

Decoding the protein-DNA recognition rules

by

Nuri Alpay Temiz

BS, Chemical Engineering, Bogazici University, 1999

MS, Chemical Engineering, Bogazici University, 2001

Submitted to the Graduate Faculty of
School of Medicine in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Nuri Alpay Temiz

It was defended on

December 3, 2008

and approved by

Dr. Philip Auron, Professor, Biology Duquesne University

Dr. Panayiotis Benos, Associate Professor, Computational Biology

Dr. Linda Jen-Jacobson, Professor, Biology

Dr. Daniel Zuckerman, Assistant Professor, Computational Biology

Dissertation Advisor: Dr. Carlos J Camacho, Associate Professor, Computational Biology

Copyright © by Nuri Alpay Temiz

2008

Decoding the protein-DNA recognition rules

Nuri Alpay Temiz, MS

University of Pittsburgh, 2008

Transcription factors (TF) are key proteins involved in gene regulation by binding to specific DNA sites. The C₂H₂ zinc finger (ZF) TFs form the largest family of DNA binding proteins in eukaryotes and are a key participant in the regulation of most genes. A major obstacle towards understanding the molecular basis of transcriptional regulation is the lack of a general recognition code for protein-DNA interactions. In this thesis, we aim to understand molecular mechanisms of DNA binding/recognition by TFs and to quantitatively estimate recognition rules for TF-DNA interactions. To this aim, we first identified key residues that play an important role in ZF-DNA binding and studied their dynamics prior to binding using molecular dynamics (MD) simulations. We found that key residues that are buried upon complexation are prealigned to conformations close to their bound state prior to binding. The bound-like behavior of some of these residues is found to be dependent on the ion concentration of the system, consistent with experimental observations of increased binding affinity with increased ionic strength in protein-DNA interactions. We identified a binding site for Cl⁻ ions located in the same pocket where DNA phosphates are found most buried in the complex structure of ZFs. Bound Cl⁻ ions constrain key side chains in conformations similar to those observed when interacting with the phosphates. These results suggest that ZFs are able to maintain bound like conformations of key residues upon encountering the DNA hinting at a general mechanism to rapidly form encounter complexes amenable for a fast readout of the DNA. Next, we develop a novel experimentally-based approach using high quality crystal structures and binding data on the promiscuous family

of C₂H₂ zinc fingers (ZF) and decode ten fundamental specific interactions responsible for protein-DNA recognition. The interactions include five hydrogen bond types, three atomic desolvation penalties, a favorable non-polar energy, and a novel water accessibility factor. We apply this code to three large data sets containing a total of 89 C₂H₂ TF mutants on the three ZFs of EGR. Guided by MD simulations of individual ZFs, we map the interactions into homology models that embody all feasible intra- and inter- molecular bonds, selecting for each sequence the structure with the lowest free energy. The interactions reproduce the change in affinity of 35 mutants of finger I ($R^2 = 0.99$), as well as two independent validation sets of 23 mutants of finger II ($R^2 = 0.97$) and 31 human ZFs on finger III ($R^2 = 0.95$). More importantly, the method predicts bound ZF-DNA complexes for all 89 mutants, decoding molecular basis of EGR-DNA specificity. Our findings reveal recognition rules that depend on DNA sequence/structure, molecular water at the interface and induced fit of the C₂H₂ TFs. Collectively, our method provides the first robust framework to decode the molecular basis of TFs binding to DNA.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	XIII
1.0 INTRODUCTION.....	1
1.1 PROTEIN – DNA RECOGNITION	1
1.1.1 Protein – DNA interactions	1
1.1.1.1 Cys₂ – His₂ Zinc finger transcription factors	2
1.1.2 Specific versus non-specific binding.....	8
1.2 EFFECTS OF COUNTER IONS IN TRANSCRIPTION FACTOR BINDING.....	10
1.3 MEASUREMENT AND PREDICTION OF BINDING SITES AND AFFINITIES OF TRANSCRIPTION FACTORS	12
1.3.1 Experimental methods to study binding.....	12
1.3.2 Computational methods to study binding	13
1.3.2.1 Structure Based Methods.....	13
1.3.2.2 Sequence based methods.	14
1.3.2.3 Additivity	15
1.4 SPECIFIC AIMS	16
1.4.1 Specific Aim 1. Identification of anchoring residues in C₂H₂ family ZF- DNA interactions.....	16

1.4.2	Specific Aim 2. Counter-ions found in the physiological environment help stabilize the anchor residues by mimicking DNA backbone phosphate groups.	16
1.4.3	Specific Aim 3. To develop an experimentally based approach to decode non-additive interactions of C ₂ H ₂ family zinc finger – DNA complexes.....	17
2.0	METHODS	18
2.1	MOLECULAR DYNAMICS SIMULATIONS	18
2.1.1	Potential energy functions.....	18
2.1.2	Computation of MD trajectories	20
2.1.3	Simulation protocol.....	20
2.1.4	Side chain root mean square deviation (rmsd).....	22
2.2	PREDICTION OF ANCHOR RESIDUES	23
2.3	MODELING ZINC FINGER – DNA INTERACTIONS	23
2.3.1	C ₂ H ₂ Zinc Finger transcription factors.....	24
2.3.2	Decoding protein-DNA interactions	25
2.3.3	Datasets of EGR mutants	27
2.3.4	Binding modes	29
2.3.5	Protein and DNA homology models	30
2.3.6	Computation of waters at the protein – DNA interface	31
3.0	RESULTS AND DISCUSSION	34
3.1	ANCHOR SIDE CHAINS IN PROTEIN-DNA INTERACTIONS	34
3.2	ROLE OF COUNTER-IONS IN BINDING	38
3.2.1	Effect of ion caging in side chain dynamics	38

3.2.2	Buried versus non-buried side chains.	40
3.2.3	Ions bind at the phosphate binding site in the protein-DNA complex structure.....	42
3.2.4	Non-specific contacts.	45
3.2.5	Summary: Counter ions act as surrogates for the backbone phosphate groups at the protein-DNA interface stabilizing the critical side chains.	51
3.3	EXPERIMENTAL BASED CONTACT ENERGIES FOR ZINC FINGER – DNA INTERACTIONS.....	52
3.3.1	Intramolecular hydrogen bonds	54
3.3.2	Recognition code for intermolecular hydrogen bonds	56
3.3.3	Minimal set of protein-DNA interactions	58
3.3.4	Protein-DNA interaction code	60
3.3.5	DNA structure and the role of water in additivity.....	65
3.3.6	Assessing water factor in binding modes	67
3.3.7	Model prediction for FI.....	70
3.3.8	Multiple complex models.....	72
3.3.9	Predicting changes in affinities due to mutations in FII and FIII.....	74
3.3.9.1	Comparison with affinity data from Segal et al. (Segal <i>et al.</i> 1999)	75
3.3.9.2	Comparison with affinity data from Bae et al.(Bae <i>et al.</i> 2003).	77
3.3.10	Comparing inter-molecular networks across different fingers	77
3.3.11	Limitations.....	80
4.0	CONCLUSIONS AND FUTURE DIRECTION	81
4.1	CONCLUSIONS.....	81

4.1.1	Anchoring residues in C ₂ H ₂ family zinc finger- DNA interactions.	81
4.1.2	Counter ions found in the physiological environment help stabilize the non-specific encounter complex by mimicking DNA backbone phosphates.	81
4.1.3	Modeling ZF-DNA interactions.	82
4.2	FUTURE DIRECTIONS.....	85
4.2.1	Effects of counter-ions in DNA recognition.....	85
4.2.2	Modeling protein-DNA interactions.....	85
	BIBLIOGRAPHY.....	87

LIST OF TABLES

Table 2.1: Experimental relative affinity data of protein mutants.	27
Table 2.2: Experimental data from Segal et al. (Segal <i>et al.</i> 1999)	32
Table 2.3: Experimental data from Bae <i>et al.</i> (Bae <i>et al.</i> 2003)	33
Table 3.1: Anchors and base contacts for zinc fingers in EGR, TFIIA and GLI	35
Table 3.2: Solvent accessible surface area (SASA) buried by phosphates.	47
Table 3.3: Residues contacting the phosphate backbone in EGR, TFIIA and GLI.....	50
Table 3.4: Look up table for amino acid – DNA hydrogen bonds.	57
Table 3.5: Optimized effective hydrogen bond potentials and desolvation penalties (kcal/mol). 63	

LIST OF FIGURES

Figure 1.1: EGR-DNA complex.	4
Figure 1.2: Cartoon representation of TFIIIA (PDB code 1TF6) bound to DNA.	6
Figure 1.3: Cartoon representation of GLI (PDB code 2GLI) bound to DNA.	7
Figure 1.4: Thermodynamic cycle of protein–DNA association under condition N.	9
Figure 2.1: Illustration of EGR FI bound to its consensus site.	28
Figure 2.2: Crystal structures of binding modes and induced fit on zinc fingers	30
Figure 3.1: RMSD profile of binding site residues of EGR at 35mM counter ion concentration.	36
Figure 3.2: Ion caging: Bound-like behavior of EGR anchor R ₊₆ in FI as a function of ion concentration.	39
Figure 3.3: Buried and exposed key side chains of EGR.	41
Figure 3.4: Ion dependence of the bound-like behavior of key side chains in EGR fingers I and III.	42
Figure 3.5: Relation between weakly bound ion and Arg ₊₆ dynamics.	44
Figure 3.6: Ions occupy the phosphate binding site in the protein–DNA complex.	46
Figure 3.7: Dynamics of phosphate contacting conserved lysines in the linker regions between fingers.	49
Figure 3.8: Distance profiles for key contacts in FI mutants and their complexes over 8 ns of MD runs	55

Figure 3.9: Sketches illustrating atom desolvation penalties and solvation effects at the protein (top)-DNA (bottom) binding interface.....	60
Figure 3.10: Predicted complex structures for 6 EGR FI mutants and 6 DNA binding site sequences.	64
Figure 3.11: Extraction of the ten fundamental interaction parameters.....	66
Figure 3.12: Rearrangement of waters at the protein-DNA interface due to cytosine to adenine mutation.	67
Figure 3.13: Models for 4 EGR FI and 3 DNA binding site sequences.	69
Figure 3.14: Predicted $\Delta\Delta G_{\text{Calc}}$ versus experimental $\Delta\Delta G_{\text{Exp}}$ changes in free energy due to protein and/or DNA mutations.....	73
Figure 3.15: Predicted complex structures for FII mutants.	76
Figure 3.16: Predicted $\Delta\Delta G_{\text{Calc}}$ versus experimental $\Delta\Delta G_{\text{Exp}}$ changes in free energy due to protein and/or DNA mutations.....	78
Figure 3.17: Predicted complex structures for FIII experiments.	79

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Carlos Camacho, for accepting me into his lab, his guidance, support and friendship throughout my studies. To my thesis committee members, Dr. Philip Auron, Dr. Takis Benos, Dr. Linda Jen-Jacobson and Dr. Daniel Zuckerman I extend my thanks for their help and constructive criticism. I would like to acknowledge members of the Camacho lab, especially Dr. Marta Bueno, Lidio Meireles and Wei Zhang and thank them for their assistance and support.

I would like to acknowledge and thank my friends Dr. Kadir Diri, Dr. Basak Isin and Dr. Chakra Chennubhotla for their constant friendship and support throughout my graduate studies.

I would like to thank my family in Turkey for giving me this opportunity, my mother, Beyhan Temiz, my brother, Gokay Temiz, I acknowledge my late father Osman Zati Temiz and late uncle Turhan Temiz and their continuing influence and lasting inspiration. I would also like to thank my family here in Pittsburgh, my fiancée and soon-to-be wife, Victoria Sieffert, her parents Mr. and Mrs. Raymond Sieffert and her sister and brother in law, Emily Clough and Dan Clough for their love and encouragement. Finally, I dedicate this work to my family.

1.0 INTRODUCTION

Gene regulation is primarily controlled by the interactions of protein complexes with DNA. Transcription factors (TFs) are key proteins involved in this regulation by binding to specific DNA sites. Understanding the structure and stability of protein-DNA complexes is a fundamental goal in structural biology. In addition, revealing the molecular basis of transcriptional regulation is critical to understand how genes are activated/repressed leading to normal cell function or to the acquisition of specific pathogenic traits.

1.1 PROTEIN – DNA RECOGNITION

1.1.1 Protein – DNA interactions

DNA binding proteins utilize a large variety of structural motifs such as, helix-turn-helix (cro repressor, mat- α 2), zinc coordinating motifs (zif268 zinc finger (ZF)), zipper type (leucine zipper, GCN4; helix-loop-helix, MyoD), β sheets (TATA box binding protein (TBP)) and enzymes (Endonuclease EcoRI and EcoRV) (Luscombe *et al.* 2001). The proteins usually utilize an α -helix or a β -sheet that protrudes towards the DNA grooves. Most of the protein – DNA complex structures contain DNA that is generally in B-form with a moderate degree of bending and deformation. In some cases, however, the DNA is significantly deformed. The most

noticeable example is the TBP – DNA complex, where the DNA is significantly bent. Other examples with DNA kinking and bending include CAP, lac repressor, EcoRV DNA complexes (Garvie and Wolberger 2001).

ZF proteins, on the other hand, bind to the major groove of DNA with little or no DNA deformation in a modular fashion. Cys₂ – His₂ (C₂H₂) family of ZF TFs are one of the most abundant DNA – binding motifs found in eukaryotes. The first such ZF identified was transcription factor IIIA (TFIIIA) from *Xenopus laevis* (Miller *et al.* 1985). ZFs are mostly involved in DNA binding and can regulate transcription of specific genes or can be part of the general transcriptional machinery (Patikoglou and Burley 1997; Wolfe *et al.* 2000; Klug 2005). Although, ZFs are mainly DNA binding proteins, they can also function as RNA binding proteins (Brown 2005) and are involved in protein – protein interactions (Gamsjaeger *et al.* 2007).

In recent years, modular and high affinity binding nature of ZFs lead to their extensive use in gene therapy and biomedical applications (Kim and Pabo 1997; Kim and Pabo 1998; Segal *et al.* 1999; Joung *et al.* 2000; Pabo *et al.* 2001; Jamieson *et al.* 2003; Bae *et al.* 2003; Magnenat *et al.* 2004; Klug 2005; Cathomen and Keith Joung 2008). In addition, the high number of available X-ray crystal structures and experimental affinity data make ZFs an ideal system to study protein-DNA recognition.

1.1.1.1 Cys₂ – His₂ Zinc finger transcription factors

The classical C₂H₂ family contains two or more ZF modular domains that work together to recognize specific DNA sequences. Individual fingers contain about 30 amino acids. The classical C₂H₂ ZF domain is composed of a ββ α fold with a hydrophobic core flanking the zinc

binding site. The zinc ion binds between the β sheet and α helix. The ZF fold is held together by a tetrahedrally coordinated zinc ion cross-linking the α -helix and the antiparallel β sheet (Pavletich and Pabo 1991; ElrodErickson *et al.* 1996). In the absence of the bound zinc ion the fingers become unfolded (Frankel *et al.* 1987).

C_2H_2 ZFs usually bind to DNA through more than one finger arranged in sequence (Figure 1.1). The N terminal half of the α -helix of each finger fits into the major groove of the DNA contacting three bases usually along one strand of DNA. The major base contacts occur through key residues at the N-terminal end of the α -helix commonly through positions -1,+2,+3 and +6 with respect to the start of the α -helix. Binding of the individual fingers leads the protein to wrap around DNA.

Mouse early growth response (EGR) factor

EGR (also called zif268) was the first ZF to be crystallized in complex with DNA (Pavletich and Pabo 1991). EGR (ElrodErickson *et al.* 1996) (Protein Data Bank (PDB) code 1AAY) protein has three zinc fingers (Figure 1.1A). The α -helices of each finger fit into the major groove of DNA, making specific contacts with DNA bases. FI binds to a GCG triplet near the 3' end of the primary DNA strand (Figure 1.1B). FII binds to a TGG triplet in the center and FIII binds to the GCG triplet near the 5' end of the primary DNA strand (Figure 1.1). The helical domains of FI and FIII have the same sequence and identical bound structure with DNA. Their critical residues are an arginine preceding the α -helix (R_{-1}), an aspartic acid on the second position of the α -helix (D_{+2}), a glutamine on the third position (E_{+3}) and an arginine on the sixth position of the α -helix (R_{+6}). FII has also an arginine immediately before the helix and an

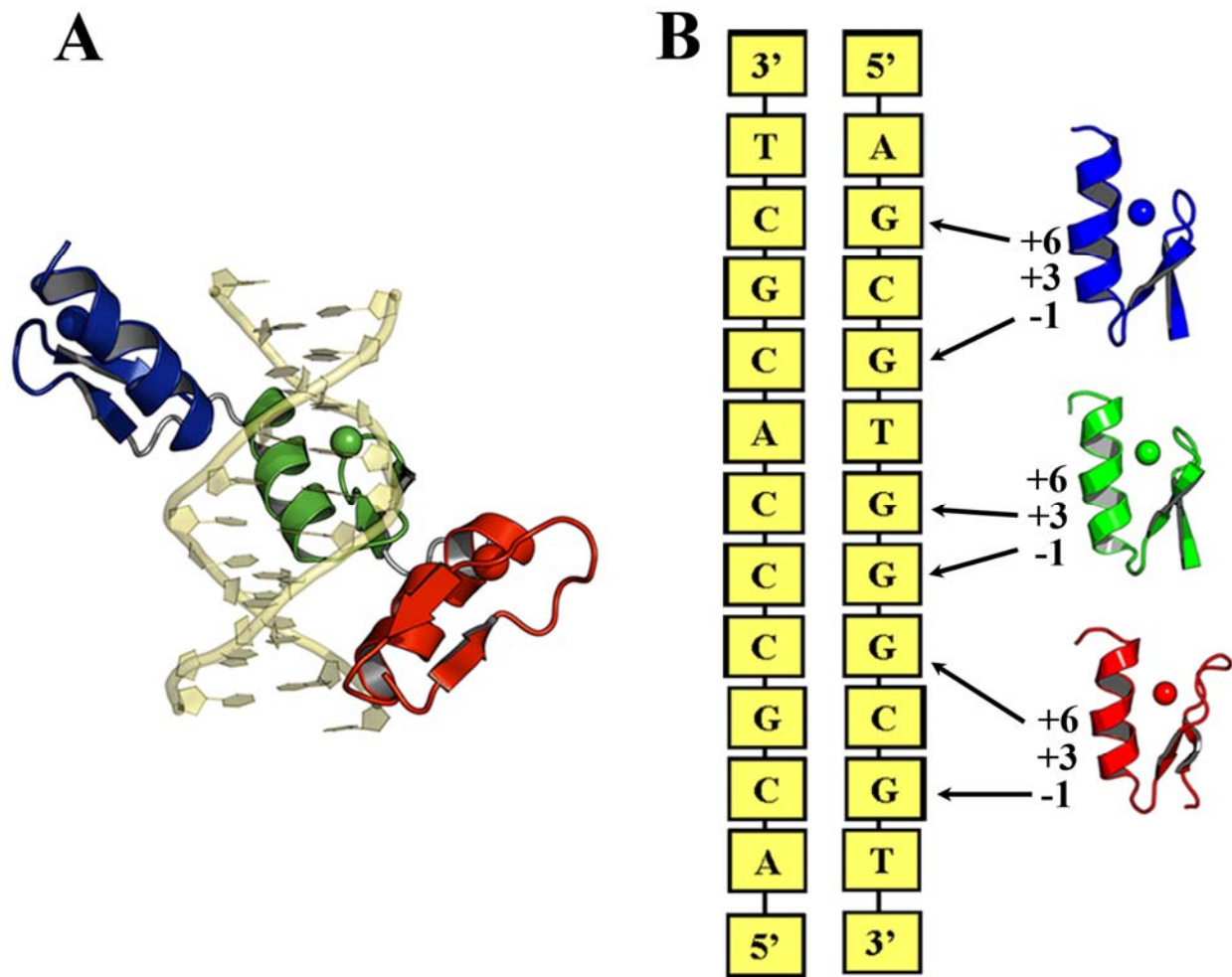


Figure 1.1: EGR-DNA complex.

A Cartoon of EGR bound to DNA (PDB code: 1AAY (ElrodErickson *et al.* 1996)). Fingers I, II and III are color red, green and blue, respectively. EGR wraps around the DNA in an anti-parallel fashion. i.e. FI binds to the 3' end of the main strand and FIII binds to the 5' end. **B** Schematics diagram showing the key side chain – base interactions between EGR and its target DNA site.

aspartic acid on position 2 of the helix. But it has a histidine on the third position of the helix (H_{+3}) and a threonine on the sixth position of the helix (T_{+6}) instead of an arginine.

Arginine residues in all three fingers make a pair of hydrogen bonds (H-bonds) with a guanine (see Figure 1.1). EGR also has several contacts with the DNA backbone. In particular, the first histidine (H_{+7}) coordinating the zinc ion forms an H-bond with a DNA sugar backbone phosphate oxygen on the primary strand. This contact is 80 % conserved (Wolfe *et al.* 2000).

Another conserved (60 %) arginine on the second β strand in each finger also contacts a DNA sugar backbone phosphate on the primary strand (Wolfe *et al.* 2000).

***Xenopus laevis* transcription factor IIIA (TFIIIA)**

Although the *Xenopus laevis* TFIIIA contains 9 fingers, the first 6 fingers are solved in the crystal structure (Nolte *et al.* 1998) (PDB code 1TF6). Figure 1.2 displays a cartoon representation of TFIIIA-DNA complex. Fingers I-III wrap around the major groove of DNA in a manner similar to the binding of EGR forming contacts with the non-coding strand of DNA. Fingers IV-VI form an open structure. Only FV have contacts with DNA in the major groove (see Figure 1.2 upper 3 fingers). Fingers I, II, III and V contact DNA through helical positions -1, +2, +3 and +6. In addition, the conserved His₊₇ – phosphate contact, similar to the sugar backbone phosphate contacts of EGR, is observed in fingers II and V. TFIIIA FIII has an additional R-G bidentate interaction at position +10 with respect to the α -helix (Nolte *et al.* 1998).

Human glioblastoma (GLI) factor

GLI (Figure 1.3) has 5 fingers in the crystal structure (Pavletich and Pabo 1993) (PDB code 2GLI). FI does not make any DNA contacts but makes extensive protein-protein interactions with finger two (Pavletich and Pabo 1993). FII has one base contact through Y₊₁ and 3 backbone phosphate contacts including H₊₇. FIII has no base contacts but it has three backbone contacts through positions -1, +5 and the packed tyrosine. It is important to note that FIII does not have the conserved backbone contact through the first zinc coordinating histidine, H₊₇. FIV has extensive base contacts through positions +1,+2,+3 and +6. In addition, it has DNA backbone contacts from positions +4,+7 and +11. FV also forms extensive contacts with the DNA through positions -2,+1,+2, +5. FV like FIII does not have the conserved H₊₇ backbone contact.

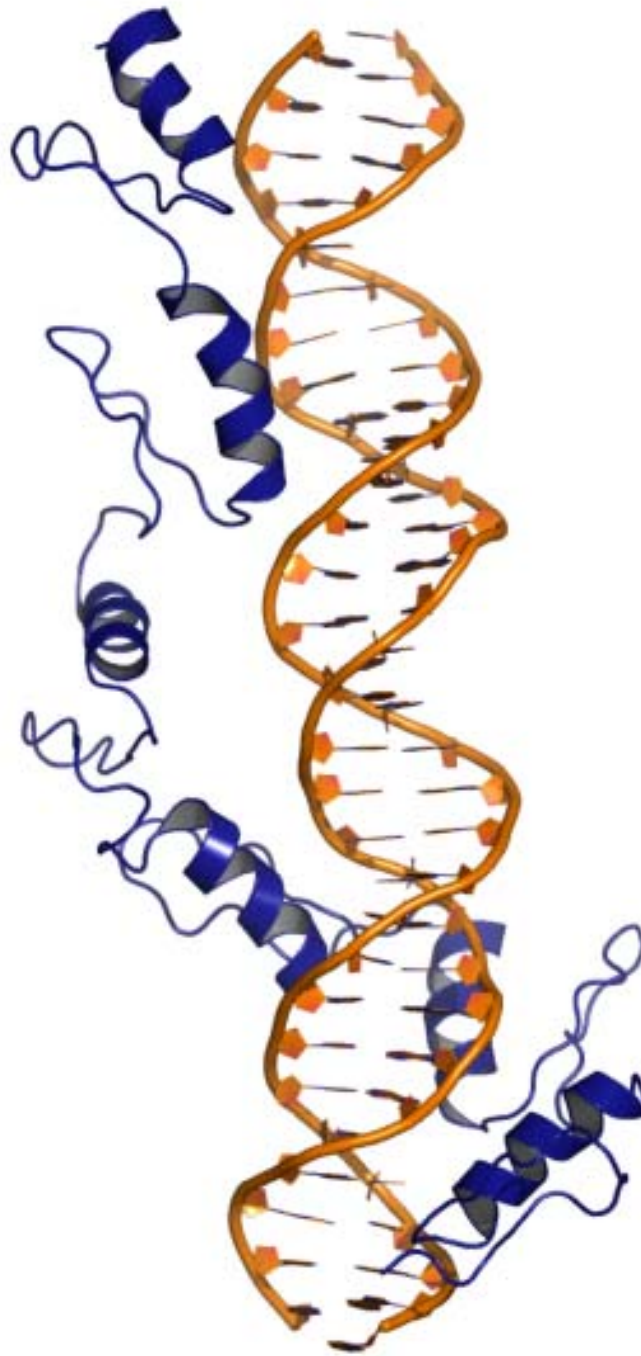


Figure 1.2: Cartoon representation of TFIIIA (PDB code 1TF6) bound to DNA. ZF binds to DNA in an anti-parallel fashion. ZF is shown in blue and DNA in orange. Fingers I, II, III and V contact DNA.

