 3O3G_A, 4P9U_F, 5U30_A,
 6T1F_B, 1J5N_A, 2HZV_A, 3O9X_A, 4PZI_A, 5UAN_B, 6T21_B, 1JFT_A, 2IRF_L, 3ON0_D, 4Q5V_E,
 5V09_Z, 6U81_A, 1JJ8_C, 2K1N_B, 3OQO_C, 4QTJ_A, 5VFX_A, 6U8V_A, 1JT0_C, 2K7F_A, 3OR3_B,
 4R4E_B, 5VL9_A, 6U9Q_A, 1L3L_A, 2KO0_A, 3OSG_A, 4R8A_F, 5VPE_C, 6UKE_X, 1LFU_P, 2KY8_A,
 3OYM_A, 4RB1_B, 5VPF_B, 6USP_A, 1LLM_D, 2LEF_A, 3PVI_B, 4RDM_B, 5VXN_D, 6V0V_A, 1LQ1_D,
 2LEX_A, 3PVP_B, 4RKG_B, 5WJR_A, 6V93_A, 1LRR_A, 2LKX_A, 3RI4_D, 4S04_B, 5WTI_Z, 6VBW_G,
 1MJ2_B, 2LT7_A, 3RMP_A, 4S0N_A, 5WVZ_B, 6VDD_A, 1MSW_D, 2LTT_B, 3SQI_A, 4WLW_A, 5X11_A,
 6W1F_A, 1NFK_A, 2MF8_A, 3SWM_B, 4WUH_B, 5X2G_A, 6W64_A, 1NKP_A, 2MRU_A, 3TAB_B,
 4X0P_C, 5X5L_H, 6WC2_D, 1NLW_D, 2MXF_A, 3U3W_B, 4X4C_C, 5XFQ_A, 6WE9_A, 1ODH_A,
 2NQJ_A, 3UBT_A, 4XVK_A, 5XH7_A, 6Y42_B, 1OWF_A, 2O49_A, 3V21_E, 4Y00_B, 5XPA_A, 6Y93_B,
 1P71_B, 2O5E_B, 3VK8_B, 4YJ0_C, 5XUT_A, 6YCQ_A, 1PAR_B, 2O6M_B, 3VKE_B, 4YO2_A, 5XVP_E,
 6Z2Y_B, 1PDN_C, 2O8K_A, 3W6V_A, 4YPR_A, 5Y3L_B, 7C9O_A, 1PP7_U, 2OEH_A, 3WAZ_A, 4Z59_A,
 5YBB_G, 7CCD_B, 1PUE_E, 2P5L_D, 3WGI_D, 4ZCF_B, 5YEJ_B, 7CE1_D, 1PUF_A, 2P7C_B,
 3WVG_C, 4ZM0_A, 5YI2_J, 1QNE_B, 2PE5_B, 3ZHM_A, 4ZPK_A, 5YIV_C, 1QRH_A, 2PI0_B, 3ZI5_A,
 4ZQ9_B, 5YWT_A, 1QRV_A, 2PJR_A, 3ZKC_A, 4ZSF_A, 5YZY_C, 1R71_B, 2QBY_A, 3ZPL_B, 5B2T_B,
 5ZAD_B, 1R8D_B, 2QBY_B, 3ZQC_A, 5B7J_A, 5ZD4_C, 1SC7_A, 2QHB_A, 3ZQL_B, 5BMZ_B, 5ZKI_B,
 1SKN_P, 2QK9_A, 3ZVK_G, 5BPD_A, 5ZR1_A, 1TAU_A, 2QSG_A, 4A12_B, 5C8E_C, 5ZUX_A, 1TC3_C,
 2RBF_A, 4A8S_B, 5CIZ_A, 5ZYU_B, 1TRO_A, 2UP1_A, 4AIK_A, 5CM3_A, 6A57_A, 1TW8_B, 2UYH_A,
 4AWL_A, 5CQQ_A, 6AMK_A, 1U8B_A, 2UZK_A, 4DA4_B, 5CZZ_A, 6ASB_F, 1U8R_B, 2VLA_A, 4DKJ_A,
 5D39_D, 6BLW_A, 1V14_B, 2VTB_A, 4DM0_A, 5D8K_B, 6C2S_C, 1VFC_A, 2VY2_A, 4DS4_D, 5D9Y_A,
 6C31_A, 1VRR_B, 2WTY_B, 4E0G_A, 5DFJ_A, 6CQ2_A, 1VTN_C, 2XKK_C, 4ED6_A, 5DNY_B,
 6DG0_A, 1WB9_A, 2XRO_E, 4EGZ_B, 5DY0_A, 6EN8_C, 1WTE_A, 2XRZ_B, 4ESJ_A, 5ED4_A,
 6ERG_E, 1WVL_A, 2YVH_A, 4FF1_B, 5EJK_E, 6FB6_B, 1XSD_A, 2Z33_A, 4FTH_A, 5EYO_C, 6FQP_B,
 1YFI_B, 2Z9O_B, 4FX4_B, 5F0Q_A, 6FWS_A, 1YFL_E, 3A01_E, 4GCL_E, 5FD3_A, 6FZT_A, 1YRN_B,
 3BDN_A, 4GDF_A, 5FFJ_B, 6G1L_A, 1YUI_A, 3BM3_A, 4GLE_A, 5FHD_A, 6GCF_B, 1Z1B_B, 3BS1_A,
 4GZN_C, 5FMF_V, 6GTC_A, 1Z9C_A, 3CRO_R, 4H10_B, 5FMP_B, 6HP7_B,

10. System-size for MD simulations of DNA-protein complexes

Sr. No	Protein	DNA	Number of atoms	Number of water molecules	Number of ions
1	Myc wild-type	Wild-type	5351	32841	34
2	Myc mutant	Mutant	5341	32832	36
3	Myc wild-type	Mutant	5355	32844	34
4	Myc mutant	Wild-type	5337	32730	36
5	Ets1 wild-type	Wild-type	2893	7872	23
6	Ets1 mutant	Mutant	2896	4659	22
7	Ets1 wild-type	Mutant	2893	4650	23
8	Ets1 mutant	Wild-type	2896	4656	22
9	EcoRV wild-type	Wild-type	8738	10905	6
10	EcoRV mutant	Mutant	8738	10905	6
11	EcoRV wild-type	Mutant	8726	10905	6
12	EcoRV mutant	Wild-type	8740	10905	6
13	Zif-268 wild-type	Wild-type	2202	4161	4
14	zif-268 MT1	MT1	2195	7398	3
15	zif-268 MT2	MT2	2193	7707	4
16	Zif-268 wild-type#	Wild-type	2119	4797	3
17	Zif-268 MT1	MT1	2105	3471	3
18	Zif-268 MT2	MT2	2105	3471	2
19	Zif-268 wild-type	MT1	2119	3345	3
20	Zif-268 wild-type	MT2	2119	3498	3
21	Zif-268 MT1	Wild-type	2081	3456	4
22	Zif-268 MT2	Wild-type	2103	4599	2

11. RMSD of wild-type and mutant proteins with DNA substrates (Å).

WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is

obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal structure. The values in bold represent the lowest RMSD for each protein.

Sr. No	Protein	DNA	Run 1	Run 2	Run 3	Average	p-value
1	Zif-268 WT	WT	1.23 +/- 0.3	1.99 +/- 0.8	1.91 +/- 0.6	1.71 +/- 0.7	-
2	Zif-268 MT1	MT1	2.54 +/- 0.7	3.34 +/- 0.9	2.72 +/- 0.9	2.87 +/- 0.9	<0.00001
3	Zif-268 MT2	MT2	2.03 +/- 0.6	3.5 +/- 0.7	3.48 +/- 1.3	3.01 +/- 1.1	<0.00001
4	Zif-268 WT #	WT	1.18 +/- 0.3	1.89 +/- 0.7	1.61 +/- 0.5	1.55 +/- 0.6	0.065
5	Zif-268 MT1 #	MT1	1.15 +/- 0.3	1.34 +/- 0.5	1.16 +/- 0.3	1.22 +/- 0.4	<0.00001
6	Zif-268 MT2 #	MT2	1.17 +/- 0.5	1.10 +/- 0.2	1.21 +/- 0.3	1.16 +/- 0.3	<0.00001
7	Zif-268 WT	MT1	1.71 +/- 0.7	0.73 +/- 0.2	1.50 +/- 0.5	1.31 +/- 0.6	0.00002
8	Zif-268 WT	MT2	1.72 +/- 0.7	1.45 +/- 0.4	1.12 +/- 0.4	1.44 +/- 0.6	0.01030
9	Zif-268 MT1	WT	0.98 +/- 0.3	1.32 +/- 0.3	1.16 +/- 0.2	1.15 +/- 0.3	<0.00001
10	Zif-268 MT2	WT	1.65 +/- 0.4	1.13 +/- 0.2	1.71 +/- 0.7	1.50 +/- 0.6	0.061
11	Myc WT	WT	4.32 +/- 1.4	5.10 +/- 1.5	4.09 +/- 1.0	4.50 +/- 1.4	-
12	Myc MT	MT	5.20 +/- 1.4	5.20 +/- 1.8	4.51 +/- 1.8	4.97 +/- 1.7	0.062
13	Myc WT	MT	3.13 +/- 1.3	6.01 +/- 2.5	3.70 +/- 1.3	4.28 +/- 2.2	0.647
14	Myc MT	WT	3.28 +/- 1.1	4.74 +/- 1.1	3.56 +/- 1.1	3.86 +/- 1.6	0.004
15	Ets1 WT	WT	0.82 +/- 0.2	0.67 +/- 0.2	0.77 +/- 0.2	0.76 +/- 0.2	-
16	Ets1 MT	MT	0.81 +/- 0.3	0.90 +/- 0.3	0.89 +/- 0.3	0.87 +/- 0.3	0.032
17	Ets1 WT	MT	0.82 +/- 0.2	0.81 +/- 0.2	0.90 +/- 0.3	0.85 +/- 0.3	0.129
18	Ets1 MT	WT	0.80 +/- 0.2	0.88 +/- 0.3	1.1 +/- 0.4	0.93 +/- 0.3	0.0001
19	EcoRV WT	WT	0.73 +/- 0.2	0.80 +/- 0.2	0.71 +/- 0.1	0.75 +/- 0.2	-
20	EcoRV MT	MT	0.76 +/- 0.2	0.90 +/- 0.2	0.77 +/- 0.1	0.81 +/- 0.2	0.06
21	EcoRV WT	MT	0.88 +/- 0.2	0.74 +/- 0.1	0.74 +/- 0.1	0.79 +/- 0.2	0.43
22	EcoRV MT	WT	0.98 +/- 0.2	0.72 +/- 0.1	0.83 +/- 0.2	0.85 +/- 0.2	0.001

12. Centroid distance between wild-type and mutant DNA-protein complexes (Å).

WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal

structure. The values in bold represent the smallest centroid distance with a DNA substrate for each protein.

Sr. No	Protein	DNA	Run 1	Run 2	Run 3	Average	p-value
1	Zif-268 WT	WT	3.10 +/- 0.3	3.65 +/- 0.6	2.45 +/- 0.3	3.06 +/- 0.7	-
2	Zif-268 MT1	MT1	3.81 +/- 0.6	2.58 +/- 0.8	3.24 +/- 0.6	3.21 +/- 0.8	0.167
3	Zif-268 MT2	MT2	3.08 +/- 0.5	2.24 +/- 0.6	4.17 +/- 0.7	3.16 +/- 1.0	0.599
4	Zif-268 WT #	WT	3.26 +/- 0.3	3.80 +/- 0.5	4.49 +/- 0.4	3.85 +/- 0.7	<0.00001
5	Zif-268 MT1 #	MT1	2.63 +/- 0.3	3.94 +/- 0.7	3.80 +/- 0.4	3.46 +/- 0.8	<0.00001
6	Zif-268 MT2 #	MT2	3.15 +/- 0.4	4.00 +/- 0.7	3.95 +/- 0.5	3.70 +/- 0.7	<0.00001
7	Zif-268 WT	MT1	3.11 +/- 0.7	1.58 +/- 0.2	2.95 +/- 0.4	2.55 +/- 0.8	<0.00001
8	Zif-268 WT	MT2	4.52 +/- 0.8	2.87 +/- 0.3	3.41 +/- 0.4	3.60 +/- 0.9	<0.00001
9	Zif-268 MT1	WT	4.33 +/- 0.5	5.01 +/- 0.5	4.36 +/- 0.3	4.57 +/- 0.5	<0.00001
10	Zif-268 MT2	WT	3.90 +/- 0.4	3.11 +/- 0.3	3.65 +/- 0.6	3.55 +/- 0.6	<0.00001
11	Myc WT	WT	18.12 +/- 0.9	16.57 +/- 0.8	15.90 +/- 0.7	16.86 +/- 1.6	-
12	Myc MT	MT	17.41 +/- 0.7	17.31 +/- 1.2	15.69 +/- 1.0	16.81 +/- 1.3	0.955
13	Myc WT	MT	17.01 +/- 0.7	18.71 +/- 1.1	18.03 +/- 0.6	17.91 +/- 1.1	<0.00001
14	Myc MT	WT	18.25 +/- 1.1	17.95 +/- 0.7	19.23 +/- 0.7	18.48 +/- 1.0	<0.00001
15	Ets1 WT	WT	24.11 +/- 1.5	22.70 +/- 1.2	22.61 +/- 0.7	22.47 +/- 1.6	-
16	Ets1 MT	MT	19.50 +/- 0.3	21.95 +/- 0.4	19.6 +/- 0.4	20.37 +/- 0.2	<0.00001
17	Ets1 WT	MT	18.95 +/- 0.3	20.83 +/- 0.6	20.28 +/- 0.5	20.02 +/- 0.9	<0.00001
18	Ets1 MT	WT	18.15 +/- 0.3	20.53 +/- 0.4	18.68 +/- 0.7	19.11 +/- 1.1	<0.00001
19	EcoRV WT	WT	9.69 +/- 0.2	11.14 +/- 0.4	10.53 +/- 0.3	10.45 +/- 0.7	-
20	EcoRV MT	MT	10.89 +/- 0.2	10.52 +/- 0.3	10.01 +/- 0.4	10.47 +/- 0.5	0.97
21	EcoRV WT	MT	10.59 +/- 0.3	10.70 +/- 0.3	9.91 +/- 0.3	10.40 +/- 0.5	0.48
22	EcoRV MT	WT	9.99 +/- 0.5	10.18 +/- 0.3	9.94 +/- 0.2	10.04 +/- 0.4	<0.00001

13. Total hydrogen bonds between different wild-type and mutant DNA-protein complexes. WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal

structure. The values in bold represent the DNA-protein complex with the highest number of hydrogen bonds.

Sr. No	Protein	DNA	Run 1	Run 2	Run 3	Average	p-value
1	Zif-268 WT	WT	48.64 +/- 5.2	36.74 +/- 9.6	26.80 +/- 6.3	37.39 +/- 11.5	-
2	Zif-268 MT1	MT1	29.46 +/- 4.2	45.68 +/- 7.7	46.34 +/- 9.3	40.39 +/- 10.7	0.008
3	Zif-268 MT2	MT2	25.50 +/- 4.2	46.24 +/- 4.3	43.26 +/- 5.6	38.33 +/- 10.3	0.850
4	Zif-268 WT #	WT	49.36 +/- 3.9	46.82 +/- 4.6	41.26 +/- 3.6	45.81 +/- 5.3	<0.00001
5	Zif-268 MT1 #	MT1	45.02 +/- 5.1	47.42 +/- 5.8	48.38 +/- 6.5	46.94 +/- 6.0	<0.00001
6	Zif-268 MT2 #	MT2	49.56 +/- 4.5	46.70 +/- 5.1	40.06 +/- 6.3	45.44 +/- 6.6	<0.00001
7	Zif-268 WT	MT1	47.26 +/- 6.5	47.00 +/- 4.3	47.00 +/- 4.3	47.09 +/- 5.1	<0.00001
8	Zif-268 WT	MT2	40.64 +/- 5.6	44.96 +/- 3.9	46.12 +/- 6.0	43.90 +/- 5.7	<0.00001
9	Zif-268 MT1	WT	43.70 +/- 6.1	43.52 +/- 4.9	44.20 +/- 5.1	43.80 +/- 5.4	<0.00001
10	Zif-268 MT2	WT	42.22 +/- 4.3	56.44 +/- 4.8	45.02 +/- 5.4	46.89 +/- 7.8	<0.00001
11	Myc WT	WT	48.28 +/- 5.9	46.40 +/- 5.0	45.30 +/- 4.9	46.66 +/- 5.4	-
12	Myc MT	MT	31.56 +/- 4.7	32.42 +/- 4.5	33.32 +/- 4.2	32.43 +/- 4.5	<0.00001
13	Myc WT	MT	38.80 +/- 4.3	36.82 +/- 4.5	43.54 +/- 4.6	39.72 +/- 5.2	<0.00001
14	Myc MT	WT	37.36 +/- 5.3	44.26 +/- 5.6	37.72 +/- 5.0	39.75 +/- 6.2	<0.00001
15	Ets1 WT	WT	26.96 +/- 3.4	30.88 +/- 3.4	30.92 +/- 3.3	29.59 +/- 3.9	-
16	Ets1 MT	MT	31.62 +/- 4.4	27.76 +/- 5.1	30.00 +/- 2.9	29.79 +/- 4.5	0.970
17	Ets1 WT	MT	34.22 +/- 2.9	29.4 +/- 2.5	31.36 +/- 3.4	31.66 +/- 3.6	0.0001
18	Ets1 MT	WT	25.34 +/- 3.9	27.58 +/- 3.2	24.74 +/- 5.8	25.85 +/- 4.6	<0.00001
19	EcoRV WT	WT	64.06 +/- 4.3	72.54 +/- 6.4	75.68 +/- 5.0	70.06 +/- 7.2	-
20	EcoRV MT	MT	63.86 +/- 6.5	74.16 +/- 5.3	64.76 +/- 7.9	70.93 +/- 7.9	0.997
21	EcoRV WT	MT	67.18 +/- 6.9	69.4 +/- 4.8	85.54 +/- 6.6	74.37 +/- 10.6	0.0047
22	EcoRV MT	WT	63.44 +/- 6.3	76.86 +/- 6.0	65.06 +/- 7.1	68.45 +/- 8.8	0.053

14. Nucleotide-specific hydrogen bonds between wild-type and mutant DNA-protein complexes.

WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal

structure. The values in bold represent the DNA-protein complex with the highest number of hydrogen bonds.

