

A COMPUTATIONAL STUDY OF DNA-PROTEIN INTERACTIONS

SHAYONI DUTTA



**DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

October 2016

©Indian Institute of Technology Delhi (IITD), New Delhi, 2016

A COMPUTATIONAL STUDY OF DNA-PROTEIN INTERACTIONS

by

SHAYONI DUTTA

Department of Biochemical Engineering and Biotechnology

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the



Indian Institute of Technology Delhi

October 2016

CERTIFICATE

This is to certify that the thesis entitled “**A computational study of DNA-Protein interactions**” being submitted by **Ms. Shayoni Dutta** to the Indian Institute of Technology Delhi for the award of the degree of “**Doctor of Philosophy**”, is a record of the bonafide research work carried out by her, which has been prepared under my supervision in conformity with the rules and regulations of the Indian Institute of Technology Delhi. The research reports and the results presented in this thesis have not been submitted for any degree or diploma in any other University or Institute.

Dr. D. Sundar

Department of Biochemical Engineering and Biotechnology
I.I.T Delhi

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. D Sundar for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Gopal Agarwal, Dr. Ritu Kulshreshtha, and Dr. Gitanjali Yadav (NIPGR, New Delhi), not only for their insightful comments and encouragement, but also for the hard questions which motivated me to widen my research from various perspectives.

I thank my fellow lab-mates especially Jaspreet, Spandan and Harsh for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four and half years. Also my special thanks to my good friend and lab mate Shashank for his innumerable contribution and support as well as brainstorming discussions, truly without whom I would have never been able to complete my Ph.D. In particular, I am grateful to Vidhi and Anjani for ensuring that I never lost hope and in keeping the lab environment full of energy, love and fun.

Last but not the least; I would like to thank my family: my parents and my brother for supporting me spiritually throughout writing this thesis and my life in general. My mother in fact has been the biggest support in my life, without whom my existence holds no worth.

Shayoni Dutta

ABSTRACT

Technology for specifically correcting 'faulty' bases or mutations within genes of humans has been the 'Holy Grail' in genetic medicine. Gene targeting using zinc finger nucleases (ZFNs) - proteins custom-designed to cut at specific DNA sequences – appears to make this possible. These “artificial” proteins combine the non-specific endonuclease activity of *FokI* restriction enzyme with the ability of zinc-finger (ZF) motifs to specifically recognize a DNA triplet sequence. The Cys₂His₂ ZF motif binds specific sequences in DNA by virtue of its unique *modular* 30 amino acid structure (stabilized by a zinc ion), the α -helix inserting into the major groove of the DNA double helix. Amino acids involved in DNA recognition within the α -helix of the ZF motifs can be changed while maintaining the remaining amino acids as a consensus backbone to generate ZF motifs with new triplet sequence specificities. Normally, three such ZF motifs are linked together in tandem to generate a ZF protein (ZFP) that binds to a 9-bp DNA target site, which is a composite of the individual DNA triplet sub-sites recognized by each of the three ZF motifs. Binding of two three-finger ZFN monomers, each recognizing a 9-bp DNA target inverted site is necessary because dimerization of the *FokI* cleavage domain is required to produce a DSB (double stranded break). Therefore, three-finger ZFNs effectively have an 18-bp recognition site, which is long enough to specify a unique genomic address in plants and mammals including humans. ZFNs thus offer a general mechanism to introduce a site-specific DSB within a plant or a mammalian genome (Berg 1993).

For ZFN-mediated gene targeting to become an efficient and powerful genome engineering tool for specific proven biological and biomedical applications, rapid

design and generation of ZFPs that determine the sequence specificity of the ZFNs is essential. The designed ZFPs appear to have the highest affinity and sequence-specificity for their targets only when the individual ZF designs are chosen in the context of their neighbouring fingers. The computational work in this thesis focuses on understanding DNA-binding specificity in zinc finger proteins (i) through analysis of the physicochemical nature behind the ZFP-DNA interactions (ii) by investigating aspects like desolvation and DNA deformation based on simulations and free energy profile data that revealed a consensus in correlating affinity and specificity as well as stability of ZFP-DNA interactions, and (iii) by development of novel prediction algorithms that were evaluated for their performance against experimental data.