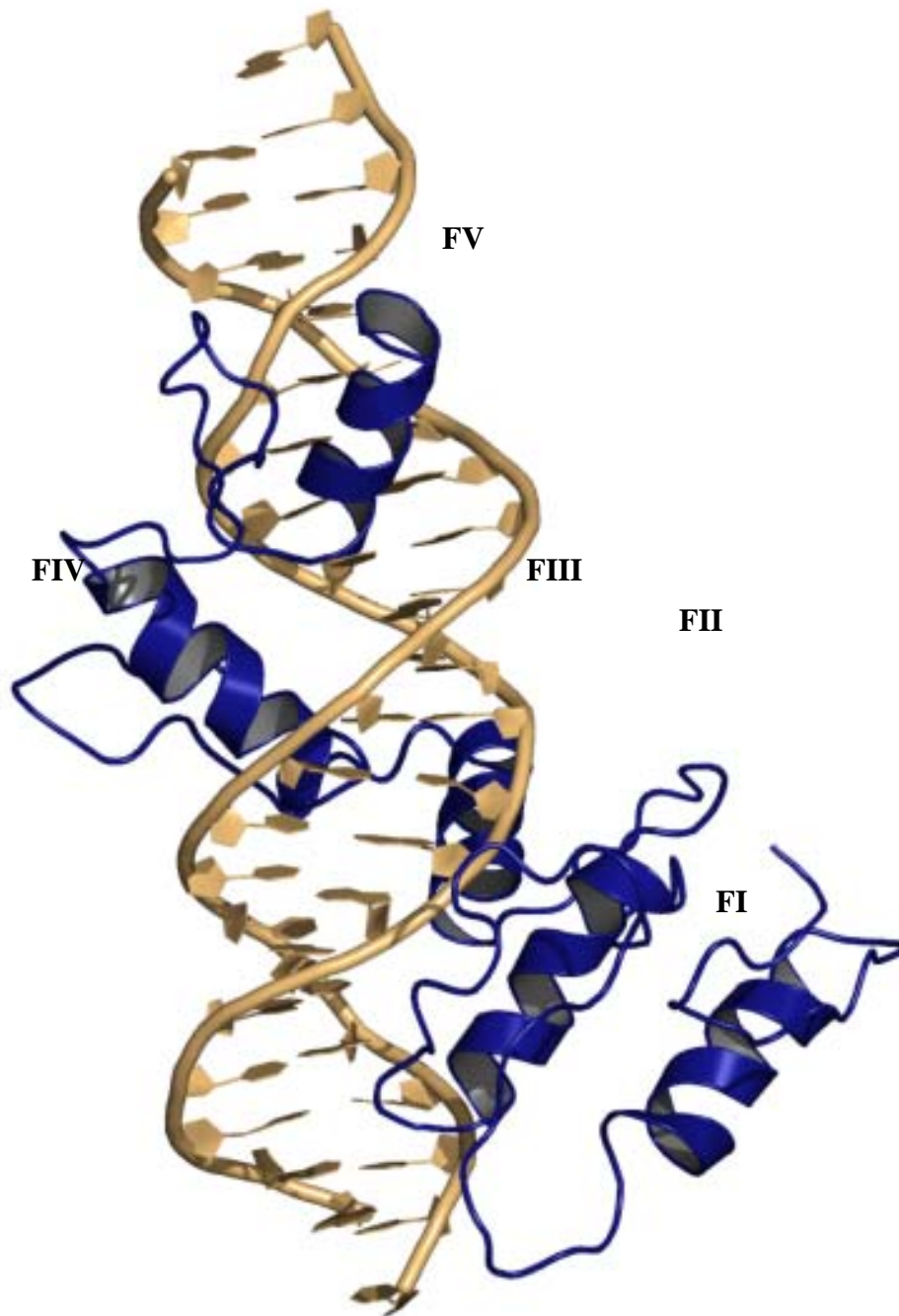


Figure 1.3: Cartoon representation of GLI (PDB code 2GLI) bound to DNA. ZF is shown in blue and DNA is shown in light orange. Fingers II-V contact DNA. FI has protein – protein interactions with FII.

1.1.2 Specific versus non-specific binding.

Different DNA binding proteins have different sequence selectivities for their binding sites. Some proteins such as transcriptional repressors and/or activators have high specificities for their target binding sites. Others such as histones and helicases bind DNA in a non-specific way to perform their respective functions.

The specificity of a protein to DNA could be defined as the ratio of its binding affinity to a specific target site and its affinity to a non-specific site. Although the description of non-specific sites are difficult, can change from protein to protein and experimental conditions (Jen-Jacobson 1997), high affinity DNA binding proteins such as ZFs usually display specificities greater than 25000 fold (Greisman and Pabo 1997; Wolfe *et al.* 1999). Specific residue – base H-bond interactions forming a network of H-bonds at the binding sites was considered enough to achieve specificity first suggested by Seeman *et al.* (Seeman *et al.* 1976) more than 30 years ago. It is now clear that in addition to the specific residue – base H-bonds, protein-DNA sugar phosphate backbone contacts, physiological environment, DNA bending and electrostatic interactions also play important roles in DNA recognition (Anderson and Record, Jr. 1995; Jayaram *et al.* 2002; Norberg 2003; Lavery 2005).

Figure 1.4 illustrates a typical thermodynamic cycle of site specific protein – DNA association under the influence of a given environmental condition. This enables one to study the effects of environmental variables such as temperature, pH, and counter-ions on specific DNA binding. Several studies have measured free energies of binding as well as the individual enthalpic and entropic contributions to the binding free energy, and changes in specific heat capacity (Record, Jr. *et al.* 1978; Jen-Jacobson *et al.* 1983; Anderson and Record, Jr. 1995; Jen-

Jacobson *et al.* 2000; Holbrook *et al.* 2001; Dragan *et al.* 2003; Dragan *et al.* 2004; Bujalowski 2006; Kozlov and Lohman 2006) under different conditions.

In this study, our approach to study molecular mechanisms of TF – DNA recognition is two-fold: one is mechanistic or phenomenological whose goal is to understand how and why TFs bind both specifically and non-specifically, another is quantitative in nature and attempts to reveal the code of protein-DNA specificity. In the next sections, we briefly introduce the effects of counter-ions in DNA binding, and then introduce the current experimental and computational approaches used to study binding affinities and to identify binding sites for ZF TFs.

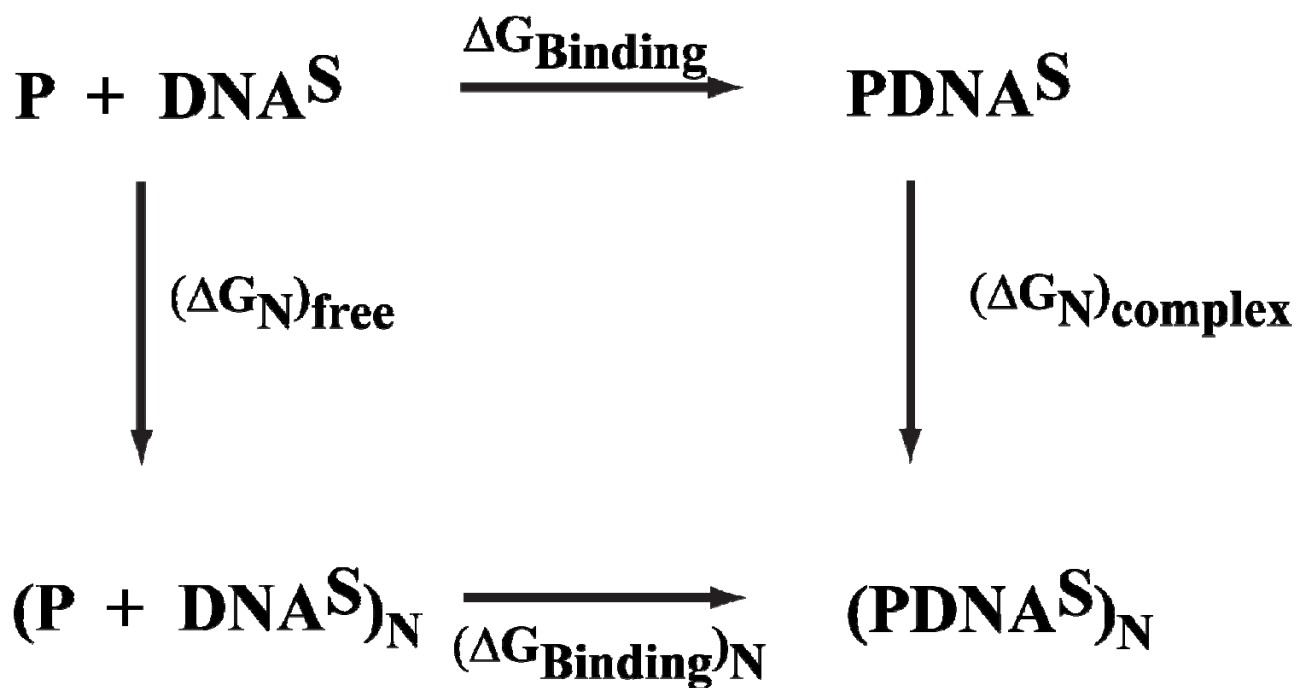


Figure 1.4: Thermodynamic cycle of protein–DNA association under condition N. P denotes the protein. S denotes the specific DNA site.

1.2 EFFECTS OF COUNTER IONS IN TRANSCRIPTION FACTOR BINDING

Despite the emergence of high quality structures in recent years, the molecular basis of the events *leading* to protein-DNA recognition and binding specificity is not yet totally understood. When DNA is involved, part of the problem is that interactions are dominated by charged and polar groups that are highly dependent on the solvent and ionic environment (Record, Jr. *et al.* 1976; Winter and von Hippel 1981; Winter *et al.* 1981; Jen-Jacobson *et al.* 1983; Ebricht *et al.* 1989; Fried and Stickle 1993; Anderson and Record, Jr. 1995; Engler *et al.* 1997; Hamilton *et al.* 1998). Positively charged counter ions associate with the negatively charged phosphate groups of nucleic acids, thus maintaining neutrality in solution. Theoretical studies of protein-DNA complexes that concentrate on the effects of counter-ions have mostly used two approaches, counter-ion condensation (CC) (Manning 1977; Record, Jr. *et al.* 1978) and Poisson-Boltzmann (PB) (Sharp and Honig 1990; Sharp and Honig 1995) theories. The main difference between these two approaches lies in the description of salt effects around the nucleic acid. CC theory considers two distinct layers of counter-ion concentration, one salt independent uniform layer around the DNA and a distant salt dependent classical ion atmosphere. PB theory, on the other hand, describes the ionic environment as a single continuous atmosphere.

In protein-protein interactions, the role of ions is described by the screening of long range electrostatic interactions consistent with classical Debye-Huckel theory (Debye and Huckel 1923). Indeed, ionic strength has been shown to tune the association rate of some highly optimized receptor-ligand systems by as much as five orders of magnitude (Schreiber and Fersht 1996). Generally, the larger the ionic strength the smaller the binding affinity. Strikingly, Jen-Jacobson and collaborators (Jen-Jacobson *et al.* 1983; Engler *et al.* 1997) and others (Fried and

Stickle 1993; Kozlov and Lohman 1998; O'Brien *et al.* 1998) have shown that the latter is not necessarily true for protein-DNA interactions, where in several instances it has been shown that *the binding affinity increases with ionic strength* (Winter and von Hippel 1981; Winter *et al.* 1981; Ebright *et al.* 1989; Romaniuk 1990; Hamilton *et al.* 1998; Moraitis *et al.* 2001; Dragan *et al.* 2006) even when the experimental conditions, the underlying effects and individual thermodynamic parameters are different. Although grouping all of the above studies with one observation (effect of the counter-ions) can be simplistic and misleading, the observation still leads to the question whether there is a common underlying role played by the negatively charged counter-ions effecting the dynamics and mechanistics of specific and non-specific binding at the molecular level.

A particularly interesting example of the role of salt in protein-DNA interactions is catabolite gene activator protein (CAP) binding to the lac promoter region. For this system, Fried and Stickle showed that for physiological relevant concentrations between 0.05-and-0.2 M there is a 5-fold *increase* in binding affinity (Fried and Stickle 1993). Similar behavior has been observed in *Escherichia Coli* lac repressor-operator complex (Winter and von Hippel 1981; Winter *et al.* 1981), EcoRI (Jen-Jacobson *et al.* 1983), EcoRV (Engler *et al.* 1997) in the low 0-0.1 M range. This phenomenological picture is able to fit the changes in binding affinity as a function of ionic strength, but no molecular/structural mechanism is revealed. Here, we postulate that counter-ions weakly interacting with the protein, prior to binding, act as surrogates for the DNA sugar backbone phosphate group and restrain critical side chains in conformations similar to those found in the complex, allowing the formation of a non-specific protein-DNA encounter complex.

1.3 MEASUREMENT AND PREDICTION OF BINDING SITES AND AFFINITIES OF TRANSCRIPTION FACTORS

1.3.1 Experimental methods to study binding

Experimental approaches to study TF binding sites are necessary to determine TFs biological functions, their interactions, and impact on gene expression and regulation. In recent years, experimental methods such as SELEX (Tuerk and Gold 1990), phage display (Rebar and Pabo 1994), yeast one-hybrid (Y1H) (Deplancke *et al.* 2004), genome wide location analysis (ChIP-chip) (Ren *et al.* 2000), DNA binding specificity by DNA immunoprecipitation with microarrays (DIP-chip) (Liu *et al.* 2005a) and protein binding microarrays (PBM) (Bulyk *et al.* 1999) are increasingly used to identify binding sites and determine specificities of individual TFs. The SELEX method is used for in vitro selection of optimal binding sites of a TF (Oliphant *et al.* 1989). ChIP-chip method is used for in vivo identification of targets for a given TF (Ren *et al.* 2000). DIP-chip (Liu *et al.* 2005a) and PBM (Bulyk *et al.* 1999) methods use DNA microarrays to identify TF binding sites. Binding data of this type usually do not give any structural information or mechanism of DNA binding. Quantitative methods such as electrophoretic mobility shift assay (EMSA), DNA footprinting and quantitative multiple fluorescence relative affinity assay (QuMFRA) (Man and Stormo 2001) are usually used to determine the association constant of binding (K_a) to DNA.

1.3.2 Computational methods to study binding

The emergence of specific protein-DNA complex structures in recent years has been instrumental in our understanding of how proteins recognize specific DNA sequences (Patikoglou and Burley 1997). Molecular dynamics (MD) simulations of protein-DNA complexes have provided insights on the dynamics of the interactions between the protein and DNA and the role of water in the complex interface (Roxstrom *et al.* 1998; Zakrzewska and Lavery 1999; Roxstrom *et al.* 2000; Tsui *et al.* 2000; Beveridge *et al.* 2004; Cheatham, III 2004). In addition to MD simulations of protein-DNA complexes (Roxstrom *et al.* 1998; Zakrzewska and Lavery 1999; Roxstrom *et al.* 2000; Tsui *et al.* 2000; Beveridge *et al.* 2004; Cheatham, III 2004), computational approaches to identify TF binding sites and study their specificities have recently been developed. These computational approaches can be divided into probabilistic and structure based methods.

1.3.2.1 Structure Based Methods

Structure based approaches of protein – DNA interactions require an experimentally determined structure or a homology model of the complex (Mandel-Gutfreund *et al.* 1995; Mandel-Gutfreund and Margalit 1998; Kono and Sarai 1999; Selvaraj *et al.* 2002; Roven and Bussemaker 2003; Endres *et al.* 2004; Paillard *et al.* 2004; Havranek *et al.* 2004; Man *et al.* 2004; Liu and Stormo 2005; Gromiha *et al.* 2005; Kaplan *et al.* 2005; Morozov *et al.* 2005; Zhang *et al.* 2005; Siggers and Honig 2007). Some of the structure based methods use a scoring function that has been developed by statistical analysis of experimentally solved protein-DNA complexes (Mandel-Gutfreund and Margalit 1998; Kono and Sarai 1999; Gromiha *et al.* 2005; Zhang *et al.* 2005; Liu *et al.* 2005b; Contreras-Moreira and Collado-Vides 2006). These

knowledge based approaches usually estimate pair wise amino acid – nucleic acid interaction energies and then predict binding sites for TFs. Other structure based approaches use molecular mechanics type potentials to predict binding specificities in protein – DNA complexes (Havranek *et al.* 2004; Morozov *et al.* 2005; Siggers and Honig 2007). For instance Morozov *et al.* (Morozov *et al.* 2005) reported $\Delta\Delta G$ predictions for FI of EGR ZF, with a correlation coefficient of 0.59. The authors used a complex energy function with van der Waals, generalized Born electrostatics, and DNA deformation terms to calculate free energies. Paillard *et al.* (Paillard *et al.* 2004) used ADAPT (Lafontaine and Lavery 2000) to calculate the free energies for FII mutants of EGR ZF. They constructed ZF-DNA complex structures based on the wild type X-ray structure (ElrodErickson *et al.* 1996) while keeping the protein backbone fixed and calculated protein–DNA pair wise contact energies based on a cut-off distance. The calculated energies show a good correlation but they are an order of magnitude higher than the experimental free energies.

1.3.2.2 Sequence based methods.

Sequence based methods are generally used to identify TF binding sites in the regulatory regions of genes. These binding sites are mostly represented by DNA motifs containing a short sequence of DNA. Probabilistic models of TF binding are typically used in the search or characterization of DNA motifs resulting in position weight matrices also called position specific scoring matrices. These matrices correspond to probabilities reflecting the likelihood of observing a specific base at a specific position (Stormo 2000; Benos *et al.* 2002a; Bulyk 2003; Siggia 2005; GuhaThakurta 2006; Bussemaker *et al.* 2007; Mahony *et al.* 2007). Binding experiments or databases such as TRANSFAC (Wingender 2008) or JASPAR (Bryne *et al.* 2008) can provide the experimentally determined position weight matrices.

1.3.2.3 Additivity

Most of the computational methods assume that the interactions between the TF and DNA are independent of each other such that total energy of binding is the sum of the energies of the individual contacts. These studies do not take into account the cooperative binding resulting from residues interacting with each other or indirect binding due to structural properties and/or deformation of DNA.

Recent studies (Benos *et al.* 2002b; Man *et al.* 2004; Liu and Stormo 2005) show that the so-called additivity assumption does not hold for protein mutations. Two recent studies (Benos *et al.* 2002b; O'Flanagan *et al.* 2005) have also addressed the additivity assumption from the DNA side. Benos *et al.* used statistical approaches to study additivity (Benos *et al.* 2002b). They studied Mnt repressor (Man and Stormo 2001) and EGR zinc finger (ZF) TF (Benos *et al.* 2002b) and concluded that the additivity assumption does not fit their experimental data but provides a good approximation in at least half of the cases. O'Flanagan *et al.* (O'Flanagan *et al.* 2005) focused on the sequence dependent flexibility of DNA deformation and used a molecular mechanics approach to study the non-additive effects of protein – DNA recognition. They studied TATA box binding protein – DNA complex and concluded that non-additive effects on the DNA site involve only dinucleotide steps.

1.4 SPECIFIC AIMS

1.4.1 Specific Aim 1. Identification of anchoring residues in C₂H₂ family ZF-DNA interactions.

This aim will test the hypothesis that buried (“anchor”) side chains in the C₂H₂ family zinc fingers are pre-oriented to their native conformation seen on the DNA-zinc finger complex prior to binding. MD simulations will be performed for each finger in explicit solvent in the absence of DNA to analyze the dynamics of critical side chains and establish whether conformations are conducive to or frustrating the formation of the complex. The role of non-specific interactions (mostly with the backbone groups) will also be analyzed, as well as the effect of the conserved linker domain.

1.4.2 Specific Aim 2. Counter-ions found in the physiological environment help stabilize the anchor residues by mimicking DNA backbone phosphate groups.

The nucleus is a highly ionic environment. The aim here is to study the role that negative charged counter ions have in the dynamics of zinc finger domains and in binding DNA. In particular, the hypothesis that ions help stabilize the anchor residues in C₂H₂ family zinc fingers prior to DNA-binding, as well as a binding mechanism where bound ions act as surrogates for DNA phosphate groups will be analyzed. Extensive MD simulations with different ionic concentrations in the absence of DNA will be performed and analyzed. These studies will provide insights on the molecular mechanism of zinc finger – DNA binding.

1.4.3 Specific Aim 3. To develop an experimentally based approach to decode non-additive interactions of C₂H₂ family zinc finger – DNA complexes.

To this aim, we will develop a novel experimentally-based approach to accurately estimate intermolecular contact free energies of hydrogen bonds and atom desolvation penalties to predict the structure and affinity of C₂H₂ ZF – DNA complexes. Homology models of mutant proteins and DNA sites will be built and the resulting protein mutants will be simulated using MD simulations in explicit solvent in the absence of DNA for systems for which there are available experimental data. The method will involve a two stage process where we first map feasible intra- and intermolecular interactions and then optimize the energy parameters according to the complex structure resulting with the lowest interaction free energy.

2.0 METHODS

2.1 MOLECULAR DYNAMICS SIMULATIONS

Molecular dynamics simulations generally use simple classical mechanics and approximate the motion of the atoms using Newton's equation of motion.

$$\mathbf{F}_i = m_i \mathbf{a}_i \quad (2.1)$$

Where \mathbf{F}_i is the force acting on atom i , m_i is the mass of the atom i and \mathbf{a}_i is its acceleration. The force \mathbf{F}_i is determined by the gradient of the potential energy function $E(\mathbf{x})$ of all coordinates. (Leach 1999)

2.1.1 Potential energy functions

A potential energy function allows for the potential energy, $E(\mathbf{x})$, of a system to be calculated as a function of its coordinates. The general formula of a potential function is

$$E = E_b + E_{ba} + E_{\text{tor}} + E_{\text{LJ}} + E_{\text{Coul}} \quad (2.2)$$

where E_b is potential for the bonded interactions, E_{ba} is the potential for bond angles, E_{tor} is the potential for torsional angles and E_{LJ} is the Lennard-Jones potential and E_{Coul} is the potential of electrostatic interactions (Schlick 2002).

The general form of the individual components usually are

$$E_b = \sum_{i,j \in S_B} S_{ij} (r_{ij} - \bar{r}_{ij})^2 \quad (2.3)$$

$$E_{ba} = \sum_{i,j,k \in S_{BA}} K_{ijk} (\cos \theta_{ijk} - \bar{\cos} \theta_{ijk})^2$$

$$E_{tor} = \sum_{i,j,k,l \in S_{Tor}} \sum_n \left(\frac{V_{nijkl}}{2} [1 \pm \cos(n\tau_{ijkl})] \right)$$

$$E_{LJ} = \sum_{i,j \in S_{NB}} \left(\frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \right)$$

$$E_{Coul} = \sum_{i,j \in S_{NB}} \left(\frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \right)$$

where S_B , S_{BA} , S_{Tor} are the sets of all bonds angles and torsions in the system and S_{NB} is the set of non-bonded interactions. Major force fields used in MD simulations are GROMOS96 (Scott *et al.* 1999), AMBER (Cornell *et al.* 1995) and CHARMM (MacKerell *et al.* 1998). Today's force fields generally use the potential function definition described above but mostly differ in the parameter optimization.

2.1.2 Computation of MD trajectories

The molecular dynamics simulates the motion of a system governed by a special force field by following molecular configurations in time according to Newton's equation of motion. One of the simplest integrators for biomolecular simulation is the Verlet algorithm (Schlick 2002). In the Verlet algorithm, the equation of motion is written as

$$M \ddot{X}(t) = F(X(t)) \quad (2.4)$$

and the force is the gradient of the potential calculated by

$$F(X(t)) = -\nabla E(X(t)) \quad (2.5)$$

Then, the Verlet algorithm is written as

$$V^{n+1} = V^n + \Delta t M^{-1} F(X(t)) \quad (2.6)$$

$$X^{n+1} = 2X^n - X^{n-1} + \Delta t^2 M^{-1} F(X(t))$$

$$V^n = (X^{n+1} - X^{n-1}) / 2\Delta t$$

Where Δt is the time step, V^n is the velocity and X^n is the position at time n (Schlick 2002).

2.1.3 Simulation protocol

Molecular dynamics simulations were performed using the MD simulation package GROMACS 3.3.1 (Van der Spoel *et al.* 2005). Individual fingers of EGR ZF TF Zif268, TFIIA, GLI and homology models of mutants of EGR FI are simulated. In all simulations, based on neutral pH

conditions, basic Arg and Lys residues were positively charged, and acidic Asp and Glu residues were negatively charged. Histidine residues coordinating the zinc ion were neutral with the hydrogen atom on the N δ atom of the His side chain, since the electronegative N ϵ atom in these histidine residues interact with the zinc ion. Each individual finger was centered in a rhombic dodecahedron box with a 15 Å minimum distance from the protein surface to the box edges. The system was solvated with simple point charge water molecules giving about 4600 waters. Then, each system was minimized by using steepest descent method with GROMOS96 (Scott *et al.* 1999) force field. In the counter-ion simulations, desired numbers of ions were added by replacing water molecules randomly with a minimum distance of 6 Å between the ions and the protein. In all other simulations, only enough ions to neutralize the systems were added using the same procedure. The metal ions in metallo-proteins have been simulated using bonded and non-bonded models (Banci 2003), or more recently using a locally polarizable model (Sakharov and Lim 2005). Since the zinc ion in ZF proteins has a structural role (Frankel *et al.* 1987), the zinc ion and the zinc coordinating residues are harmonically constrained to keep the tetrahedral coordination using a force constant of 2.4 kCal/mol/Å². This constraint is also consistent with recent NMR solution structures of various C2-H2 ZFs which show that the ZF is highly stable (Lee *et al.* 1989; Omichinski *et al.* 1990; Lu and Klug 2007), changing little from the unbound to the bound structure. Hence, we have also harmonically constrained the N and C atoms of each protein. The temperature in each simulation was coupled to a bath of 300K with a coupling time constant of 0.1 ps. The pressure was coupled to 1 Bar using 0.5 ps time constant and 4.5 10⁻⁵ Bar⁻¹ compressibility. A cut-off radius of 10 Å was used in the simulations for non-bonded interactions. Initial velocities were generated randomly from a Maxwell distribution at 300K. A 2 fs time step was used in the simulations. 3 or more independent simulations were performed

for each individual ZF. We performed 5 ns and 9 ns long runs for all fingers. 4 or more independent runs were performed of reach finger with at least 16 ns of aggregate simulation time. Independent runs were started with different initial velocities. The counter-ion simulations started with different counter-ion concentrations. Mutation and homology model simulations only included the necessary number of counter ions to neutralize each simulation system. Coordinates were saved every picosecond. Initial equilibration is set to 1 ns followed by 4 or 8 ns of production runs. Aggregate simulation time was 0.94 μ s for the counter-ion simulations and 0.1 μ s for the simulations of FI mutants.

2.1.4 Side chain root mean square deviation (rmsd)

The side-chain dynamics are analyzed by extracting snapshots from each MD trajectory. The snapshots are overlapped with the bound crystal structure of each finger using the α -helix C^α atoms. The α -carbons of each side-chain are further translated to coincide with the α -carbon of the side-chain in the crystal structure. The RMSDs are calculated with respect to the crystal structure using the side-chain heavy atoms starting from C^β atoms. Large side-chains like, arginine, histidine, tyrosine, lysine and tryptophan are considered bound-like if the RMSD is under 2 \AA . 2 \AA is chosen for the large residues because it allows for flexibility at the tip of these side chains. For example, RMSD of C^β , C^γ , C^δ and N^ϵ atoms of the large residue arginine is less than 0.5 \AA when the total RMSD of the side chain is actually close to 2 \AA from the reference conformation. Following the 2 \AA cut-off for large amino acids like arginine: medium sized side chains of glutamate and glutamine are considered bound-like if the side-chain RMSD is less than 1.5 \AA ; smaller side chains such as aspartate, asparagine, and leucine are considered bound-like

with RMSD less than 1.25 Å and finally threonine is considered bound-like when RMSD is less than 1 Å.

2.2 PREDICTION OF ANCHOR RESIDUES

Residues important for recognition, or “anchors”, are residues that bury largest amounts of solvent accessible surface area (SASA) upon binding. Computationally, one can predict anchors as those side chains that in the complex structure bury more than 80% of their total SASA and bury the largest amount of SASA upon binding (80 \AA^2 or more), while being more than 60% solvated in the free state (Rajamani *et al.* 2004). Selecting residues with more than 80 % of their accessible surface area buried in the complex and more than 60 % exposed in the free protein as anchor candidates ensures that the binding process itself is mostly responsible in burying the anchor residue. Solvent accessible surface areas are computed using the software NACCESS (Hubbard *et al.* 1991) with a water radius of 1.4 Å.

2.3 MODELING ZINC FINGER – DNA INTERACTIONS

The binding reaction of a ZF to a specific DNA sequence D is defined as



The affinity of the ZF to specific DNA sequence D can be expressed in terms of dissociation constant K_d :

$$K_d = \frac{[ZF][D]}{[ZFD]} = \frac{k_{off}}{k_{on}} = e^{\Delta G / RT}, \quad (2.8)$$

where ΔG is the free energy of binding, R is the gas constant and T is temperature.

