THE PROTEIN BINDING POTENTIAL OF C2H2 ZINC FINGER DOMAINS

by

KATHRYN JO BRAYER

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHARMACOLOGY AND TOXICOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2008

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation
prepared by Kathryn Jo Brayer
entitled The Protein Binding Potential of C2H2 Zinc Finger Domains
and recommend that it be accepted as fulfilling the dissertation requirement for the
Degree of Doctor of Philosophy

Dr. David Segal Date: 02/18/08

Dr. Richard Vaillancourt Date: 02/18/08

Dr. John Regan Date: 02/18/08

Dr. Matt Cordes Date: 02/18/08

Dr. Nancy Horton Date: 02/18/08

Final approval and acceptance of this dissertation is contingent upon the candidate's
submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and
recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: Dr. David Segal Date: 02/18/08

Dissertation Director: Dr. Richard Vaillancourt Date: 02/18/08

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Kathryn Jo Brayer

ACKNOWLEDGEMENTS

I would like to thank the many people who made my PhD experience enjoyable. First, I would like to thank my advisor, Dr. David Segal, for the enthusiasm, the support, and the guidance. Your faith helped me complete the journey, and your mentorship helped me grow as a scientist. I would also like to thank Dr. Richard Vaillancourt, Dr. John Regan, Dr. Nancy Horton and Dr. Matt Cordes for serving on my committee. Your timely advice helped me to focus my project and achieve success. Thanks also to Dr. Michael Wright, Dr. Ian Korf and Dr. Mark Beida, each of who helped me refine my methods and improve my research.

I am also very grateful for all members of the Segal lab, past and present. Thanks for the fun, the conversation, the toys, and the support through the successes and the failures. In particular thanks, to Aik Ooi, who was there at the start, and to Mital Bhakta, Yuka Shimizu, and Sarah Lockwood, who were there at the end.

Thanks also to the students, faculty and staff in the College of Pharmacy at the University of Arizona and in the Genome Science Department at the University of California at Davis, for your assistance, support and friendship. Thanks also to the IGERT-genomics program, the Pharm/Tox program, and ADCRC for the opportunity and the financial support.

Finally, thanks to my friends and family who supported me through the ups and downs of the last few years. Without you, I truly don't think I could have done it. Thanks Mom, thanks Dad, thanks Erin.

TABLE OF CONTENTS

LIST OF FIGURES	9
LIST OF TABLES	11
ABSTRACT.....	12
INTRODUCTION	13
CHAPTER I: BACKGROUND.....	15
1.1 Zinc Finger Proteins.....	15
1.2 C2H2 Zinc Finger Domain	15
1.2.1 Structure	18
1.2.2 Function	20
1.3 DNA Binding by C2H2 Zinc Finger Domains	20
1.3.1 α -helix DNA Binding Surface: Canonical.....	21
1.3.2 α -helix DNA Binding Surface: Non-Canonical	27
1.4 Protein Binding by C2H2 Zinc Finger Domains	31
1.4.1 α -Helix Binding Using DNA Binding Residues	31
1.4.2 α -Helix Binding, Non-DNA Binding Residues.....	39
1.4.3 β -Sheet Binding	42
1.4.4 Incompletely Annotated Domains	48
1.5 Assessing the Binding Potential of C2H2 Domains	48
CHAPTER II: DISTINGUISHING PROTEIN BINDING AND DNA BINDING C2H2 DOMAINS.....	50
2.1 Introduction.....	50
2.2 Linker Characterization	51
2.3 Regional Spacing Variation	55

2.3.1	Background	55
2.3.2	Results	55
2.3.3	Summary	61
2.4	Amino Acid Characterization	63
2.4.1	Background	63
2.4.2	Results	63
2.4.3	Summary	72
2.5	Discussion	72
CHAPTER III: PREDICTING BINDING FUNCTION		75
3.1	Introduction	75
3.2	Hidden Markov Modeling	77
3.1.1	Background	77
3.1.2	Results	78
3.1.3	Summary	81
3.1.4	Methods	82
3.2	Binding Surface Discrimination by Cluster Analysis	82
3.2.1	Background	82
3.2.2	Results	83
3.2.3	Summary	89
3.2.4	Methods	90
3.3	Discussion	91
CHAPTER IV: DNA BINDING BY ZINC FINGERS		93
4.1	Introduction	93
4.2	Results	99