Sr. No	Protein	DNA	Run1	Run2	Run3	Average	p-value
1	Zif-268 WT	WT	16.64 +/- 2.0	11.06 +/- 3.2	11.92 +/- 2.9	13.21 +/- 3.7	-
2	Zif-268 MT1	MT1	14.56 +/- 1.8	12.96 +/- 3.0	15.56 +/- 2.4	14.36 +/- 2.6	0.001
3	Zif-268 MT2	MT2	9.94 +/- 2.1	13.24 +/- 2.6	12.56 +/- 1.5	11.91 +/- 2.5	0.001
4	Zif-268 WT #	WT	17.38 +/- 2.1	15.74 +/- 1.6	10.30 +/- 1.8	14.47 +/- 3.5	0.00006
5	Zif-268 MT1 #	MT1	14.40 +/- 2.1	16.46 +/- 2.4	16.80 +/- 2.7	15.89 +/- 2.6	<0.00001
6	Zif-268 MT2 #	MT2	17.78 +/- 2.2	17.46 +/- 2.3	14.72 +/- 2.3	16.65 +/- 2.7	<0.00001
7	Zif-268 WT	MT1	13.38 +/- 2.0	14.86 +/- 1.6	14.44 +/- 1.8	14.23 +/- 1.9	0.006
8	Zif-268 WT	MT2	16.58 +/- 1.8	16.48 +/- 1.7	16.80 +/- 1.9	16.62 +/- 1.8	<0.00001
9	Zif-268 MT1	WT	15.54 +/- 2.4	14.92 +/- 1.6	13.50 +/- 1.3	14.65 +/- 2.0	0.00005
10	Zif-268 MT2	WT	12.74 +/- 2.1	17.28 +/- 1.8	14.74 +/- 2.2	14.92 +/- 2.7	<0.00001
11	Myc WT	WT	7.44 +/- 1.6	5.80 +/- 2.2	6.10 +/- 2.3	6.45 +/- 2.2	-
12	Myc MT	MT	5.52 +/- 2.1	5.34 +/- 1.5	4.80 +/- 1.6	5.22 +/- 1.8	<0.00001
13	Myc WT	MT	7.32 +/- 2.0	6.32 +/- 1.7	8.62 +/- 2.5	7.42 +/- 2.3	0.0002
14	Myc MT	WT	3.72 +/- 2.3	9.36 +/- 2.4	4.6 +/- 1.7	5.89 +/- 3.3	0.083
15	Ets1 WT	WT	7.66 +/- 0.7	7.48 +/- 1.0	7.52 +/- 0.9	7.55 +/- 3.8	-
16	Ets1 MT	MT	6.42 +/- 0.9	5.53 +/- 1.5	6.4 +/- 1.1	6.11 +/- 4.5	<0.00001
17	Ets1 WT	MT	8.62 +/- 0.9	6.52 +/- 1.0	8.2 +/- 1.3	7.78 +/- 3.6	0.391
18	Ets1 MT	WT	4.72 +/- 1.5	5.18 +/- 1.1	6.02 +/- 1.5	5.31 +/- 4.6	<0.00001
19	EcoRV WT	WT	18.64 +/- 1.6	18.82 +/- 1.7	20.06 +/- 1.8	19.17 +/- 1.8	-
20	EcoRV MT	MT	19.10 +/- 2.2	21.60 +/- 1.5	17.74 +/- 2.9	19.48 +/- 2.8	0.629
21	EcoRV WT	MT	18.14 +/- 3.3	20.02 +/- 1.8	26.94 +/- 3.4	21.7 +/- 4.8	<0.0001
22	EcoRV MT	WT	17.50 +/- 2.1	19.44 +/- 1.4	20.12 +/- 2.5	19.02 +/- 2.3	0.942

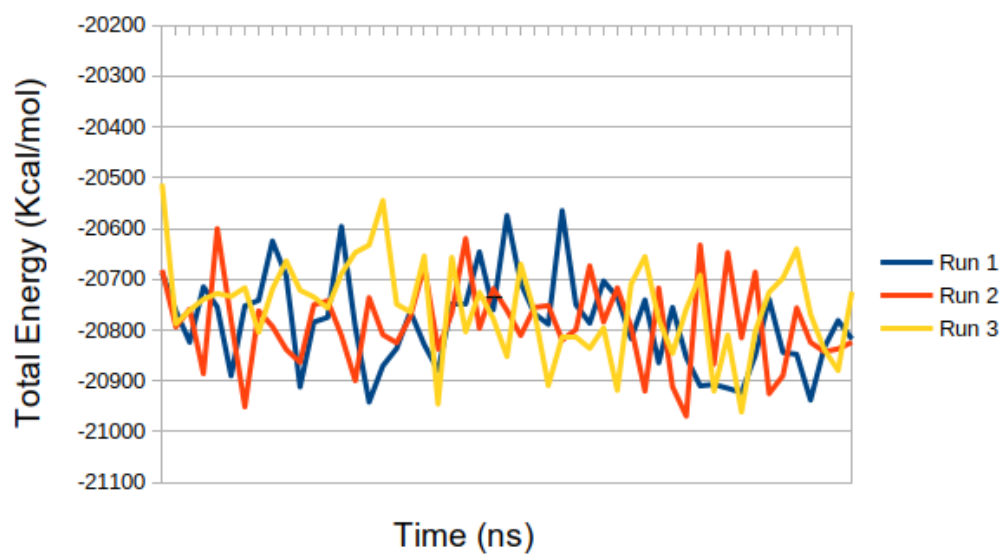
15. Binding affinities of wild-type and mutant proteins with DNA substrates (kcal/mol).

WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal structure. The values in bold represent the best binding energy for each protein.

Sr. No	Protein	DNA	Run 1	Run 2	Run 3	Average	p-value
1	Zif-268 WT	WT	-60.8 +/- 9.3	-70.8 +/- 8.0	-55.6 +/- 9.3	-62.4 +/- 10.9	-
2	Zif-268 MT1	MT1	-85.3 +/- 7.7	-53.3 +/- 10.8	-47.2 +/- 11.1	-61.9 +/- 19.5	0.6614
3	Zif-268 MT2	MT2	-131.2 +/- 8.3	-85.6 +/- 8.9	-72.9 +/- 7.3	-96.6 +/- 26.4	<0.00001
4	Zif-268 WT #	WT	-60.9 +/- 7.6	-63.6 +/- 9.4	-74.9 +/- 9.4	-66.5 +/- 10.8	<0.00001
5	Zif-268 MT1 #	MT1	-48.8 +/- 10.0	-78.5 +/- 6.2	-56.0 +/- 6.9	-61.1 +/- 14.9	0.06000
6	Zif-268 MT2 #	MT2	-98.9 +/- 8.1	-80.4 +/- 6.9	-57.3 +/- 7.6	-78.9 +/- 18.6	<0.00001
7	Zif-268 WT	MT1	-70.5 +/- 7.7	-95.3 +/- 8.4	-84.8 +/- 8.1	-83.5 +/- 13.0	<0.00001
8	Zif-268 WT	MT2	-40.2 +/- 6.0	-52.9 +/- 7.1	-49.5 +/- 7.6	-47.6 +/- 8.8	<0.00001
9	Zif-268 MT1	WT	-49.3 +/- 5.9	-80.1 +/- 7.7	-82.6 +/- 8.9	-70.7 +/- 16.9	<0.00001
10	Zif-268 MT2	WT	-68.1 +/- 11.6	-108.3 +/- 7.8	-57.4 +/- 7.0	-77.9 +/- 23.7	<0.00001
11	Myc WT	WT	-89.9 +/- 17.7	-121.4 +/- 10.8	-111.6 +/- 10.6	-107.6 +/- 18.8	-
12	Myc MT	MT	-76.4 +/- 10.6	-58.9 +/- 7.4	-62.5 +/- 9.8	-65.9 +/- 12.0	<0.00001
13	Myc WT	MT	-94.4 +/- 7.8	-70.2 +/- 6.7	-85.3 +/- 10.0	-83.3 +/- 13.0	<0.00001
14	Myc MT	WT	-49.5 +/- 9.3	-99.6 +/- 14.2	-97.7 +/- 9.7	-82.3 +/- 25.8	<0.00001
15	Ets1 WT	WT	-60.5 +/- 4.9	-71.5 +/- 6.9	-75.5 +/- 7.8	-69.2 +/- 9.2	-
16	Ets1 MT	MT	-75.5 +/- 8.6	-53.5 +/- 6.5	-69.5 +/- 7.1	-66.2 +/- 11.9	<0.00001
17	Ets1 WT	MT	-93.7 +/- 6.2	-59.8 +/- 7.2	-67.6 +/- 9.0	-73.7 +/- 16.4	<0.00001
18	Ets1 MT	WT	-72.4 +/- 7.1	-65.7 +/- 8.2	-49.9 +/- 5.9	-62.7 +/- 11.8	<0.00001
19	EcoRV WT	WT	-97.7 +/- 12.0	-140.8 +/- 12.0	-129.0 +/- 14.9	-122.5 +/- 22.4	-
20	EcoRV MT	MT	-103.2 +/- 13.6	-123.4 +/- 11.1	-83.6 +/- 12.2	-103.4 +/- 20.4	<0.00001
21	EcoRV WT	MT	-112.8 +/- 11.7	-109.9 +/- 14.6	-148.5 +/- 14.7	-123.7 +/- 22.3	0.24288
22	EcoRV MT	WT	-94.5 +/- 9.3	-135.5 +/- 10.6	-83.3 +/- 9.4	-104.4 +/- 24.5	<0.00001

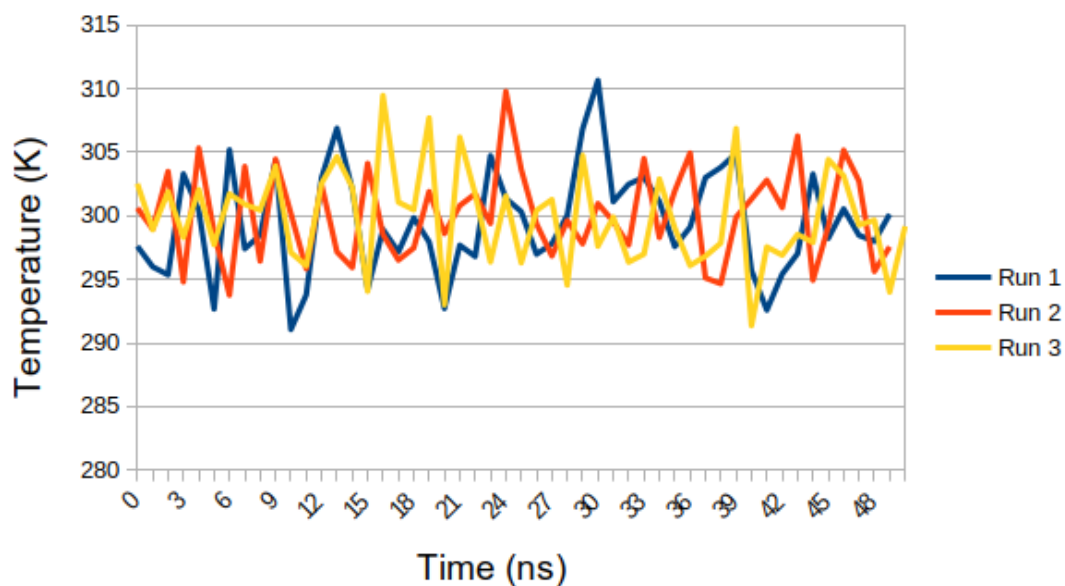
16. Total energy of the Zif-268 wild-type protein with wild-type DNA.

The x-axis is the time in ns and y-axis is the energy in kcal/mol. The three runs are shown in blue, orange and yellow.



17. Temperature of the system with Zif-268 wild-type DNA-protein complex.

The x-axis is the time in ns and y-axis is the temperature in K. The three runs are shown in blue, orange and yellow.



18. Retention of hydrogen bonds present in the crystal structure of the proteins.

The values are average of all hydrogen bonds and the maximum value is 50 (Number of frames). WT and MT stand for wild-type and mutant, respectively. # indicates that the models of these proteins were used instead of the crystal structure. The p-value in the last column is obtained through one-way ANOVA test with respect to the wild-type DNA-protein crystal structure. The values in bold represent the best binding energy for each protein.

Sr. No	Protein	DNA	Run1	Run2	Run3	Average	P-value
1	Zif-268 WT	WT	40.1 +/- 13.0	22.8 +/- 19.1	29.1 +/- 18.3	30.6 +/- 18.1	-
2	Zif-268 MT1	MT1	31.8 +/- 18.7	25.1 +/- 17.5	30.4 +/- 17.8	29.1 +/- 18.3	0.99
3	Zif-268 MT2	MT2	28.9 +/- 21.6	34.9 +/- 18.0	30.4 +/- 22.2	31.3 +/- 20.3	0.99
4	Zif-268 WT #	WT	35.8 +/- 16.7	32.9 +/- 18.8	22.7 +/- 23.0	30.5 +/- 20.1	0.99
5	Zif-268 MT1 #	MT1	32.1 +/- 17.5	40.0 +/- 14.3	35.7 +/- 17.7	35.9 +/- 16.5	0.42
6	Zif-268 MT2 #	MT2	43.4 +/- 9.6	43.5 +/- 10.4	39.6 +/- 13.3	43.5 +/- 11.1	0.06
7	Zif-268 WT	MT1	25.1 +/- 21.7	29.9 +/- 22.0	25.9 +/- 21.4	26.9 +/- 21.4	0.88
8	Zif-268 WT	MT2	32.6 +/- 18.3	34.6 +/- 20.0	35.1 +/- 17.6	33.2 +/- 18.3	0.9
9	Zif-268 MT1	WT	35.7 +/- 18.1	36.4 +/- 16.8	33.9 +/- 19.6	35.3 +/- 17.8	0.57
10	Zif-268 MT2	WT	34.2 +/- 18.0	37.6 +/- 17.0	40.7 +/- 14.1	37.5 +/- 16.3	0.15
11	Myc WT	WT	22.1 +/- 19.4	21.4 +/- 18.8	22.2 +/- 17.3	21.9 +/- 17.9	-
12	Myc MT	MT	13.1 +/- 18.5	10.2 +/- 13.1	16.4 +/- 15.3	13.2 +/- 15.4	0.17
13	Myc WT	MT	26.8 +/- 19.1	24.2 +/- 16.9	27.7 +/- 17.7	26.2 +/- 17.4	0.73
14	Myc MT	WT	9.9 +/- 14.4	12.0 +/- 15.8	12.4 +/- 14.8	11.4 +/- 14.5	0.07
15	Ets1 WT	WT	42.3 +/- 17.8	43.6 +/- 12.6	45.6 +/- 9.6	43.8 +/- 13.1	-
16	Ets1 MT	MT	29.6 +/- 20.8	17.1 +/- 20.1	16.3 +/- 21.0	21.0 +/- 20.5	0.00008
17	Ets1 WT	MT	43.3 +/- 9.2	36.6 +/- 18.7	44.0 +/- 12.8	41.3 +/- 13.8	0.96
18	Ets1 MT	WT	25.0 +/- 20.6	28.1 +/- 22.5	24.0 +/- 20.0	25.7 +/- 20.0	0.004
19	EcoRV WT	WT	34.2 +/- 19.8	36.7 +/- 14.4	37.1 +/- 16.3	36.0 +/- 16.7	-
20	EcoRV MT	MT	32.3 +/- 17.2	32.7 +/- 19.6	24.3 +/- 18.3	29.7 +/- 18.5	0.74
21	EcoRV WT	MT	26.6 +/- 18.8	34.4 +/- 18.2	40.4 +/- 15.5	32.2 +/- 19.6	0.99
22	EcoRV MT	WT	28.3 +/- 19.7	34.5 +/- 16.9	31.6 +/- 14.1	31.4 +/- 17.0	0.96