The change in binding free energy due to a point mutation on either the DNA site or the ZF can be described as the ratio of the affinities of the mutant (Mut) complex and the reference state wild type (WT) complex

$$\frac{K_{dMut}}{K_{dWt}} = e^{\Delta \Delta G / RT}, \quad (2.9)$$

where

$$\Delta \Delta G = \Delta G_{Mut} - \Delta G_{WT}. \quad (2.10)$$

2.3.1 C₂H₂ Zinc Finger transcription factors

The classical ZF domain is composed of a $\beta\beta\alpha$ fold that typically interacts with three to four base-pairs of DNA using key residues in the N-terminal part of its α -helix to make the contacts. The classical ZF EGR has three fingers that wrap around DNA (Elrod-Erickson *et al.* 1996), with the α -helices fitting into the major groove (Figure 2.1). FI binds to a GCG triplet near the 3' end

of the primary DNA strand. FII binds to the TGG triplet in the center and FIII binds to the GCG triplet near the 5' end of the primary DNA strand. Figure 2.1A and Figure 2.1B shows a cartoon and sketch of the intra- and inter-molecular H-bonds for each finger. Note that although the binding site residues and nucleotides of fingers I and III are identical, an Arg preceding the α -helix (R_{-1} , where number is relative to the first residue of α -helix), an aspartic acid on the second position (Pos. +2) of the α -helix (D_{+2}), a glutamic acid at Pos. +3 (E_{+3}) and an Arg at Pos. +6 (R_{+6}), R_{-1} and R_{+6} are not symmetric in their exposure to solvent. In what follows, all fingers in the text are named using the amino acids at positions -1, +2, +3 and +6 of the recognition helix (i.e. EGR FI is RDER).

2.3.2 Decoding protein-DNA interactions

The basic assumption is that changes in the affinity of a complex due to mutations are uniquely determined by changes in effective contact free energies and solvation factors between the different structures. Hence, the scheme to define the potentials is as follows:

- (i) Build homology models of mutant TF based on templates from known complex structures.
- (ii) Perform MD simulations of the homology models in the absence of DNA in explicit solvent to readily identify strong intra-molecular H-bonds.
- (iii) Bonds are established based on distance thresholds obtained from MD of mutants in the absence of DNA, and superimposed into the models of the complex. Then, all plausible intra-and-inter molecular H-bond networks are built into the homology models of each complex.

- (iv) Effective free energies are assigned to all gained and lost H-bonds relative to a reference state, usually the WT complex: ε_{ij} to inter-molecular H-bonds, δ_i to atomic desolvation penalties of unmatched H-bond donors or acceptors at the binding interface and buried hydrophobic residues. These interactions are further modulated by a water factor λ_w that is applied depending on the number of water molecules contacting the H-bonds (see below Results Section for more details). Thus, given a model, these assignments allow us to compute the change of binding free energy as:

$$\Delta\Delta G_{\text{Calc}} = \sum_k (f(\lambda_w) \times \varepsilon_k + f(\lambda_w) \times \delta_k), \quad 2.11$$

where $f(\lambda_w) = 1$ (default), $=1 - \lambda_w$ (if k contacts extra waters), and $=1/(1 - \lambda_w)$ (if k contacts less waters than default).

- (v) Then, using Eqn. 2.9 and $\Delta\Delta G_{\text{Calc}}$ one can trivially relate biochemical binding data with structural models.
- (vi) Using Eqn. 2.11, minimize

$$\text{Argmin} \left(\sum_{ij} \frac{\Delta\Delta G_{\text{Exp}ij} - \Delta\Delta G_{\text{Calc}ij}}{\Delta\Delta G_{\text{Exp}ij}} \right), \quad 2.12$$

for relevant mutants, obtaining inter-molecular H-bonds that best fit the available experimental data.

- (vi) Since we have more mutants than interactions, Eqn. 2.12 is only used as a measure of the quality of the predictions.

2.3.3 Datasets of EGR mutants

Liu and Stormo (Liu and Stormo 2005) mutated FI α -helix positions -1 and +3 resulting in 3 single (RDNR, QDER, DDER) and 2 double (QDNR and DDNR) mutants of EGR FI. They reported 36 binding affinity measurements of these 5 mutants and the wild type protein binding to the consensus DNA site GCG and its mutants GCA, GCC, GAG, GAA and GAC using a quantitative binding assay (Man and Stormo 2001) (Table 2.1 lists the relative binding affinities).

DNA binding site trinucleotides are numbered using the middle base as the reference point from 5' to 3' (e.g., 3'- G₊₁C₀G₋₁ -5'), and nucleotides in the complementary strand are denoted with a "prime symbol" in their subscript (e.g., C_{+1'}).

Table 2.1: Experimental relative affinity data of protein mutants¹.

	RDER	RDNR	QDER	QDNR	DDER	DDNR
GCG	1 ²	5.01	135.3	54.9	162.4	338
GCA	3.7	13.5	90.2	162.4	145	290
GCC	12.3	12.3	63.4	104.1	86.4	131
GAG	6.4	3.9	104.1	12.3	112.8	23.9
GAA	15.6	7.3	101.5	22.6	109.7	84.6
GAC	25.4	5.3	71.2	19.3	73.8	29

¹ The mutants are QDER,QDNR,DDER,DDNR (positions -1,+2,+3 and +6 with respect to the start of α -helix) to different binding sites from Liu and Stormo (Liu and Stormo 2005)

² Reference state, the WT complex. Relative affinities are calculated against the WT finger I – DNA complex, RDER/GCG.

Two completely independent affinity measurement datasets of FII mutants and human ZFs fused to FIII of EGR are from Segal *et al.* (Segal *et al.* 1999) and Bae *et al.* (Bae *et al.* 2003), respectively. Segal *et al.* (Segal *et al.* 1999) used phage display selection, randomizing FII α -helix positions -1, +1, +2, +3, +5 and +6 and reported affinity measurements of 23 FII mutants using mobility shift assays of the purified proteins. Bae *et al.* (Bae *et al.* 2003) utilized yeast one hybrid system to select ZF domains amplified from human genome fused to EGR instead of FIII and reported affinity measurements of 32 selected domains against the selected DNA binding sites.

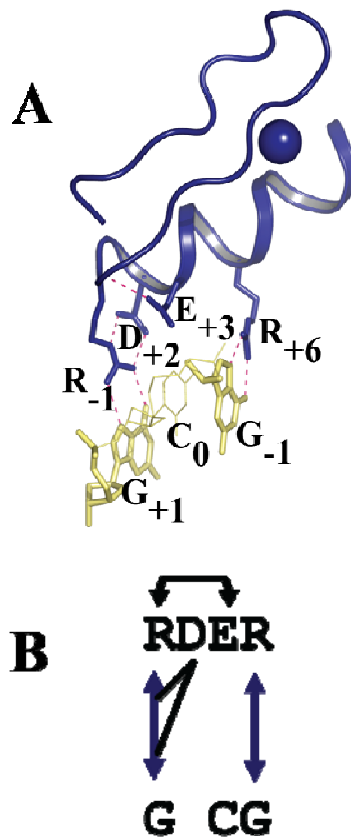


Figure 2.1: Illustration of EGR FI bound to its consensus site.

A. Binding mode of finger I of EGR **B.** Diagram of interaction network of FI. Arrows indicate H-bonds. Colors correspond to a classification scheme detailed in Table 3.5. Black arrows indicate intra-molecular H-bonds, those drawn above/below protein sequence correspond to sc-bb/sc-sc bonds.

2.3.4 Binding modes

Structural insights revealed from the complex crystal structures of EGR (Elrod-Erickson *et al.* 1996) and four mutants (Elrod-Erickson *et al.* 1998) allow us to identify five binding modes for FI, resulting in specific amino acid – base H-bond patterns. Representative structures of these binding modes are shown in Figure 2.2. They are: (a) WT (default) mode (Figure 2.2A) EGR that allows Arg residues to form a bidentate interaction at Pos. -1 (pdb code: 1AAY) (Elrod-Erickson *et al.* 1996), (b) Q mode (Figure 2.2B) from QGSR/GCA mutant (PDB code: 1A1H) shows that Q₋₁ can reach closer to the DNA forming a bond with A₊₁ if there is also a single matching bond at Pos. +3 (e.g., S₊₃-C₀) (Elrod-Erickson *et al.* 1998), (c) D mode (Figure 2.2C) from DSNR/GAC mutant (PDB code: 1A1F), which can reach even closer than Q mode if N₊₃ forms two H-bonds with A₀. Furthermore, two H-bond configurations between R₋₁ and the DNA- backbone (bb) phosphate has been resolved in two different structures (Elrod-Erickson *et al.* 1998): (d) in the BB1 mode (Figure 2.2D) from RDER/GCA mutant (PDB code: 1A1L) the R₋₁.NH₂ group found on the surface (i.e., partially solvated) contacts the C₀ phosphate group, while E₊₃ forms an intra-molecular H-bond with the buried NH₂; and, (e) BB2 mode (Figure 2.2E) based on the mutant RADR/GCG (PDB code: 1A1J), in which R₋₁ contacts the DNA-bb phosphate through the buried NH₂ group, while the second NH₂ is fully solvated. In this complex, D₊₃ prevents a full water attack of the R₋₁ sc by forming an intra-molecular H-bond with HE of R₋₁.

It is important to emphasize that, as shown in Figure 2.2F, crystal structures suggest that ZFs do most of the induced fit upon complexation. This induced fit is in response to well defined attractive interactions that become stronger upon bending. Thus, predicted homology models are

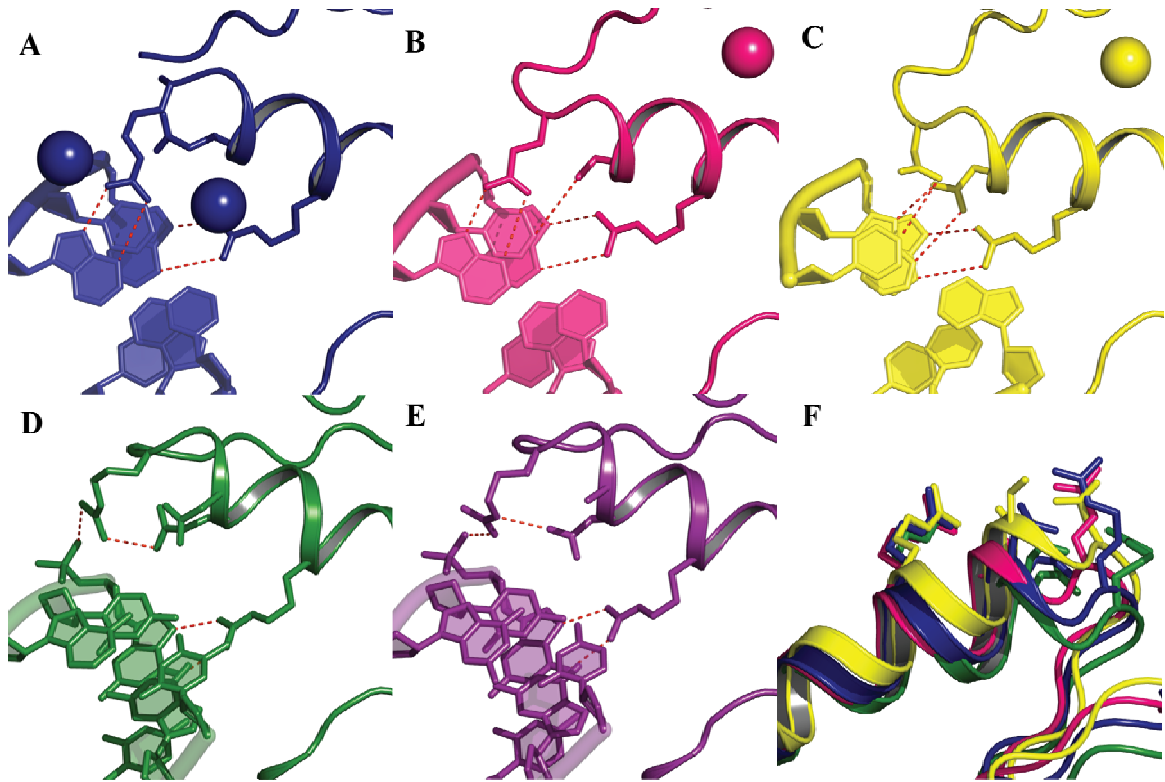


Figure 2.2: Crystal structures of binding modes and induced fit on zinc fingers

A. Wild type (RDER) EGR with GCG site (blue). B. QGSR mutant with GCA site (yellow). C. DSNR mutant with GAC site (pink). D. WT with mutant GCA site (green). E. RADR mutant with GCG site (purple). Hydrogen bonds between the side chains and the bases are shown as dashed lines. F. Superimposition of the α -helices of the four modes after aligning DNA-bb's. Note that α -helices of Q and D modes are closer to DNA than WT mode.

restricted to those that satisfy the complementarity observed in the aforementioned structures, and no new backbones are postulated for either protein or DNA.

2.3.5 Protein and DNA homology models

Mutants of FI, II and III are built using the corresponding finger structure in the EGR crystal (ElrodErickson *et al.* 1996) (pdb code:1AAY) as the template. All fingers in the text are named using the amino acids at positions -1, +2, +3 and +6 of the recognition helix (i.e. EGR FI is RDER). These protein models are then used as the starting structures in the MD simulations described in Section 2.1.3. DNA binding site triplets are taken from, wild type1AAY

(Elrod-Erickson *et al.* 1996), 1A1F (Elrod-Erickson *et al.* 1998), 1A1H (Elrod-Erickson *et al.* 1998), 1A1L/1A1J (Elrod-Erickson *et al.* 1998), 1MEY (Kim and Berg 1996), 1MDY (Ma *et al.* 1994), 2I13 (Segal *et al.* 2006), TFIIIA 1TF6 (Nolte *et al.* 1998) and ETS domain DNA complex 1MDM (Garvie *et al.* 2002). To assure the continuity of the DNA chain the triplets are simply superimposed to the backbone of the appropriate binding mode.

2.3.6 Computation of waters at the protein – DNA interface

In order to have a rough estimate of the number of waters that fit at the binding interface, modeled ZF–DNA complexes are solvated in a box of 1.4 Å radius water molecules, removing waters that overlap with the protein and DNA. These waters are then compared to crystal waters in order to assess the likelihood for models to trap an excess of waters at the interface relative to WT.

Table 2.2: Experimental data from Segal et al. (Segal *et al.* 1999)

FII helix position^a	FII subsite^b	Relative Affinity	$\Delta\Delta G_{\text{Exp}}$ (kcal/mol)^c
RDHT ^d	TGG	1	0.00
RDHR	GGG	0.04	-1.90
RDKR	GGG	0.6	-0.30
QAHR	GGA	0.3	-0.71
TGHR	GGT	1.5	0.24
DGHR	GGC	4	0.82
RDNR	GAG	0.1	-1.36
RDNR	GGG	4.5	0.89
QSNR	GAA	0.05	-1.77
TGNR	GAT	0.3	-0.71
DGNR	GAC	0.3	-0.71
DGNR	GCC	9	1.30
RDSR	GTG	0.3	-0.71
RDER	GTG	1.5	0.24
RDER	GAG	3	0.65
QSSR	GTA	2.5	0.54
QSSR	GTG	100	2.72
TGSR	GTT	0.5	-0.41
DGAR	GTC	4	0.82
RDDR	GCG	0.9	-0.06
RDDR	GAG	0.6	-0.30
QGDR	GCA	0.2	-0.95
QGDR	GCT	1	0.00
TGER	GCT	6.5	1.10
DRDR	GCC	8	1.23

Table 2.3: Experimental data from Bae *et al.* (Bae *et al.* 2003)

FIII helix position	FIII subsite	Relative Affinity	$\Delta\Delta G_{\text{Exp}}$(kcal/mol)
RDER	GCG	1.00	0.00
HSNK	GAC	178	3.1
HSSR	GTT	0.8	-0.2
KSNR	GAG	1.7	0.3
QGNR	GAA	1.2	0.1
QSHR5	GGA	2	0.40
QSHT	AGA	17	1.7
QSNR1	GAA	1	0.03
QSNK	GAA	2.7	0.6
QSSR1	GTA	0.8	-0.1
QSTR	GTA	0.9	-0.1
QTHR1	GGA	1.6	0.3
RDHR1	GGG	0.9	-0.1
RDHT	AGG	6.8	1.1
RDKR	GGG	4.5	0.9
RSNR	GAG	0.7	-0.20
RSHR	GGG	0.2	-1.0
SSNR	GAG	1.3	0.2
QAHR	GGA	8.4	1.3
QTHQ	CGA	0.6	-0.3
DSAR	GTC	0.4	-0.6
CSNR	GAC	2.3	0.5
DSCR	GCC	3.9	0.8
ISNR	GAT	0.1	-1.3
QFNR	GAG	66.1	2.5
QSHV	CGA	64.3	2.5
QSN1	CAA	51.8	2.3
QSNV	CAA	4.1	0.8
VSNV	AAT	2.5	0.5
VSSR	GTG	2.1	0.4
VSTR	GCT	14.5	1.6
WSNR	GGT	1.3	0.2

3.0 RESULTS AND DISCUSSION

3.1 ANCHOR SIDE CHAINS IN PROTEIN-DNA INTERACTIONS

Understanding molecular mechanisms of DNA binding/recognition by TFs is one of the fundamental questions in molecular biology. To this aim, we identified key residues that play important roles in TF-DNA binding by adopting the so-called anchor theory of protein-protein interactions to protein-DNA interactions. The anchor theory (Rajamani *et al.* 2004) of protein-protein interactions predicts that in the absence of their binding partner anchor side chains sample bound-like conformations for a significant amount of time (~30% or more) in order to rapidly form a bound-like intermediate as the first step towards the formation of the high affinity complex. Table 3.1 lists the predicted anchors (highlighted in red) and base contacting residues for three C₂H₂ ZF proteins, as well as the percent of time these side chains are in a conformation close to their bound structure in a solvent without ions and at physiological ion concentrations. From the table, one can readily identify (see Section 2.2) R₊₆ and R₋₁ in FI and FIII of EGR, respectively, as the main anchor residues in this complex. Consistent with the theory (Rajamani *et al.* 2004), molecular dynamics simulations confirm that, *in the absence of DNA*, these side chains spend a significant amount of time (> 30%; see Table 3.1 and Figure 3.1) in conformations similar to those found in the bound structure with DNA. Figure 3.1 shows the RMSD profiles at physiological ion concentrations (150-160mM) of binding site residues of EGR listed in

Table 3.1: Anchors and base contacts for zinc fingers in EGR, TFIIIA and GLI

			Δ SASA (\AA^2) ¹	Buried Free ² (%)	Buried complex ³ (%)	Bound- like(ions) ⁴ (%)	Bound-like (no ion) ⁵ (%)
EGR	FI	R118 (-1)	97	27	75	23±5	27±16
		E121 (+3)	44	64	97	76±1	94±1
		R124 (+6)	129	20	84	63±9	24±6
	FII	R146 (-1)	113	38	94	41±4	41±5
		H149 (+3)	74	42	93	52±7	40±9
		T152 (+6)	41	69	75	88±2	77±3
	FIII	R174 (-1)	134	27	93	69±8	82±6
		E177 (+3)	42	61	93	87±7	64±9
		R180 (+6)	99	13	62	46±14	28±17
TFIIIA	FI	K26 (-1)	85	38	90	63±0.2	55±11
		W28 (+2)	107	25	76	43±14	66±13
	FII	H58 (+2)	62	29	71	54±24	66±21
		H59 (+3)	60	52	93	77±20	73±19
		R62 (+6)	116	35	93	12±5	11±8
	FV	N89 (+3)	46	53	97	73±19	91±9
		K92 (+6)	63	42	81	27±1	19±2
		R96 (+10)	86	26	69	0	0
	GLI	FII	Y155(+2)	104	25	83	28±3
FIV		D216(+3)	46	37	82	71±4	64±6
		K219(+6)	59	48	84	93±6	29±3

¹ Change in Solvent accessible surface areas (SASA) upon complexation. SASA are calculated using the program NACCESS (Hubbard *et al.* 1991).

² Fraction of buried area in the free protein with respect to the tri-peptide Ala-X-Ala.

³ Fraction of buried area in the protein-DNA complex with respect to the tri-peptide Ala-X-Ala

⁴ Fraction of bound-like conformations (rmsd less than 2 Å) for each side-chain in the **presence** of counter-ions at physiological concentrations (150 – 160 mM). See methods.

⁵ Fraction of bound-like conformations for each side-chain in the **absence** of counter-ions (rmsd less than 2 Å). See methods.

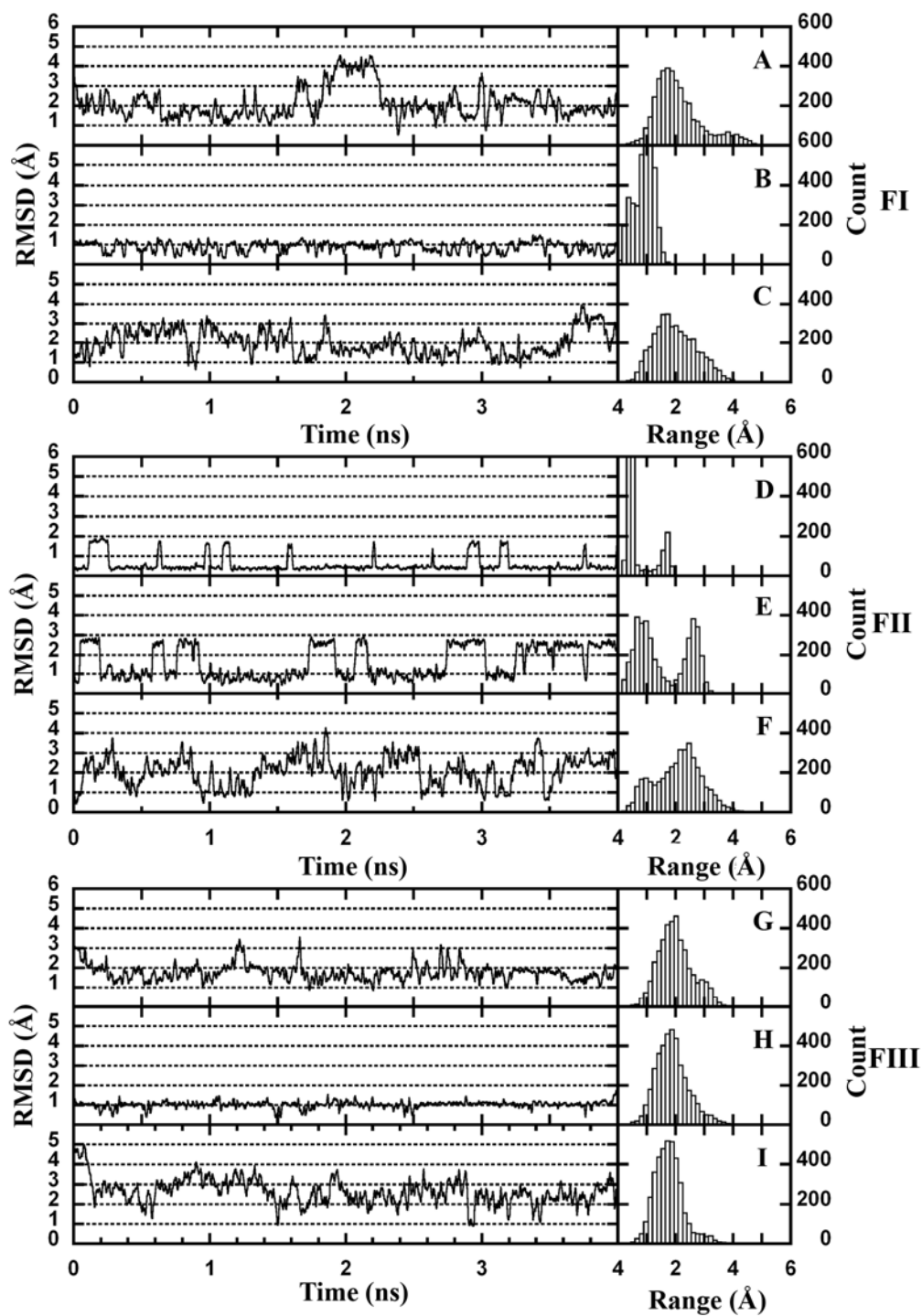


Figure 3.1: RMSD profile of binding site residues of EGR at 35mM counter ion concentration. **A.** EGR FI R₋₁. **B.** EGR FI E₊₃. **C.** EGR FI R₊₆. **D.** EGR FII T₊₆. **E.** EGR FII H₊₃. **F.** EGR FII R₋₁. **G.** EGR FIII R₋₁. **H.** EGR FIII E₊₃. **I.** EGR FIII R₊₆. Side panels show the histograms of the RMSDs of each residue.

Table 3.1 with FI residues in Figure 3.1A-C, FII residues in D-F and FIII residues in G-I. R₋₁ in FI (Figure 3.1A) is 28% bound-like, E₊₃ (Figure 3.1B) is 98%, R₊₆ (anchor residue in FI in Figure 3.1C) 50% bound-like. Figure 3.1B and Figure 3.1H show rmsd profiles of E+3 in fingers I and III, respectively. 100% bound-like rmsd profiles of these two residues may lead one to ask why these glutamic acids at Pos. +3 are not also anchors? The main reason is that both residues are already highly buried in the free, unbound fingers (65% and 61% buried in FI and FIII, respectively) and the change in SASA upon complexation is only about 43 Å. Strikingly, we found that the bound-like behavior for some of these residues is highly dependent on ion concentration.

One might argue that the widespread bound-like behavior of anchor side chains in Table 3.1 is due to the harmonic constraints of the N and C to the bound structure. However, these constraints are consistent with structural evidence showing that the ZF fold does not depend much in sequence (see Figure 1 in Ref. (Lu and Klug 2007)) and does not change before and after binding DNA (Pavletich and Pabo 1991; Lu and Klug 2007). Moreover, the different behavior observed in solvated anchor residues as a function of ionic strength (see Figure 3.2 and Figure 3.4) clearly reflect the critical role that explicit solvent has in the dynamics of these side chains. In Table 3.1, we see that K₋₁ in FI of TFIIIA, Y₊₂ in FII of GLI are also predicted and confirmed as anchor residues. R₊₆ (Arg62) in FII of TFIIIA, on the other hand, does not agree with the theory since it failed to extensively sample bound-like conformations. We should point out though that uncertainties in the modeling of H₊₂ and H₊₃ might play a role in the dynamics of this side chain.

3.2 ROLE OF COUNTER-IONS IN BINDING

3.2.1 Effect of ion caging in side chain dynamics

We emphasized that all MD simulations are performed for each individual zinc finger domain in the absence of DNA (see Section 2.1), as a function of the number of Cl^- ions in the water box. Then, as a function of ion concentration, we find that solvated side chains important for both specific and non-specific association become increasingly bound-like. For instance, each panel in Figure 3.2 shows the RMSD of the main anchor of EGR, R_{+6} in FI, with respect to its bound conformation as a function of time during 4 ns MD simulations and increasing number of Cl^- ions in the water box. Figure 3.2A shows the RMSD of R_{+6} with no ions present in the simulation box. In this case, the side chain is found in a bound-like conformation (i.e., less than 2 Å RMSD from the bound structure) 29% of the time (average of 4 independent simulations is 23 ± 6 %). Addition of counter ions increases this percentage to as much as 79% (Figure 3.2F). The histograms in the right insets clearly show how as a function of Cl^- ion concentration the distribution of R_{+6} conformations shifts to low RMSDs with respect to the bound structure.