CONTENTS

Title	Page No.
List of Figures -----	I
List of Tables -----	II
List of Equations -----	III
List of Abbreviations -----	IV
Chapter 1: Introduction and Objectives -----	1
1.1 Introduction	2
1.2 Genome editing tools: ZFP	4
1.3 Engineering the genome	6
1.3.1 The need for genome engineering	6
1.3.2 Zinc fingers are instrumental in mediating gene therapy	7
1.3.3 Applications of engineered ZFP	8
1.4 Zinc finger Proteins	11
1.4.1 History of zinc finger proteins	11
1.4.2 Different types of zinc fingers	11
1.4.3 Crystal structure of Zif-268	12
1.4.4 Recognition code of zinc finger proteins	14
1.4.5 Custom design and selection strategy of zinc finger proteins	15
1.5 Prediction tools for engineering customized zinc finger proteins	17
1.6 Motivation	18
1.7 Definition of the problem	19
1.8 Objectives	20

Chapter 2: Literature Acquisition and Analysis	21
2.1 Discovery and current knowledge of ZFP binding to target DNA.	22
2.2 Zif-268-DNA-binding patterns	23
2.3 Zinc finger engineering	26
2.3.1 Why Zif-268.....	27
2.3.2 Design strategies	28
2.4 Different databases & prediction tools	32
2.5 Issue of affinity versus specificity	35
2.6 Analysis of factors affecting the binding affinity of ZFPs	36
2.6.1 Amino acid pool dominating the interaction interface of ZFP-DNA.....	37
2.6.2 Effect of zinc on ZFP-DNA binding affinity and consequent stability.....	38
2.6.3 DNA distortion and effect of major groove upon ZFP-DNA binding	40
2.6.4 Direct and indirect interactions dominating the interaction interface.....	40
2.6.5 <i>Modular</i> and <i>Synergistic</i> modes of ZFP binding	41
2.7 Conclusion	44
Chapter 3: Development of novel computational algorithm for predicting DNA binding specificity in zinc finger proteins	47
3.1 Physico-chemical versus computational approaches for the prediction of ZFP-DNA interactions.....	48
3.2 Research methodology.....	49
3.3 Approach1: <i>Modular</i> binding of DNA targets with Zif-268 mutants derived from a small pool of amino acids.....	52
3.4 Approach 2: <i>Synergistic</i> binding of DNA targets with Zif-268 mutants derived from a small pool of amino acids.....	57
3.5 Approach 3: <i>Modular</i> binding of DNA targets with all Zif-268 mutants.....	60
3.6 Analysis of Predictions and Validation with experimental data.....	62
3.6.1 Approach 1: Assumes <i>Modular</i> mode of binding and mutations from a pool of amino acids	63

3.6.2 Approach 2: Assumes <i>Synergistic</i> mode of binding and mutations from a pool of amino acids	64
3.6.3 Approach 3: Assumes <i>modular</i> mode of binding and mutations using all 20 amino acids	66
3.6.4 Development of webserver: Interfacial Hydrogen Bond Energy	68
3.7 A comparative analysis of the three approaches	68
3.8 Factors affecting zinc finger binding specificities.....	72
3.8.1 Amino Acid preference	72
3.8.2 Positional dependence of DNA codon	73
3.9 Conclusion.....	76
Chapter 4: Indirect factors affecting ZFP-DNA binding specificity -----	79
4.1 Introduction	80
4.2 Background.....	81
4.3 Research methodology	83
4.3.1 Starting structures, models and docking studies.....	83
4.3.2 Molecular Dynamics simulation procedure.....	85
4.3.3 Procedure to evaluate DNA deformation upon complexation	86
4.4 Results and Discussion	88
4.4.1 Binding affinity determined by docking scores and respective K_d values.....	88
4.4.2 Direct correlation between binding affinity and stability of complex determined by RMSD plots	89
4.4.3 Indirect interactions of Zif-268 with DNA targets of the type 5' GNN-GNN-GNN 3' demonstrating the varying binding strength	90
4.5 Conclusion.....	95
Chapter 5: Statistical modelling based computational approaches to predict ZFP-DNA interactions -----	98
5.1 Introduction	99
5.2 Statistical and predictive modelling based on Neural network	99
5.2.1 Background.....	101