4.3	Discussion	103
4.4	Materials and Methods.....	104
4.4.1	Plasmid Construction	104
4.4.2	Protein Purification	105
4.4.3	Cyclical Amplification and Selection of Targets (CAST) assay	105
CHAPTER V: PROTEIN BINDING BY ZINC FINGERS		107
5.1	Introduction.....	107
5.2	Results.....	112
5.2.1	Protein Interactions Detected by Yeast Two-Hybrid Assay	112
5.2.2	Verification of Select Protein Interactions by MBP Pull-down Assay...	117
5.2.3	Verification of Select Protein Interactions by Mammalian Co-IP.....	121
5.2.4	Observed Protein Interactions are Dependent on Zinc	127
5.2.5	Observed Protein Interactions are Not Dependent on DNA.....	129
5.3	Discussion	129
5.4	Materials and Methods.....	133
CHAPTER VI: PREDICTING BINDING TYPE FOR INDIVIDUAL C2H2 DOMAINS		138
6.1	Background	138
6.2	Results.....	141
6.3	Discussion	143

CHAPTER VII: SUMMARY	147
APPENDIX A: OTHER PROTEIN BINDING DOMAINS	151
APPENDIX B: SUPPLEMENTARY MATERIAL	157
APPENDIX C: PUBLISHED MANUSCRIPTS	164
REFERENCES	165

LIST OF FIGURES

Figure 1.1: C2H2 zinc finger domain structure.	19
Figure 1.2: DNA binding by C2H2 zinc finger proteins.	22
Figure 1.3: Canonical DNA binding by C2H2 zinc finger domains.....	25
Figure 1.4: DNA binding by Gli1 C2H2 domains.....	29
Figure 1.5: Non-canonical base contacts.	30
Figure 1.6: Comparison of C2H2 and C2HC zinc finger domains.....	34
Figure 1.7: Protein binding surface of the third C2H2 domain in FOG1.	35
Figure 1.8: Critical amino acids in Eos-mediated protein interactions.....	38
Figure 1.9: α -helix inter-finger protein contacts.	41
Figure 1.10: Amino acids in EEA1 important for dimerization with Rab5.....	43
Figure 1.11: Location of PPI binding surfaces in Zac1.	45
Figure 1.12: Intra-protein binding surface of MBP1.	47
Figure 2.1: Analysis of variation in the number of amino acids between the conserved cysteine pairs and histidine pairs in C2H2 domains.	59
Figure 2.2: Deviation from expected values for each category.	60
Figure 2.3: Location of variable amino acids in C2H2 domains.	66
Figure 2.4: Sequence logos for DNA binding, protein binding and random C2H2 domains.	68
Figure 3.1: C2H2 domains clusters reveal amino acid traits related to binding type.	86
Figure 3.2: Splitting Group 1 into two sub-groups reveals distribution of negatively charged amino acids along “DNA binding face” of the α -helix.....	88
Figure 4.1: Overview of CAST Assay Steps.	96
Figure 4.2: Schematic representation of human OAZ protein.	97
Figure 4.3: CAST analysis of specific DNA-binding activity.....	100

LIST OF FIGURES *continued*

Figure 4.4: Results of PCR amplification of recovered DNA to produce the library for the next round of CAST.....	101
Figure 4.5: The binding site selected by Cluster 1 in the CAST assay.....	102
Figure 5.1: Overview of the yeast two-hybrid method.....	109
Figure 5.2: Schematic representation of human Zif268.....	111
Figure 5.3: Representative MBP pull-down experiment results.....	119
Figure 5.4: Representative mammalian co-immunoprecipitation results.....	125
Figure 5.5: Protein-protein interactions mediated by Cluster 6 of hOAZ and Cluster Z of Zif268.....	126
Figure 5.6: Protein-protein interactions mediated by Clusters 1 and 6.....	128
Figure 6.1: C2H2 Domains in human OAZ.....	139
Figure 6.2: C2H2 Domains in human Zif268.....	140
Figure B.1: Alignment of DNA-binding C2H2 domains used in amino acid characterization.....	157
Figure B.2: Alignment of protein-binding C2H2 domains used in amino acid characterization.....	158
Figure B.3: Alignment of random C2H2 domains used in amino acid characterization.....	160
Figure B.4: Alignment of domains used to build the protein HMM profile.....	161
Figure B.5: Alignment of domains used to build the DNA HMM profile.....	162
Figure B.6: Alignment of domains used in cluster analysis.....	163