This correlation between counter-ions and bound-like behavior was also observed in other specificity determinant anchor side chains, for example, in EGR, R_{+6} in FIII (Figure 3.5) and H_{+3} in FII (Table 3.1). The latter improves from 40% with no ions to around 55% with more than 5 ions in solution. A similar correlation has also been observed for GLI (Pavletich and Pabo 1993) and TFIIA (Nolte *et al.* 1998) anchors (see Table 3.1), as well as for side chains involved in non-specific binding. Some examples are discussed below. In all cases, the fraction of

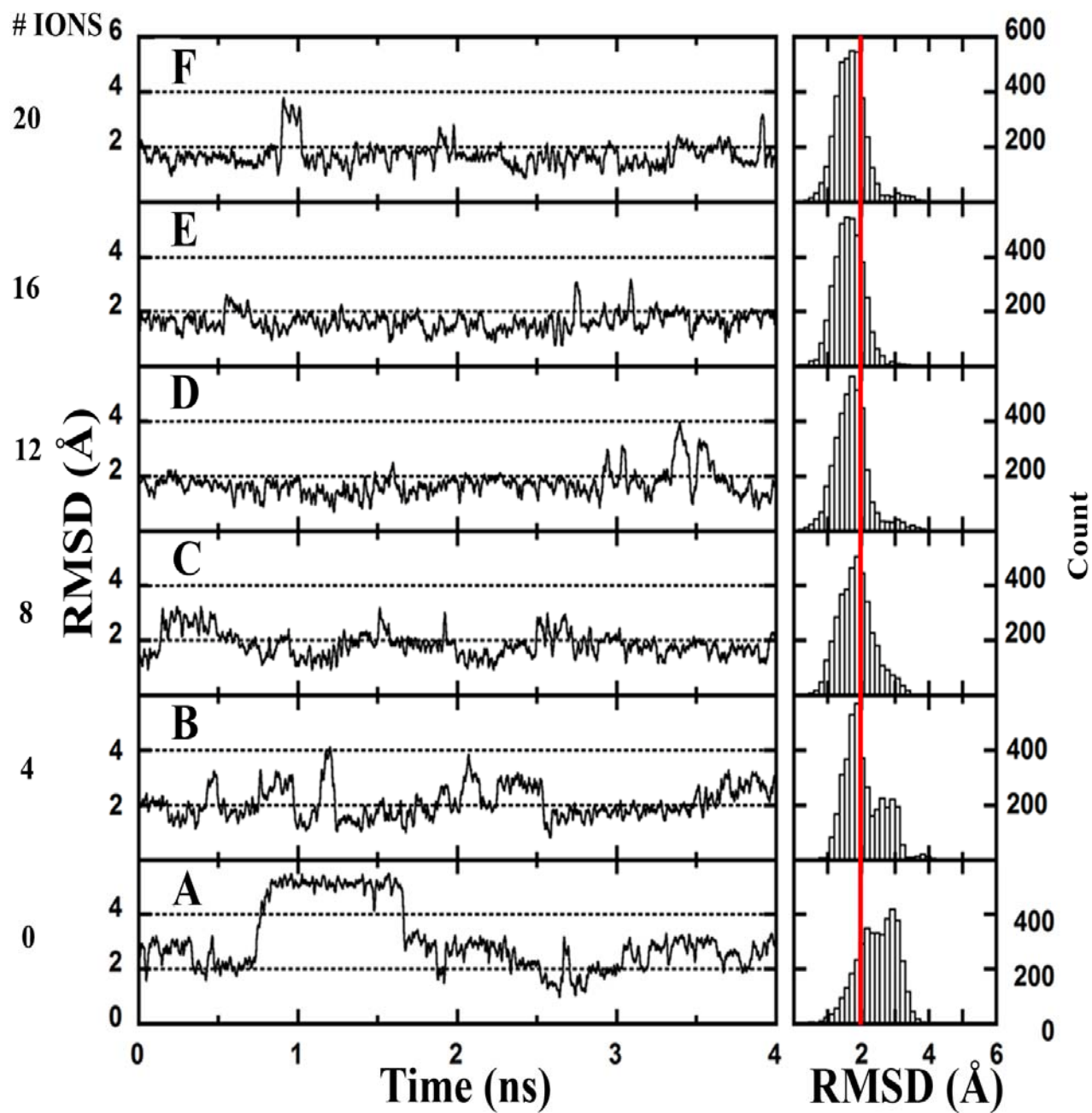


Figure 3.2: Ion caging: Bound-like behavior of EGR anchor R_{+6} in FI as a function of ion concentration. Change in RMSD of R_{+6} with respect to its bound conformation as a function of time and increasing number of counter ions in the simulation box: A No ions, B 4 ions, C 8 ions, D 12 ions, E 16 ions, and F 20 ions. The histograms show the distribution of the RMSDs in each simulation. Red line indicates the 2 Å cut-off used to determine fraction of bound-like conformations.

bound-like conformations of side chains observed in the simulations saturates at an effective ionic strength close to the physiological relevant value of 150 mM. The glutamate residues at Pos. +3 in fingers I and III are two exceptions. They become less bound-like when the counter ions (Table 3.1 second and eighth lines) are at higher (physiological) concentrations. The main reason for this behavior is that both glutamic acid and the weakly bound Cl⁻ ions are negatively charged and repel each other. Note that these glutamic acids are not anchor residues and do not contact DNA.

3.2.2 Buried versus non-buried side chains.

The principle behind the different roles predicted for buried (“anchors”) and non-buried (“latches”) side chains (Rajamani *et al.* 2004) in protein-DNA recognition is that misfolded chains at the core of the encounter complex will not easily rearrange, whereas at the periphery (i.e. solvent exposed) conformations can rearrange to optimize complementary interactions. This is clearly reflected in Figure 3.1 (A and G) and Figure 3.4 (A and B), where the dynamics of the two side chains that bury the largest amount of SASA in EGR, R₊₆ and R₋₁ in FI and FIII, respectively, is more than 60% bound-like. Figure 3.3 shows these buried side chains in EGR in cyan spheres. It also shows the partially exposed side chains capping the N-and-C terminals. These side chains are R₋₁ of FI shown as the lower dark blue spheres, and R₊₆ of FIII shown as the upper dark blue spheres. The dynamics of these partially exposed side chains, on the other hand, is between 20-and-50% bound-like. Hence, despite the fact that the two domains have an identical helical binding sequence (i.e., RDER) and bound crystal structure, the dynamics of their key side chains appear to have evolved differently satisfying the buried versus non-buried paradigm.

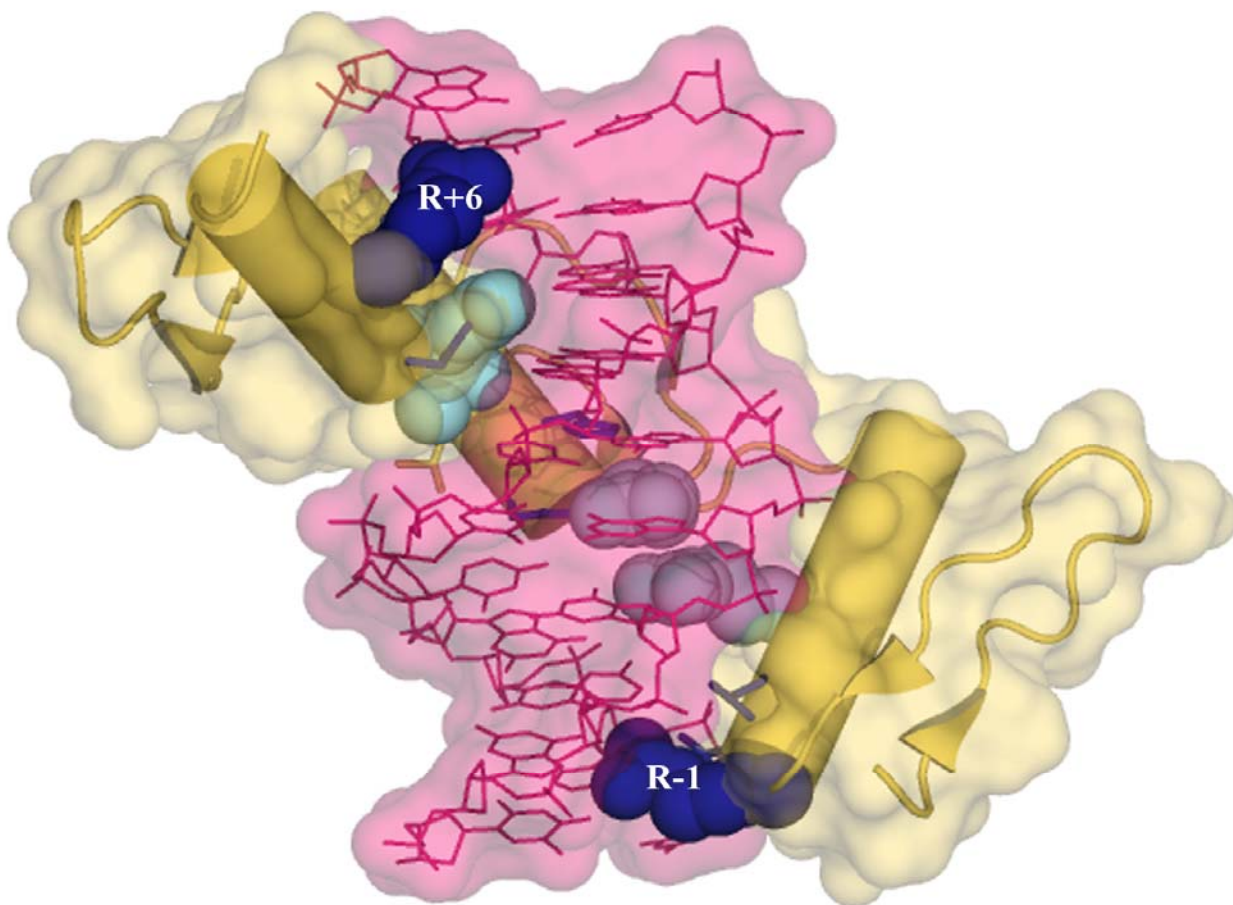


Figure 3.3: Buried and exposed key side chains of EGR.

EGR is shown in cartoon representation and transparent surface in yellow. DNA is shown as sticks and transparent surface in pink. Two exposed side chains at the binding sites of fingers I and III are shown as blue spheres. Buried key side chains are shown as cyan spheres.

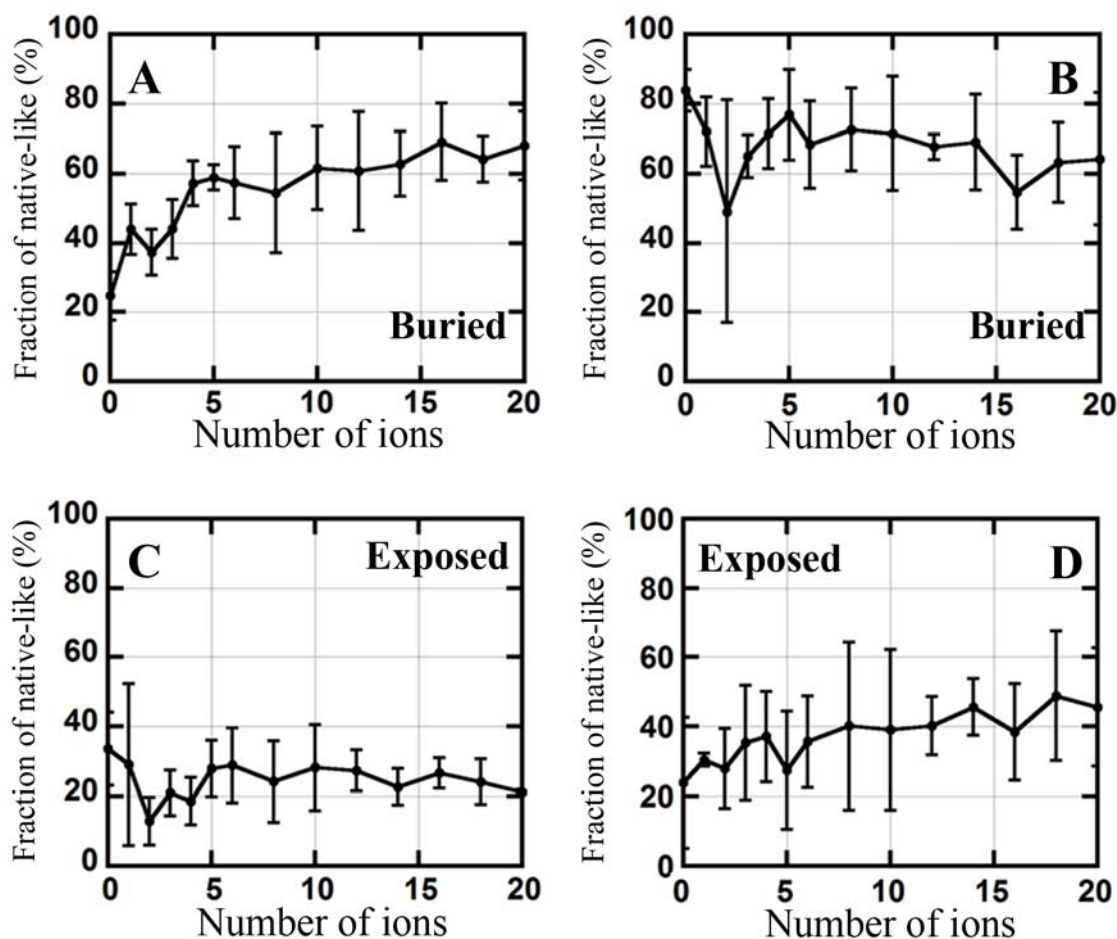


Figure 3.4: Ion dependence of the bound-like behavior of key side chains in EGR fingers I and III. The effective ionic strength in mM corresponds to the number of ions \times 11.5. The conformations are considered bound-like if the RMSD from the crystal structure conformation is under 2 Å. **A.** Fraction of bound-like conformations for buried Arg+6 in FI. **B.** Fraction of bound-like conformations for buried Arg-1 in FIII. **C.** Fraction of bound-like conformations for exposed R₁ in FI. **D.** Fraction of bound-like conformations for exposed R₊₆ in FIII. Error bars are the direct standard deviation from three or more independent 4 nanosecond MD simulations.

3.2.3 Ions bind at the phosphate binding site in the protein-DNA complex structure.

The correlation between ion concentration and bound-like behavior is no accident, but in fact it is due to a weak binding site for Cl⁻ ions located next to the Zn⁺² ion and one of its coordinating histidines (+7). Figure 3.5 shows this observed correlation from two representative simulations,

one with low ion concentration (Figure 3.5A and Figure 3.5B), approximately 35 mM, and another with a high ion concentration of about 210 mM (Figure 3.5C and Figure 3.5D). The conserved backbone phosphate contact formed by H₊₇ is made through its N^δ atom acting as the donor. Figure 3.5A and Figure 3.5C display the minimum distances of an ion to this N^δ atom of H₊₇ during the simulations with different ion concentrations. Figure 3.5B and D show the side chain dynamics of the anchor residue R₊₆ in the two simulations, respectively. Note that when the ion is within interaction distance to H₊₇ (i.e. under 5 Å distance to H₊₇), R₊₆ samples bound-like conformations. This Cl⁻ binding site is complemented by two additional basic groups, one on the α-helix (R₊₆ of FI and FIII of EGR, Figure 3.6A and Figure 3.6C) and the other one on the second β strand usually before the conserved core residue phenylalanine. These two basic groups on either side of the Zn⁺² coordinating H₊₇ play an active role in binding either by contacting a base on the α-helix side or by interacting with DNA backbone phosphate. If bound, ions interact strongly with these nearby side chains. Note that FII of EGR does not have a basic group at Pos. +6, but a polar threonine that does not contact DNA. The loss of this key basic group prevents FII from binding a Cl⁻ ion as strong as FI and FIII. The key observation here is that the binding site of the Cl⁻ ions on the surface of the ZF corresponds to the same locus where phosphate groups bury the largest amount of SASA upon complexation (see Table 3.2 and Figure 3.6). Hence, the “ordering” of R₊₆ and other side chains into bound-like conformations in the absence of DNA reflects in part the phosphate-like electrostatic interactions mimicked by the Cl⁻ ions.

The Cl⁻-phosphate binding site is quite robust. Indeed, for EGR, TFIIIA and GLI, we find that those phosphates that bury the largest amount of SASA in the complex are always mimicked by a Cl⁻ ion (Table 3.2). Note that in FIII of EGR the corresponding phosphate is missing from the crystal structure. The percent of simulation time that Cl⁻ ions spend in the phosphate binding

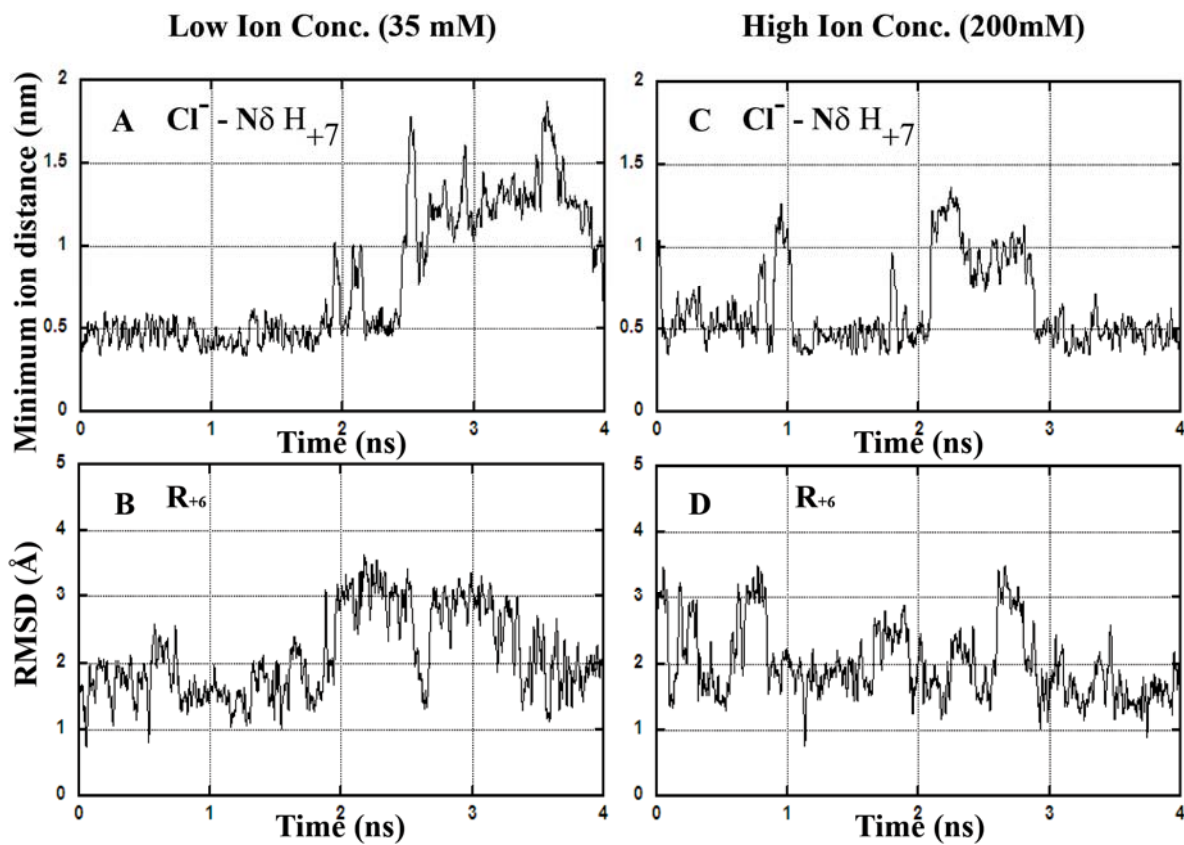


Figure 3.5: Relation between weakly bound ion and Arg+6 dynamics.

A. Minimum Cl^- ion distance to $\text{N}\delta$ atom of the conserved H_{+7} in EGR in a low ion concentration simulation. **FI.** **B.** Side chain RMSD profile of R_{+6} (anchor) from the low ion concentration simulation shown in panel A. **C.** Cl^- ion distance to $\text{N}\delta$ atom of the conserved H_{+7} in EGR in a high ion concentration simulation. **FI.** **D.** Side chain RMSD profile of R_{+6} (anchor) from the high ion concentration simulation shown in panel C.

(i.e., within 3 \AA of the phosphorus atom) site is 20% or more, compared to 1-3% for phosphates that bury 50 \AA^2 or less of SASA. The 3 \AA clustering radius around the phosphorus position is quite stringent considering the four coordinating oxygens nearby and their van der Waals radiuses. Indeed, as shown in the histograms in Figure 3.6, weakly bound ions are never much further than 5 \AA . For comparison, Figure 3.6A also shows the histogram of Cl^- ions around the second most buried phosphate of FI in EGR.

It is important to emphasize that our results show that phosphates that do not bury large amounts of their SASA, do not have a mimicking counter ion at the binding interface. In fact, only phosphates with a surrogate Cl⁻ bind in the classical geometry of EGR (Figure 1.1), burying the phosphate within 4 Å of the conserved N^δ of H₊₇. FIV of TFIIA and FI of GLI, which do not bind DNA at all, do not attract a counter ion. From the eight stable ions observed on the ZF binding site, only the G26 phosphate site of TFIIA do not interact directly with H₊₇ but with an OH group of Tyr24. In addition, the fingers that do not have the conserved backbone contact through H₊₇, fingers III and V in GLI, do not have weakly bound counter ions. Our results show that fingers whose phosphate buries approximately 55 Å² or more SASA usually have a weakly bound Cl⁻ counter ion acting as surrogate (see Table 3.2).

3.2.4 Non-specific contacts.

It is well known that protein-DNA interactions have a subtle interplay between specific and non-specific binding (Winter *et al.* 1981; von Hippel and Berg 1986; Halford and Marko 2004; von Hippel 2004). So far, we have described how side chains that contact DNA bases have conformations suitable for recognition of their consensus sequence, i.e., they acquire structural motifs similar to those observed in their bound state. In what follows we ask the following question: Do side chains involved in non-specific contacts with the DNA backbone have preferred bound-like conformations?

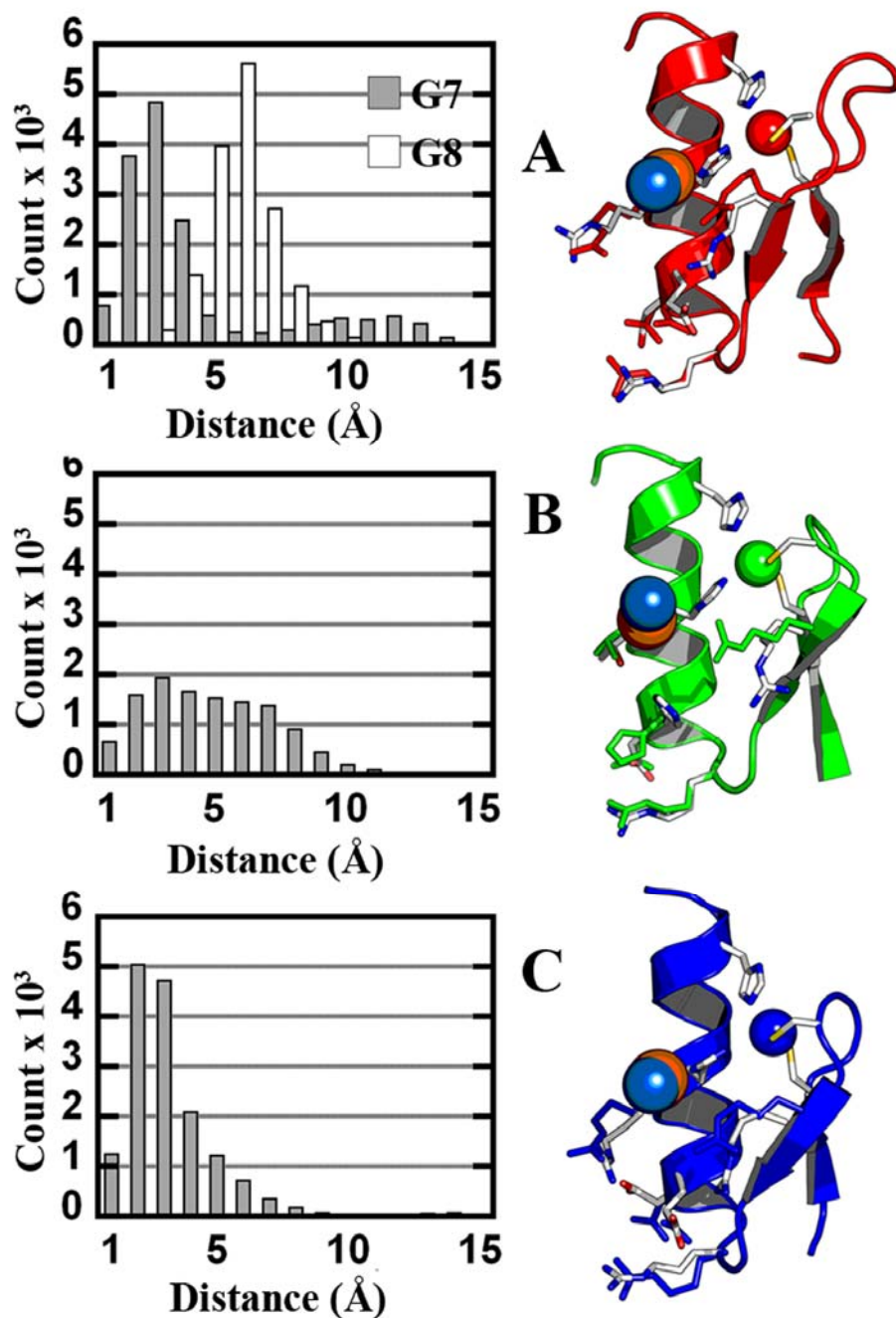


Figure 3.6: Ions occupy the phosphate binding site in the protein–DNA complex.

Counter ion positions in fingers I-III of EGR and their distribution with respect to the phosphorus atom position. The phosphorus atom is the binding partner of the first metal coordinating histidine on the α -helix. The cartoon representations of each finger are on the right: DNA phosphorus atom shown in orange spheres; Cys, His residues and binding motif residues are shown as sticks; CPK colored sticks show the crystal conformation; and, colored sticks show a snapshot of the position of key side-chains. A blue sphere shows for each finger the position of the Cl⁻ in the simulation that had the largest residence time using a 3 Å clustering radius. **A** FI (red), for comparison we show the distribution with respect to a phosphate binding site that is occupied by a Cl⁻ ion, G7 (filled column), and one in which is empty (empty column), G8. **B**. FII (green). **C**. Finger III (blue).

Table 3.2: Solvent accessible surface area (SASA) buried by phosphates.