19. Distance of N and O atoms with DNA in the PDNA-285 datasets

Distance	Frequency	Frequency (%)
0.0	0	0.0

0.5	2	0.0
1.0	19	0.2
1.5	83	0.8
2.0	223	2.1
2.5	500	4.7
3.0	1455	13.7
3.5	2842	26.7
4.0	5521	51.9

20. Accuracy of prediction of amino acid specific for a sequence after three-fold cross-validation.

Amino Acid	Training			Testing		
	Top1	Top2	Top3	Top1	Top2	Top3
C	0+/- 0	0+/- 0	0+/- 0	0+/- 0	0+/- 0	0+/- 0
DE	20+/- 6.5	75.0+/-4.6	81.8+/-3.	20.2+/-92	79.5+/-14.6	79.8+/-14.7
H	3.4+/- 5.9	16.0+/-4.9	56.0+/-5.7	0+/- 0	25.5+/-22.1	52.3+/-8.3
K	5.7+/- 1.8	67.0+/-1.0	87.1+/-2.7	3.6+/-3.1	73.5+/-3.2	90.5+/-3.2
NQ	37.0 +/- 7.4	70.8+/-3.9	82.9+/-07	34.1+/-3.1	63.8+/-9.0	75.6+/-106
R	84.3+/- 3.3	90.5+/-2.2	92.0+/-1.3	82.5+/-3.7	88.4+/-3.2	88.9+/-2/7
ST	10.2+/-2.9	51.5+/-2.8	78.5+/-0.4	16.0+/-9.3	58.5+/-7.9	79.8+/-3.6
W	0+/- 0	0+/- 0	0+/- 0	0+/- 0	0+/- 0	0+/- 0
Y	26.1+/-5.3	46.9+/-8.9	69.2+/-1.4	27.8+/-25.5	47.8+/-13.5	61.7+/-12.6
Mc	48.9+/-9.0	72.2+/-1.7	85.1+/-3.3	41.2+/-15.6	70.5+/-5.5	85.0+/-2.6

21. Comparison of intra and intermolecular potential energy

Sr. No	Protein	DNA	Intermolecular Energy	Intramolecular Energy
1	Zif-268 WT	WT	-62.4 +/- 10.9	-3820.78 +/- 45.4
2	Zif-268 MT1	MT1	-61.9 +/- 19.5	-3579.79 +/- 32.0
3	Zif-268 MT2	MT2	-96.6 +/- 26.4	-3432.16 +/- 39.7
4	Zif-268 WT #	WT	-66.5 +/- 10.8	-3581.17 +/- 33.3
5	Zif-268 MT1 #	MT1	-61.1 +/- 14.9	-3316.04 +/- 40.0
6	Zif-268 MT2 #	MT2	-78.9 +/- 18.6	-3217.29 +/- 29.2

7	Zif-268 WT	MT1	-83.5 +/- 13.0	-3582.31 +/- 36.6
8	Zif-268 WT	MT2	-47.6 +/- 8.8	-3577.32 +/- 31.9
9	Zif-268 MT1	WT	-70.7 +/- 16.9	-3086.26 +/- 32.8
10	Zif-268 MT2	WT	-77.9 +/- 23.7	-3188.07 +/- 34.8
11	Myc WT	WT	-107.6 +/- 18.8	-6777.41 +/- 50.9
12	Myc MT	MT	-65.9 +/- 12.0	-6364.50 +/- 49.7
13	Myc WT	MT	-83.3 +/- 13.0	-6765.37 +/- 54.5
14	Myc MT	WT	-82.3 +/- 25.8	-6373.51 +/- 52.5
15	Ets1 WT	WT	-69.2 +/- 9.2	-1909.67 +/- 32.3
16	Ets1 MT	MT	-66.2 +/- 11.9	-2152.48 +/- 34.1
17	Ets1 WT	MT	-73.7 +/- 16.4	-1895.81 +/- 32.7
18	Ets1 MT	WT	-62.7 +/- 11.8	-2151.14 +/- 33.1
19	EcoRV WT	WT	-122.5 +/- 22.4	-8383.87 +/- 69.7
20	EcoRV MT	MT	-103.4 +/- 20.4	-8482.77 +/- 68.7
21	EcoRV WT	MT	-123.7 +/- 22.3	-8397.09 +/- 67.2
22	EcoRV MT	WT	-104.4 +/- 24.5	-8485.94 +/- 70.7

The Zinc finger MT2 protein with its cognate substrate had the highest binding energy. The intramolecular potential energy of the proteins was highest for the wild-type protein. Thus the binding to DNA and intermolecular potential energy determines the overall affinity of the complex.

For Ets1, the best binding energy was obtained for Ets1 wild-type protein with mutant DNA. While when we look at the protein potential energies, they are invariant for the protein regardless of the substrate used. However, the mutant had slightly more potential energy than the wild-type protein.

For Myc protein, the wild-type protein with the wild-type DNA had the largest binding energy. This is also reflected in the protein potential energy. The potential energy of the wild-type proteins is better than that of the mutant proteins.

EcoRV wild-type protein with wild-type DNA had the best intermolecular binding energy. However, the intramolecular potential energy was similar for all DNA-protein complexes.

References

- [1] A. Kornberg, T.A. Baker, DNA Replication, 2nd ed., W. H. Freeman and Company, 1992.
- [2] S. Waga, B. Stillman, The DNA replication fork in eukaryotic cells, *Annu. Rev. Biochem.* 67 (1998) 721–751. doi:10.1146/ANNUREV.BIOCHEM.67.1.721.
- [3] J.A. Carrodeguas, K.G. Pinz, D.F. Bogenhagen, DNA Binding Properties of Human pol γ B *, *J. Biol. Chem.* 277 (2002) 50008–50014. doi:10.1074/JBC.M207030200.
- [4] I. Frouin, A. Montecucco, S. Spadari, G. Maga, DNA replication: a complex matter, *EMBO Rep.* 4 (2003) 666. doi:10.1038/SJ.EMBOR.EMBOR886.
- [5] A. Sancar, L.A. Lindsey-Boltz, K. Ünsal-Kaçmaz, S. Linn, Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints, [Http://Dx.Doi.Org/10.1146/Annurev.Biochem.73.011303.073723](http://Dx.Doi.Org/10.1146/Annurev.Biochem.73.011303.073723). 73 (2004) 39–85. doi:10.1146/ANNUREV.BIOCHEM.73.011303.073723.
- [6] T.T. Huang, A.D. D'Andrea, Regulation of DNA repair by ubiquitylation, *Nat. Rev. Mol. Cell Biol.* 2006 75. 7 (2006) 323–334. doi:10.1038/nrm1908.
- [7] K. M, V.H. B, Rad4 recognition-at-a-distance: Physical basis of conformation-specific anomalous diffusion of DNA repair proteins, *Prog. Biophys. Mol. Biol.* 127 (2017) 93–104. doi:10.1016/J.PBIOMOLBIO.2016.12.004.
- [8] S.N. Powell, L.A. Kachnic, Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation, *Oncogene* 2003 2237. 22 (2003) 5784–5791. doi:10.1038/sj.onc.1206678.
- [9] F. González, D. Georgieva, F. Vanoli, Z.D. Shi, M. Stadtfeld, T. Ludwig, M. Jasin, D. Huangfu, Homologous Recombination DNA Repair Genes Play a Critical Role in Reprogramming to a Pluripotent State, *Cell Rep.* 3 (2013) 651–660. doi:10.1016/J.CELREP.2013.02.005.
- [10] J.M. Short, Y. Liu, S. Chen, N. Soni, M.S. Madhusudhan, M.K.K. Shivji, A.R. Venkitaraman, High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy., *Nucleic Acids Res.* 44 (2016) 9017–9030. doi:10.1093/nar/gkw783.
- [11] E. Pires, P. Sung, C. Wiese, Role of RAD51AP1 in homologous recombination DNA repair and carcinogenesis, *DNA Repair (Amst).* 59 (2017) 76–81. doi:10.1016/J.DNAREP.2017.09.008.
- [12] T. Obsil, V. Obsilova, Structural basis for DNA recognition by FOXO proteins., *Biochim. Biophys. Acta.* 1813 (2011) 1946–53. doi:10.1016/j.bbamcr.2010.11.025.
- [13] Y.N. Chirgadze, V.S. Sivozhelezov, R. V Polozov, V.A. Stepanenko, V. V Ivanov, Recognition Rules for Binding of Homeodomains to Operator DNA, *J. Biomol. Struct. Dyn.* 29 (2012) 739–1102. <http://www.jbsdonline.com>.
- [14] V.R. Yella, M. Bansal, DNA structural features of eukaryotic TATA-containing and TATA-less promoters, *FEBS Open Bio.* (2016). doi:10.1002/2211-5463.12166.
- [15] N.M. Luscombe, S.E. Austin, H.M. Berman, J.M. Thornton, H.M. Berman, J.M. Thornton, An overview of the structures of protein-DNA complexes., *Genome Biol.* 1 (2000) REVIEWS001. doi:10.1186/gb-2000-1-1-reviews001.
- [16] C. Grandori, S.M. Cowley, L.P. James, R.N. Eisenman, The Myc/Max/Mad network and the transcriptional control of cell behavior, *Annu. Rev. Cell Dev. Biol.* 16 (2000) 653–699. doi:10.1146/ANNUREV.CELLBIO.16.1.653.
- [17] A. Rizzino, The Sox2-Oct4 Connection: Critical players in a much larger interdependent network integrated at multiple levels, *Stem Cells.* 31 (2013) 1033. doi:10.1002/STEM.1352.
- [18] M.F. Wernet, C. Desplan, Homothorax and Extradenticle alter the transcription factor network in *Drosophila* ommatidia at the dorsal rim of the retina, *Dev.* 141 (2014) 918–928. doi:10.1242/DEV.103127/-/DC1.

- [19] D.S. Latchman, Transcription-factor mutations and disease, *N. Engl. J. Med.* 334 (1996) 28–33. doi:10.1056/NEJM199601043340108.
- [20] J.J. Schott, D.W. Benson, C.T. Basson, W. Pease, G.M. Silberbach, J.P. Moak, B.J. Maron, C.E. Seidman, J.G. Seidman, Congenital heart disease caused by mutations in the transcription factor NKX2-5, *Science* (80-.). 281 (1998) 108–111. doi:10.1126/SCIENCE.281.5373.108/ASSET/65C1731F-0E22-40D3-9294-1BAAF764CEAF/ASSETS/GRAPHIC/SE2886651003.JPEG.
- [21] B.A. Bunnell, R.A. Morgan, Gene therapy for infectious diseases, *Clin. Microbiol. Rev.* 11 (1998) 42–56. doi:10.1128/CMR.11.1.42/ASSET/68AA4AC7-EA41-40B1-A833-CFDBF815595F/ASSETS/GRAPHIC/CM0180006003.JPEG.
- [22] D.M. O'Connor, N.M. Boulis, Gene therapy for neurodegenerative diseases, *Trends Mol. Med.* 21 (2015) 504–512. doi:10.1016/J.MOLMED.2015.06.001.
- [23] P. Arbuthnot, Engineering Sequence-Specific DNA Binding Proteins for Antiviral Gene Editing, in: *Gene Ther. Viral Infect.*, Elsevier, 2015: pp. 63–94. doi:10.1016/B978-0-12-410518-8.00003-X.
- [24] C. Roma-rodrigues, L. Rivas-garcía, P. V. Baptista, A.R. Fernandes, Gene Therapy in Cancer Treatment: Why Go Nano?, *Pharm.* 2020, Vol. 12, Page 233. 12 (2020) 233. doi:10.3390/PHARMACEUTICS12030233.
- [25] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity, *Science* (80-.). 337 (2012) 816–821. doi:10.1126/SCIENCE.1225829.
- [26] A. V. Wright, J.K. Nuñez, J.A. Doudna, Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering, *Cell.* 164 (2016) 29–44. doi:10.1016/J.CELL.2015.12.035.
- [27] I.M. Slaymaker, L. Gao, B. Zetsche, D.A. Scott, W.X. Yan, F. Zhang, Rationally engineered Cas9 nucleases with improved specificity, *Science* (80-.). 351 (2016) 84–88. doi:10.1126/SCIENCE.AAD5227/SUPPL_FILE/SLAYMAKER-SM.PDF.
- [28] C.A. Vakulskas, D.P. Dever, G.R. Rettig, R. Turk, A.M. Jacobi, M.A. Collingwood, N.M. Bode, M.S. McNeill, S. Yan, J. Camarena, C.M. Lee, S. Hyun Park, V. Wiebking, R.O. Bak, N. Gomez-Ospina, M. Pavel-Dinu, W. Sun, G. Bao, M.H. Porteus, M.A. Behlke, A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells, *Nat. Med.* (n.d.). doi:10.1038/s41591-018-0137-0.
- [29] A. Casini, M. Olivieri, G. Petris, C. Montagna, G. Reginato, G. Maule, F. Lorenzin, D. Prandi, A. Romanel, F. Demichelis, A. Inga, A. Cereseto, A highly specific SpCas9 variant is identified by in vivo screening in yeast, *Nat. Biotechnol.* 2018 363. 36 (2018) 265–271. doi:10.1038/nbt.4066.
- [30] N.C. Seeman, J.M. Rosenberg, A. Rich, Sequence-specific recognition of double helical nucleic acids by proteins, *Biophysics (Oxf).* 73 (1976) 804–808.
- [31] A.C. Cheng, W.W. Chen, C.N. Fuhrmann, A.D. Frankel, Recognition of Nucleic Acid Bases and Base-pairs by Hydrogen Bonding to Amino Acid Side-chains, *J. Mol. Biol.* 327 (2003) 781–796. doi:10.1016/S0022-2836(03)00091-3.
- [32] M.M. Hoffman, M.A. Khrapov, J.C. Cox, J. Yao, L. Tong, A.D. Ellington, AANT: The Amino Acid-Nucleotide Interaction Database, *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh128.
- [33] N.M. Luscombe, R.A. Laskowski, J.M. Thornton, Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level, *Nucleic Acids Res.* 29 (2001) 2860–2874.
- [34] A. Sarai, H. Kono, PROTEIN-DNA RECOGNITION PATTERNS AND PREDICTIONS, *Annu. Rev. Biophys. Biomol. Struct.* 34 (2005) 379–98. doi:10.1146/annurev.biophys.34.040204.144537.
- [35] Y. Choo, A. Klug, Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage, *Proc. Natl. Acad. Sci. U. S. A.* 91 (1994) 11163–11167. doi:10.1073/pnas.91.23.11163.