5.2.2 Research methodology105

5.2.3 Results and Discussion110

5.2.4 Conclusion115

5.3 Improving prediction accuracy using convex optimization.....118

5.3.1 Computational prediction of DNA binding specificity.....118

5.3.2 Hydrogen Bonds and Energy119

5.3.3 Research methodology121

5.3.4 Results.....125

5.3.5 Analysis.....127

5.3.6 Conclusion130

Chapter 6: Conclusion ----- 131

References ----- 145

Appendix 1: Supplementary Figures----- 159

Appendix 2: Supplementary Tables----- 170

Appendix 3: Softwares used ----- 184

List of Publications ----- 187

Resume of the Author----- 190

LIST OF FIGURES

Fig. No.	Figure Title	Page No.
1.1	Structure of zinc finger proteins -----	5
1.2	Genome manipulation using ZFNs -----	7
1.3	Applications of ZFPs -----	10
2.1	A zinc finger protein binding to its specific target DNA site-----	25
2.2	Key interacting residues of the α -helix of ZFP -----	38
2.3.1	A 5'GGGGGGGGG 3' DNA sequence with Zif-268 RMSD trajectory in the presence and absence of zinc -----	39
2.3.2	RMSD trajectory individually for DNA and protein due to zinc ion -----	39
2.4	Studying the effect of DNA distortion upon ZFP-DNA complexation -----	41
2.5	Interactions dominating the interaction interface of ZFP-DNA-----	42
2.6	A schematic representation of DNA-zinc finger protein interaction depicting the two possible modes of binding-----	44
3.1	Comparative analysis and validation of all the three prediction approaches -----	52
3.2	A schematic pipeline of the three prediction approaches -----	59
3.3	Amino acid propensity w.r.t key residue positions of the ZFP helix for each finger -----	74
4.1	Desolvation -----	83
4.2	RMSD versus time plot for target DNA complexed to Zif-268 -----	91
4.3	H-bond variation over simulation trajectory of entire target DNA sequences complexed to Zif-268-----	93
4.4	DNA deformation as a function of binding strength-----	94
4.5	Correlation between binding strength, docking score and stability (RMSD) of sample targets -----	96
4.6	Correlation between solvation energy and binding affinity-----	96
4.7	Correlation between FEP , docking score and solvation energy -----	97
5.1	Pipeline to predict optimal ZFPs for any 9bp target DNA -----	103
5.2	RMSD scatter plot for validating rotamer database-----	125

LIST OF TABLES

Table No.	Table Title	Page No.
2.1	Web Tools for predicting DNA-binding specificity in zinc finger proteins -----	31
3.1	Validation of predictions made by assuming <i>modular</i> mode of binding and mutations from a small consensus pool of amino acids (Approach 1)-----	64
3.2	Validation of predictions by assuming <i>Synergistic</i> mode of binding and mutations from a small consensus pool of amino acids (Approach 2)-----	65
3.3	Comparative analysis of predictions based on different modes of binding against experimental data -----	67
3.4	Effect of DNA sub-site position on ZFP binding specificity -----	70
4.1	Sample set of eight 9 bp DNA targets -----	84
4.2	Free energy perturbation and docking score data for our sample of 6 GNNGNNGNN target DNA bound to Zif-268 protein sequence -----	89
5.1	DNA Sequences used for training and testing of micro neural network model-----	106
5.2	Accuracy of micro neural network model for both the training and testing datasets (Sequence Identity and BLAST e-value scores) -----	112
5.3	Comparison of ZifNN predictions with other tools reported in literature -----	116
5.4	Substrate specificity and affinity for finger-2 of Zif-268-----	129

LIST OF EQUATIONS

Eq. No.	Equation Title	Page No.
1	Interfacial hydrogen bond energy equation	57
2	OPLS_2005 Intermolecular interaction energy	87
3	Non-linear transformation function of μ NN	109
4	Ansatz – distance based fit function	120
5	Law of triangle equation	122

LIST OF ABBREVIATIONS

ZFP	zinc finger proteins
ZFN	zinc finger nucleases
ns	nanosecond
μ NN	micro neural network
IHBE	interfacial hydrogen bond energy
FEP	free energy of perturbation
MD	molecular dynamic simulations
CoDA	context-dependent assembly
OPEN	oligomerized pool engineering
SVM	support vector machines
RMSD	root mean square deviation
ANN	artificial neural network
BN	Bayesian network