		DNA	Δ SASA ⁶ (Å ²)	Ion residence ⁷ (%)
EGR	FI	G7	72	70±12
	FI	G8	69	2±1
	FII	G4	60	35±4
	FII	T5	39	4±2
	FIII	Mod	-	68±7
	FIII	G2	36	0
TFIIIA	FI	G26	53	36±8
	FI	G34	51	4±1
	FI	A27	37	15±3
	FII	T23	77	19±4
	FII	A22	35	1
	FV	T8	66	20±4
	FV	C7	49	0
	FV	A53	27	2±1
	FIII	G20	54	2±1
	FIII	C40	40	1
	FIII	A22	13	0
	GLI	FII	A65	77
FII		G66	50	1
FIV		C59	68	24±5
FIV		C7	40	4±1
FIV		C58	40	0
FIV		T8	22	4±1
FV		G56	55	7±1
FV		A57	45	1
FIII		C62	47	6±1
FIII		A64	34	1
FIII		T6	20	3±1

⁶ Change in SASA upon complexation. Surface areas of phosphate groups are calculated in the absence and presence of the protein.

⁷ The fraction of ion residence time corresponds to the fraction of the simulation time a counter-ion is observed within a 3 Å radius sphere from the position of the phosphorus atom.

Table 3.3 lists all backbone contacting side-chains for EGR, TFIIIA and GLI, with 22 out of 30 contacting side chains showing a significant amount of bound-like behavior (greater than 30% of simulation time (Rajamani *et al.* 2004)). For EGR all DNA backbone contacts are dynamically selected to conformations that are more than 50% of the time in bound-like conformations, with the exception of Arg142 in FII and other side chains that remain partially solvated after docking. We recall that based on the fact that Arg142 and 3 other side chains remain around 50% solvated in the bound structure (not an anchor). Thus, they are predicted to undergo induced fit *after* the formation of the encounter complex (a “latch” side chain) (Rajamani *et al.* 2004). For these side chains, MD trajectories show that they are consistently away from the DNA, not interfering with the binding pocket (see section 3.2.3). Several of these side chains are also partially buried in the free state, and therefore significantly constrained to move to a drastically different conformation from the bound state, e.g., H₊₇. Other side chains benefit from the phosphate mimicking Cl⁻ ions. Arg114 in FI of EGR improves its bound-like behavior from 30% to 60% as a function of ionic strength. Two side chains are partially buried and their dynamics are not bound-like, and one solvated side chain Arg183 of FIII of GLI does not behave as predicted.

The lysine in the conserved linker sequence (TG-E/Q-KP) in ZF proteins also contact and stabilize the protein-DNA complex (Choo and Klug 1993; Wuttke *et al.* 1997; Laity *et al.* 2000). These hinge regions play a critical role capping the helical domains and become rigid upon DNA binding (Laity *et al.* 2000). Figure 3.7 shows the side chain dynamics of conserved linker Lys residues from simulations of consecutive ZFs, FI-FII and FII-FIII in EGR (using the same simulation protocol). The linker lysine between FI and FII (Figure 3.7A) is 84% of the time in a

bound-like conformation and the one between FII and FIII is 100% of the time in bound-like conformations.

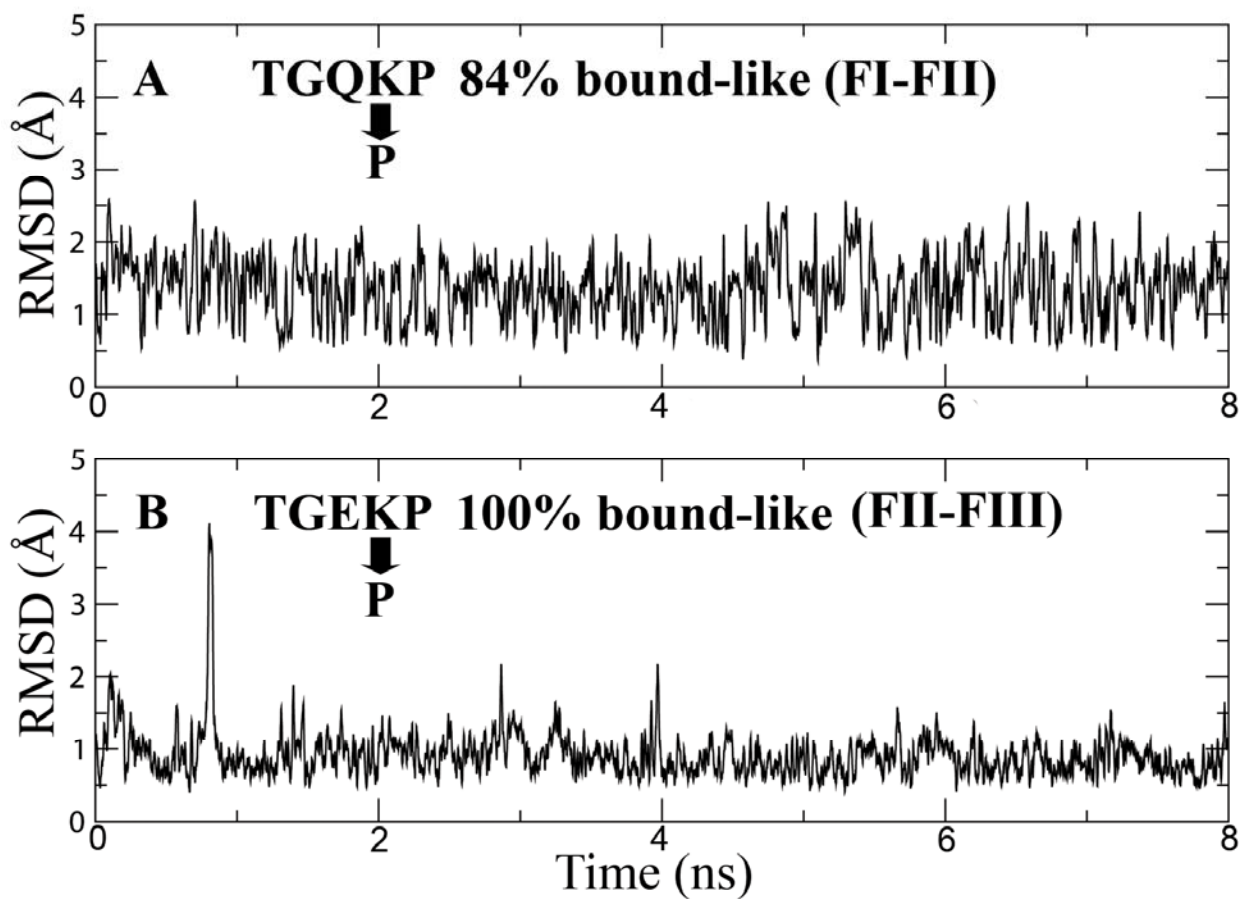


Figure 3.7: Dynamics of phosphate contacting conserved lysines in the linker regions between fingers. DNA sugar backbone phosphate contacting residues are indicated with an arrow. **A.** RMSD profile of Lys between fingers I and II of EGR. **B.** RMSD profile of Lys between fingers II and III of EGR.

Table 3.3: Residues contacting the phosphate backbone in EGR, TFIIIA and GLI.

		DNA contact	Residue	Δ SASA (\AA^2)	Buried Free (%)	Buried complex (%)	Native-like (ions)	Native-like (no ion)	
EGR	I	G8	R103	40	36	56	52±8	42±8	
		G7	R114	43	35	57	58±9	30±9	
		G7	H125(+7)	19	84	97	100	100	
	II	T5	R142	53	23	48	0	0	
		G4	H153(+7)	24	77	93	100	100	
	III	G2	R170	53	32	59	96±2	98±0.5	
		-	His181 (+7)	0	84	84	100	100	
	TFIIIA	I	G34	R12	65	34	66	0	0
			G34	Y13	44	48	73	100	100
G26			Y24	44	68	92	100	100	
A27			K29 (+3)	54	63	96	0	0	
II		T23	H63 (+7)	14	86	95	100	100	
		A22	T66 (+10)	43	43	86	29±7	27±7	
III		A22	T85	88	8	94	46±19	52±5	
		C40	K87 (+1)	36	28	50	0	0	
V		T8	K144	45	30	58	30±6	0	
		T8	H155 (+7)	17	82	93	100	100	
		C7	V158 (+10)	51	20	65	21±4	0	
GLI		II	G66	R146	37	66	85	0	0
			G66	K152	56	20	54	0	0
	A65		H160 (+7)	16	89	100	100	100	
	III	C62	Y181	57	64	96	100	100	
		A64	R183	142	12	83	0	0	
		T6	K188(+5)	29	51	68	97	100	
	IV	C7	Y200	78	39	83	77±4	80±8	
		T8	R217(+4)	24	68	80	71±4	88±9	
		C59	H220 (+7)	24	79	95	100	100	
		C58	T224(+11)	45	37	81	84±0.2	85±9	
	V	G56	Y242	51	69	98	100	98	
		A57	T243	108	0	96	94±3	92±9	

The robust bound-like behavior of backbone contacting side chains suggests a dominant role of phosphate groups in protein-DNA recognition. These “built in” motifs in the ZF domains are consistent with a binding mechanism that first associates non-specifically to the DNA’s backbone to then diffuse along the 1-D DNA sequence, as first suggested by Winter, Berg, and Von Hippel (Winter *et al.* 1981).

3.2.5 Summary: Counter ions act as surrogates for the backbone phosphate groups at the protein-DNA interface stabilizing the critical side chains.

As the counter ion bound ZF approach DNA, the negative charged DNA phosphates will push the weakly bound counter ion out of the binding pocket. Hence, upon ion removal, bound-like side chains rapidly grab the phosphate forming the non-specific encounter complex. The formation of the encounter complex, in turn, should force the release of DNA bound counterions. We expect the “ordering” of key side chains to increase up to around 150 mM (a 150 mM concentration corresponds to about 14 ions in Figure 3.4), where it saturates, leading to a decrease in side chain entropy loss upon binding. Based on side chain entropy estimates (Lee *et al.* 1994), side chain “ordering” could easily account for one kcal/mol difference, more than enough to explain the observed 4-fold increase in affinity at low ionic strength. At ionic strengths larger than 0.2 M, electrostatic screening starts to sharply curtail long range electrostatic steering, resulting in the decrease of the TF-DNA association rate observed experimentally (Fried and Stickle 1993).

This mechanism is not only fully consistent with our observations regarding ion/phosphate binding in well defined sites on the surface of TF, but can also rationalize a

kinetically efficient process consistent with association rates close to the diffusion limit (Winter and von Hippel 1981; Romaniuk 1990; Fried and Stickle 1993; Engler *et al.* 1997; Hamilton *et al.* 1998). We note that simulations with acetate have shown similar propensities for protein association than Cl⁻. Therefore, we predict that counter ion association is a general mechanism for DNA-ion destabilization and non-specific binding. Binding experiments with different salts might provide a quantitative estimate on the effect of counter ion interactions in the binding affinity as a function of ionic strength.

3.3 EXPERIMENTAL BASED CONTACT ENERGIES FOR ZINC FINGER – DNA INTERACTIONS

Understanding the molecular basis and specificity of transcriptional regulation is an important step not only to learn how cells normally function, but also for understanding how mutations affect the binding specificity. Today, most methods to detect DNA binding or regulatory sites rely on a combination of sequence information, conservation patterns, genome annotations, and affinity data (Stormo 2000; Bulyk 2003; Siggia 2005; GuhaThakurta 2006). However, the short length of binding sites and intrinsically degenerate nature of DNA leads to a high number of false positives. Since the under-prediction and, more significantly, the over-prediction of protein-DNA interactions is the current bottleneck for understanding regulatory networks, it is of prime importance to develop new methods to eliminate as many as possible the relatively large number of false positive predictions.

Here, we use a comprehensive analysis of high quality binding experiments from Liu and Stormo (Liu and Stormo 2005) and crystal structures solved by Pabo and collaborators (Elrod-Erickson *et al.* 1996; Elrod-Erickson *et al.* 1998) to decode a minimal set of ten fundamental interactions that allow us to predict the affinity and complex structures of 89 different EGR-like C₂H₂ TFs. The interactions account for a novel classification of inter-molecular hydrogen bonds (H-bonds) and atom desolvation penalties, as well as a water accessibility factor that mediates these interactions. To predict the change of binding affinity for each mutant, we use the EGR crystal structure (Elrod-Erickson *et al.* 1996) to build homology models of all possible intra-and-inter molecular H-bonds allowed in the different binding modes resolved for this complex (Elrod-Erickson *et al.* 1998), and then select the model with the lowest free energy. Three independent data sets of 35 mutants of FI (Liu and Stormo 2005), 23 mutants of FII (Segal *et al.* 1999), and 31 different FIII proteins (Bae *et al.* 2003) are predicted with correlation coefficients R² of 0.998, 0.96 and 0.94, respectively. It is worth noting that FIII proteins are ZF domains amplified from the human genome, i.e. the sequence identity between human ZFs and EGR is minimal. Our approach also selects the lowest free energy structure as the most likely structural model for each protein. This information is quite valuable since only two structures out of the 90 ZFs considered here have been resolved experimentally. Specific interactions show little or no contribution from long range interactions or water mediated H-bonds. However, solvent at the interface modulates the strength of inter-molecular interactions. The good agreement between predicted and experimental data provided by the interaction and recognition code developed here suggests that DNA deformations impose important constraints in both the allowed H-bond network and the number of water molecules present at the binding interface. Moreover, homology models and known crystals suggest that most of the induced fit occurs from the

protein side steered by short range inter-molecular H-bonds. Desolvation penalties account for buried donor and/or acceptor side chain (sc) groups that do not form an H-bond with the bb of protein or DNA, referred here as free or unmatched polar groups. Our approach highlights that the full assessment of protein-DNA interactions is intimately related to detailed predictions of the loci of water molecules at the binding interface.

3.3.1 Intramolecular hydrogen bonds

As described in the Methods section, we run 9 ns long MD simulations of EGR FI and its mutants to sample the intra-molecular hydrogen bonds that are formed within each protein domain. Consistent with properties already observed in protein-protein interactions (Rajamani *et al.* 2004), the MDs reveal that key structural motifs observed in the co-crystals are also observed in the dynamics of individual fingers. For instance, a key feature is that side chains R₊₆ in FI, R₋₁ in FII, and R₋₁ in FIII that are found buried in the complex, already behave very much bound-like in the absence of DNA (not shown). More interestingly, we find that the H-bond between the donor backbone N at Pos. -1 and acceptor sc at Pos. +3 is quite stable for almost all protein sequences: WT, QDER, DDER, QDNR, and RDNR for 97%, 89%, 62%, 40% and 24% ($\pm 5\%$) of the simulation time, respectively. On the other hand, repulsive interactions between Asp side chains forbid this bond in DDNR. Therefore, unless other constraints are present, this bond will not be allowed for this sequence. A strong side chain-side chain (sc-sc) H-bond is observed between Asp at Pos. -1 and Asn at Pos. +3 in the double mutant DDNR (79% of the simulation time). Also, in QDNR, D at Pos. +2 is forming a bond with either Q₋₁ or N₊₃ for about 42% each; and, in QDER, Q forms a bond with D₊₂ for 26% of the time. These bonds prove very important to validate possible inter-molecular bonds in homology models.

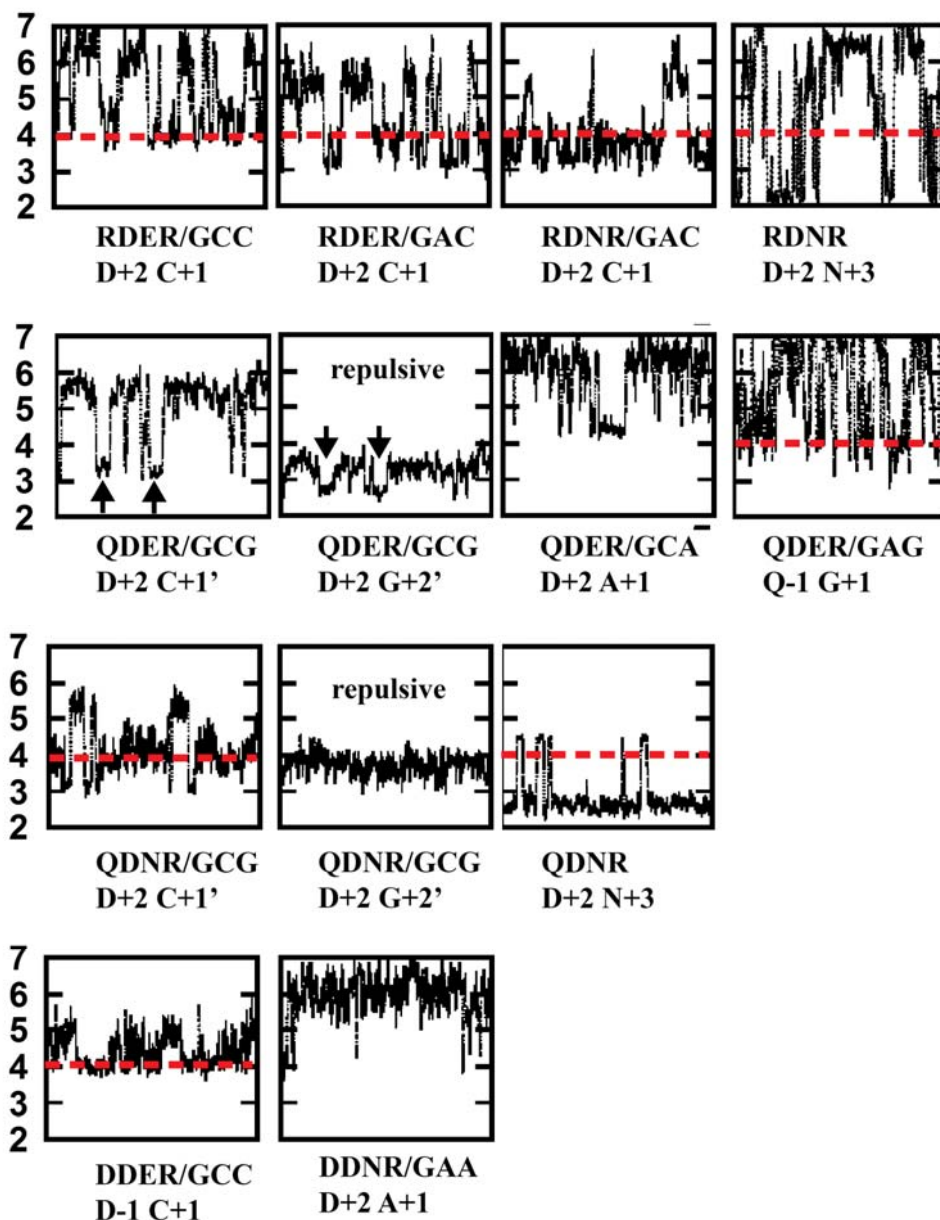


Figure 3.8: Distance profiles for key contacts in FI mutants and their complexes over 8 ns of MD runs . The distance profiles are running averages over 25 ps of the MD trajectories. The protein-DNA mutant complexes (e.g., RDER/GCC) and the protein mutants are shown below each panel for inter-molecular and intra-molecular contact profiles, respectively. Residues and bases involved in each contact are also indicated below each panel. H-bonds are feasible if and only if the distance between the acceptor and the hydrogen is less than 4 Å (as indicated by those plots with a dashed line). Note that a quick transition such as in DDNR/GAA at very short times is not considered stable enough to form the bond. Also, note that in QDER/GCG, repulsive interactions of $D_{+2}-G_{+2}'$ are less than 3 Å whenever D_{+2} is within contact distance with C_{+1}' (pointed by black arrows), preventing $D_{+2}-C_{+1}'$ H-bond to be formed. Whereas, in QDNR/GCG, $D_{+2}-C_{+1}'$ can be made, where the intra-molecular H-bond $D_{+2}-N_{+3}$ pull D_{+2} further away from the repulsive G_{+2}' (an additional 1 Å separation compared to QDER/GCG).

3.3.2 Recognition code for intermolecular hydrogen bonds

Using homology models of the different protein and DNA sequences, we search for all possible inter-molecular H-bonds allowed for the appropriate binding mode in Figure 2.2. H-bonds are assumed to be formed if the distance between hydrogen and acceptor atoms is less than or equal to 4 Å (see sample of distances for key contacts in Figure 3.8). This distance is larger than that of a typical hydrogen bond (1.8-to-3.0 Å), since it assumes a small 1 Å induced fit (or error) in our models.

The key observation here is that the superposition of tri-nucleotides in the DNA backbone imposes nontrivial distance constraints between protein and DNA molecules. For instance, in WT mode clashes prevent E₊₃ from forming a bond with the middle nucleotide and N₊₃ from reaching C₀ in GCG/GCA. In addition, D₋₁ does not reach GCG/GCA but can reach GCC/GAC/GTA. These constraints, listed in Table 3.4, are at the core of the recognition rules for C₂H₂ ZF-DNA interactions. The list can be assumed to be incomplete, since one cannot rule out the existence of binding modes not yet revealed by crystallographic efforts. Nevertheless, it implements currently validated inter-molecular H-bond networks. Finally, binding modes observed for EGR and its mutants only show inter-molecular contacts between nucleotide bases and side chains at positions -1,+2,+3 and +6 of the α -helix; a limited number of possible DNA-bb contacts are also considered. Water mediated H-bonds are implicit in the desolvation penalties but otherwise neglected.

Table 3.4: Look up table for amino acid – DNA hydrogen bonds.⁸

	Pos. -1	Pos. +3	Pos. +6	Mode
G/A-C-X	R ₋₁ -G ₊₁		R/K ₊₆ -G ₋₁	WT
	D ₋₁ -C ₊₁		R/K ₊₆ -G ₋₁	WT
		N ₊₃ -C ₀	R/K ₊₆ -G ₋₁	WT
	Q/H ₋₁ -X ₊₁	D/N/S ₊₃ -C ₀	R/K ₊₆ -G ₋₁	Q
G/A-T-X	R ₋₁ -G ₊₁		R/K ₊₆ -G ₋₁	WT
	H/Q/S/T ₋₁ -X ₊₁	S/T ₊₃ -T ₀	R/K ₊₆ -G ₋₁	Q
	Q ₋₁ -A ₊₁	S/T ₊₃ -T ₀	R ₊₆ -DNA-bb	Q
G/A-A-X	R ₋₁ -G ₊₁		R/K ₊₆ -G ₋₁	WT
	R ₋₁ -G ₊₁	N ₊₃ -A ₀	R/K ₊₆ -G ₋₁	WT
	C/D/I/T/V ₋₁ -X ₊₁	N ₊₃ =A ₀	R/K ₊₆ -G ₋₁	D
	Q ₋₁ -X ₊₁	S/D/N ₊₃ -A ₀	R/K ₊₆ -G ₋₁	Q
	Q ₋₁ -X ₊₁	S/D/N ₊₃ -A ₀		Q
G/A-G-X	Q/R ₋₁ -G ₊₁	H ₊₃ -G ₀	R/K ₊₆ -G ₋₁	WT
	Q/R ₋₁ -X ₊₁	H ₊₃ -G ₀		WT

⁸ An “=” sign means a double hydrogen bond and a “-“ sign means a single hydrogen bond.

3.3.3 Minimal set of protein-DNA interactions

The set of interactions capable of modeling the EGR mutants encompass a novel group of five H-bond categories, three atomic desolvation penalties, a hydrophobic desolvation energy, and a water factor that accounts for water accessibility at the binding interface. Chemically similar bonds are assumed to scale according to the relative partial charge of the atoms involved, as established by the AMBER force field (Cornell *et al.* 1995). The origin of each of these interactions is well founded on successful empirical free energies of protein-protein interactions (Lazaridis and Karplus 2000; Davis and Baker 2009), as well as in careful consideration of the modular interactions that characterize the classical C₂H₂ ZF-DNA complex. Thus, the ten relevant interactions are:

- *Five hydrogen bond categories:* (i) The bidentate H-bond interactions between Arg and Guanine, $R=G$, which is also assumed to be twice the strength of a single $K-G$ H-bond, as well as that of any sc H-bond to the bb ; (ii) the bidentate H-bond interaction $Q=A$, assumed to have the same strength as $N=A$, while the strength of individual H-bonds for these side chains are partitioned according to their partial charges; (iii) the $S-C$ H-bond, used to extrapolate Ser, Thr and Cys H-bonds (e.g., $S-T$, $T-T$, $T-C$) and related interactions; (iv) the $D-C$ H-bond, used to estimate all bonds involving Asp side chains. For instance, the strength of $D-A$ is $0.97 \times D-C$, where the ratio of AMBER partial charges of donors of C and A is $C.N_4H/A.N_6H = 0.42/0.43 = 0.97$; and, (v) the $H-G$ bond that also determines all His H-bonds with other DNA bases.
- *Three atomic desolvation penalties* (Figure 3.9A): Polar groups buried at the binding interface trigger costly desolvation penalties if their H-bonds are not

properly matched. These desolvation penalties are: (vi) δ_{OD} for a free sc-oxygen at the binding interface or an unmatched sc-oxygen from Gln or Asn; (vii) δ_{NH_2} for unmatched NH_2 sc-groups, and half this penalty $\delta_{NH} = \delta_{NH_2}/2$ for unmatched NH sc-groups; and, (viii) δ_{HB} for burying a sc-sc H-bond between any two interface residues at positions -1, +2, +3 or +6 leaving at least one oxygen unmatched. This penalty is different from atomic desolvations because of the extra entropy loss of trapping two side chains. It is also worth noting that unless solvated by crystal waters sc-sc H-bonds are highly penalized in protein-protein interfaces as well (Bueno and Camacho 2007).

- *Hydrophobic desolvation*: If a non-polar group is buried at the binding interface, an attractive (ix) desolvation energy δ_{NP} is assumed.
- *Water factor* (Figure 3.9B): Water accessibility at the binding interface is modeled by a unique (x) water factor λ_w , corresponding to the fraction by which the transition state of H-bonds exposed to a few extra waters is decreased. Note that this factor only applies to partially solvated bonds, fully solvent exposed H-bonds do not contribute to the binding free energy at all. Hence, the strength of an H-bond exposed to extra waters is reduced to $(1 - \lambda_w)$, whereas an H-bond that gets desolvated is strengthened by $1/(1-\lambda_w)$. The same factor λ_w is used for all H-bonds, as well as for the desolvation penalties that are weakened in the presence of extra solvent.