- [36] Y. Choo, A. Klug, Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions, *Proc. Natl. Acad. Sci. U. S. A.* 91 (1994) 11168–11172. doi:10.1073/PNAS.91.23.11168.
- [37] M. Isalan, A. Klug, Y. Choo, Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers, *Biochemistry*. 37 (1998) 12026–12033. doi:10.1021/bi981358z.
- [38] J. Boch, H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, U. Bonas, Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors, *Science* (80-.). 326 (2009). <http://science.sciencemag.org/content/326/5959/1509.full> (accessed April 18, 2017).
- [39] M.J. Moscou, A.J. Bogdanove, A simple cipher governs DNA recognition by TAL effectors, *Science*. 326 (2009) 1501. doi:10.1126/SCIENCE.1178817.
- [40] M.N. Nguyen, K.P. Tan, M.S. Madhusudhan, CLICK—topology-independent comparison of biomolecular 3D structures, *Nucleic Acids Res.* 39 (2011) W24–W28. doi:10.1093/NAR/GKR393.
- [41] M.N. Nguyen, M.S. Madhusudhan, Biological insights from topology independent comparison of protein 3D structures, *Nucleic Acids Res.* 39 (2011) e94–e94. doi:10.1093/nar/gkr348.
- [42] Nature Publishing Company., John Wiley & Sons., *Encyclopedia of life sciences.*, Nature Pub. Group ;Wiley., London ;;New York ;Vols. 21-32 ;;Chichester West Sussex U.K., 2002.
- [43] T. Lundbäck, T. Hård, Sequence-specific DNA-binding dominated by dehydration, *Proc. Natl. Acad. Sci. U. S. A.* 93 (1996) 4754–4759. doi:10.1073/PNAS.93.10.4754.
- [44] R.G. Tap, H. Duong, K. Zakrzewska, The Role of DNA-Protein Salt Bridges in Molecular Recognition: A Model Study, *Biopoly.* 49 (1999) 313–327. doi:10.1002/(SICI)1097-0282(19990405)49:4.
- [45] M.M. Gromiha, C. Santhosh, M. Suwa, Influence of cation –p interactions in protein –DNA complexes, (n.d.). doi:10.1016/j.polymer.2003.10.069.
- [46] P.H. Von Hippel, O.G. Berg, On the Specificity of DNA--Protein Interactions, *Source Proc. Natl. Acad. Sci. United States Am.* 83 (1986) 1608–1612. <http://www.jstor.org>.
- [47] J. Dolenc, U. Borštnik, M. Hodošček, J. Koller, D. Janežič, An ab initio QM/MM study of the conformational stability of complexes formed by netropsin and DNA. The importance of van der Waals interactions and hydrogen bonding, *J. Mol. Struct. THEOCHEM.* 718 (2005) 77–85. doi:10.1016/j.theochem.2004.12.019.
- [48] R.E. Hubbard, M. Kamran Haider, R.E. Hubbard, M. Kamran Haider, Hydrogen Bonds in Proteins: Role and Strength, in: *Encycl. Life Sci.*, John Wiley & Sons, Ltd, Chichester, UK, 2010. doi:10.1002/9780470015902.a0003011.pub2.
- [49] N.C. Seeman, J.M. Rosenberg, A. Rich, J.R.A.R. NC Seeman, N.C. Seeman, J.M. Rosenberg, A. Rich, Sequence-specific recognition of double helical nucleic acids by proteins., *Proc. Natl. Acad. Sci.* 73 (1976) 804–808. doi:10.1073/pnas.73.3.804.
- [50] M.A. Schumacher, A. Glasfeld, H. Zalkin, R.G. Brennan, The x-ray structure of the PurR-guanine-purF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity, *J. Biol. Chem.* 272 (1997) 22648–22653. doi:10.1074/jbc.272.36.22648.
- [51] C.K. Reddy, A. Das, B. Jayaram, Do Water Molecules Mediate Protein-DNA Recognition?, (n.d.). doi:10.1006/jmbi.2001.5154.
- [52] D.S. Wilson, B. Guenther, C. Desplan, J. Kuriyan, High Resolution Crystal Structure of a Paired (Pax) Class Cooperative Homeodomain Dimer on DNA, *Cell.* 82 (1995) 709–719.
- [53] C.G. Kalodimos, N. Biris, A.M.J.J. Bonvin, M.M. Levandoski, M. Guennuegues, R. Boelens, R. Kaptein, Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes, *Science*. 305 (2004) 386–389. doi:10.1126/SCIENCE.1097064.
- [54] Y. Mandel-Gutfreund, H. Margalit, R.L. Jernigan, V.B. Zhurkin, A role for CH...O interactions in protein-DNA recognition., (n.d.).

- [55] M.T. Record, W. Zhang, C.F. Anderson, Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: a practical guide to recognizing and interpreting polyelectrolyte effects, Hofmeister effects, and osmotic effects of salts, *Adv. Protein Chem.* 51 (1998) 281–353. doi:10.1016/S0065-3233(08)60655-5.
- [56] P.L. Privalov, A.I. Dragan, C. Crane-Robinson, Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components, *Nucleic Acids Res.* 39 (2011) 2483. doi:10.1093/NAR/GKQ984.
- [57] M.P. Horvath, V.L. Schweiker, J.M. Bevilacqua, J.A. Ruggles, S.C. Schultz, Crystal structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA, *Cell.* 95 (1998) 963–974. doi:10.1016/S0092-8674(00)81720-1.
- [58] E. Kamau, N.D. Tsihlis, L.A. Simmons, A. Grove, Surface salt bridges modulate the DNA site size of bacterial histone-like HU proteins, *Biochem. J.* 390 (2005) 49. doi:10.1042/BJ20050274.
- [59] T.D. Vo, A.L. Schneider, W. David, W.* Ab, G.M.K. Poon, Salt bridge dynamics in protein/DNA recognition: a comparative analysis of Elk1 and ETV6, *Phys. Chem. Chem. Phys.* 23 (2021) 13490–13502. doi:10.1039/d1cp01568k.
- [60] B.P. Dimitrijević, S.Z. Borozan, S.D. Stojanović, π – π and cation– π interactions in protein–porphyrin complex crystal structures, *RSC Adv.* 2 (2012) 12963–12972. doi:10.1039/C2RA21937A.
- [61] K.A. Wilson, J.L. Kellie, S.D. Wetmore, DNA-protein π -interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar, *Nucleic Acids Res.* 42 (2014) 6726–6741. doi:10.1093/nar/gku269.
- [62] P. Mignon, S. Loverix, J. Steyaert, P. Geerlings, Influence of the π - π interaction on the hydrogen bonding capacity of stacked DNA/RNA bases, *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki317.
- [63] M.M. Gromiha, C. Santhosh, M. Suwa, Influence of cation– π interactions in protein–DNA complexes, *Polymer (Guildf).* 45 (2004) 633–639. doi:10.1016/J.POLYMER.2003.10.069.
- [64] J.C. Ma, D.A. Dougherty, The Cation– π Interaction, *Chem. Rev.* 97 (1997) 1303–1324. doi:10.1021/CR9603744.
- [65] D.A. Dougherty, Cation- π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp, *Science.* 271 (1996) 163–168. doi:10.1126/SCIENCE.271.5246.163.
- [66] J.P. Gullivan, D.A. Dougherty, Cation- π interactions in structural biology, *Chem. Biochem.* 96 (1999) 9459–9464.
- [67] D.A. Dougherty, The Cation- π Interaction, *Acc. Chem. Res.* 46 (2013) 885. doi:10.1021/AR300265Y.
- [68] J. Labahn, O.D. Schärer, A. Long, K. Ezaz-Nikpay, G.L. Verdine, T.E. Ellenberger, Structural Basis for the Excision Repair of Alkylation-Damaged DNA, *Cell.* 86 (1996) 321–329. doi:10.1016/S0092-8674(00)80103-8.
- [69] A. D. McNaught and A. Wilkinson, IUPAC. Compendium of Chemical Terminology, (the “Gold Book”), 2nd ed., Blackwell Scientific Publications, Oxford, 1997. doi:https://doi.org/10.1351/goldbook.
- [70] G.M. Spitzer, B. Wellenzohn, P. Markt, J. Kirchmair, T. Langer, K.R. Liedl, Hydrogen-bonding patterns of minor groove-binder - DNA complexes reveal criteria for discovery of new scaffolds, *J. Chem. Inf. Model.* (2009). doi:10.1021/ci800455f.
- [71] R. Ivarie, Thymine methyls and DNA-protein interactions., *Nucleic Acids Res.* 15 (1987) 9975. doi:10.1093/NAR/15.23.9975.
- [72] C.W. Garvie, C. Wolberger, Recognition of specific DNA sequences, *Mol. Cell.* (2001). doi:10.1016/S1097-2765(01)00392-6.
- [73] A.E. Maris, M.R. Sawaya, M. Kaczor-Grzeskowiak, M.R. Jarvis, S. Bearson, M.L. Kopka, I. Schröder, R.P. Gunsalus, R.E. Dickerson, Dimerization allows DNA target site recognition by the NarL response regulator, *Nat. Struct. Biol.* 9 (2002) 771–778.
- [74] K. Roy, S. Kar, R.N. Das, K. Roy, S. Kar, R.N. Das, Chapter 12 – Future Avenues, 2015.

- [75] Y. Kim, J.H. Geiger, S. Hahn, P.B. Sigler, Crystal structure of a yeast TBP/TATA-box complex, *Nature*. 365 (1993) 512–520. doi:10.1038/365512A0.
- [76] J.L. Kim, D.B. Nikolov, S.K. Burley, Co-crystal structure of TBP recognizing the minor groove of a TATA element, *Nat.* 1993 3656446. 365 (1993) 520–527. doi:10.1038/365520a0.
- [77] H. Robinson, Y.G. Gao, B.S. McCrary, S.P. Edmondson, J.W. Shriver, A.H.J. Wang, The hyperthermophile chromosomal protein Sac7d sharply kinks DNA., *Nature*. 392 (1998) 202–205. doi:10.1038/32455.
- [78] Y.G. Gao, S.Y. Su, H. Robinson, S. Padmanabhan, L. Lim, B.S. McCrary, S.P. Edmondson, J.W. Shriver, A.H.J. Wang, The crystal structure of the hyperthermophile chromosomal protein Sso7d bound to DNA, *Nat. Struct. Biol.* 5 (1998) 782–786. doi:10.1038/1822.
- [79] A. Glasfeld, A.N. Koehler, M.A. Schumacher, R.G. Brennan, The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions, *J. Mol. Biol.* 291 (1999) 347–361. doi:10.1006/JMBI.1999.2946.
- [80] J.W. Chase, K.R. Williams, SINGLE-STRANDED DNA BINDING PROTEINS REQUIRED FOR DNA REPLICATION, (1986). www.annualreviews.org (accessed March 16, 2022).
- [81] C. Iftode, Y. Daniely, J.A. Borowiec, Replication Protein A (RPA): The Eukaryotic SSB, [Http://Dx.Doi.Org/10.1080/10409239991209255](http://Dx.Doi.Org/10.1080/10409239991209255). 34 (2008) 141–180. doi:10.1080/10409239991209255.
- [82] D. Coverley, M.K. Kenny, D.P. Lane, R.D. Wood, A role for the human single-Stranded DNA binding protein HSSB/RPA in an early stage of nucleotide excision repair, *Nucleic Acids Res.* 20 (1992) 3873–3880. doi:10.1093/NAR/20.15.3873.
- [83] G. Xu, H. Lu, L. Wang, H. Chen, Z. Xu, Y. Hu, B. Tian, Y. Hua, DdrB stimulates single-stranded DNA annealing and facilitates RecA-independent DNA repair in *Deinococcus radiodurans*, *DNA Repair (Amst)*. 9 (2010) 805–812. doi:10.1016/J.DNAREP.2010.04.006.
- [84] O. Mortusewicz, B. Evers, T. Helleday, PC4 promotes genome stability and DNA repair through binding of ssDNA at DNA damage sites, *Oncogene* 2016 356. 35 (2015) 761–770. doi:10.1038/onc.2015.135.
- [85] G. Petukhova, S.A. Stratton, P. Sung, Single Strand DNA Binding and Annealing Activities in the Yeast Recombination Factor Rad59 *, *J. Biol. Chem.* 274 (1999) 33839–33842. doi:10.1074/JBC.274.48.33839.
- [86] T. Yadav, B. Carrasco, A.R. Myers, N.P. George, J.L. Keck, J.C. Alonso, Genetic recombination in *Bacillus subtilis*: a division of labor between two single-strand DNA-binding proteins, *Nucleic Acids Res.* 40 (2012) 5546–5559. doi:10.1093/NAR/GKS173.
- [87] B. Souquet, E. Abby, R. Hervé, F. Finsterbusch, S. Tourpin, R. Le Bouffant, C. Duquenne, S. Messiaen, E. Martini, J. Bernardino-Sgherri, A. Toth, R. Habert, G. Livera, MEIOB Targets Single-Strand DNA and Is Necessary for Meiotic Recombination, *PLOS Genet.* 9 (2013) e1003784. doi:10.1371/JOURNAL.PGEN.1003784.
- [88] T. Hollis, J.M. Stattel, D.S. Walther, C.C. Richardson, T. Ellenberger, Structure of the gene 2.5 protein, a single-stranded DNA binding protein encoded by bacteriophage T7, *Proc. Natl. Acad. Sci. U. S. A.* 98 (2001) 9557. doi:10.1073/PNAS.171317698.
- [89] D.I. Pretto, S. Tsutakawa, C.A. Brosey, A. Castillo, M.E. Chagot, J.A. Smith, J.A. Tainer, W.J. Chazin, Structural dynamics and ssDNA binding activity of the three N-terminal domains of the large subunit of Replication Protein A from small angle X-ray scattering, *Biochemistry*. 49 (2010) 2880. doi:10.1021/BI9019934.
- [90] T. Wakamatsu, Y. Kitamura, Y. Kotera, N. Nakagawa, S. Kuramitsu, R. Masui, Structure of RecJ Exonuclease Defines Its Specificity for Single-stranded DNA, *J. Biol. Chem.* 285 (2010) 9762. doi:10.1074/JBC.M109.096487.

- [91] A.H. Marceau, S. Bahng, S.C. Massoni, N.P. George, S.J. Sandler, K.J. Marians, J.L. Keck, Structure of the SSB–DNA polymerase III interface and its role in DNA replication, *EMBO J.* 30 (2011) 4236. doi:10.1038/EMBOJ.2011.305.
- [92] R.J. Evans, D.R. Davies, J.M. Bullard, J. Christensen, L.S. Green, J.W. Guiles, J.D. Pata, W.K. Ribble, N. Janjic, T.C. Jarvis, From the Cover: Structure of PolC reveals unique DNA binding and fidelity determinants, *Proc. Natl. Acad. Sci. U. S. A.* 105 (2008) 20695. doi:10.1073/PNAS.0809989106.
- [93] W. Wang, J. Liu, X. Zhou, Identification of single-stranded and double-stranded dna binding proteins based on protein structure, *BMC Bioinformatics.* 15 (2014) 1–9. doi:10.1186/1471-2105-15-S12-S4.
- [94] W. Wang, L. Sun, S. Zhang, H. Zhang, J. Shi, T. Xu, K. Li, Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences, *BMC Bioinformatics.* 18 (2017) 300. doi:10.1186/s12859-017-1715-8.
- [95] C. Tan, T. Wang, W. Yang, L. Deng, PredPSD: A Gradient Tree Boosting Approach for Single-Stranded and Double-Stranded DNA Binding Protein Prediction, *Molecules.* 25 (2019) 98. doi:10.3390/molecules25010098.
- [96] A.H.J. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. Van Boom, G. Van Der Marel, A. Rich, Molecular structure of a left-handed double helical DNA fragment at atomic resolution, *Nat.* 1979 2825740. 282 (1979) 680–686. doi:10.1038/282680a0.
- [97] S. Arnott, R. Chandrasekaran, D.L. Birdsall, A.G.W. Leslie, R.L. Ratliff, Left-handed DNA helices, *Nat.* 1980 2835749. 283 (1980) 743–745. doi:10.1038/283743a0.
- [98] Z. Shakked, D. Rabinovich, The effect of the base sequence on the fine structure of the DNA double helix, *Prog. Biophys. Mol. Biol.* 47 (1986) 159–195. doi:10.1016/0079-6107(86)90013-1.
- [99] H.C.M. Nelson, J.T. Finch, B.F. Luisi, A. Klug, The structure of an oligo(dA)·oligo(dT) tract and its biological implications, *Nature Publishing Group*, 1987. doi:10.1038/330221a0.
- [100] A.A. Gorin, V.B. Zhurkin, W.K. Olson, B-DNA twisting correlates with base-pair morphology, *J. Mol. Biol.* 247 (1995) 34–48. doi:10.1006/JMBI.1994.0120.
- [101] G. Guzikevich-Guerstein, Z. Shakked, A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein, *Nat. Struct. Biol.* 1996 31. 3 (1996) 32–37. doi:10.1038/nsb0196-32.
- [102] H.O. Bertrand, T. Ha-Duong, S. Femandjian, B. Hartmann, Flexibility of the B-DNA backbone: Effects of local and neighbouring sequences on pyrimidine-purine steps, *Nucleic Acids Res.* 26 (1998) 1261–1267. doi:10.1093/NAR/26.5.1261.
- [103] W.K. Olson, A.A. Gorin, X.-J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, *Biophys. Commun.* 95 (1998) 11163–11168.
- [104] D.R. Mack, T.K. Chiu, R.E. Dickerson, Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: A comparative analysis of T-A and A-T steps within A-tracts, *J. Mol. Biol.* 312 (2001) 1037–1049. doi:10.1006/JMBI.2001.4994.
- [105] V.B. Zhurkin, M.Y. Tolstorukov, F. Xu, A. V. Colasanti, W.K. Olson, Sequence-Dependent Variability of B-DNA, *DNA Conform. Transcr.* (2005) 18–34. doi:10.1007/0-387-29148-2_2.
- [106] J. Dekker, GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p, *Genome Biol.* 8 (2007) 1–14. doi:10.1186/GB-2007-8-6-R116/FIGURES/7.
- [107] J.M. Passner, H.D. Ryoo, L. Shen, R.S. Mann, A.K. Aggarwal, Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex, *Nat.* 1999 3976721. 397 (1999) 714–719. doi:10.1038/17833.
- [108] D. Golovenko, B. Br€, P. Vyas, T.E. Haran, H. Rozenberg, Z. Shakked, New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins, *Structure.* 26 (2018). doi:10.1016/j.str.2018.06.006.