LIST OF TABLES

Table 1.1: Distribution and Abundance of C2H2 Zinc Finger Domains	17
Table 1.2: DNA Contacting Residues in C2H2 Domains With Solved Structures.	23
Table 1.3: Binding Surfaces for Protein-Protein Interactions Involving C2H2 Domains.	33
Table 2.1: Linker Regions for Protein Interacting C2H2 Domains	53
Table 2.2: TGEKP Linker in Multiple Proteomes	54
Table 2.3: Distribution of C2H2 ZF Domains in Each Category	58
Table 3.1: HMM Profile Scores.....	80
Table 5.1: Yeast Two-Hybrid Results	114
Table 5.2: Potential Interaction Partners for OAZ Zinc Finger Clusters	115
Table 5.3: MBP Pull-Down Assay Results.....	120
Table 5.4: Mammalian Co-IP Results.....	122
Table 6.1: Single C2H2 Domain Prediction	142
Table A.1: Proteins Containing C2H2 Domains Implicated in Protein-Protein Interactions.....	151

ABSTRACT

Cys2-His2 (C2H2) zinc finger domains were originally identified as DNA binding domains, and uncharacterized domains are typically assumed to bind DNA. However, a growing body of evidence suggests an important and widespread role for these domains in protein binding. Over 100 C2H2 zinc finger-protein interactions have been described. This study uses common bioinformatics tools to identify sequence features that predict a DNA- or protein-binding function. Several issues, including uncertainties about the full functional capabilities of the zinc fingers, complicated these efforts. Therefore, an unbiased approach which directly examined the potential for zinc fingers to facilitate DNA or protein interactions was used to determine the full functional capabilities of the C2H2 domains in two model proteins, human OLF-1/EBF associated zinc finger (OAZ) protein and Zif268. OAZ contains 30 zinc fingers in six clusters, some of which have been previously indicated in DNA or protein interactions. Zif268 is a well-known DNA binding protein with three C2H2 domains. DNA binding was assessed using a target site selection (CAST) assay, and protein binding was assessed using a yeast two-hybrid assay. Results indicate that clusters known to bind DNA could facilitate specific protein interactions, but clusters known to bind protein did not facilitate specific DNA interactions, indicating that DNA binding is a more restricted function of zinc fingers than has previously been recognized. These results also suggest that the role of C2H2 zinc finger domains in protein interactions has probably been underestimated. The implication of these findings for the prediction of zinc finger function is discussed.

INTRODUCTION

There are over 10,000 Cys2-His2 (C2H2) zinc finger domains distributed among more than 1000 zinc finger proteins in the human genome. Although C2H2 domains are well known as the DNA binding domain found in many transcription factors, there is a growing body of evidence suggesting they have an important and widespread role in mediating protein-protein interactions. Over 100 C2H2 zinc finger-protein interactions have been described; however, only a handful of those have been examined in detail with the express purpose of understanding the C2H2 domain's role in the interaction. As a result, it is difficult to determine how common protein binding by C2H2 domains is, or even whether this represents an additional or an alternative function for the domain. This study examines the protein binding potential of C2H2 domains.

Chapter I reviews current knowledge of C2H2 domains, including both protein and DNA binding capabilities and mechanisms. In Chapter II, a bioinformatics based analysis was used to examine the C2H2 domain for distinguishing features common to previously characterized DNA or protein binding domains in order to develop a predictive algorithm (Chapter III). Several issues, including uncertainties about the full functional capabilities of the zinc fingers, complicated this effort. Therefore, an unbiased approach was used to directly examine the potential for clusters of zinc fingers to facilitate DNA or protein interactions in two model proteins (Chapters IV and V, respectively). Chapter VI examines the amino acid sequence for zinc fingers within each cluster and predicts the contribution of individual C2H2 domains to known protein and DNA interactions. hOAZ contains 30 zinc fingers in 6 clusters, some of which have been

previously indicated as DNA or protein binding domains. Zif268 is a well-known DNA binding protein containing 3 C2H2 domains. DNA binding was assessed using a target site selection (CAST) assay, and protein binding was assessed using a yeast two-hybrid assay. Results indicate that clusters known to bind DNA could facilitate specific protein interactions, but clusters known to bind protein did not facilitate specific DNA interactions.