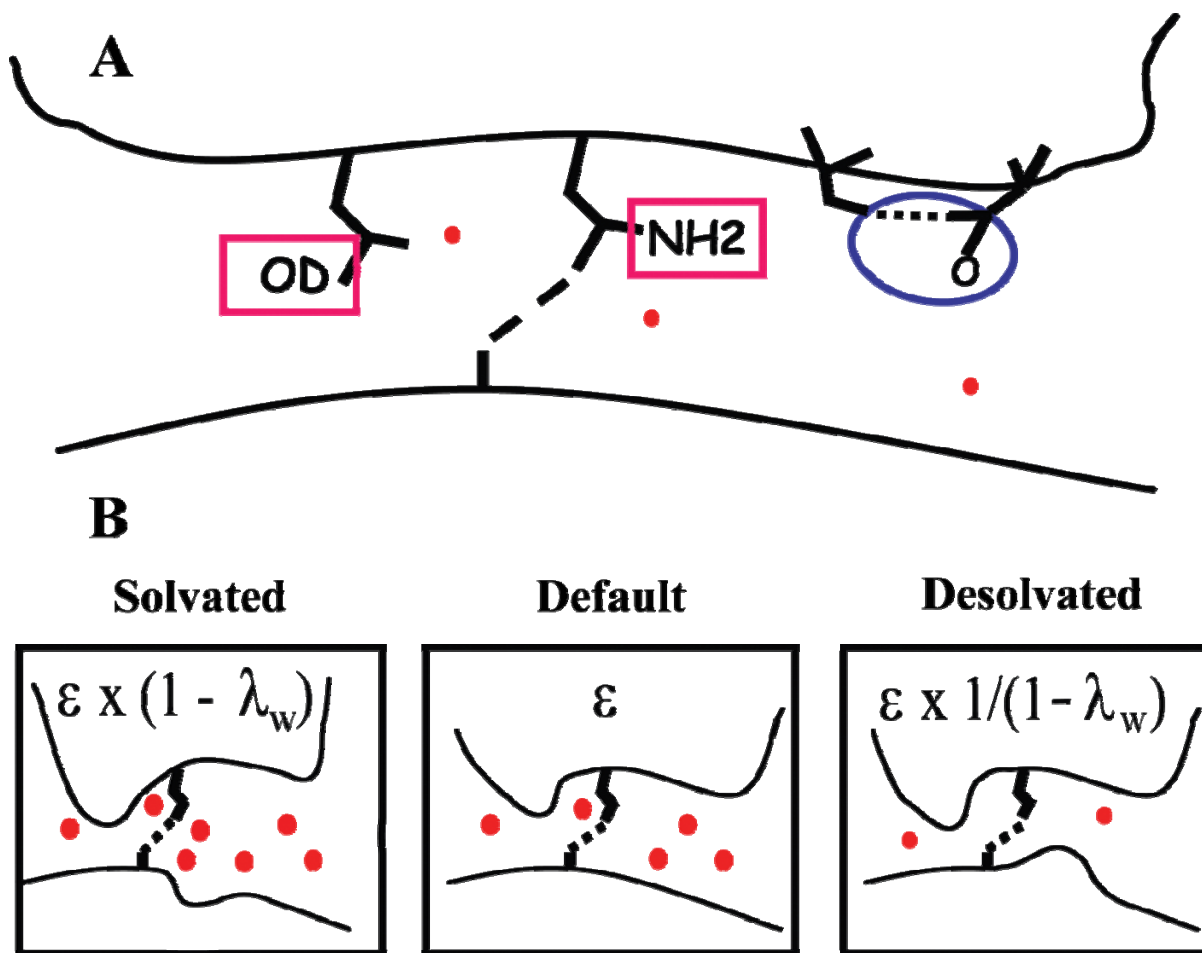


Figure 3.9: Sketches illustrating atom desolvation penalties and solvation effects at the protein (top)-DNA (bottom) binding interface.

H-bonds are indicated as dashed lines and filled spheres correspond to water molecules. **A.** From left to right, oxygen (δ_O) and NH_2 (δ_{NH_2}) desolvation penalty arises when atom does not form an H-bond with protein or DNA. Intra-molecular H-bond desolvation penalty (δ_{HB}) is assessed when oxygen groups are left unmatched. **B.** Effect of solvation on the strength of inter-molecular H-bonds. **Center.** Default binding interface with ϵ as the effective H-bond strength. The illustration also reflects the fact that bonds required a surface to lay on. **Left.** Solvated binding interface. Competing water molecules are weakening the inter-molecular H-bond by a factor of λ_w . **Right.** Desolvated binding interface increases H-bond strength by a factor of $1/(1-\lambda_w)$.

3.3.4 Protein-DNA interaction code

Based on the inter-molecular models for FI mutants in (Liu and Stormo 2005), the interaction code, listed in Table 3.5, is decoded using Eqn. 4. These interactions then determine the lowest free energy models for all mutants sketched in Figure 3.10 (see Model Prediction section).

Namely:





















- Comparing the RDNR/GCG mutant with WT FI defines the strength of δ_{NH_2} as 0.95 kcal/mol, such that $e^{(\delta_{\text{NH}_2}/kT)}$ matches the observed 5 fold drop in affinity.
- The $R=G$ bidentate H-bond is decoded from QDER/GCA as 2.66 kcal/mol matching the 90 fold decrease in affinity with respect to WT.
- There are several models that trigger a δ_{OD} desolvation penalty. We chose DDER/GCA to quantify this bond, since MD shows that D_{+2} in DDER does not form bonds with other atoms, whereas H-bond interactions between Q_{-1} and D_{+2} in QDER are likely to affect the strength of the desolvation penalty. Of course, these subtle dynamic differences are not quantified here.
- The $Q-C$ and $D-C$ bonds are now easily extracted from QDER/GCC and DDER/GCC, respectively. Moreover, the similar chemistry of Q and N side chains led us to assume that the bidentate interactions $Q=A$ and $N=A$ had the same strength.
- The penalty for burying an H-bond is based on QDNR/GCC.
- The water factor was defined by the RDER/GAG model (see below).
- The $S-C$ bond was decoded based on QGSR/GCA crystal structure (Elrod-Erickson *et al.* 1998) and affinity measurements of Kang (Kang 2007). Based on the relative affinity of this mutant and the wild type protein, $\epsilon_{S-C} = 0.93$ kcal/mol. Similarly, based on partial charges, the ratio of the strength between $S-A$ and $S-C$ H-bonds is 0.93, resulting in $\epsilon_{S-A} = 0.87$ kcal/mol. Threonine bonds with A/G/C are also extrapolated based on $S-C$.

- The $H-G$ bond was predicted based on FIII mutant RDHR/GGG, which differs in an additional $H-G$ bond, the loss of a bb-phosphate contact at Pos. +1 and the removal of extra waters weakening one of the H-bonds between R_{+6} and G_{-1} by the stacking $H-G$ bond. The relative affinity between these two sequences results in $\epsilon_{H-G} = 0.31$ kcal/mol for $H-G$.
- FIII includes unique mutants involving up to 5 different possible hydrophobic contacts, and two aromatic residues in the recognition helix. Here, we assume a single parameter δ_{NP} to describe the burying of a non-polar group at the binding interface. Comparing ISNR/GAT and QSNR/GAA in FIII, we predict $\delta_{NP} = -0.61$ kcal/mol.
- Finally, a reported mutation (Bae *et al.* 2003) in WT FIII of Lys to Asn at Pos. +5, leads to a 1.4 fold decrease in affinity with respect to WT, or a penalty of $\delta_{N+5} = 0.2$ kcal/mol.

It is important to emphasize that we read the values of these specific H-bond parameters and atom desolvation penalties directly from the experimental data points mentioned above without fitting to the whole data. Figure 3.11 sketches the data points used to extract the parameters.

As expected, the $R=G$ bidentate H-bond results in the strongest protein-DNA interaction, followed by the $N=A/Q=A$, $S-C$, $C-C$, $H-G$, and $D-C$. A striking validation for these interactions is that an unconstrained minimization of FI models with arbitrary energies failed to improve the error function in Eqn. 2.12.

Table 3.5: Optimized effective hydrogen bond potentials and desolvation penalties (kcal/mol).

	Energy (kcal/mol)			
Desolvation Penalties		δ_{OD}	0.28	
		δ_{NH_2}	0.95	 δ_{NH} 0.48
		δ_{HB}	0.57	
	+	δ_{NP}	0.61	
		δ_{+5}	0.19	applies to only K+5N mutation in FIII
		λ_W	0.41	
Hydrogen Bonds		R=G	2.67	
		R-G	1.33	 K-G
			1.33	 Sc-bb(DNA)
		Solvated	0.78	 Sc-bb: $(1-\lambda_W)$ R-G
		Q=A	0.96	 Q-G N-A N-C
		N=A		
		S-C	0.93	 S-A S-T S-G
				 T-T C-C
		H-G	0.31	 H-T
		D-C	0.31	 D-A

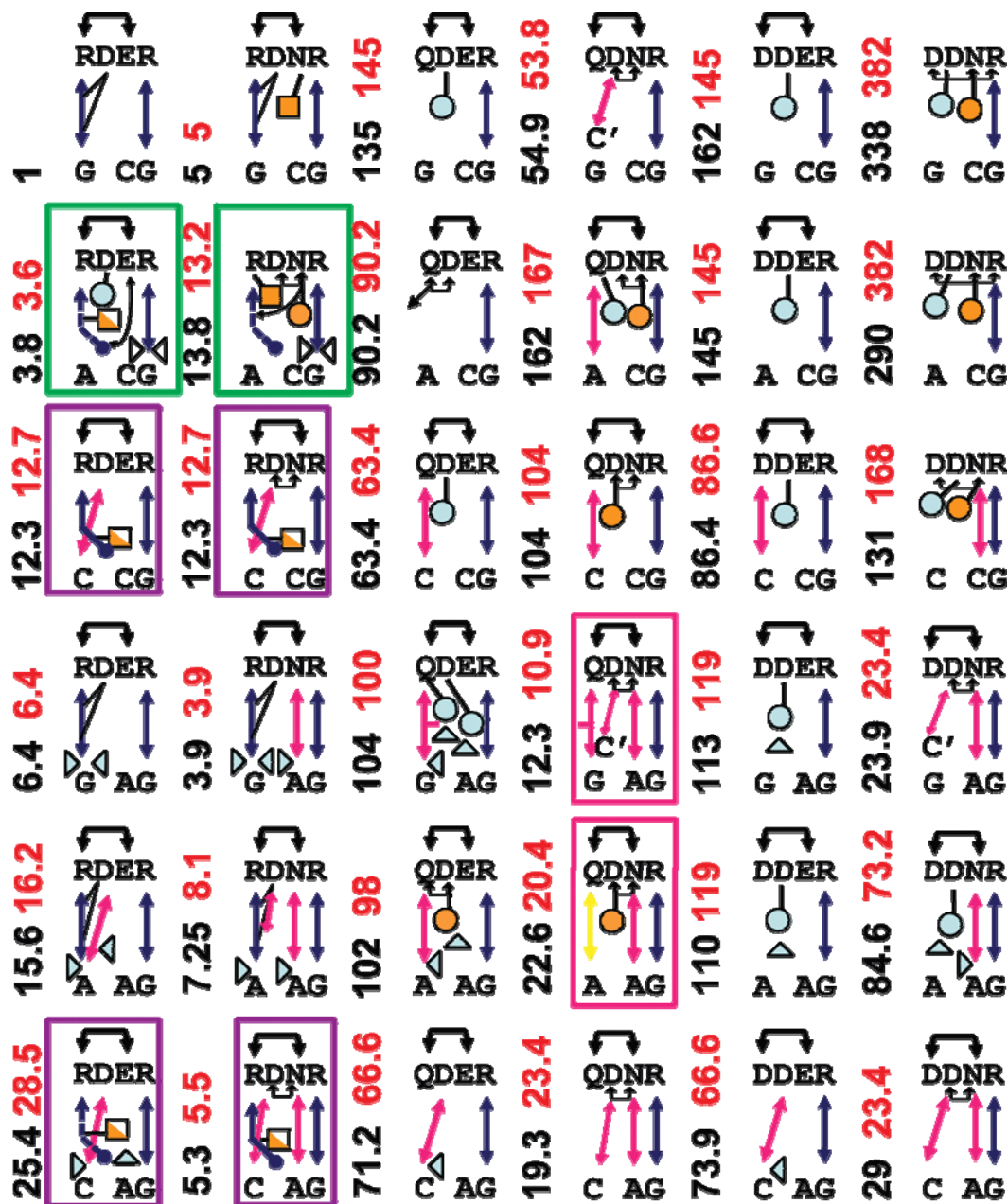


Figure 3.10: Predicted complex structures for 6 EGR FI mutants and 6 DNA binding site sequences. Arrows indicate H-bonds, and dashed arrows denote H-bonds to bb phosphates. Intra-molecular H-bonds are indicated by back arrows/lines. Blue spheres show the desolvation penalties for sc oxygens (δ_o). Orange spheres show the desolvation penalty for intra-molecular H-bonds (δ_{HB}). Rectangles are the desolvation penalties for NH_2 groups (δ_{NH_2}). Filled/open triangles point to the interaction that is solvated/desolvated at the binding interface. The numbers on the left of each model indicate the experimental (black) and predicted (red) change in affinity with respect to RDER/GCG wild type structure shown in upper left corner. Predictions can easily be reproduced by decoding interactions using Table 3.5 and Eqn. 2.11. All complexes are built on top of the WT FI crystal, unless shown inside a rectangle. Red/green/magenta rectangles denote those complexes whose homology models were superimposed to Q/BB1/BB2 binding modes, respectively.

3.3.5 DNA structure and the role of water in additivity

From a biophysical point of view, the most important contribution of this work is the quantitative prediction of the water factor mediating protein interactions. This prediction was borne out of the detailed analysis of the middle cytosine (C_0) mutation to adenine (A_0) resulting in a 5'-GAG-3' tri-nucleotide bound to WT RDER. Despite the apparent neutral character of this mutation, which should still result in the same inter-molecular H-bonds as WT complex, the observed 6.4 fold decrease in affinity says otherwise. Careful analysis of the predicted model shows that the only difference between these structures is a larger cavity on the GAG mutant that accommodates at least two more water molecules in the binding interface of RDER next to the $R_{-1}=G_{+1}$ bond between helix positions -1 and +3 (see Figure 3.12). Consistent with the notion that water molecules weaken H-bonds, the extra waters of partially solvated bonds are modeled by weakening the corresponding bond by a water factor $\lambda_w = 0.41$ —e.g., $R_{-1}=G_{+1}$ (extra waters) $\equiv (1 - \lambda_w) \times R_{-1}=G_{+1}$ (WT), leading to the experimental 6.4 fold decrease in affinity.

Further analysis of our models showed that any two purines at DNA positions -1 and 0 build a cavity, which might be filled by either protein or water. For instance, an $H-G$ or $N=A$ H-bonds at Pos. +3 or a sc-sc H-bond between D_{+2} and N_{+3} block the presence of extra waters (not shown). Although our modeling of water molecules is crude, the assumption that cavities large enough to fit water molecules will do so is well founded (Ernst *et al.* 1995).































	δ_{NH2}	5	5			FI
	R=G	90.2	90.2			FI
	δ_{OD}	162	145			FI
	Q-C	104	104			FI
	D-C	86.4	88.6			FI
	δ_{HB}	104	104			FI
	λ_{W}	6.4	6.4			FI
	S-C	0.04	0.04			QGSR/GCA crystal and Kang affinity data
	H-G	0.9	0.9			FIII
	δ_{NP}	0.1	1.1			FIII

Figure 3.11: Extraction of the ten fundamental interaction parameters.

Each row shows the specific parameter extracted from the experimental value, the experimental data point and lowest energy models used. The first column shows the extracted parameter. 2nd and 3rd columns show the compared and reference models. Last column shows the binding data used. FI denotes finger I data points from Liu and Stormo (Liu and Stormo 2005) and FII denotes finger III data points from Bae et al. (Bae *et al.* 2003). For all interaction parameters only one relative affinity data is used to read each individual parameter. ISNR is compared to QSNR since the only difference is in Pos. -1. Symbols and color code follow Table 3.5.

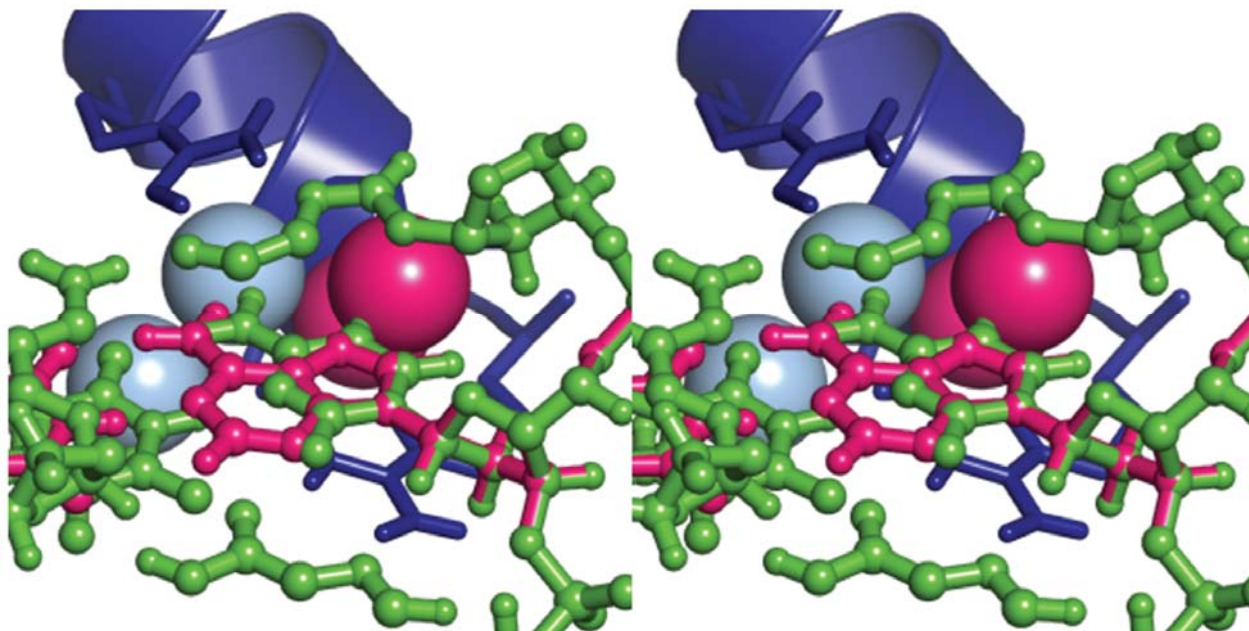


Figure 3.12: Rearrangement of waters at the protein-DNA interface due to cytosine to adenine mutation. FI of EGR is shown in dark blue. Green ball and sticks show crystal GCG triplet. Mutated A_0 is shown as pink ball and sticks. Cyan spheres are the waters at the interface found in the crystal of WT (GCG) complex. Pink spheres are modeled extra waters at the interface of EGR FI-GAG complex. Note the shift in the base due to C->A mutation allowing waters to fit in.

3.3.6 Assessing water factor in binding modes

Relative to WT, FI mutants DSNR/GAC (D mode), QGSR/GAC (Q mode), and RDER/GCA (BB1 mode) have been shown to be remarkably more stable than expected, i.e., -1.7, -1.9, and 0.4 kcal/mol, respectively (Kang 2007). This extra stability is fully rationalized by the missing crystal waters observed relative to WT in Pabo's crystal structures (Elrod-Erickson *et al.* 1998). Specifically, a water molecule that sits below the $R_{+6}=G_{-1}$ H-bonds in WT is not present in any of these mutants. Consistent with our water factor, the desolvation of the $R_{+6}=G_{-1}$ bonds increases the strength of the bonds by $1/(1-\lambda_w)$ to 4.54 kcal/mol. For BB1 the key water is instead

coordinated between the C_0 and A_{+1} bases, suggesting that nucleotides that disrupt these bonds should not be able to sequester this critical water away from the $R=G$ H-bond. The latter is consistent with our prediction that RDER/GCC and RDNR/GCC do not enhance the $R_{+6}=G_{-1}$ bond. Similarly, D and Q modes show that the middle bonds $N_{+3}.OD-A_0$ and $S_{+3}-C_0$, respectively, are also desolvated with respect to WT, triggering a $1/(1-\lambda_w)$ bond enhancement as well. Hence, based on Table 3.5, we predict $\Delta\Delta G_{\text{Calc}} = -2.05$, -1.92 , and 0.21 kcal/mol for D, Q and BB1 binding modes, respectively. Figure 3.12 shows a diagram of models and affinities for 12 mutants, including these binding modes. Interestingly, the desolvated/enhanced bonds in Q mode are canceled out by the extra waters entailed by dinucleotide purine steps as in GAG.

In summary, we use λ_w as the only factor regulating the excess or decrease of waters trapped at the interface.

Figure 3.13 sketches H-bond networks and shows a direct comparison for the affinities between several FI binding modes. A point of caution is that different experimental conditions can lead to different affinities. Indeed, experiments on the same dataset by (Kang 2007) and (Rebar and Pabo 1994) resulted in some different binding free energies. We chose to compare against the more recent dataset in Ref. (Kang 2007). It is important to point out that both of these experiments have a key mutation with respect to Liu and Stormo (Liu and Stormo 2005) that we predict has a role on the solvation of the $R_{-1}-G_{+1}.N_7$ H-bond. Specifically, beyond the GCG consensus sequence Refs. (Kang 2007), (Rebar and Pabo 1994) have a $C_{+2}A_{+3}$ compared to $C_{+2}T_{+3}$ in Ref. (Liu and Stormo 2005). Structural models suggest that A_{+3} , (complementary strand) protects the $R_{-1}-G_{+1}$ bond better than T_{+3} , preventing waters from clustering around the bond. The predicted models match well Pabo's crystal structures (Elrod-Erickson *et al.* 1998), with the exceptions of RADR/GCG and RADR/GAC (cases for which the experiments in Ref

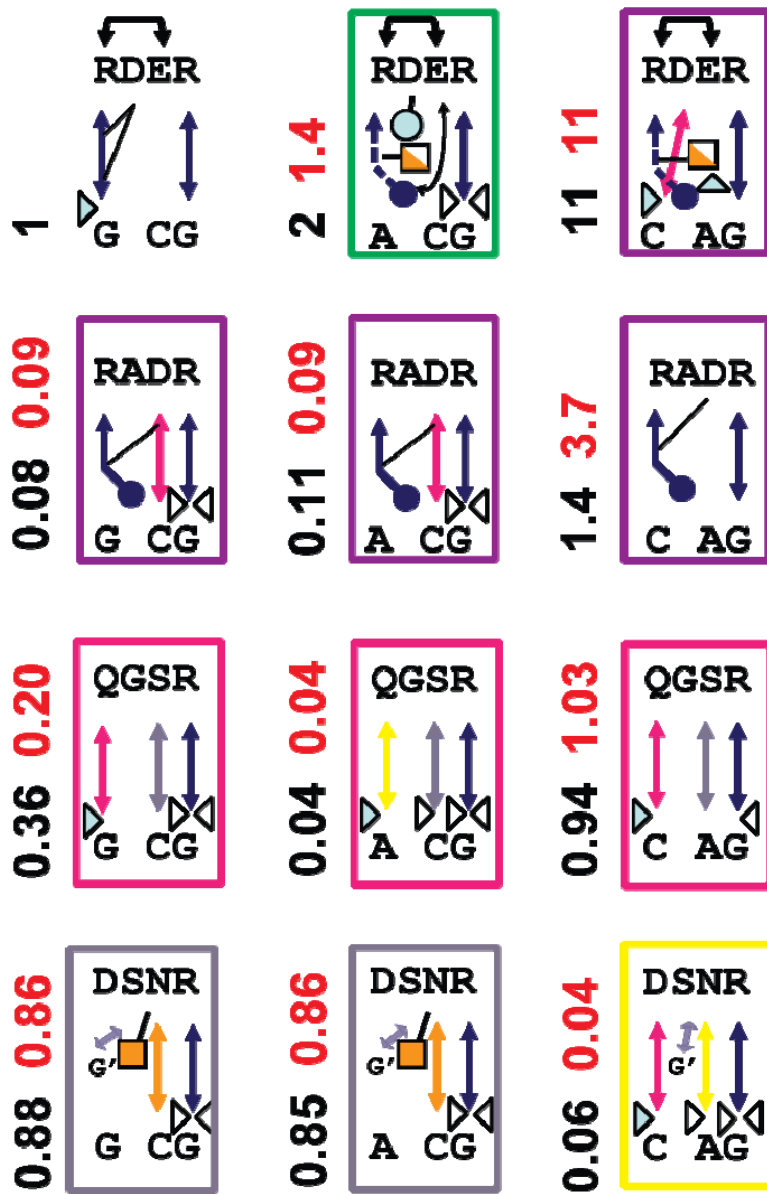


Figure 3.13: Models for 4 EGR FI and 3 DNA binding site sequences.

Arrows and symbols are based on the interactions shown in Table 3.5. Filled/open triangles point to the interaction that is solvated/desolvated at the binding interface. The numbers on the left of each model indicate the experimental (black) (Kang 2007) and predicted (red) change in affinity with respect to RDER/GCG wild type structure shown in upper left corner. Yellow/red/green/magenta rectangles denote complexes in D/Q/BB1/BB2 binding modes, respectively. Models are based on the crystal structures of RDER/GCG (ElrodErickson *et al.* 1996) (wild type), D mode DSNR/GAC (Elrod-Erickson *et al.* 1998), Q mode QGSR/GCA (Elrod-Erickson *et al.* 1998), BB1 mode RDER/GCA (Elrod-Erickson *et al.* 1998), BB2 mode RADR/GCG (Elrod-Erickson *et al.* 1998) and DSNR/GCG (violet boxes; no sequence matched this mode) (Elrod-Erickson *et al.* 1998). Note that $R_{-1}=G_{+1}$ in wild type is partially solvated. This is due to a CA dinucleotide following the EGR binding site (i.e. GCGTGGGCG-CA), which is CT in Liu and Stormo (Liu and Stormo 2005) .

(Kang 2007) and (Rebar and Pabo 1994) also do not agree, and crystals show relatively high B-factors for key sc). For instance, in RADR/GCG, we predict the same binding mode as RADR/GCA or RDER/GCA, i.e., a desolvated $R_{+6}=G_{-1}$ bond leading to a predicted energy of $\Delta\Delta G_{\text{Calc}} = -1.4$ kcal/mol compared to $\Delta\Delta G_{\text{Exp}} = -1.5$ kcal/mol. The problem here is that the RADR/GCG crystal (Elrod-Erickson *et al.* 1998) does not show a desolvated R_{+6} H-bond. Arguably, differences in the crystallization and binding assay conditions might be responsible for this inconsistency. Otherwise, our code simply cannot reconcile this crystal with $\Delta\Delta G_{\text{Exp}}$.

3.3.7 Model prediction for FI

Figure 3.10 shows the corresponding lowest free energy structures, binding affinities and binding modes predicted by the interactions in Table 3.5. If an H-bond is not formed, it is either farther apart than 4 Å, or lead to a higher energy. For instance, as expected, the widely reported conserved *intra*-molecular interaction $R_{-1}=D_{+2}$ (Fairall *et al.* 1993; ElrodErickson *et al.* 1996; Kim and Berg 1996; Segal *et al.* 1999) plays a critical role stabilizing the *inter*-molecular $R_{-1}=G_{+1}$ H-bond. From a physical point of view, D_{+2} protects R_{-1} from a water attack. This complementarity is enforced by the fact that if R_{-1} is not stabilized by D_{+2} , then the unmatched HE hydrogen will trigger a δ_{NH} penalty. We note that R_{+6} is matched by a highly coordinated crystal water.

Most models come down to a straightforward optimal pairing of inter-molecular bonds. Nevertheless, some observations are in order:

- The strong intra-molecular bonds suggested by MD are present in almost all the models. For instance, N or E at Pos. +3 often forms an intra-molecular bond with the bb at Pos. -1 as observed in the WT complex.