- [109] R. Joshi, J.M. Passner, R. Rohs, R. Jain, A. Sosinsky, M.A. Crickmore, V. Jacob, A.K. Aggarwal, B. Honig, R.S. Mann, Functional specificity of a Hox protein mediated by the recognition of minor groove structure, *Cell*. 131 (2007) 530. doi:10.1016/J.CELL.2007.09.024.
- [110] R. Rohs, S.M. West, A. Sosinsky, P. Liu, R.S. Mann, B. Honig, The role of DNA shape in protein-DNA recognition, *Nature*. 461 (2009) 1248–1253. doi:10.1038/nature08473.
- [111] M. Kisiala, M. Kowalska, M. Pastor, H.J. Korza, H. Czapinska, M. Bochtler, Restriction endonucleases that cleave RNA/DNA heteroduplexes bind dsDNA in A-like conformation, *Nucleic Acids Res.* 48 (2020) 6954–6969. doi:10.1093/NAR/GKAA403.
- [112] D.C. Chiang, Y. Li, S.K. Ng, The Role of the Z-DNA Binding Domain in Innate Immunity and Stress Granules, *Front. Immunol.* 11 (2021). doi:10.3389/FIMMU.2020.625504.
- [113] R.R. Sinden, Introduction to the Structure, Properties, and Reactions of DNA, *DNA Struct. Funct.* (1994) 1–57. doi:10.1016/B978-0-08-057173-7.50006-7.
- [114] X.J. Lu, Z. Shakked, W.K. Olson, A-form conformational motifs in ligand-bound DNA structures, *J. Mol. Biol.* 300 (2000) 819–840. doi:10.1006/JMBI.2000.3690.
- [115] T. Schwartz, J. Behlke, K. Lowenhaupt, U. Heinemann, A. Rich, Structure of the DLM-1–Z-DNA complex reveals a conserved family of Z-DNA-binding proteins, *Nat. Struct. Biol.* 8 (2001) 761–765. doi:10.1038/nsb0901-761.
- [116] C.O. Pabo, R.T. Sauer, PROTEIN-DNA RECOGNITION, *Ann. Rev. Biochem.* 53 (1984) 293–321. doi:10.1146/ANNUREV.BI.53.070184.001453.
- [117] B.W. Matthews, Protein-DNA interaction. No code for recognition, *Nature*. 335 (1988) 294–295. doi:10.1038/335294A0.
- [118] S. Ahmad, M.M. Gromiha, A. Sarai, A. S. G. MM, S. A, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, *20 (2004) 477–486*. <https://pubmed.ncbi.nlm.nih.gov/14990443/> (accessed October 22, 2021).
- [119] G. Paillard, R. Lavery, Analyzing Protein-DNA Recognition Mechanisms, *Structure*. (2004). doi:10.1016/j.str.2003.11.022.
- [120] H. Kono, A. Sarai, Structure-Based Prediction of DNA Target Sites by Regulatory Proteins, *Proteins*. 35 (1999) 114–31.
- [121] M.M. Gromiha, J.G. Siebers, S. Selvaraj, H. Kono, A. Sarai, Role of inter and intramolecular interactions in protein-DNA recognition, in: *Gene*, 2005. doi:10.1016/j.gene.2005.07.022.
- [122] T. Sera, C. Uranga, Rational design of artificial zinc-finger proteins using a nondegenerate recognition code table., *Biochemistry*. 41 (2002) 7074–81. <http://www.ncbi.nlm.nih.gov/pubmed/12033941> (accessed April 24, 2017).
- [123] S.A. Wolfe, R.A. Grant, M. Elrod-Erickson, C.O. Pabo, Beyond the “Recognition Code”: Structures of Two Cys2His2 Zinc Finger/TATA Box Complexes, *Structure*. 9 (2001) 717–723. doi:10.1016/S0969-2126(01)00632-3.
- [124] Y.-G. Kim, J. Cha, S. Chandrasegaran, Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain (Flavobacterium okeanokoites/chimeric restriction endonuclease/protein engineering/recognition and cleavage domains), *Biochemistry*. 93 (1996) 1156–1160. <http://www.pnas.org/content/93/3/1156.full.pdf> (accessed April 26, 2017).
- [125] M. Bibikova, D. Carroll, D.J. Segal, J.K. Trautman, J. Smith, Y.-G. Kim, S. Chandrasegaran, Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases, *Mol. Cell. Biol.* 21 (2001) 289. doi:10.1128/MCB.21.1.289-297.2001.
- [126] M. Bibikova, M. Golic, K.G. Golic, D. Carroll, Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases., *Genetics*. 161 (2002) 1169. doi:10.1093/genetics/161.3.1169.

- [127] M. Bibikova, K. Beumer, J.K. Trautman, D. Carroll, Enhancing gene targeting with designed zinc finger nucleases, *Science*. 300 (2003) 764. doi:10.1126/SCIENCE.1079512.
- [128] M.H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells, *Science*. 300 (2003) 763. doi:10.1126/SCIENCE.1078395.
- [129] J.C. Miller, S. Tan, G. Qiao, K.A. Barlow, J. Wang, D.F. Xia, X. Meng, D.E. Paschon, E. Leung, S.J. Hinkley, G.P. Dulay, K.L. Hua, I. Ankoudinova, G.J. Cost, F.D. Urnov, H.S. Zhang, M.C. Holmes, L. Zhang, P.D. Gregory, E.J. Rebar, A TALE nuclease architecture for efficient genome editing, *Nat. Biotechnol.* 29 (2011) 143–150. doi:10.1038/NBT.1755.
- [130] J.K. Joung, J.D. Sander, TALENs: a widely applicable technology for targeted genome editing, *Nat. Rev. Mol. Cell Biol.* 2012 141. 14 (2012) 49–55. doi:10.1038/nrm3486.
- [131] Q. Ding, Y.K. Lee, E.A.K. Schaefer, D.T. Peters, A. Veres, K. Kim, N. Kuperwasser, D.L. Motola, T.B. Meissner, W.T. Hendriks, M. Trevisan, R.M. Gupta, A. Moisan, E. Banks, M. Friesen, R.T. Schinzel, F. Xia, A. Tang, Y. Xia, E. Figueroa, A. Wann, T. Ahfeldt, L. Daheron, F. Zhang, L.L. Rubin, L.F. Peng, R.T. Chung, K. Musunuru, C.A. Cowan, A TALEN genome-editing system for generating human stem cell-based disease models, *Cell Stem Cell*. 12 (2013) 238–251. doi:10.1016/J.STEM.2012.11.011.
- [132] J.M. Rogers, L.A. Barrera, D. Reyon, J.D. Sander, M. Kellis, J. Keith Joung, M.L. Bulyk, Context influences on TALE–DNA binding revealed by quantitative profiling, *Nat. Commun.* 2015 61. 6 (2015) 1–10. doi:10.1038/ncomms8440.
- [133] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Res.* 33 (2005) 2302–2309. doi:10.1093/NAR/GKI524.
- [134] A.S. Konagurthu, J.C. Whisstock, P.J. Stuckey, A.M. Lesk, MUSTANG: A multiple structural alignment algorithm, *Proteins Struct. Funct. Bioinforma.* 64 (2006) 559–574. doi:10.1002/PROT.20921.
- [135] H. Braberg, B.M. Webb, E. Tjioe, U. Pieper, A. Sali, M.S. Madhusudhan, SALIGN: a web server for alignment of multiple protein sequences and structures, *Bioinformatics*. 28 (2012) 2072–2073. doi:10.1093/BIOINFORMATICS/BTS302.
- [136] L. Holm, Using Dali for Protein Structure Comparison, *Methods Mol. Biol.* 2112 (2020) 29–42. doi:10.1007/978-1-0716-0270-6_3.
- [137] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH – a hierarchic classification of protein domain structures, *Structure*. 5 (1997) 1093–1109. doi:10.1016/S0969-2126(97)00260-8.
- [138] C.A. Orengo, A.M. Martin, G. Hutchinson, S. Jones, D.T. Jones, A.D. Michie, M.B. Swindells, J.M. Thornton, Classifying a protein in the CATH database of domain structures, *Acta Crystallogr. D. Biol. Crystallogr.* 54 (1998) 1155–1167. doi:10.1107/S09074444998007501.
- [139] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol.* 2002 99. 9 (2002) 646–652. doi:10.1038/nsb0902-646.
- [140] W.D. Cornell, P. Cieplak, / Christopher, I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, 1995. <https://pubs.acs.org/sharingguidelines>.
- [141] P.K. Weiner, P.A. Kollman, AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions, *J. Comput. Chem.* 2 (1981) 287–303. doi:10.1002/JCC.540020311.
- [142] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, Development and Testing of a General Amber Force Field, 2004.
- [143] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* 4 (1983) 187–217. doi:10.1002/JCC.540040211.

- [144] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A.D. Mackerell, CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields, *J Comput Chem.* 31 (2010) 671–690. doi:10.1002/jcc.21367.
- [145] W.L. Jorgensen, J. Tirado-Rives, The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin, *Annu. Rev. Biophys. Bioeng.* 110 (1988) 1657–1666. <https://pubs.acs.org/sharingguidelines> (accessed March 31, 2022).
- [146] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc.* 118 (1996) 11225–11236. doi:10.1021/JA9621760/SUPPL_FILE/JA11225.PDF.
- [147] W.F. Van Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hunenberger, P. Kruger, A.E. Mark, W.R.P. Scott, I.G. Tironi, The GROMOS96 manual and user guide, Biomos, Zurich, Switz. (1996).
- [148] C. Oostenbrink, A. Villa, A.E. Mark, W.F. Van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6, *J. Comput. Chem.* 25 (2004) 1656–1676. doi:10.1002/JCC.20090.
- [149] J.C. Phillips, D.J. Hardy, J.D.C. Maia, J.E. Stone, J. V. Ribeiro, R.C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M.C.R. Melo, B.K. Radak, R.D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, E. Tajkhorshid, Scalable molecular dynamics on CPU and GPU architectures with NAMD, *J. Chem. Phys.* 153 (2020). doi:10.1063/5.0014475.
- [150] H.J.C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun.* 91 (1995) 43–56. doi:10.1016/0010-4655(95)00042-E.
- [151] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindah, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX.* 1–2 (2015) 19–25. doi:10.1016/J.SOFTX.2015.06.001.
- [152] W.R.P. Scott, P.H. Hünenberger, I.G. Tironi, A.E. Mark, S.R. Billeter, J. Fennen, A.E. Torda, T. Huber, P. Krüger, W.F. Van Gunsteren, The GROMOS Biomolecular Simulation Program Package, *J. Phys. Chem. A.* 103 (1999) 3596–3607. doi:10.1021/JP984217F.
- [153] R. Salomon-Ferrer, D.A. Case, R.C. Walker, An overview of the Amber biomolecular simulation package, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3 (2013) 198–210. doi:10.1002/WCMS.1121.
- [154] S.A. Adcock, J.A. McCammon, Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins, *Chem. Rev.* 106 (2006) 1589–1615. doi:10.1021/CR040426M.
- [155] A. Hospital, J.R. Goñi, M. Orozco, J.L. Gelpí, Molecular dynamics simulations: advances and applications, *Adv. Appl. Bioinforma. Chem.* 8 (2015) 37–47. doi:10.2147/AABC.S70333.
- [156] S.A. Hollingsworth, R.O. Dror, Molecular Dynamics Simulation for All, *Neuron.* 99 (2018) 1129–1143.
- [157] R. Lazim, D. Suh, S. Choi, Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems, *Int. J. Mol. Sci.* 21 (2020) 1–20. doi:10.3390/IJMS21176339.
- [158] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, D.A. Case, Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices, (1998). doi:10.1021/JA981844.
- [159] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models, *Acc. Chem. Res.* 33 (2000) 889–897. doi:10.1021/AR000033J.

- [160] V. Tsui, D.A. Case, Calculations of the Absolute Free Energies of Binding between RNA and Metal Ions Using Molecular Dynamics Simulations and Continuum Electrostatics, *J. Phys. Chem. B.* 105 (2001) 11314–11325. doi:10.1021/JP011923Z.
- [161] I. Massova, P.A. Kollman, Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding, *Perspect. Drug Discov. Des.* 2000 181. 18 (2000) 113–135. doi:10.1023/A:1008763014207.
- [162] A. Vergara-Jaque, J. Comer, L. Monsalve, F.D. González-Nilo, C. Sandoval, Computationally efficient methodology for atomic-level characterization of dendrimer-drug complexes: A comparison of amine- and acetyl-terminated PAMAM, *J. Phys. Chem. B.* 117 (2013) 6801–6813. doi:10.1021/JP4000363/SUPPL_FILE/JP4000363_SI_002.PDF.
- [163] S. Genheden, U. Ryde, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities, *Expert Opin. Drug Discov.* 10 (2015) 449. doi:10.1517/17460441.2015.1032936.
- [164] Z. Cournia, B. Allen, W. Sherman, Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations, *J. Chem. Inf. Model.* 57 (2017) 2911–2937. doi:10.1021/ACS.JCIM.7B00564/ASSET/IMAGES/ACS.JCIM.7B00564.SOCIAL.JPEG_V03.
- [165] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J.Z.H. Zhang, T. Hou, End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design, *Chem. Rev.* 119 (2019) 9478–9508. doi:10.1021/ACS.CHEMREV.9B00055/ASSET/IMAGES/ACS.CHEMREV.9B00055.SOCIAL.JPEG_V03.
- [166] R. Hall, T. Dixon, A. Dickson, On Calculating Free Energy Differences Using Ensembles of Transition Paths, *Front. Mol. Biosci.* 7 (2020) 106. doi:10.3389/FMOLB.2020.00106/BIBTEX.
- [167] M.A. Martí-Renom, A.C. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Šali, Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000) 291–325. doi:10.1146/ANNUREV.BIOPHYS.29.1.291.
- [168] B. Webb, A. Sali, Comparative Protein Structure Modeling Using MODELLER, *Curr. Protoc. Bioinforma.* 54 (2016) 5.6.1–5.6.37. doi:10.1002/CPBI.3.
- [169] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions, *J. Mol. Biol.* 268 (1997) 209–225. doi:10.1006/JMBI.1997.0959.
- [170] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein Structure Prediction Using Rosetta, *Methods Enzymol.* 383 (2004) 66–93. doi:10.1016/S0076-6879(04)83004-0.
- [171] P. Bradley, K.M.S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins, *Science* (80-.). 309 (2005) 1868–1871. doi:10.1126/SCIENCE.1113801/SUPPL_FILE/BRADLEY_SOM.PDF.
- [172] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nat.* 2021 5967873. 596 (2021) 583–589. doi:10.1038/s41586-021-03819-2.
- [173] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, *Nat. Protoc.* 2010 54. 5 (2010) 725–738. doi:10.1038/nprot.2010.5.
- [174] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction, *Nat. Methods.* 12 (2015) 7. doi:10.1038/NMETH.3213.
- [175] C. Kemena, C. Notredame, Upcoming challenges for multiple sequence alignment methods in the high-throughput era, *Bioinformatics.* 25 (2009) 2455–2465. doi:10.1093/BIOINFORMATICS/BTP452.