As discussed in greater detail in Chapter VII, this study found DNA binding to be a more restricted function of zinc fingers, while their potential for mediating protein interactions is likely greater. These results suggest that the role of C2H2 zinc finger domains in protein interactions has probably been underestimated. The implication of these findings for the prediction of zinc finger function is discussed.

CHAPTER I: BACKGROUND

1.1 *Zinc Finger Proteins*

Zinc finger domains (ZF) are protein structures stabilized by the coordinated binding of a zinc ion between conserved amino acids, generally cysteines and/or histidines. This structurally diverse superfamily encompasses over 40 different domains ranging from simple folds that coordinate the binding of a single zinc ion to complex folds comprised of multiple, interleaved, beta sheets and alpha helices which bind multiple zinc ions (Krishna et al., 2003; Laity et al., 2001; Matthews and Sunde, 2002). Capable of binding DNA, RNA, protein, lipids, and small molecules, ZF domains are as functionally diverse as they are structurally varied; in many cases a single type of domain can interact with multiple types of biomolecules (Gamsjaeger et al., 2007). As a result of their ability to interact with a variety of substrates, ZFs are involved in numerous cellular processes including chromatin remodeling, gene transcription, translation, RNA trafficking and packaging, protein folding, cellular organization, cell adhesion, and zinc sensing (Brown, 2005; Fox et al., 1999; Hall, 2005; Honma et al., 1999; Kelley et al., 1998; Laity et al., 2001; Lee et al., 2006; Lu et al., 2003; Morgan et al., 1997; Pelham and Brown, 1980; Perdomo et al., 2000; Simpson et al., 2004; Sun et al., 1996).

1.2 *C2H2 Zinc Finger Domain*

The “classical,” or Cys₂-His₂ (C2H2) zinc finger is the best-characterized type of zinc finger domain. Since their discovery in 1985, C2H2 ZF proteins (ZFP) have grown to be recognized as an important class of genomic regulators. Upon completing the

rough draft of the human genome sequence, Venter et al. (2001) predicted that there were 4500 C2H2 domains in the human genome, significantly outnumbering any other type of protein domain (Venter et al., 2001). Those estimates are likely an underestimation. Current predictions estimate the number of C2H2 domains at ~10,000 (Letunic et al., 2002), or approximately 3% of the human genome (Kersey et al., 2005; Mackay and Crossley, 1998; Matthews and Sunde, 2002).

As additional whole genome sequencing projects are completed, it has become clear that distribution of the C2H2 domain is as broad as it is abundant. Found throughout all kingdoms, the C2H2 domain is not only ubiquitous, but is also one of the most common protein domains found within many eukaryotic proteomes (Table 1.1). While not found in all species, the C2H2 domain is clearly an important protein fold. Understanding the functionalities that ZFPs impart therefore provides insights into one of the largest protein superfamilies.

Table 1.1: Distribution and Abundance of C2H2 Zinc Finger Domains

Taxon	Number of Proteins with C2H2 Domains^{1,2}	Ranking within Proteome²	Coverage (%)³
Archaea	68		
Bacteria	153		
Eukaryotes	13617		
<i>Arabidopsis thaliana</i>	164	50	0.5
<i>Caenorhabditis elegans</i>	216	13	1.0
<i>Danio rerio</i>	216	9	1.8
<i>Drosophila melanogaster</i>	349	2	2.1
<i>Gallus gallus</i>	74	18	1.4
<i>Homo sapiens</i>	1055	2	2.8
<i>Mus musculus</i>	837	4	2.5
<i>Rattus norvegicus</i>	183	8	1.5
<i>Saccharomyes cerevisiae</i>	47	19	0.8
Viruses	74		

¹Reference (Mulder et al., 2007)²Reference (Kersey et al., 2005)³Percentage of proteome containing proteins with C2H2 domains (Kersey et al., 2005)