- MDs also provide clues for the complementarity of the H-bond network. For example, in QDNR and DDNR, the sc of D_{+2} and D_{+1} , respectively, are the ones forming an H-bond with N_{+3} . Also, D_{+2} can form an H-bond with C_{+1} , in some models but not others. For DDNR/GCG/GCA, MDs show that the strong repulsion between the negatively charged aspartic acids forbids D_{+2} from forming a bond with N_{+3} and, therefore, it cannot reach C_{+1} , as well. The only exceptions are DDNR/GAG/GAC, where the bond between N_{+3} - A_0 stabilizes the intramolecular bond between D_{+2} and N_{+3} , and only then D_{+2} can reach C_{+1} / C_{+1} . Finally, in DDNR/GAC, D_{+2} can reach within 3 Å of C_{+1} .
- In QDER, D_{+2} samples two configurations: One where it is buried deep against A_{+3} , bb, the other reaches to C_{+1} , in GCG. However, this configuration is even closer to the repulsive O_6 group of G_{+2} , resulting in the desolvation penalty shown in the GCG complex. On the other hand, in GCA, while not reaching close enough to form a bond with A_{+1} , it can stay pointing out towards the solvent through its interaction with Q_{-1} , resulting in no desolvation penalty.
- In QDER/GAG, Q makes a bond with $G.O_6$ forcing OE to be pointing inwards. However, with the extra support of an N - A bond in QDNR/GAG, this mutant can further rearrange into the Q binding mode with Q forming a bond with $G.N_7$. This bond now allows the free OE to rotate outwards to the solvent, canceling the OE desolvation penalty.
- Hydrogen bonds to the DNA-bb would have been difficult to predict *de novo*. For FI, the BB1 and BB2 binding modes provide the necessary insight to unravel the

contacts, e.g., one having the bb bond partially solvated (BB1) while the other with normal strength (BB2).

Figure 3.14 shows the predicted binding free energies ($\Delta\Delta G_{\text{Calc}}$) of 35 mutants versus experimental relative affinities ($\Delta\Delta G_{\text{Exp}}$) (Liu and Stormo 2005). The straight line represents the exact match, i.e. $y=x$. Figure 3.10 also highlights the different binding modes used for the different models, which are consistent with crystal contacts of available EGR-like structures. The agreement is quite remarkable considering that Figure 3.10 involves 15 different inter-molecular H-bonds (not counting desolvation or intra-molecular bonds), which here are modeled with only seven decoded energy terms. Interestingly, the largest deviations in the binding energy come from complexes whose desolvation penalties are difficult to assess. For instance, in DDNR/GCG/GCA/GCC complexes, the OD (δ_{OD}) desolvation penalty of D_{+2} is likely to be more solvated in GCA/GCC than in GCG, as also reflected by the relatively weaker affinity of the DDER/GCG mutant. Two subtle observations from the models are: First, the D - A bond in RDNR/GAA is protected from water behind the N - A bond; and, second, the δ_{HB} penalty in Q-mode QDNR/GAA is forced by the close proximity of D and N, leaving no room to break this bond.

3.3.8 Multiple complex models

Given the interactions listed in Table 3.4, Figure 3.10 depicts the H-bonds of the homology models with the lowest binding free energy. Alternative models were also considered (not shown), but resulted in higher free energies. For example, for RDER/GCA, besides the BB1 binding mode, we also predicted a model where R_{-1} forms an inter-molecular H-bond with

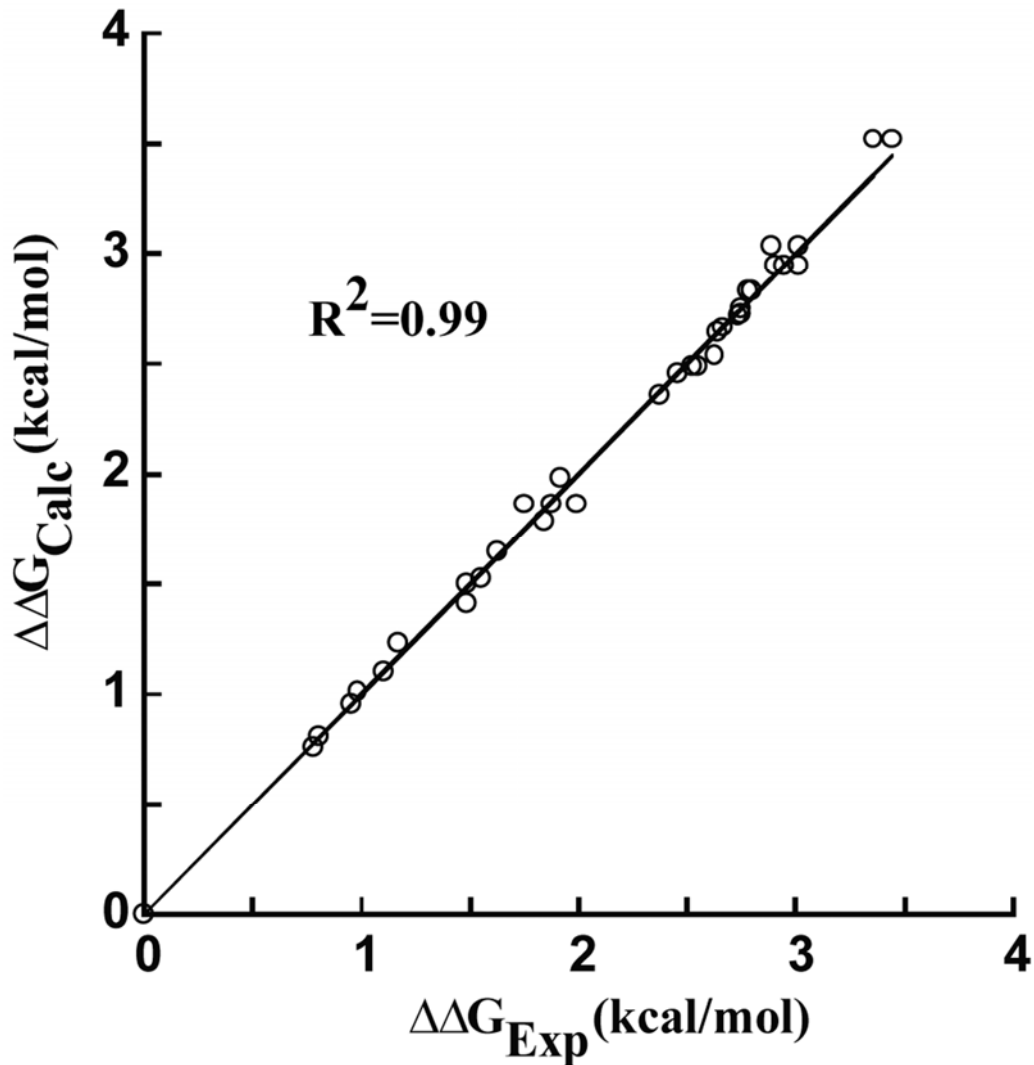


Figure 3.14: Predicted $\Delta\Delta G_{\text{Calc}}$ versus experimental $\Delta\Delta G_{\text{Exp}}$ changes in free energy due to protein and/or DNA mutations. $\Delta\Delta G$ s are computed using Eqn. 2.9. Solid line corresponds the $y=x$ line. Since interaction code is predicted based on experiments, the same error bars apply to both.

$A_{+1}.N_7$, with D_{+2} matching $R_{-1}.HE$ and $R_{-1}.NH_2$ (as in WT) while forming a bond with $A_{+1}.N_6$. The relative affinity of this complex is predicted to be 4.7, higher than the one predicted in Figure 3.10. Interestingly, the crystal structure (Elrod-Erickson *et al.* 1998) also shows a second configuration similar to our model, but with an unusual clash of the hydrogens from R_{-1} and A_{+1} .

3.3.9 Predicting changes in affinities due to mutations in FII and FIII

The binding modes and inter-molecular H-bond networks resolved for FI are assumed to also apply to FII and FIII. One important caveat is that the distribution of water molecules and sc distances to the DNA-bb are not the same. For FII and FIII, the crystal shows a significant number of extra waters at both Pos. +3 and Pos. +6. These waters weakened the bidentate $R_{+6}=G_{-1}$ H-bond interactions, and the solvent exposed H-bond at Pos. +3 (unless either the bond involves a His residue with its ring structure protecting the H-bond from water, or there is a large aromatic ring next to the H-bond that blocks the waters). Indeed, the specificity role of R_{-1} and R_{+6} are somewhat reversed with respect to FI. This can be seen in the structure of FII, where as in FI Pos. +6, the $R_{-1}=G_{+1}$ bonds are protected from a water attack by a highly coordinated (4 bonds) group of crystal waters, preventing the solvation of these bonds. For FIII, $G_{+1}.O_6$ does not have this protection, and a purine sequence solvates the bond between $R_{-1}.NH_2$ and $G_{+1}.O_6$ but not the H-bond to $G_{+1}.N_7$. In Q or D binding mode, the resulting desolvation of the interface in FI (see modes in Figure 3.13) translates into bringing only one of the $R_{+6}-G_{-1}$ bonds to normal strength (the second bond remains partially solvated), as well as desolvating/strengthening H-bonds to $A.N/G.O_6$ or $C.N/T.O_4$. Note that the latter is not desolvated either in FI because the solvent is on the solvent side of Pos. -1, or in a purine sequence that brings extra water next to Pos. -1.

For WT FII, T_{+6} in RDHT appears fully solvated in a cluster of at least 10 water molecules, hence, no desolvation penalty is assessed to this polar group. Similarly, homology models indicate that without an inter-molecular H-bond between Pos. +3 and a base at Pos. 0, providing a contact area, an R_{+6} sc will be surrounded by water molecules destabilizing any possible H-bond since repulsions with R_{-1} of FIII prevent close contacts to the other side of the

pocket. If the middle bond is formed, then we estimate that R_{+6} .HE is always matched by a water molecule (see, e.g., HOH221 in FII and HOH226 in FIII of the WT crystal (ElrodErickson *et al.* 1996)). Note in the WT FIII crystal (ElrodErickson *et al.* 1996), R_{+6} - G_{-1} bonds are partially desolvated by A₋₂. Indeed, we predict that a sequence with either T/C₋₂ should destabilize this bond. Finally, MD simulations strongly suggest that R_{+6} - G_{-1} is destabilized by two consecutive hydroxyl residues (Ser/Thr) at Pos. +2 and +3.

In what follows, we use the recognition code in Table 3.4 and interaction code in Table 3.5 to predict changes in affinities in two independent data sets of EGR mutants: 23 mutants of FII (Segal *et al.* 1999) and 32 human ZF domains swapped with FIII of EGR (Bae *et al.* 2003).

3.3.9.1 Comparison with affinity data from Segal et al. (Segal *et al.* 1999)

Figure 3.15 shows the predicted models for FII mutants. The recognition rules lead to a correlation coefficient of $R^2 = 0.96$ of experimental versus calculated $\Delta\Delta G$ (Figure 3.16A). A strong support for our code is the good agreement obtained for complexes with stacking interactions. Complexes that break this symmetry are somewhat less reliable since it is difficult to fully appreciate the role of solvent. For instance, based on the distance constraints of our models, four K/R mutants at Pos. +1 and +5 are predicted to form partially solvated H-bonds with DNA phosphate groups of strength $(1-\lambda_w) \times R-G = 0.78$ kcal/mol, while HE is expected to be fully solvated. Needless to say, there is no structural validation for these bonds. The above notwithstanding, given that the predicted models are based on *feasible minimum energy configurations* the predicted $\Delta\Delta G$ s should be considered an upper bound on the experimental $\Delta\Delta G$ s.

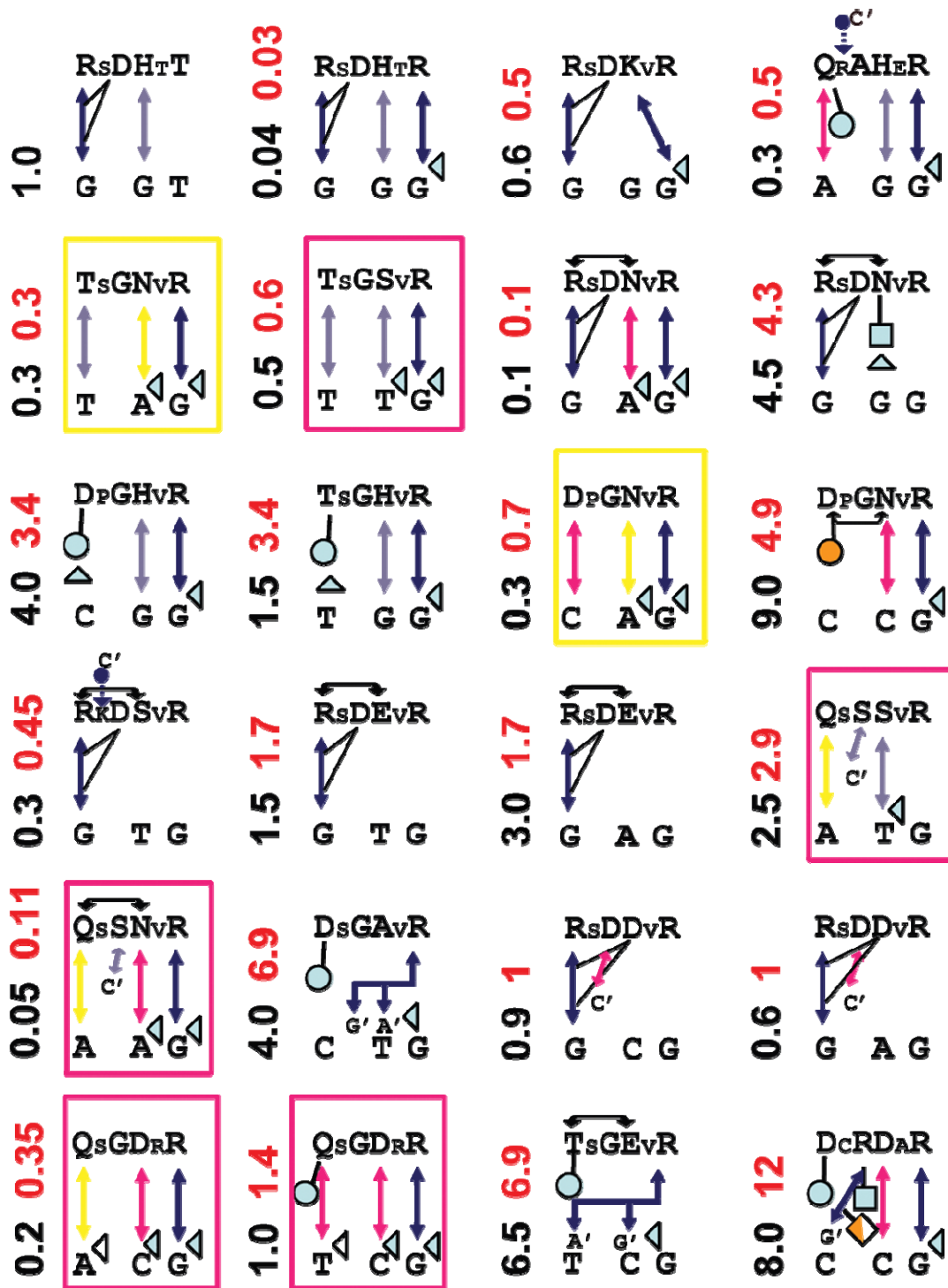


Figure 3.15: Predicted complex structures for FII mutants.

Figure 3.10 and Table 3.5. Homology models built on D binding mode are indicated by a yellow rectangle. WT is indicated in upper left corner.

3.3.9.2 Comparison with affinity data from Bae et al. (Bae et al. 2003).

Many mutations on the human zinc fingers swapped with FIII in EGR do not interact directly with DNA and, therefore, are ignored despite the fact that they might have an indirect effect in affinity. We identified three sites whose mutations can change $\Delta\Delta G$. These are Ala, Ser, and Lys at positions -2, +1 and +5, respectively. Mutations of A₋₂ to an H-bond donor can form a DNA-bb bond; S₊₁ forms an H-bond with DNA-bb in WT, but K/R/H₊₁ sc mutations are too long and are predicted to be fully solvated; and, a K₊₅ mutation to N₊₅ has been experimentally shown to decrease the affinity of WT by 0.195 kcal/mol (Bae et al. 2003). Besides these unique bonds and already mentioned solvation caveats, the allowed inter-molecular networks are the same as FII and FI.

Figure 3.17 shows the predicted complexes for FIII proteins, resulting in $R^2 = 0.94$ (Figure 3.16B). Predictions are similar to FII, with the caveat that FIII adds a new class of mutants involving hydrophobic and aromatic residues. We model this new bond with a single parameter, δ_{NP} , to account for non-polar buried residues (see Table 3.5).

3.3.10 Comparing inter-molecular networks across different fingers

The small structural differences between the three fingers are ignored. This allows us to apply the same models to all ZFs. The robustness of the recognition code to screen ZF interactions is then best portrayed by its consistency across ZFs. For example, RDNR/GAG has the same H-bond network in FI and FII, but the role of solvent is reversed between Pos. -1 and +6; all D modes (shown in a yellow box in Figure 3.15 and Figure 3.17) have the same H-bond network but solvations are different —e.g., a mutation of S₊₂ to G and/or a purine sequence can cancel the desolvation of the bond at Pos. -1; all Q modes shown in a red box cancel the desolvation if G₊₂

is mutated to S₊₂; see, also, the similar networks between FII and FIII of RDHR/GGG, QAHR/GGA, QSNR/GAA, QSSR/GTA (in FIII it can reach a DNA-bb phosphate), and so on.

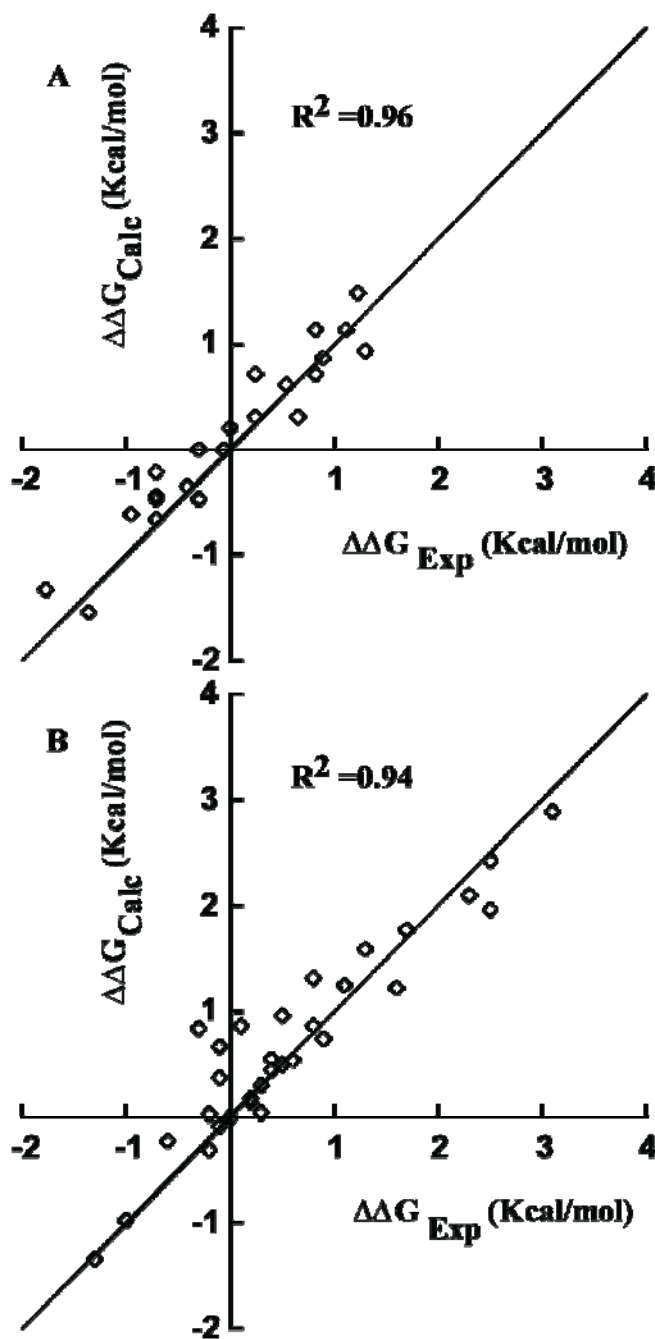


Figure 3.16: Predicted $\Delta\Delta G_{\text{Calc}}$ versus experimental $\Delta\Delta G_{\text{Exp}}$ changes in free energy due to protein and/or DNA mutations. (A) FII (Segal *et al.* 1999) and (B) FIII (Bae *et al.* 2003). As expected, minimum energy models typically resulted in an upper bound of $\Delta\Delta G_{\text{Exp}}$, suggesting the possibility of yet more subtle models or solvent conditions for some sequences.

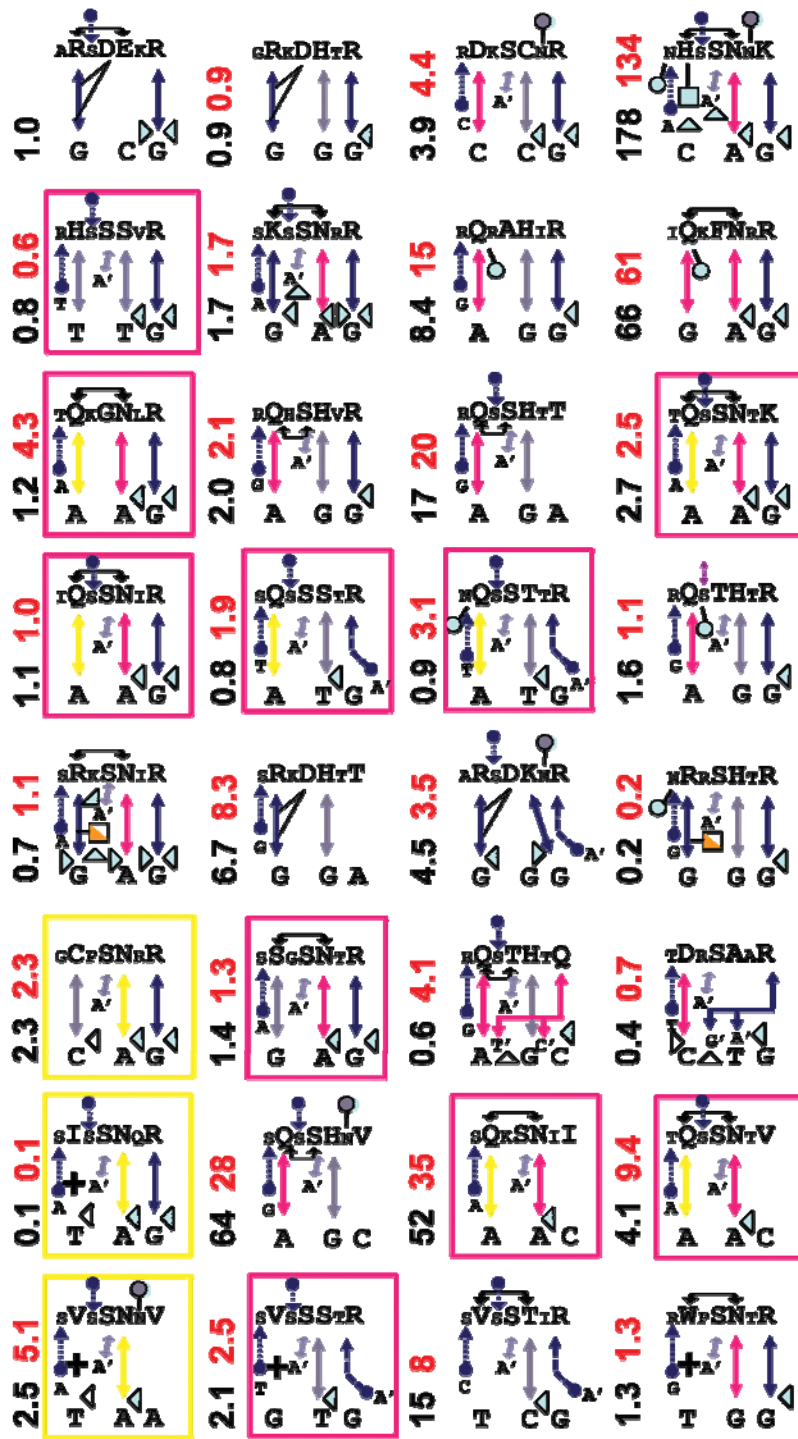


Figure 3.17: Predicted complex structures for FIII experiments.

Symbols are the same as in Figure 3.10 and Table 3.5. Plus signs show desolvation of hydrophobic groups (δ_{NP}). Purple spheres show the desolvation penalty for N_{+5} (δ_{N+5}). WT is in upper left corner.

3.3.11 Limitations

The major shortcomings of the recognition rules described here are: (a) Lack of an accurate tool to model molecular waters at the binding interface limits the application of the water factor in *de novo* H-bond networks; (b) The interaction code is so far well matched to single point mutations of hydrophilic side chains from FI, but it is less clear whether a simple extrapolation of partial charges is well suited to capture the full contribution of side chains that also have non-polar groups on them. Specifically, a few side chains not present in the sequences studied so far, such as Met, Tyr, and even Phe, Trp and Cys that only appear in a few human FIII, have yet to be fully cross-checked. (c) The structural code applies only to ZFs that bind in the classical EGR mode. Hence, we do not address the problem of C₂H₂ ZFs that bind in non-classical modes (see, e.g., Ref. (Pabo and Nekludova 2000)). (d) Induced fit was assumed based on crystal information, but there is no proof that crystals have revealed every possible binding mode. (e) Finally, it is worth mentioning that experimental assays depend on buffer conditions (ions), length of the DNA target, equilibration and so on (Hamilton *et al.* 1998; Segal *et al.* 1999; Bae *et al.* 2003; Liu and Stormo 2005; Kang 2007). Hence, interaction energies might require some re-scaling depending on experimental conditions.

4.0 CONCLUSIONS AND FUTURE DIRECTION

4.1 CONCLUSIONS

4.1.1 Anchoring residues in C₂H₂ family zinc finger- DNA interactions.

Our findings show that, as in protein-protein recognition (Rajamani *et al.* 2004), buried side chains that bury the largest amount of SASA in the bound state are found in rotamer conformations similar to those of the complex *before* encountering the DNA. Hence, they are structurally predisposed to recognize their consensus sequence. Exposed side chains, on the other hand, are not prealigned to their bound conformations and act as latches rearranging to optimize complementary interactions.

4.1.2 Counter ions found in the physiological environment help stabilize the non-specific encounter complex by mimicking DNA backbone phosphates.

We show that the bound-like behavior of the key side chains is intimately related to the ionic environment in which DNA is found. Namely, prior to the formation of the non-specific encounter complex, counter ions take on the role of DNA backbone phosphates on the surface of ZFs ordering side chains responsible for both non-specific and specific binding. The existence of bound-like backbone contacts through phosphate mimicking counter ions rapidly stabilize non-

specific complexes avoiding the large desolvation energy barrier entailed by a random encounter complex. Our evidence suggests that this mimicking is a *necessary condition* for ZFs to bind in the classical EGR geometry, since lacking a counter ion next to the conserved His+7 prevents phosphates from desolvating a large surface area. The latter is consistent with the non-conventional binding modes observed in TFIIA and GLI TFs.

Our results are fully consistent with a binding mechanism that rapidly associates non-specifically proteins and DNA (Iwahara *et al.* 2006) by counter ion specific mediated interactions. This efficient mechanism reconciles association rate constants on the order of the diffusion limit, $10^9 \text{ M}^{-1}\text{s}^{-1}$ for EGR (Hamilton *et al.* 1998).

The observed pre-disposition of bound-like backbone contacts in more than one finger suggest that non-specific encounter complexes might involve more than one ZF, which could allow for partial dissociations and rapid reattachments of individual ZF as they diffuse from phosphate-to-phosphate along the DNA. This simple mechanism reconciles the “sliding” of transcription factors along DNA (Winter *et al.* 1981; von Hippel and Berg 1986; Halford and Marko 2004; von Hippel 2004) by means of non-specific extended desolvation states, while bound-like specificity determinant side chains are ready to stall the 1-D diffusion process at their consensus sequence.