- [176] J.D. Thompson, B. Linard, O. Lecompte, O. Poch, A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives, *PLoS One*. 6 (2011) e18093. doi:10.1371/JOURNAL.PONE.0018093.
- [177] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* 215 (1990) 403–410. doi:10.1016/S0022-2836(05)80360-2.
- [178] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402. doi:10.1093/NAR/25.17.3389.
- [179] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Mol. Syst. Biol.* 7 (2011) 539. doi:10.1038/MSB.2011.75.
- [180] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.* 22 (1994) 4673–4680. doi:10.1093/NAR/22.22.4673.
- [181] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* 32 (2004) 1792–1797. doi:10.1093/NAR/GKH340.
- [182] C. Zambrano-Vega, B. Oviedo, R. Villamar-Torres, M. Botto-Tobar, M. Barros-Rodríguez, An Overview of Multiple Sequence Alignment Methods Applied to Transmembrane Proteins, *Commun. Comput. Inf. Sci.* 895 (2019) 410–419. doi:10.1007/978-3-030-05532-5_30.
- [183] T. Warnow, Revisiting Evaluation of Multiple Sequence Alignment Methods, *Methods Mol. Biol.* 2231 (2021) 299–317. doi:10.1007/978-1-0716-1036-7_17.
- [184] Y. Zhang, Q. Zhang, J. Zhou, Q. Zou, A survey on the algorithm and development of multiple sequence alignment, *Brief. Bioinform.* (2022). doi:10.1093/BIB/BBAC069.
- [185] R.K. Saiki, S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science*. 230 (1985) 1350–1354. doi:10.1126/SCIENCE.2999980.
- [186] R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, H.A. Erlich, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science*. 239 (1988) 487–491. doi:10.1126/SCIENCE.2448875.
- [187] N. Arnheim, H. Erlich, POLYMERASE CHAIN REACTION STRATEGY, *Annu. Rev. Biochem.* 61 (1992) 131–156. www.annualreviews.org (accessed March 31, 2022).
- [188] M.A.A. Valones, R.L. Guimarães, L.A.C. Brandão, P.R.E. De Souza, A. De Albuquerque Tavares Carvalho, S. Crovela, Principles and applications of polymerase chain reaction in medical diagnostic fields: a review, *Brazilian J. Microbiol.* 40 (2009) 1. doi:10.1590/S1517-83822009000100001.
- [189] S. Atawodi, J. Atawodi, A. Dzikwi, Polymerase chain reaction: Theory, practice and application: A review, *Sahel Med. J.* 13 (2011). doi:10.4314/smj2.v13i2.64834.
- [190] F. Van Den Ent, J. Löwe, RF cloning: A restriction-free method for inserting target genes into plasmids, *J. Biochem. Biophys. Methods*. 67 (2006) 67–74. doi:10.1016/J.JBBM.2005.12.008.
- [191] M. Fried, D.M. Crothers, Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis, *Nucleic Acids Res.* 9 (1981) 6505–6525. doi:10.1093/NAR/9.23.6505.
- [192] L.M. Hellman, M.G. Fried, Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nucleic Acid Interactions, *Nat. Protoc.* 2 (2007) 1849. doi:10.1038/NPROT.2007.249.
- [193] N.S. Holden, C.E. Tacon, Principles and problems of the electrophoretic mobility shift assay, *J. Pharmacol. Toxicol. Methods*. 63 (2011) 7–14. doi:10.1016/J.VASCN.2010.03.002.

- [194] C. Ludwig, Diffusion zwischen ungleich erwärmten Orten gleich zusammengesetzter Lösung, Aus Der KK Hof-Und Staatsdruckerei, Comm. Bei W. Braumüller, Buchhändler Des KK Hofes Und Der K. Akad. Der Wissenschaften. 20 (1856) 539–540.
- [195] M. Jerabek-Willemsen, C.J. Wienken, D. Braun, P. Baaske, S. Duhr, Molecular interaction studies using microscale thermophoresis, *Assay Drug Dev. Technol.* 9 (2011) 342–353. doi:10.1089/ADT.2011.0380/ASSET/IMAGES/LARGE/FIGURE7.JPEG.
- [196] K. Zillner, M. Jerabek-Willemsen, S. Duhr, D. Braun, G. Längst, P. Baaske, Microscale Thermophoresis as a Sensitive Method to Quantify Protein: Nucleic Acid Interactions in Solution, *Methods Mol. Biol.* 815 (2012) 241–252. doi:10.1007/978-1-61779-424-7_18.
- [197] S.A.I. Seidel, P.M. Dijkman, W.A. Lea, G. van den Bogaart, M. Jerabek-Willemsen, A. Lazic, J.S. Joseph, P. Srinivasan, P. Baaske, A. Simeonov, I. Katritch, F.A. Melo, J.E. Ladbury, G. Schreiber, A. Watts, D. Braun, S. Duhr, Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions, *Methods.* 59 (2013) 301–315. doi:10.1016/J.YMETH.2012.12.005.
- [198] M. Jerabek-Willemsen, T. André, R. Wanner, H.M. Roth, S. Duhr, P. Baaske, D. Breitsprecher, MicroScale Thermophoresis: Interaction analysis and beyond, *J. Mol. Struct.* 1077 (2014) 101–113. doi:10.1016/J.MOLSTRUC.2014.03.009.
- [199] B. López-Méndez, B. Baron, C.A. Brautigam, T.A. Jowitt, S.H. Knauer, S. Uebel, M.A. Williams, A. Sedivy, O. Abian, C. Abreu, M. Adamczyk, W. Bal, S. Berger, A.K. Buell, C. Carolis, T. Daviter, A. Fish, M. Garcia-Alai, C. Guenther, J. Hamacek, J. Holková, J. Houser, C. Johnson, S. Kelly, A. Leech, C. Mas, D. Matulis, S.H. McLaughlin, R. Montserret, R. Nasreddine, R. Nehmé, Q. Nguyen, D. Ortega-Alarcón, K. Perez, K. Pirc, G. Piszczek, M. Podobnik, N. Rodrigo, J. Rokov-Plavec, S. Schaefer, T. Sharpe, J. Southall, D. Staunton, P. Tavares, O. Vanek, M. Weyand, D. Wu, Reproducibility and accuracy of microscale thermophoresis in the NanoTemper Monolith: a multi laboratory benchmark study, *Eur. Biophys. J.* 50 (2021) 411–427. doi:10.1007/S00249-021-01532-6/TABLES/5.
- [200] C.G. Alexander, R. Wanner, C.M. Johnson, D. Breitsprecher, G. Winter, S. Duhr, P. Baaske, N. Ferguson, Novel microscale approaches for easy, rapid determination of protein stability in academic and commercial settings, *Biochim. Biophys. Acta - Proteins Proteomics.* 1844 (2014) 2241–2250. doi:10.1016/J.BBAPAP.2014.09.016.
- [201] M.C. Lo, A. Aulabaugh, G. Jin, R. Cowling, J. Bard, M. Malamas, G. Ellestad, Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery, *Anal. Biochem.* 332 (2004) 153–159. doi:10.1016/J.AB.2004.04.031.
- [202] S.H. Kim, H.J. Yoo, E.J. Park, D.H. Na, Nano Differential Scanning Fluorimetry-Based Thermal Stability Screening and Optimal Buffer Selection for Immunoglobulin G, *Pharmaceuticals (Basel).* 15 (2021). doi:10.3390/PH15010029.
- [203] K. Gao, R. Oerlemans, M.R. Groves, Theory and applications of differential scanning fluorimetry in early-stage drug discovery, *Biophys. Rev.* 2020 121. 12 (2020) 85–104. doi:10.1007/S12551-020-00619-2.
- [204] J. Wen, H. Lord, N. Knutson, M. Wikström, Nano differential scanning fluorimetry for comparability studies of therapeutic proteins, *Anal. Biochem.* 593 (2020). doi:10.1016/J.AB.2020.113581.
- [205] A. Real-Hohn, M. Groznica, N. Löffler, D. Blaas, H. Kowalski, nanoDSF: In vitro Label-Free Method to Monitor Picornavirus Uncoating and Test Compounds Affecting Particle Stability, *Front. Microbiol.* 11 (2020) 1442. doi:10.3389/FMICB.2020.01442/BIBTEX.
- [206] H.M. Berman, J.D. Westbrook, Z. Feng, G.L. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, The Protein Data Bank, 28 235–42. doi:10.1107/S0907444902003451.

- [207] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*. 26 (2010) 680–682. doi:10.1093/bioinformatics/btq003.
- [208] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.H. Hsieh, A.R. Srinivasan, B. Schneider, The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids, *Biophys. J.* 63 (1992) 751–759. doi:10.1016/S0006-3495(92)81649-1.
- [209] UniProt: the universal protein knowledgebase, *Nucleic Acids Res.* 45 (2017) D158–D169. doi:10.1093/nar/gkw1099.
- [210] S. Kosugi, M. Hasebe, M. Tomita, H. Yanagawa, Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs., *Proc. Natl. Acad. Sci. U. S. A.* 106 (2009) 10171–6. doi:10.1073/pnas.0900604106.
- [211] M. Brameier, A. Krings, R.M. MacCallum, NucPred Predicting nuclear localization of proteins, *Bioinformatics*. 23 (2007) 1159–1160. doi:10.1093/bioinformatics/btm066.
- [212] R.G. Tap, H. Duong, K. Zakrzewska, The Role of DNA–Protein Salt Bridges in Molecular Recognition: A Model Study, (n.d.).
- [213] N.C. Seeman, J.M. Rosenberg, A. Rich, Sequence-specific recognition of double helical nucleic acids by proteins., *Proc. Natl. Acad. Sci.* 73 (1976) 804–808. doi:10.1073/pnas.73.3.804.
- [214] R. Wintjens, J. Liévin, M. Rooman, E. Buisine, Contribution of cation- π interactions to the stability of protein-DNA complexes, *J. Mol. Biol.* 302 (2000) 393–408. doi:10.1006/jmbi.2000.4040.
- [215] N.M. Luscombe, Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level, *Nucleic Acids Res.* 29 (2001) 2860–2874. doi:10.1093/nar/29.13.2860.
- [216] C. Cardelli, V. Bianco, L. Rovigatti, F. Nerattini, L. Tubiana, C. Dellago, I. Coluzza, The role of directional interactions in the designability of generalized heteropolymers, *Sci. Reports* 2017 71. 7 (2017) 1–10. doi:10.1038/s41598-017-04720-7.
- [217] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612. doi:10.1002/JCC.20084.
- [218] G.B. McGaughey, M. Gagné, A.K. Rappé, π -Stacking Interactions: ALIVE AND WELL IN PROTEINS *, *J. Biol. Chem.* 273 (1998) 15458–15463. doi:10.1074/JBC.273.25.15458.
- [219] J.P. Gallivan, D.A. Dougherty, Cation- π interactions in structural biology., *Proc. Natl. Acad. Sci. U. S. A.* 96 (1999) 9459–9464. doi:10.1073/PNAS.96.17.9459.
- [220] K.P. Tan, K. Singh, A. Hazra, M.S. Madhusudhan, Peptide bond planarity constrains hydrogen bond geometry and influences secondary structure conformations, *Curr. Res. Struct. Biol.* 3 (2021) 1–8. doi:10.1016/J.CRSTBI.2020.11.002.
- [221] D.G. Renter, J.J. Higgins, J.M. Sargeant, PERFORMANCE OF THE EXACT AND CHI-SQUARE TESTS ON SPARSE CONTINGENCY TABLES, *Conf. Appl. Stat. Agric.* (2000). doi:10.4148/2475-7772.1253.
- [222] M.J. Aukerman, R.J. Schmidt, B. Burr, F.A. Burr, An arginine to lysine substitution in the bZIP domain of an opaque-2 mutant in maize abolishes specific DNA binding., *Genes Dev.* 5 (1991) 310–20. <http://www.ncbi.nlm.nih.gov/pubmed/1899843> (accessed April 24, 2017).
- [223] G. Komazin-Meredith, W.L. Santos, D.J. Filman, J.M. Hogle, G.L. Verdine, D.M. Coen, The positively charged surface of herpes simplex virus UL42 mediates DNA binding., *J. Biol. Chem.* 283 (2008) 6154–61. doi:10.1074/jbc.M708691200.
- [224] E.R. Girden, ANOVA: Repeated measures, Sage, 1992.
- [225] Q. Liu, Z.Q. Xia, C.C. Case, Validated zinc finger protein designs for all 16 GNN DNA triplet targets, *J. Biol. Chem.* 277 (2002) 3850–3856. doi:10.1074/JBC.M110669200.

- [226] E. Wingender, P. Dietze, H. Karas, R. Knüppel, TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites, *Nucleic Acids Res.* 24 (1996) 238–241. doi:10.1093/NAR/24.1.238.
- [227] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics.* 26 (2010) 680–682. doi:10.1093/bioinformatics/btq003.
- [228] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta - Protein Struct.* 405 (1975) 442–451. doi:10.1016/0005-2795(75)90109-9.
- [229] A.A. Roy, A.S. Dhawanjewar, P. Sharma, G. Singh, M.S. Madhusudhan, Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein–protein interactions, *Nucleic Acids Res.* 47 (2019) W331–W337. doi:10.1093/NAR/GKZ368.
- [230] P. Prabakaran, J. An, M.M. Gromiha, S. Selvaraj, H. Uedaira, H. Kono, A. Sarai, Thermodynamic database for protein-nucleic acid interactions (ProNIT), *Bioinformatics.* 17 (2001) 1027–1034. doi:10.1093/BIOINFORMATICS/17.11.1027.
- [231] J.H. Carra, P.L. Privalov, Energetics of folding and DNA binding of the MAT alpha 2 homeodomain, *Biochemistry.* 36 (1997) 526–535. doi:10.1021/BI962206B.
- [232] S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, *J. Comput. Chem.* 29 (2008) 1859–1865. doi:10.1002/JCC.20945.
- [233] J. Lee, X. Cheng, J.M. Swails, M.S. Yeom, P.K. Eastman, J.A. Lemkul, S. Wei, J. Buckner, J.C. Jeong, Y. Qi, S. Jo, V.S. Pande, D.A. Case, C.L. Brooks, A.D. MacKerell, J.B. Klauda, W. Im, CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field, *J. Chem. Theory Comput.* 12 (2016) 405–413. doi:10.1021/ACS.JCTC.5B00935/SUPPL_FILE/CT5B00935_SI_001.PDF.
- [234] W. Humphrey, A. Dalke, K. Schulten, {VMD} -- {V}isual {M}olecular {D}ynamics, *J. Mol. Graph.* 14 (1996) 33–38.
- [235] N.P. Pavletich, C.O. Pabo, Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å, *Science.* 252 (1991) 809–817. doi:10.1126/SCIENCE.2028256.
- [236] M. Elrod-Erickson, M.A. Rould, L. Nekludova, C.O. Pabo, Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions, *Structure.* 4 (1996) 1171–1180. doi:10.1016/S0969-2126(96)00125-6.
- [237] M. Elrod-Erickson, T.E. Benson, C.O. Pabo, High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition, *Structure.* 6 (1998) 451–464. doi:10.1016/S0969-2126(98)00047-1.
- [238] C. V. Dang, MYC on the Path to Cancer, *Cell.* 149 (2012) 22. doi:10.1016/J.CELL.2012.03.003.
- [239] S.K. Madden, A.D. de Araujo, M. Gerhardt, D.P. Fairlie, J.M. Mason, Taking the Myc out of cancer: toward therapeutic strategies to directly inhibit c-Myc, *Mol. Cancer* 2021 201. 20 (2021) 1–18. doi:10.1186/S12943-020-01291-6.
- [240] M.J. Duffy, S. O'Grady, M. Tang, J. Crown, MYC as a target for cancer treatment, *Cancer Treat. Rev.* 94 (2021) 102154. doi:10.1016/J.CTRV.2021.102154.
- [241] E. Wolf, M. Eilers, Targeting MYC Proteins for Tumor Therapy, <https://doi.org/10.1146/Annurev-Cancerbio-030518-055826>. 4 (2020) 61–75. doi:10.1146/ANNUREV-CANCERBIO-030518-055826.
- [242] L.A. Jung, A. Gebhardt, W. Koelmel, C.P. Ade, S. Walz, J. Kuper, B. von Eyss, S. Letschert, C. Redel, L. d'Artista, A. Biankin, L. Zender, M. Sauer, E. Wolf, G. Evan, C. Kisker, M. Eilers, OmoMYC blunts promoter invasion by oncogenic MYC to inhibit gene expression characteristic of MYC-dependent tumors, *Oncogene* 2017 3614. 36 (2016) 1911–1924. doi:10.1038/onc.2016.354.
- [243] K.I. Zeller, X.D. Zhao, C.W.H. Lee, P.C. Kuo, F. Yao, J.T. Yustein, S.O. Hong, Y.L. Orlov, A. Shahab, C.Y. How, Y.T. Fu, Z. Weng, V.A. Kuznetsov, W.K. Sung, Y. Ruan, C. V. Dang, C.L. Wei, Global

mapping of c-Myc binding sites and target gene networks in human B cells, *Proc. Natl. Acad. Sci. U. S. A.* 103 (2006) 17834. doi:10.1073/PNAS.0604129103.