1.2.1 Structure

The C2H2 domain self-folds to form a simple $\beta\beta\alpha$ structure, and is so named for the conserved N-terminal cysteines and C-terminal histidines which coordinate the binding of a single zinc ion (Figure 1.1; (Frankel et al., 1987; Lee et al., 1989; Parraga et al., 1988; Pavletich and Pabo, 1991)). C2H2 domains are typically defined by the consensus sequence (F/Y)-X-C-X₂₋₅-C-X₃-(F/Y)-X₅-Ψ-X₂-H-X₃₋₅-H, where X is any amino acid and Ψ is any hydrophobic residue (Wolfe et al., 2000). Natural variants that contain cysteine as the final zinc-chelating residue (C2HC) also fold into the same structure (Liew et al., 2000). The tetrahedral binding of a single zinc ion by the cysteine and histidine residues provides most of the stability to the fold (Wolfe et al., 2000). Hydrophobic residues, which are also well conserved (although their spacing within the domain changes), provide additional stability by packing into the core of the fold (Michael et al., 1992; Miller et al., 1985; Page et al., 1987; Shi and Berg, 1995; Weiss and Keutmann, 1990).

Proteins containing zinc finger domains are extremely varied. Zinc finger proteins (ZFP) only rarely contain a single domain, and may contain up to 40 C2H2 domains (Iuchi, 2001). Multi-fingered proteins generally contain tandem repeats of the domains separated by short linker sequences. In addition, proteins may contain multiple clusters of repeated fingers, separated by longer stretches of unstructured protein (Letunic et al., 2002). Finally, in addition to containing variable numbers of C2H2 domains, ZFPs frequently contain a wide variety of other types of domains; KRAB, BTB, and SCAN being the most common types (Sander et al., 2003).

Figure 1.1: C2H2 zinc finger domain structure.

Ribbon diagram of the third C2H2 domain from TFIIIA in *Xenopus laevis* (PDB accession number 1TF3) showing canonical stabilization of the $\beta\beta\alpha$ fold by the coordinated binding of a zinc ion (yellow) by two cysteine (green) and two histidine (blue) residues.

1.2.2 Function

C2H2 domains were initially identified as the sequence specific DNA binding domain of **Transcription Factor IIIA (TFIIIA)** in *Xenopus laevis* (Miller et al., 1985). Today, DNA binding by C2H2 domains is well characterized, and the many C2H2 ZFPs that have not yet been functionally characterized are generally assumed to have DNA binding capabilities. However, C2H2 domains have proven to be as diverse functionally as they are abundant, having been shown capable of interacting with RNA and protein, although these functions are not as well described (Brown, 2005; Hall, 2005; Lee et al., 2006; Lu et al., 2003).

1.3 DNA Binding by C2H2 Zinc Finger Domains

Over the past twenty years, numerous studies have examined the DNA-protein interactions mediated by C2H2 domains. Today there are structures for roughly 22 proteins interacting with their DNA targets; however only 6 of these are naturally occurring proteins (Berman et al., 2000). DNA recognition requires repetition of C2H2 domains, and typically requires a minimum of two domains (Wolfe et al., 2000). Multi-fingered proteins are oriented such that the protein is “antiparallel” to the DNA with the N-terminal finger contacting the 3' end of the primary DNA strand, and successive fingers wrapping through the major groove with the final finger contacting the 5' end of the DNA sequence (Figure 1.2; (Pavletich and Pabo, 1991; Wolfe et al., 2000). As discussed in greater detail below, all contact with the nucleotides is mediated by specific amino acids in the N-terminal region of the zinc finger (Wolfe et al., 2000).

Occasionally, additional amino acids outside of this region will contact the phosphate backbone of the DNA (Wolfe et al., 2000). Importantly, the interaction between C2H2 fingers and the DNA is virtually identical with only slight variations between “canonical” and “non-canonical” domains.