4.1.3 Modeling ZF-DNA interactions.

Understanding the molecular basis and specificity of transcriptional regulation is one of the most important problems in molecular/structural biology. In combination with structural insights from the H-bond networks allowed by the primary sequence of C₂H₂ ZFs and DNA, we developed an

experimentally based methodology to decode the strength of H-bonds and atomic desolvation free energies for protein-DNA interactions. We apply this code to a set of 89 mutants of FI, FII and FIII of EGR, predicting both bound structures and changes in binding affinities. Our results are in good agreement with experiments (Rebar and Pabo 1994; Segal *et al.* 1999; Bae *et al.* 2003; Liu and Stormo 2005; Kang 2007) and known crystals (ElrodErickson *et al.* 1996; Elrod-Erickson *et al.* 1998), and compares well with known approaches.

Based on sequence alone, our approach decoded nine novel interactions and a water modulation factor. All the parameters are experimentally calibrated free energies in kcal/mol. The excellent agreement with experiments strongly supports the basic assumptions of the interaction code. Namely: (a) Short-range interactions ($< 4 \text{ \AA}$) are dominant, suggesting that long-range electrostatics do not play an important role in protein-DNA specificity. Note that this is certainly not the case for non-specific interactions (von Hippel 2007). (b) Desolvation of free polar sc-groups contributes negatively to the binding free energy, but rigid groups such as protein-bb or DNA do not. Crystals suggest that water molecules always patch free DNA H-bonds donors and acceptors. (c) Water screens both electrostatically attractive and repulsive desolvation interactions. (d) Our code does not require an explicit contact energy for water mediated H-bonds.

Induced fit plays a critical role in resolving the recognition code. Our results indicate that binding of ZFs have a relatively larger impact on the protein side (see Figure 2.2), more often in the context of at least three inter-molecular bonds. One exception is the C_0 base in FI DSNR/GCG, whose DNA configuration is clearly shifted by almost 1 \AA relative to GCG in WT. This shift allows N_3 to turn and form an H-bond with C_0 , something that is not possible in

GCG/GCA WT sequences (see Figure 3.10). The above notwithstanding, the structure of DNA alone has a strong dependence on binding by regulating water accessibility.

Our analysis also reveals a novel decomposition of desolvation penalties. Besides atomic desolvation of acceptor (δ_{OD}) and donors (δ_{NH_2}), we find that sc-sc H-bonds that do not match all acceptors carry an extra non-trivial penalty (δ_{HB}). This penalty is consistent with the extra side chain entropy loss entailed by such a bond (Bueno and Camacho 2007). The water factor λ_w is a simple approximation that allows us, for the first time, to quantify the role of molecular water at the binding interface.

Finally, the simplicity of the interaction code motivated us to develop a diagrammatic scheme to represent C₂H₂ zinc fingers interactions with DNA. The scheme depicts physical interactions with symbols that allow a direct reading of the free energies. Hence, researchers not only can reproduce our changes in free energy estimates by subtracting the reference state (i.e., WT interactions), but they can also challenge, improve, disprove the resulting models for each complex.

Applying this code to a set of 89 EGR mutants unveils detailed recognition rules for ZF-DNA complexes and their free energies relative to wild type complex. Some of the rules depend on nucleotides that are +2 nucleotides away from the traditional tri-nucleotide consensus sequence, suggesting that there is still much to be accomplished before revealing all possible protein-DNA interaction networks. Nevertheless, our methodological approach of predicting energies based on *realistic structural models* significantly limits the number of false positives, leaving the door open to further structural refinements. One cannot stress enough the valuable insights that detailed crystallographic studies and careful experiments provided here, which in

combination with molecular modeling resulted in a novel rational approach to decode the recognition code of protein-DNA specific interactions.

4.2 FUTURE DIRECTIONS

4.2.1 Effects of counter-ions in DNA recognition

MD simulations of individual fingers of EGR, TFIIIA and GLI showed the role of counter ions in the dynamics of the critical side chains at the protein – DNA interface. The Cl⁻ ions were found to weakly bind to an electrostatically hot spot formed by three residues that contact DNA. This spot interestingly corresponds to a pocket where DNA phosphates are found most buried in the complex. Binding of the Cl⁻ orders these residues to their bound-like conformations by mimicking the DNA phosphates prior to binding. The next step is to perform similar MD simulations of the remaining ZF-DNA complex structures human YY1 (Houbaviy *et al.* 1996), tramtrack protein (Fairall *et al.* 1993), ZF-TATA box complex (Wolfe *et al.* 2001), designed ZF (Kim and Berg 1996) and designed dimeric ZF chimera (Wolfe *et al.* 2003) to further analyze the effects of counter ions in the formation of the protein-DNA encounter complex by mimicking DNA phosphate groups and directing the protein towards the DNA.

4.2.2 Modeling protein-DNA interactions

In this study, we developed and optimized an experimentally based potential of effective hydrogen bond energies and atom desolvation penalties to predict changes in binding affinities of

ZF TFs. The optimization of the potential was based on the experimental relative affinity data of EGR FI and its mutants binding to 6 different DNA sites. The optimized five H-bond, three atom desolvation, one hydrophobic desolvation and the water solvation factors are universal and can be used in any protein – DNA interaction. Additional H-bond energies can be estimated if needed. One direction would be to predict ZF-DNA complex structures for all the engineered/selected fingers in Zinc Finger Database (Fu *et al.* 2008) using our approach. Another possibility is to design ZFs with a targeted specificity or predict binding sites of ZF TFs based on the simple recognition rules described here.

BIBLIOGRAPHY

- Anderson C.F. and Record M.T., Jr. 1995. Salt-nucleic acid interactions. *Annu.Rev.Phys.Chem.* 46: 657-700.
- Bae K.H., Do Kwon Y., Shin H.C., Hwang M.S., Ryu E.H., Park K.S., Yang H.Y., Lee D.k., Lee Y., Park J., Sun Kwon H., Kim H.W., Yeh B.I., Lee H.W., Hyung Sohn S., Yoon J., Seol W. and Kim J.S. 2003. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat Biotech* 21: 275-280.
- Banci L. 2003. Molecular dynamics simulations of metalloproteins. *Current Opinion in Chemical Biology* 7: 143-149.
- Benos P.V., Lapedes A.S. and Stormo G.D. 2002a. Probabilistic code for DNA recognition by proteins of the EGR family. *Journal of Molecular Biology* 323: 701-727.
- Benos P.V., Bulyk M.L. and Stormo G.D. 2002b. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Research* 30: 4442-4451.
- Beveridge D.L., Dixit S.B., Byun K.S., Barreiro G., Thayer K.M. and Ponomarev S. 2004. Molecular dynamics of DNA and protein-DNA complexes: Progress on sequence effects, conformational stability, axis curvature, and structural bioinformatics. pp.13-64.
- Brown R.S. 2005. Zinc finger proteins: getting a grip on RNA. *Curr.Opin.Struct.Biol.* 15: 94-98.
- Bryne J.C., Valen E., Tang M.H., Marstrand T., Winther O., da P., I, Krogh A., Lenhard B. and Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 36: D102-D106.
- Bueno M. and Camacho C.J. 2007. Acidic groups docked to well defined wetted pockets at the core of the binding interface: a tale of scoring and missing protein interactions in CAPRI. *Proteins* 69: 786-792.
- Bujalowski W. 2006. Thermodynamic and kinetic methods of analyses of protein-nucleic acid interactions. From simpler to more complex systems. *Chem.Rev.* 106: 556-606.

- Bulyk M.L., Gentalen E., Lockhart D.J. and Church G.M. 1999. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnology* 17: 573-577.
- Bulyk M. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biology* 5: 201.
- Bussemaker H.J., Foat B.C. and Ward L.D. 2007. Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules. *Annual Review of Biophysics and Biomolecular Structure* 36: 329-347.
- Cathomen T. and Keith Joung J. 2008. Zinc-finger Nucleases: The Next Generation Emerges. *Mol Ther* 16: 1200-1207.
- Cheatham T.E., III 2004. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* 14: 360-367.
- Choo Y. and Klug A. 1993. A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA. *Nucleic Acids Res.* 21: 3341-3346.
- Contreras-Moreira B. and Collado-Vides J. 2006. Comparative footprinting of DNA-binding proteins. *Bioinformatics* 22: e74-e80.
- Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W. and Kollman P.A. 1995. A 2Nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* 117: 5179-5197.
- Davis I.W. and Baker D. 2009. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *Journal of Molecular Biology* 385: 381-392.
- Debye P. and Huckel E. 1923. Zur Theorie der Elektrolyte. *Physik.Z.* 24: 185-206.
- Deplancke B., Dupuy D., Vidal M. and Walhout A.J.M. 2004. A Gateway-Compatible Yeast One-Hybrid System. *Genome Research* 14: 2093-2101.
- Dragan A.I., Li Z., Makeyeva E.N., Milgotina E.I., Liu Y., Crane-Robinson C. and Privalov P.L. 2006. Forces driving the binding of homeodomains to DNA. *Biochemistry* 45: 141-151.
- Dragan A.I., Liggins J.R., Crane-Robinson C. and Privalov P.L. 2003. The energetics of specific binding of AT-hooks from HMGA1 to target DNA. *J.Mol.Biol.* 327: 393-411.
- Dragan A.I., Read C.M., Makeyeva E.N., Milgotina E.I., Churchill M.E., Crane-Robinson C. and Privalov P.L. 2004. DNA binding and bending by HMG boxes: energetic determinants of specificity. *J.Mol.Biol.* 343: 371-393.
- Ebright R.H., Ebright Y.W. and Gunasekera A. 1989. Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher

affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Res.* 17: 10295-10305.

- Elrod-Erickson M., Benson T.E. and Pabo C.O. 1998. High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger DNA recognition. *Structure with Folding & Design* 6: 451-464.
- Elrod-Erickson M., Rould M.A., Nekludova L. and Pabo C.O. 1996. Zif268 protein-DNA complex refined at 1.6 angstrom: A model system for understanding zinc finger-DNA interactions. *Structure* 4: 1171-1180.
- Endres R.G., Schulthess T.C. and Wingreen N.S. 2004. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins-Structure Function and Bioinformatics* 57: 262-268.
- Engler L.E., Welch K.K. and Jen-Jacobson L. 1997. Specific binding by EcoRV endonuclease to its DNA recognition site GATATC. *J.Mol.Biol.* 269: 82-101.
- Ernst J.A., Clubb R.T., Zhou H.X., Gronenborn A.M. and Clore G.M. 1995. Demonstration of Positionally Disordered Water Within A Protein Hydrophobic Cavity by Nmr. *Science* 267: 1813-1817.
- Fairall L., Schwabe J.W., Chapman L., Finch J.T. and Rhodes D. 1993. The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* 366: 483-487.
- Frankel A.D., Berg J.M. and Pabo C.O. 1987. Metal-Dependent Folding of a Single Zinc Finger from Transcription Factor IIIA. *Proceedings of the National Academy of Sciences* 84: 4841-4845.
- Fried M.G. and Stickle D.F. 1993. Ion-exchange reactions of proteins during DNA binding. *Eur.J.Biochem.* 218: 469-475.
- Fu F., Sander J.D., Maeder M., Thibodeau-Beganny S., Joung J.K., Dobbs D., Miller L. and Voytas D.F. 2008. Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res.*
- Gamsjaeger R., Liew C.K., Loughlin F.E., Crossley M. and Mackay J.P. 2007. Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends in Biochemical Sciences* 32: 63-70.
- Garvie C.W., Pufall M.A., Graves B.J. and Wolberger C. 2002. Structural analysis of the autoinhibition of Ets-1 and its role in protein partnerships. *J.Biol.Chem.* 277: 45529-45536.
- Garvie C.W. and Wolberger C. 2001. Recognition of Specific DNA Sequences. *Molecular Cell* 8: 937-946.

- Greisman H.A. and Pabo C.O. 1997. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* 275: 657-661.
- Gromiha M.M., Siebers J.G., Selvaraj S., Kono H. and Sarai A. 2005. Role of inter and intramolecular interactions in protein-DNA recognition. *Gene* 364: 108-113.
- GuhaThakurta D. 2006. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Research* 34: 3585-3598.
- Halford S.E. and Marko J.F. 2004. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* 32: 3040-3052.
- Hamilton T.B., Borel F. and Romaniuk P.J. 1998. Comparison of the DNA Binding Characteristics of the Related Zinc Finger Proteins WT1 and EGR1. *Biochemistry* 37: 2051-2058.
- Havranek J.J., Duarte C.M. and Baker D. 2004. A simple physical model for the prediction and design of protein-DNA interactions. *Journal of Molecular Biology* 344: 59-70.
- Holbrook J.A., Tsodikov O.V., Saecker R.M. and Record M.T. 2001. Specific and non-specific interactions of integration host factor with DNA: Thermodynamic evidence for disruption of multiple IHF surface salt-bridges coupled to DNA binding. *Journal of Molecular Biology* 310: 379-401.
- Houbaviy H.B., Usheva A., Shenk T. and Burley S.K. 1996. Cocystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl. Acad. Sci. U.S.A* 93: 13577-13582.
- Hubbard S.J., Campbell S.F. and Thornton J.M. 1991. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* 220: 507-530.
- Iwahara J., Zweckstetter M. and Clore G.M. 2006. NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proceedings of the National Academy of Sciences of the United States of America* 103: 15062-15067.
- Jamieson A.C., Miller J.C. and Pabo C.O. 2003. Drug discovery with engineered zinc-finger proteins. *Nature Reviews Drug Discovery* 2: 361-368.
- Jayaram B., McConnell K., Dixit S.B., Das A. and Beveridge D.L. 2002. Free-energy component analysis of 40 protein-DNA complexes: A consensus view on the thermodynamics of binding at the molecular level. *Journal of Computational Chemistry* 23: 1-14.
- Jen-Jacobson L. 1997. Protein-DNA recognition complexes: conservation of structure and binding energy in the transition state. *Biopolymers* 44: 153-180.

- Jen-Jacobson L., Engler L.E. and Jacobson L.A. 2000. Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure* 8: 1015-1023.
- Jen-Jacobson L., Kurpiewski M., Lesser D., Grable J., Boyer H.W., Rosenberg J.M. and Greene P.J. 1983. Coordinate ion pair formation between EcoRI endonuclease and DNA. *J.Biol.Chem.* 258: 14638-14646.
- Joung J.K., Ramm E.I. and Pabo C.O. 2000. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7382-7387.
- Kang J.S. 2007. Correlation between functional and binding activities of designer zinc-finger proteins. *Biochemical Journal* 403: 177-182.
- Kaplan T., Friedman N. and Margalit H. 2005. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput.Biol.* 1: e1.
- Kim C.A. and Berg J.M. 1996. A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. *Nat.Struct.Biol.* 3: 940-945.
- Kim J.S. and Pabo C.O. 1997. Transcriptional repression by zinc finger peptides - Exploring the potential for applications in gene therapy. *Journal of Biological Chemistry* 272: 29795-29800.
- Kim J.S. and Pabo C.O. 1998. Getting a handhold on DNA: Design of poly-zinc finger proteins with femtomolar dissociation constants. *Proceedings of the National Academy of Sciences of the United States of America* 95: 2812-2817.
- Klug A. 2005. Towards therapeutic applications of engineered zinc finger proteins. *FEBS Lett.* 579: 892-894.
- Kono H. and Sarai A. 1999. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* 35: 114-131.
- Kozlov A.G. and Lohman T.M. 1998. Calorimetric studies of E-coli SSB protein single-stranded DNA interactions. Effects of monovalent salts on binding enthalpy. *Journal of Molecular Biology* 278: 999-1014.
- Kozlov A.G. and Lohman T.M. 2006. Effects of monovalent anions on a temperature-dependent heat capacity change for Escherichia coli SSB tetramer binding to single-stranded DNA. *Biochemistry* 45: 5190-5205.
- Lafontaine I. and Lavery R. 2000. ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers* 56: 292-310.
- Laity J.H., Dyson H.J. and Wright P.E. 2000. DNA-induced [alpha]-helix capping in conserved linker sequences is a determinant of binding affinity in Cys2-His2 zinc fingers. *Journal of Molecular Biology* 295: 719-727.

- Lavery R. 2005. Recognizing DNA. *Q.Rev.Biophys.* 38: 339-344.
- Lazaridis T. and Karplus M. 2000. Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology* 10: 139-145.
- Leach A.R. 1999. *Molecular Modeling Principles and Applications*. Pearson Education Limited, Essex.
- Lee K.H., Xie D., Freire E. and Amzel L.M. 1994. Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins* 20: 68-84.
- Lee M.S., Gippert G.P., Soman K.V., Case D.A. and Wright P.E. 1989. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* 245: 635-637.
- Liu J.J. and Stormo G.D. 2005. Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *Bmc Bioinformatics* 6: 176.
- Liu X., Noll D.M., Lieb J.D. and Clarke N.D. 2005a. DIP-chip: Rapid and accurate determination of DNA-binding specificity. *Genome Research* 15: 421-427.
- Liu Z., Mao F., Guo J.T., Yan B., Wang P., Qu Y. and Xu Y. 2005b. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Res* 33: 546-558.
- Lu D. and Klug A. 2007. Invariance of the zinc finger module: A comparison of the free structure with those in nucleic-acid complexes. *Proteins-Structure Function and Bioinformatics* 67: 508-512.
- Luscombe N.M., Laskowski R.A. and Thornton J.M. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research* 29: 2860-2874.
- Ma P.C.M., Rould M.A., Weintraub H. and Pabo C.O. 1994. Crystal structure of MyoD bHLH domain-DNA complex: Perspectives on DNA recognition and implications for transcriptional activation. *Cell* 77: 451-459.
- MacKerell A.D., Bashford D., Bellott M., Dunbrack R.L., Evanseck J.D., Field M.J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau F.T.K., Mattos C., Michnick S., Ngo T., Nguyen D.T., Prodhom B., Reiher W.E., Roux B., Schlenkrich M., Smith J.C., Stote R., Straub J., Watanabe M., Wiorkiewicz-Kuczera J., Yin D. and Karplus M. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* 102: 3586-3616.
- Magenat L., Blancafort P. and Barbas C.F. 2004. In vivo selection of combinatorial libraries and designed affinity maturation of polydactyl zinc finger transcription factors for

ICAM-1 provides new insights into gene regulation. *Journal of Molecular Biology* 341: 635-649.

- Mahony S., Auron P.E. and Benos P.V. 2007. DNA familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. *Plos Computational Biology* 3: 578-591.
- Man T.K., Yang J.S.W. and Stormo G.D. 2004. Quantitative modeling of DNA-protein interactions: effects of amino acid substitutions on binding specificity of the Mnt repressor. *Nucleic Acids Research* 32: 4026-4032.
- Man T.K. and Stormo G.D. 2001. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research* 29: 2471-2478.
- Mandel-Gutfreund Y. and Margalit H. 1998. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res* 26: 2306-2312.
- Mandel-Gutfreund Y., Schueler O. and Margalit H. 1995. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol. Biol.* 253: 370-382.
- Manning G.S. 1977. Limiting laws and counterion condensation in polyelectrolyte solutions : IV. The approach to the limit and the extraordinary stability of the charge fraction. *Biophysical Chemistry* 7: 95-102.
- Miller J., McLachlan A.D. and Klug A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*. *EMBO J.* 4: 1609-1614.
- Moraitis M.I., Xu H. and Matthews K.S. 2001. Ion Concentration and Temperature Dependence of DNA Binding: Comparison of PurR and LacI Repressor Proteins. *Biochemistry* 40: 8109-8117.
- Morozov A.V., Havranek J.J., Baker D. and Siggia E.D. 2005. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33: 5781-5798.
- Nolte R.T., Conlin R.M., Harrison S.C. and Brown R.S. 1998. Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc. Natl. Acad. Sci. U.S.A* 95: 2938-2943.
- Norberg J. 2003. Association of protein-DNA recognition complexes: electrostatic and nonelectrostatic effects. *Arch. Biochem. Biophys.* 410: 48-68.
- O'Brien R., DeDecker B., Fleming K.G., Sigler P.B. and Ladbury J.E. 1998. The effects of salt on the TATA binding protein-DNA interaction from a hyperthermophilic archaeon. *Journal of Molecular Biology* 279: 117-125.

- O'Flanagan R.A., Paillard G., Lavery R. and Sengupta A.M. 2005. Non-additivity in protein-DNA binding. *Bioinformatics* 21: 2254-2263.
- Oliphant A.R., Brandl C.J. and Struhl K. 1989. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and Cellular Biology* 9: 2944-2949.
- Omichinski J.G., Clore G.M., Appella E., Sakaguchi K. and Gronenborn A.M. 1990. High-resolution three-dimensional structure of a single zinc finger from a human enhancer binding protein in solution. *Biochemistry* 29: 9324-9334.
- Pabo C.O. and Nekludova L. 2000. Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *Journal of Molecular Biology* 301: 597-624.
- Pabo C.O., Peisach E. and Grant R.A. 2001. Design and selection of novel Cys(2)His(2) zinc finger proteins. *Annual Review of Biochemistry* 70: 313-340.
- Paillard G., Deremble C. and Lavery R. 2004. Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.* 32: 6673-6682.
- Patikoglou G. and Burley S.K. 1997. Eukaryotic transcription factor-DNA complexes. *Annu.Rev.Biophys.Biomol.Struct.* 26: 289-325.
- Pavletich N.P. and Pabo C.O. 1993. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* 261: 1701-1707.
- Pavletich N.P. and Pabo C.O. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 252: 809-817.
- Rajamani D., Thiel S., Vajda S. and Camacho C.J. 2004. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences* 101: 11287-11292.
- Rebar E.J. and Pabo C.O. 1994. Zinc-Finger Phage - Affinity Selection of Fingers with New Dna-Binding Specificities. *Science* 263: 671-673.
- Record M.T., Jr., Anderson C.F. and Lohman T.M. 1978. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Q.Rev.Biophys.* 11: 103-178.
- Record M.T., Jr., Lohman M.L. and De H.P. 1976. Ion effects on ligand-nucleic acid interactions. *J.Mol.Biol.* 107: 145-158.
- Ren B., Robert F., Wyrick J.J., Aparicio O., Jennings E.G., Simon I., Zeitlinger J., Schreiber J., Hannett N., Kanin E., Volkert T.L., Wilson C.J., Bell S.P. and Young R.A. 2000.

Genome-Wide Location and Function of DNA Binding Proteins. *Science* 290: 2306-2309.

- Romaniuk P.J. 1990. Characterization of the equilibrium binding of *Xenopus* transcription factor IIIA to the 5 S RNA gene. *J. Biol. Chem.* 265: 17593-17600.
- Roven C. and Bussemaker H.J. 2003. REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res* 31: 3487-3490.
- Roxstrom G., Paulino M. and Tapia O. 2000. Recognition determinants in a T4 \leftarrow G4 mutant derived from a 5'-GCGTGGGCGT-3' oligomer in a zinc finger 268-DNA complex. A molecular dynamics study of the fully charged complex in water. *Theoretical Chemistry Accounts* 104: 96-108.
- Roxstrom G., Velazquez I., Paulino M. and Tapia O. 1998. Molecular dynamics simulation of a Zif268-DNA complex in water. Spatial patterns and fluctuations sensed from a nanosecond trajectory. *Journal of Physical Chemistry B* 102: 1828-1832.
- Sakharov D.V. and Lim C. 2005. Zn Protein Simulations Including Charge Transfer and Local Polarization Effects. *Journal of the American Chemical Society* 127: 4921-4929.
- Schlick T. 2002. *Molecular Modeling and Simulation*. Springer-Verlag, New York.
- Schreiber G. and Fersht A.R. 1996. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* 3: 427-431.
- Scott W.R.P., Hunenberger P.H., Tironi I.G., Mark A.E., Billeter S.R., Fennen J., Torda A.E., Huber T., Kruger P. and van Gunsteren W.F. 1999. The GROMOS biomolecular simulation program package. *Journal of Physical Chemistry A* 103: 3596-3607.
- Seeman N.C., Rosenberg J.M. and Rich A. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U.S.A* 73: 804-808.
- Segal D.J., Crotty J.W., Bhakta M.S., Barbas C.F. and Horton N.C. 2006. Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA. *Journal of Molecular Biology* 363: 405-421.
- Segal D.J., Dreier B., Beerli R.R. and Barbas C.F., III 1999. Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proceedings of the National Academy of Sciences* 96: 2758-2763.
- Selvaraj S., Kono H. and Sarai A. 2002. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J Mol. Biol.* 322: 907-915.

- Sharp K.A. and Honig B. 1995. Salt Effects on Nucleic-Acids. *Current Opinion in Structural Biology* 5: 323-328.
- Sharp K.A. and Honig B. 1990. Electrostatic Interactions in Macromolecules: Theory and Applications. *Annual Review of Biophysics and Biophysical Chemistry* 19: 301-332.
- Siggers T.W. and Honig B. 2007. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res* 35: 1085-1097.
- Siggia E.D. 2005. Computational methods for transcriptional regulation. *Current Opinion in Genetics & Development* 15: 214-221.
- Stormo G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16: 16-23.
- Tsui V., Radhakrishnan I., Wright P.E. and Case D.A. 2000. NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex. *Journal of Molecular Biology* 302: 1101-1117.
- Tuerk C. and Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249: 505-510.
- Van der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A.E. and Berendsen H.J.C. 2005. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* 26: 1701-1718.
- von Hippel P.H. 2004. Biochemistry. Completing the view of transcriptional regulation. *Science* 305: 350-352.
- von Hippel P.H. and Berg O.G. 1986. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. U.S.A* 83: 1608-1612.
- von Hippel P.H. 2007. From Simple DNA-Protein Interactions to the Macromolecular Machines of Gene Expression. *Annual Review of Biophysics and Biomolecular Structure* 36: 79-105.
- Wingender E. 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* 9: 326-332.
- Winter R.B., Berg O.G. and von Hippel P.H. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* 20: 6961-6977.
- Winter R.B. and von Hippel P.H. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor--operator interaction: equilibrium measurements. *Biochemistry* 20: 6948-6960.

- Wolfe S.A., Grant R.A., Elrod-Erickson M. and Pabo C.O. 2001. Beyond the "recognition code": Structures of two Cys(2)His(2) zinc finger/TATA box complexes. *Structure* 9: 717-723.
- Wolfe S.A., Grant R.A. and Pabo C.O. 2003. Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry* 42: 13401-13409.
- Wolfe S.A., Greisman H.A., Ramm E.I. and Pabo C.O. 1999. Analysis of zinc fingers optimized via phage display: Evaluating the utility of a recognition code. *Journal of Molecular Biology* 285: 1917-1934.
- Wolfe S.A., Nekludova L. and Pabo C.O. 2000. DNA recognition by Cys(2)His(2) zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure* 29: 183-212.
- Wuttke D.S., Foster M.P., Case D.A., Gottesfeld J.M. and Wright P.E. 1997. Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *Journal of Molecular Biology* 273: 183-206.
- Zakrzewska K. and Lavery R. 1999. Modelling Protein-DNA Interactions. In: Leszczynski J (ed), *Computational Molecular Biology* pp. 441-483. Elsevier.
- Zhang C., Liu S., Zhu Q. and Zhou Y. 2005. A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes. *Journal of Medicinal Chemistry* 48: 2325-2335.