- [244] A. Sabò, B. Amati, Genome Recognition by MYC, *Cold Spring Harb. Perspect. Med.* 4 (2014). doi:10.1101/CSHPERSPECT.A014191.
- [245] L.A. Garrett-Sinha, Review of Ets1 structure, function, and roles in immunity, *Cell. Mol. Life Sci.* 70 (2013) 3375. doi:10.1007/S00018-012-1243-7.
- [246] J. Dittmer, The Biology of the Ets1 Proto-Oncogene, *Mol. Cancer* 2003 21. 2 (2003) 1–21. doi:10.1186/1476-4598-2-29.
- [247] J.A. Nye, J.M. Petersen, C. V. Gunther, M.D. Jonsen, B.J. Graves, Interaction of murine ets-1 with GGA-binding sites establishes the ETS domain as a new DNA-binding motif, *Genes Dev.* 6 (1992) 975–990. doi:10.1101/GAD.6.6.975.
- [248] M. Shiina, K. Hamada, T. Inoue-Bungo, M. Shimamura, A. Uchiyama, S. Baba, K. Sato, M. Yamamoto, K. Ogata, A Novel Allosteric Mechanism on Protein–DNA Interactions underlying the Phosphorylation-Dependent Regulation of Ets1 Target Gene Expressions, *J. Mol. Biol.* 427 (2015) 1655–1669. doi:10.1016/J.JMB.2014.07.020.
- [249] I. Schildkraut, C.D.B. Banner, C.S. Rhodes, S. Parekh, The cleavage site for the restriction endonuclease EcoRV is 5'-GAT/ATC-3', *Gene.* 27 (1984) 327–329. doi:10.1016/0378-1119(84)90078-7.
- [250] V. Pingoud, H. Geyer, R. Geyer, E. Kubareva, J.M. Bujnicki, A. Pingoud, Identification of base-specific contacts in protein-DNA complexes by photocrosslinking and mass spectrometry: a case study using the restriction endonuclease SsoII, *Mol. Biosyst.* 1 (2005) 135–141. doi:10.1039/B503091A.
- [251] V. Thielking, U. Selent, E. Koehler, A. Landgraf, H. Wolfes, J. Alves, A. Pingoud, Magnesium(2+) confers DNA binding specificity to the EcoRV restriction endonuclease, *Biochemistry.* 31 (1992) 3727–3732. doi:10.1021/bi00130a001.
- [252] M. Zahran, I. Daidone, J.C. Smith, P. Imhof, Mechanism of DNA Recognition by the Restriction Enzyme EcoRV, *J. Mol. Biol.* 401 (2010) 415–432. doi:10.1016/J.JMB.2010.06.026.
- [253] M. Oda, K. Furukawa, K. Ogata, A. Sarai, H. Nakamura, Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain, *J. Mol. Biol.* 276 (1998) 571–590. doi:10.1006/JMBI.1997.1564.
- [254] F.D. Urnov, J.C. Miller, Y.L. Lee, C.M. Beausejour, J.M. Rock, S. Augustus, A.C. Jamieson, M.H. Porteus, P.D. Gregory, M.C. Holmes, Highly efficient endogenous human gene correction using designed zinc-finger nucleases, *Nature.* 435 (2005) 646–651. doi:10.1038/NATURE03556.
- [255] R. Werther, J.P. Hallinan, A.R. Lambert, K. Havens, M. Pogson, J. Jarjour, R. Galizi, N. Windbichler, A. Crisanti, T. Nolan, B.L. Stoddard, Crystallographic analyses illustrate significant plasticity and efficient recoding of meganuclease target specificity, *Nucleic Acids Res.* 45 (2017) 8621. doi:10.1093/NAR/GKX544.
- [256] A.J. Bogdanove, A. Bohm, J.C. Miller, R.D. Morgan, B.L. Stoddard, Engineering altered protein–DNA recognition specificity, *Nucleic Acids Res.* 46 (2018) 4845. doi:10.1093/NAR/GKY289.
- [257] L. Rockah-Shmuel, D.S. Tawfik, Evolutionary transitions to new DNA methyltransferases through target site expansion and shrinkage, *Nucleic Acids Res.* 40 (2012) 11627–11637. doi:10.1093/NAR/GKS944.
- [258] R.A. Ikeda, C.M. Ligan, S. Warshamana, T7 promoter contacts essential for promoter activity in vivo., *Nucleic Acids Res.* 20 (1992) 2517. doi:10.1093/NAR/20.10.2517.
- [259] L.E. Macdonald, R.K. Durbin, J.J. Dunn, W.T. McAllister, Characterization of two types of termination signal for bacteriophage T7 RNA polymerase, *J. Mol. Biol.* 238 (1994) 145–158. doi:10.1006/JMBI.1994.1277.

- [260] C.A. Schneider, W.S. Rasband, K.W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, *Nat. Methods*. 9 (2012) 671–675. doi:10.1038/NMETH.2089.
- [261] A. Mishra, P. Pokhrel, M.T. Hoque, StackDPPred: a stacking based prediction of DNA-binding protein from sequence, *Bioinformatics*. 35 (2019) 433–441. doi:10.1093/BIOINFORMATICS/BTY653.
- [262] T.C. Mueser, J.M. Hinerman, J.M. Devos, R.A. Boyer, K.J. Williams, Structural analysis of bacteriophage T4 DNA replication: a review in the *Virology Journal* series on bacteriophage T4 and its relatives, *Viol. J.* 2010 71. 7 (2010) 1–16. doi:10.1186/1743-422X-7-359.
- [263] A.C. Mason, R.P. Rambo, B. Greer, M. Pritchett, J.A. Tainer, D. Cortez, B.F. Eichman, A structure-specific nucleic acid-binding domain conserved among DNA repair proteins, *Proc. Natl. Acad. Sci. U. S. A.* 111 (2014) 7618–7623. <https://pubmed.ncbi.nlm.nih.gov/24821763/> (accessed October 22, 2021).
- [264] M. Kong, B. Van Houten, Rad4 recognition-at-a-distance: Physical basis of conformation-specific anomalous diffusion of DNA repair proteins, *Prog. Biophys. Mol. Biol.* 127 (2016) 93–104. doi:10.1016/j.pbiomolbio.2016.12.004.
- [265] G. Badis, M.F. Berger, A.A. Philippakis, S. Talukder, A.R. Gehrke, S.A. Jaeger, E.T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D.E. Newburger, Q. Morris, T.R. Hughes, M.L. Bulyk, B. G, B. MF, P. AA, T. S, G. AR, J. SA, C. ET, M. G, V. A, C. X, K. H, W. CF, C. D, N. DE, M. Q, H. TR, B. ML, Diversity and complexity in DNA recognition by transcription factors, *Science*. 324 (2009) 1720–1723. doi:10.1126/SCIENCE.1162327.
- [266] R. Gordân, K.F. Murphy, R.P. McCord, C. Zhu, A. Vedenko, M.L. Bulyk, Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights, *Genome Biol.* 12 (2011) 1–18. doi:<https://doi.org/10.1186/gb-2011-12-12-r125>.
- [267] M. Levine, T. Hoey, Homeobox proteins as sequence-specific transcription factors, *Cell*. 55 (1988) 537–540. doi:10.1016/0092-8674(88)90209-7.
- [268] W. Shen, D. Chrobak, K. Krishnan, H.J. Lawrence, C. Largman, HOXB6 protein is bound to CREB-binding protein and represses globin expression in a DNA binding-dependent, PBX interaction-independent process, *J Biol Chem*. 279 (2004) 39895–39904. doi:10.1074/jbc.m404132200.
- [269] P.J. Troy, G.S. Daftary, C.N. Bagot, H.S. Taylor, Transcriptional repression of peri-implantation EMX2 expression in mammalian reproduction by HOXA10, *Mol Cell Biol*. 23 (2003) 1–13. doi:10.1128/mcb.23.1.1-13.2003.
- [270] E. Lewis, A gene complex controlling segmentation in *Drosophila*, *Nature*. 276 (1978) 565–570. doi:10.1038/276565a0.
- [271] G.A. Grammatopoulos, E. Bell, L. Toole, A. Lumsden, A.S. Tucker, Homeotic transformation of branchial arch identity after *Hoxa2* overexpression, *Development*. 127 (2000) 5355–5365.
- [272] A. Tkatchenko, R. Visconti, L. Shang, T. Papenbrock, N. Pruett, T. Ito, M. Ogawa, A. Awgulewitsch, Overexpression of *Hoxc13* in differentiating keratinocytes results in downregulation of a novel hair keratin gene cluster and alopecia, *Development*. 128 (2001) 1547–1558.
- [273] Y. Wu, M. Moser, V.L. Bautch, C. Patterson, *HoxB5* is an upstream transcriptional switch for differentiation of the vascular endothelium from precursor cells, *Mol Cell Biol*. 23 (2003) 5680–5691. doi:10.1128/mcb.23.16.5680-5691.2003.
- [274] T.J. Richmond, C.A. Davey, The structure of DNA in the nucleosome core, *Nat.* 2003 4236936. 423 (2003) 145–150. doi:10.1038/nature01595.
- [275] C.R. Kissinger, B. Liu, E. Martin-Blanco, T.B. Kornberg, C.O. Pabo, Crystal structure of an engrailed homeodomain-DNA complex at 2.8Å resolution: a framework for understanding homeodomain-DNA interactions, *Cell*. 63 (1990) 579–590. doi:10.1016/0092-8674(90)90453-I.

- [276] D.E. Piper, A.H. Batchelor, C.-P. Chang, M.L. Cleary, C. Wolberger, Structure of a HoxB1-Pbx1 heterodimer bound to DNA: Role of the hexapeptide and a fourth homeodomain helix in complex formation, *Cell*. 96 (1999) 587–597. doi:10.1016/s0092-8674(00)80662-5.
- [277] E.A. Niskanen, O. Kalliollinna, T.O. Ihalainen, M. Häkkinen, M. Vihinen-Ranta, Mutations in DNA Binding and Transactivation Domains Affect the Dynamics of Parvovirus NS1 Protein, *J. Virol.* 87 (2013) 11762–11774. doi:10.1128/JVI.01678-13.
- [278] G.-N. Gao, M. Wang, N. Yang, Y. Huang, R.-M. Xu, Structure of Zeste–DNA Complex Reveals a New Modality of DNA Recognition by Homeodomain-Like Proteins, *J. Mol. Biol.* 427 (2015) 3824–3833. doi:10.1016/j.jmb.2015.10.008.
- [279] T.A. Hilder, J.M. Hodgkiss, Molecular Mechanism of Binding between 17 β -Estradiol and DNA, *Comput. Struct. Biotechnol. J.* 15 (2017) 91–97. doi:10.1016/j.csbj.2016.12.001.
- [280] A.N. Bullock, A.R. Fersht, Rescuing the function of mutant p53, *Nat. Rev. Cancer.* 1 (2001) 68–76. doi:10.1038/35094077.
- [281] M.D. Ilsley, S. Huang, G.W. Magor, M.J. Landsberg, K.R. Gillinder, A.C. Perkins, Corrupted DNA-binding specificity and ectopic transcription underpin dominant neomorphic mutations in KLF/SP transcription factors, *BMC Genomics* 2019 201. 20 (2019) 1–12. doi:10.1186/S12864-019-5805-Z.
- [282] Z. Xiao, L. Guo, Y. Zhang, L. Cui, Y. Dai, Z. Lan, Q. Zhang, S. Wang, W. Liu, Structural analysis of missense mutations occurring in the DNA-binding domain of HSF4 associated with congenital cataracts, *J. Struct. Biol. X.* 4 (2020) 100015. doi:10.1016/J.YJSBX.2019.100015.
- [283] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank., *Nucleic Acids Res.* 28 (2000) 235–42. <http://www.ncbi.nlm.nih.gov/pubmed/10592235> (accessed April 24, 2017).
- [284] P. Aggarwal, M. Das Gupta, A.P. Joseph, N. Chatterjee, N. Srinivasan, U. Nath, Identification of specific DNA binding residues in the TCP family of transcription factors in Arabidopsis, *Plant Cell.* 22 (2010) 1174–1189. doi:10.1105/TPC.109.066647.
- [285] H. Geyer, R. Geyer, V. Pingoud, A novel strategy for the identification of protein-DNA contacts by photocrosslinking and mass spectrometry, *Nucleic Acids Res.* 32 (2004). doi:10.1093/NAR/GNH131.
- [286] T.S. Furey, ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions, *Nat. Rev. Genet.* 2012 1312. 13 (2012) 840–852. doi:10.1038/nrg3306.
- [287] M. Adli, B.E. Bernstein, Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq, *Nat. Protoc.* 2011 610. 6 (2011) 1656–1668. doi:10.1038/nprot.2011.402.
- [288] H.S. Rhee, B.F. Pugh, Comprehensive Genome-wide Protein-DNA Interactions Detected at Single Nucleotide Resolution, *Cell.* 147 (2011) 1408. doi:10.1016/J.CELL.2011.11.013.
- [289] K. Markham, Y. Bai, G. Schmitt-Ulms, Co-immunoprecipitations revisited: An update on experimental concepts and their implementation for sensitive interactome investigations of endogenous proteins, *Anal. Bioanal. Chem.* 389 (2007) 461–473. doi:10.1007/S00216-007-1385-X/FIGURES/5.
- [290] K.K. Andrienas, A. Penvose, T. Siggers, Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes, *Brief. Funct. Genomics.* 14 (2015) 17–29. doi:10.1093/BFGP/ELU046.
- [291] M.L. Bulyk, Protein Binding Microarrays for the Characterization of Protein-DNA Interactions, *Adv. Biochem. Eng. Biotechnol.* 104 (2007) 65–85. doi:10.1007/10_025.
- [292] D.J. Galas, A. Schmitz, DNase footprinting: a simple method for the detection of protein-DNA binding specificity, *Nucleic Acids Res.* 5 (1978) 3157–3170. doi:10.1093/NAR/5.9.3157.
- [293] A.R. Oliphant, C.J. Brandl, K. Struhl, Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein, *Mol. Cell. Biol.* 9 (1989) 2944–2949. doi:10.1128/MCB.9.7.2944-2949.1989.

- [294] L. Wang, S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, *Nucleic Acids Res.* 34 (2006) W243–W248. doi:10.1093/NAR/GKL298.
- [295] L. Wang, C. Huang, M.Q. Yang, J.Y. Yang, BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, *BMC Syst. Biol.* 2010 41. 4 (2010) 1–9. doi:10.1186/1752-0509-4-S1-S3.
- [296] Y.-H. Zhu, J. Hu, X.-N. Song, D.-J. Yu, DNAPred: Accurate Identification of DNA-Binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines, *J. Chem. Inf. Model.* 59 (2019) 3057–3071. doi:10.1021/ACS.JCIM.8B00749.
- [297] J. Si, Z. Zhang, B. Lin, M. Schroeder, B. Huang, MetaDBSite: a meta approach to improve protein DNA-binding sites prediction, *BMC Syst. Biol.* 5 (2011) 1–7. doi:10.1186/1752-0509-5-S1-S7.
- [298] W.-Y. Chu, Y.-F. Huang, C.-C. Huang, Y.-S. Cheng, C.-K. Huang, Y.-J. Oyang, ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors, *Nucleic Acids Res.* 37 (2009) W396–W401. doi:10.1093/NAR/GKP449.
- [299] X. Yang, J. Wang, J. Sun, R. Liu, SNBRFinder: A Sequence-Based Hybrid Algorithm for Enhanced Prediction of Nucleic Acid-Binding Residues, *PLoS One.* 10 (2015) e0133260. doi:10.1371/JOURNAL.PONE.0133260.
- [300] H. Su, M. Liu, S. Sun, Z. Peng, J. Yang, Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods, *Bioinformatics.* 35 (2019) 930–936. doi:10.1093/BIOINFORMATICS/BTY756.
- [301] J. Hu, Y. Li, M. Zhang, X. Yang, H.-B. Shen, D.-J. Yu, Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-based Features and Boosting Multiple SVMs, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14 (2017) 1389–1398. doi:10.1109/TCBB.2016.2616469.
- [302] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, *BMC Bioinforma.* 2005 61. 6 (2005) 1–6. doi:10.1186/1471-2105-6-33.
- [303] Y. Ofra, V. Mysore, B. Rost, Prediction of DNA-binding residues from sequence, *Bioinformatics.* 23 (2007) i347–i353. doi:10.1093/BIOINFORMATICS/BTM174.
- [304] Z. Peng, L. Kurgan, High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder, *Nucleic Acids Res.* 43 (2015) e121. doi:10.1093/NAR/GKV585.
- [305] Y. Zhang, S. Qiao, S. Ji, N. Han, D. Liu, J. Zhou, Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information, *Eng. Appl. Artif. Intell.* 79 (2019) 58–66. doi:10.1016/J.ENGAPPAI.2019.01.003.
- [306] J. Zhou, Q. Lu, R. Xu, L. Gui, H. Wang, CNNsite: Prediction of DNA-binding residues in proteins using Convolutional Neural Network with sequence features, *Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016.* (2016) 78–85. doi:10.1109/BIBM.2016.7822496.
- [307] L. Wang, M.Q. Yang, J.Y. Yang, Prediction of DNA-binding residues from protein sequence information using random forests, *BMC Genomics* 2009 101. 10 (2009) 1–9. doi:10.1186/1471-2164-10-S1-S1.
- [308] C. Yan, M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs, V. Honavar, Predicting DNA-binding sites of proteins from amino acid sequence, *BMC Bioinforma.* 2006 71. 7 (2006) 1–10. doi:10.1186/1471-2105-7-262.
- [309] J. Zhou, Q. Lu, R. Xu, Y. He, H. Wang, EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM Relation Transformation, *BMC Bioinforma.* 2017 181. 18 (2017) 1–16. doi:10.1186/S12859-017-1792-8.
- [310] B.P. Nguyen, Q.H. Nguyen, G.-N. Doan-Ngoc, T.-H. Nguyen-Vo, S. Rahardja, iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks, *BMC Bioinforma.* 2019 2023. 20 (2019) 1–12. doi:10.1186/S12859-019-3295-2.

- [311] X. Ma, J. Guo, H.-D. Liu, J.-M. Xie, X. Sun, Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 9 (2012) 1766–1775. doi:10.1109/TCBB.2012.106.
- [312] S. Hwang, Z. Gou, I.B. Kuznetsov, DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, *Bioinformatics.* 23 (2007) 634–636. doi:10.1093/BIOINFORMATICS/BTL672.
- [313] I.B. Kuznetsov, Z. Gou, R. Li, S. Hwang, Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins, *Proteins.* 64 (2006) 19–27. doi:10.1002/PROT.20977.
- [314] Y. Jing, L. Kurgan, DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues, *Nucleic Acids Res.* 45 (2017) e84. doi:10.1093/NAR/GKX059.
- [315] A. Amirkhani, M. Kolahdoozi, C. Wang, L.A. Kurgan, Prediction of DNA-Binding Residues in Local Segments of Protein Sequences with Fuzzy Cognitive Maps, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17 (2020) 1372–1382. doi:10.1109/TCBB.2018.2890261.
- [316] J. Zhang, Z. Ma, L. Kurgan, Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains, *Brief. Bioinform.* 20 (2019) 1250–1268. doi:10.1093/BIB/BBX168.
- [317] M.B. Carson, R. Langlois, H. Lu, NAPS: a residue-level nucleic acid-binding prediction server, *Nucleic Acids Res.* 38 (2010) W431–W435. doi:10.1093/NAR/GKQ361.
- [318] C. Shen, Y. Ding, J. Tang, J. Song, F. Guo, Identification of DNA–protein Binding Sites through Multi-Scale Local Average Blocks on Sequence Information, *Mol. A J. Synth. Chem. Nat. Prod. Chem.* 22 (2017) 2079. doi:10.3390/MOLECULES22122079.
- [319] D.J. Yu, J. Hu, J. Yang, H. Bin Shen, J. Tang, J.Y. Yang, Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 10 (2013) 994–1008. doi:10.1109/TCBB.2013.104.
- [320] Q. Wu, Z. Peng, Y. Zhang, J. Yang, COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking, *Nucleic Acids Res.* 46 (2018) W438–W442. doi:10.1093/NAR/GKY439.
- [321] M. Gao, J. Skolnick, DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions, *Nucleic Acids Res.* 36 (2008) 3978–3992. doi:10.1093/NAR/GKN332.
- [322] I. Dror, S. Shazman, S. Mukherjee, Y. Zhang, F. Glaser, Y. Mandel-Gutfreund, Predicting nucleic acid binding interfaces from structural models of proteins, *Proteins Struct. Funct. Bioinforma.* 80 (2012) 482–489. doi:10.1002/PROT.23214.
- [323] S. Jones, H.P. Shanahan, H.M. Berman, J.M. Thornton, Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, 31 (2003) 7189–7198. <https://pubmed.ncbi.nlm.nih.gov/14654694/> (accessed October 22, 2021).
- [324] Y. Tsuchiya, K. Kinoshita, H. Nakamura, PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces, *Bioinformatics.* 21 (2005) 1721–1723. doi:10.1093/BIOINFORMATICS/BTI232.
- [325] B.-Q. Li, K.-Y. Feng, J. Ding, Y.-D. Cai, Predicting DNA-binding sites of proteins based on sequential and 3D structural information, *Mol. Genet. Genomics.* 289 (2014) 489–499. doi:10.1007/S00438-014-0812-X.
- [326] X. Zhu, S.S. Ericksen, J.C. Mitchell, DBSI: DNA-binding site identifier, *Nucleic Acids Res.* 41 (2013) e160. doi:10.1093/NAR/GKT617.
- [327] T. Li, Q.-Z. Li, S. Liu, G.-L. Fan, Y.-C. Zuo, Y. Peng, PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information, *Bioinformatics.* 29 (2013) 678–685. doi:10.1093/BIOINFORMATICS/BTT029.

- [328] J. Zhou, R. Xu, Y. He, Q. Lu, H. Wang, B. Kong, PDNAsite: Identification of DNA-binding Site from Protein Sequence by Incorporating Spatial and Sequence Context, *Sci. Reports* 2016 61. 6 (2016) 1–15. doi:10.1038/srep27653.
- [329] H. Tjong, H.-X. Zhou, DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces, *Nucleic Acids Res.* 35 (2007) 1465–1477. doi:10.1093/NAR/GKM008.
- [330] Y. Xia, C.Q. Xia, X. Pan, H. Bin Shen, GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues, *Nucleic Acids Res.* 49 (2021) e51–e51. doi:10.1093/NAR/GKAB044.
- [331] P. Ozbek, S. Soner, B. Erman, T. Haliloglu, DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues, *Nucleic Acids Res.* 38 (2010) W417–W423. doi:10.1093/NAR/GKQ396.
- [332] J. Segura, P.F. Jones, N. Fernandez-Fuentes, A holistic in silico approach to predict functional sites in protein structures, *Bioinformatics.* 28 (2012) 1845–1850. doi:10.1093/BIOINFORMATICS/BTS269.
- [333] Y.C. Chen, J.D. Wright, C. Lim, DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry, *Nucleic Acids Res.* 40 (2012) W249–W256. doi:10.1093/NAR/GKS481.
- [334] F. Corsi, R. Lavery, E. Laine, A. Carbone, Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties, *PLOS Comput. Biol.* 16 (2020) e1007624. doi:10.1371/JOURNAL.PCBI.1007624.
- [335] Z. Miao, E. Westhof, Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score, *Nucleic Acids Res.* 43 (2015) 5340–5351. doi:10.1093/NAR/GKV446.
- [336] L. Deng, J. Pan, X. Xu, W. Yang, C. Liu, H. Liu, PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine, *BMC Bioinforma.* 2018 1919. 19 (2018) 135–145. doi:10.1186/S12859-018-2527-1.
- [337] R. Liu, J. Hu, DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches, *Proteins.* 81 (2013) 1885–1899. doi:10.1002/PROT.24330.
- [338] T. Werner, Computer-assisted analysis of transcription control regions. MatInspector and other programs, *Methods Mol. Biol.* 132 (2000) 337–349. doi:10.1385/1-59259-192-2:337.
- [339] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, T. Werner, MatInspector and beyond: promoter analysis based on transcription factor binding sites, *Bioinformatics.* 21 (2005) 2933–2942. doi:10.1093/BIOINFORMATICS/BTI473.
- [340] A.E. Kel, E. Gößling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, E. Wingender, MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res.* 31 (2003) 3576–3579. doi:10.1093/NAR/GKG585.
- [341] A. Sandelin, W.W. Wasserman, B. Lenhard, ConSite: web-based prediction of regulatory elements using cross-species comparison, *Nucleic Acids Res.* 32 (2004) W249–W452. doi:10.1093/NAR/GKH372.
- [342] E. Berezikov, V. Guryev, E. Cuppen, CONREAL web server: identification and visualization of conserved transcription factor binding sites, *Nucleic Acids Res.* 33 (2005) W447–W450. doi:10.1093/NAR/GKI378.
- [343] A. de Jong, H. Pietersma, M. Cordes, O.P. Kuipers, J. Kok, PePPER: a webserver for prediction of prokaryote promoter elements and regulons, *BMC Genomics.* 13 (2012) 1–10. doi:10.1186/1471-2164-13-299/FIGURES/4.
- [344] X. Messeguer, R. Escudero, D. Farré, O. Núñez, J. Martínez, M.M. Albà, PROMO: detection of known transcription regulatory elements using species-tailored searches, *Bioinformatics.* 18 (2002) 333–334. doi:10.1093/BIOINFORMATICS/18.2.333.

- [345] D. Farré, R. Roset, M. Huerta, J.E. Adsuará, L. Roselló, M.M. Albà, X. Messeguer, Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN, *Nucleic Acids Res.* 31 (2003) 3651. doi:10.1093/NAR/GKG605.
- [346] L.J. Gearing, H.E. Cumming, R. Chapman, A.M. Finkel, I.B. Woodhouse, K. Luu, J.A. Gould, S.C. Forster, P.J. Hertzog, CiiDER: A tool for predicting and analysing transcription factor binding sites, *PLoS One.* 14 (2019) e0215495. doi:10.1371/JOURNAL.PONE.0215495.
- [347] D. Cirillo, T. Botta-Orfila, G.G. Tartaglia, By the company they keep: interaction networks define the binding ability of transcription factors, *Nucleic Acids Res.* 43 (2015) e125. doi:10.1093/NAR/GKV607.
- [348] G. Fan, Q. Li, K. Yang, TFBSs: A web server for predicting transcription factor binding sites, 2012 IEEE Int. Conf. Cyber Technol. Autom. Control. Intell. Syst. (2013) 65–68. doi:10.1109/CYBER.2012.6319888.
- [349] B. Park, J. Im, N. Tuvshinjargal, W. Lee, K. Han, Sequence-based prediction of protein-binding sites in DNA: Comparative study of two SVM models, *Comput. Methods Programs Biomed.* 117 (2014) 158–167.
- [350] C.-K. Lin, C.-Y. Chen, PiDNA: Predicting protein-DNA interactions with structural models., *Nucleic Acids Res.* 41 (2013) W523–W530. doi:10.1093/nar/gkt388.
- [351] R. Gabdoulline, D. Eckweiler, A. Kel, P. Stegmaier, 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations, *Nucleic Acids Res.* 40 (2012) W180–W185. doi:10.1093/NAR/GKS551.
- [352] A. Farrel, J. tao Guo, An efficient algorithm for improving structure-based prediction of transcription factor binding sites, *BMC Bioinformatics.* 18 (2017) 1–11. doi:10.1186/S12859-017-1755-0/FIGURES/6.
- [353] T.-Y. Chien, C.-K. Lin, C.-W. Lin, Y.-Z. Weng, C.-Y. Chen, D.T.-H. Chang, DBD2BS: connecting a DNA-binding protein with its binding sites, *Nucleic Acids Res.* 40 (2012) W173–W179. doi:10.1093/NAR/GKS564.
- [354] C.-Y. Chen, T.-Y. Chien, C.-W.C.-K. Lin, C.-W.C.-K. Lin, Y.-Z. Weng, D.T.-H. Chang, Predicting Target DNA Sequences of DNA-Binding Proteins Based on Unbound Structures, *PLoS One.* 7 (2012) e30446. doi:10.1371/JOURNAL.PONE.0030446.
- [355] A. Mathelier, X. Zhao, A.W. Zhang, F. Parcy, R. Worsley-Hunt, D.J. Arenillas, S. Buchman, C.Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, W.W. Wasserman, JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles, *Nucleic Acids Res.* 42 (2014) D142–D147. doi:10.1093/nar/gkt997.
- [356] A. Šali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* 234 (1993) 779–815. doi:10.1006/JMBI.1993.1626.
- [357] Y. Yan, D. Zhang, P. Zhou, B. Li, S.Y. Huang, HDock: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy, *Nucleic Acids Res.* 45 (2017) W365–W373. doi:10.1093/NAR/GKX407.
- [358] Y. Yan, S.Y. Huang, Modeling Protein–Protein or Protein–DNA/RNA Complexes Using the HDock Webserver, *Methods Mol. Biol.* 2165 (2020) 217–229. doi:10.1007/978-1-0716-0708-4_12.
- [359] O. Tange, GNU Parallel 2018, (2018). doi:10.5281/ZENODO.1146014.
- [360] S. Poddar, D. Chakravarty, P. Chakrabarti, Structural changes in DNA-binding proteins on complexation, *Nucleic Acids Res.* 46 (2018) 3298–3308. doi:10.1093/NAR/GKY170.
- [361] M. Andrabi, K. Mizuguchi, S. Ahmad, A. M, M. K, A. S, Conformational changes in DNA-binding proteins: relationships with precomplex features and contributions to specificity and stability, *Proteins.* 82 (2014) 841–857. <https://pubmed.ncbi.nlm.nih.gov/24265157/> (accessed April 18, 2017).
- [362] A. V. Persikov, M. Singh, De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins, *Nucleic Acids Res.* 42 (2014) 97–108. doi:10.1093/NAR/GKT890.

- [363] D.R. Whelan, T.J. Hiscox, J.I. Rood, K.R. Bambery, D. McNaughton, B.R. Wood, Detection of an en masse and reversible B- to A-DNA conformational transition in prokaryotes in response to desiccation, *J. R. Soc. Interface.* 11 (2014). doi:10.1098/RSIF.2014.0454.
- [364] F. DiMaio, X. Yu, E. Rensen, M. Krupovic, D. Prangishvili, E.H. Egelman, A Virus that Infects a Hyperthermophile Encapsidates A-Form DNA, *Science.* 348 (2015) 914. doi:10.1126/SCIENCE.AAA4181.
- [365] D. Svozil, J. Kalina, M. Omelka, B. Schneider, DNA conformations and their sequence preferences, *Nucleic Acids Res.* 36 (2008) 3690. doi:10.1093/NAR/GKN260.
- [366] C. Oguey, N. Foloppe, B. Hartmann, Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions, *PLoS One.* 5 (2010) e15931. doi:10.1371/JOURNAL.PONE.0015931.
- [367] E.N. Nikolova, G.D. Bascom, I. Andricioaei, H.M. Al-Hashimi, Probing sequence-specific DNA flexibility in A-tracts and pyrimidine-purine steps by nuclear magnetic resonance ¹³C relaxation and molecular dynamics simulations, *Biochemistry.* 51 (2012) 8654–8664. doi:10.1021/BI3009517/SUPPL_FILE/BI3009517_SI_001.PDF.
- [368] A. V. Morozov, J.J. Havranek, D. Baker, E.D. Siggia, Protein-DNA binding specificity predictions with structural models, 33 (2005) 5781–5798. <https://pubmed.ncbi.nlm.nih.gov/16246914/> (accessed December 30, 2021).
- [369] J.D. Blanco, L. Radusky, H. Climente-González, L. Serrano, FoldX accurate structural protein–DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1), *Nucleic Acids Res.* 46 (2018) 3852–3863. doi:10.1093/NAR/GKY228.
- [370] W. Yang, L. Deng, PreDBA: A heterogeneous ensemble approach for predicting protein-DNA binding affinity, *Sci. Reports* 2020 101. 10 (2020) 1–11. doi:10.1038/s41598-020-57778-1.
- [371] J.R. Desjarlais, T.M. Handel, De novo design of the hydrophobic cores of proteins., *Protein Sci.* 4 (1995) 2006. doi:10.1002/PRO.5560041006.
- [372] G. Dantas, B. Kuhlman, D. Callender, M. Wong, D. Baker, A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins, *J. Mol. Biol.* 332 (2003) 449–460. doi:10.1016/S0022-2836(03)00888-X.
- [373] A. Leaver-Fay, M. Tyka, S.M. Lewis, O.F. Lange, J. Thompson, R. Jacak, K. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, I.W. Davis, S. Cooper, A. Treuille, D.J. Mandell, F. Richter, Y.E.A. Ban, S.J. Fleishman, J.E. Corn, D.E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popović, J.J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J. Gray, B. Kuhlman, D. Baker, P. Bradley, Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules, *Methods Enzymol.* 487 (2011) 545–574. doi:10.1016/B978-0-12-381270-4.00019-6.
- [374] A. Pandurangan and M. Topf, RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps, *Bioinformatics* 28 (2012) 2391–2393. doi:10.1093/bioinformatics/bts446

Publications

[1] **S. Nair**, M. Madhusudhan “JEDII: Juxtaposition Enabled DNA-binding Interface Identifier” Biorxiv Submitted (Manuscript ID: BIORXIV/2022/492702)

[2] T. R. Kanitkar, N. Sen, **S. Nair**, N. Soni, K. Amritkar, Y. Ramtirtha, and M. Madhusudhan, “Methods for Molecular Modelling of Protein Complexes,” *Methods in Molecular Biology* (Clifton, NJ), vol. 2305, pp. 53–80, 2021.

[3] N. Sen, T. R. Kanitkar, A. A. Roy, N. Soni, K. Amritkar, S. Supekar, **S. Nair**, G. Singh, and M. S. Madhusudhan, “Predicting and designing therapeutics against the Nipah virus,” *PLOS Neglected Tropical Diseases*, vol. 13, pp. 1–23, 2019.

[4] A. Roy*, **S. Nair***, N. Sen, N. Soni, and M. Madhusudhan, “In silico methods for design of biological therapeutics,” *Methods*, vol. 131, pp. 33 – 65, 2017.

[5] N. Farheen, N. Sen, **S. Nair**, K. P. Tan, and M. Madhusudhan, “Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations,” *Progress in Biophysics and Molecular Biology*, vol. 128, pp. 14 – 23, 2017.

*Indicates equal contribution