1.3.1 α -helix DNA Binding Surface: Canonical

Structures of C2H2 domains and their DNA target are available for 6 proteins, or a total of 16 fingers. The best-characterized DNA-ZF interaction is that of **Early Growth Response Factor 1 (aka Zif268)**. Primarily recognized to function as a transcriptional regulator of development and morphogenesis of cardiac, neural, and myeloid cells (Rebhan et al., 1997), some evidence suggests Zif268 may function as a tumor suppressor (Huang et al., 1995). Containing three C2H2 domains, Zif268 binds to GC-rich DNA sequences, the consensus of which is 5'-GCG(G/T)GGGCG-3' (Christy and Nathans, 1989). Originally crystallized by Pavletich and Pabo (1991), there are now almost 20 structures of Zif268, and its variants, bound to their target DNA sequences (Berman et al., 2000).

Figure 1.2: DNA binding by C2H2 zinc finger proteins.

Ribbon diagram of the three C2H2 domains from Zif268 in *Mus musculus* (shown in color) bound to DNA sequence (gray; PDB accession number 1ZAA). C2H2 domains bind DNA in an anti-parallel fashion, with finger 1 binding to the 3' end of the primary DNA strand, and successive fingers wrapping through the major groove. Zinc ions shown as blue spheres.

Table 1.2: DNA Contacting Residues in C2H2 Domains With Solved Structures.

Protein	Finger	Position of DNA Contacting Residues	Type of Binding	PDB Accession	Ref
GAGA	F1	2, 3, 6	Non-Canonical		(Omichinski et al., 1997)
GLI	F2	6	Non-Canonical	2GLI	(Pavletich and Pabo, 1993)
	F4	1, 2, 3, 6	Non-Canonical		
	F5	-1, 2, 3, 5, 6	Non-Canonical		
TFIIIA	F1	-1, 2	Non-Canonical	1TF3, 1TF6	(Nolte et al., 1998; Wuttke et al., 1997)
	F2	2, 3, 6	Non-Canonical		
	F3	2, 3, 6, 10	Canonical		
	F5	-1, 2, 3, 6	Non-Canonical		
TTK	F1	2, 3, 6	Non-Canonical	2DRP	(Fairall et al., 1993)
	F2	-1, 2, 3	Canonical		
YY1	F1	6	Non-Canonical	1UBD	(Houbaviy et al., 1996)
	F2	-1, 2, 3, 6	Non-Canonical	1UBD	
	F3	-1, 2, 3	Canonical		
	F4	-1, 2, 3	Non-Canonical		
WT1	F2	-1, 6	Non-Canonical		(Stoll et al., 2007)
	F3	-1, 2, 3	Non-Canonical		
	F4	-1, 6	Non-Canonical		
Zif268	F1	-1, 2, 6	Canonical	1AAY	(Pavletich and Pabo, 1991)
	F2	-1, 2, 3	Canonical		
	F3	-1, 2, 6	Canonical		

As shown in figure 1.3, Zif268's C2H2 domains interact with DNA by inserting the N-terminal portion of the α -helix into the DNA's major groove (Pavletich and Pabo, 1991). Key residues include the arginine immediately preceding the start of the α -helix, henceforth referred to as position -1, as well as the glutamic acid and the arginine in positions 3 and 6, respectively, of the α -helix (Figure 1.3). The arginines in position -1 and 6 form hydrogen bonds with the 3' guanine and 5' guanine of the primary strand. Hydrophobic interactions mediate the contacts between the glutamic acid in position 3 and primarily strand's cytosine (Figure 1.3). Additional contacts are made by the aspartic acid in position 2 to the 5' adenine in the complementary DNA strand (Figure 1.3; (Pavletich and Pabo, 1991)).

Residues in the α -helix also make contact with the DNA backbone, although the significance of these interactions is not well understood. While the interactions do not contribute to nucleotide specificity, they contribute to the overall binding energy of the interaction (Wolfe et al., 2000). The most conserved of these contacts is between the first conserved histidine residue, position 7 in the α -helix, and the phosphate backbone of the primary strand (Pavletich and Pabo, 1991; Wolfe et al., 2000). Observed in all three fingers of Zif268, this interaction seems to be important because it brings N-terminal portion of the α -helix into close proximity with the primary DNA strand.

Figure 1.3: Canonical DNA binding by C2H2 zinc finger domains.

Ribbon diagram of the first C2H2 domain from Zif268 in *Mus musculus* bound to its target DNA sequence (PDB accession 1ZAA). Specific contacts are made between four amino acids in α -helix and the nucleotides. The arginine (R) in position 6 relative to the start of the α -helix (blue) forms hydrogen bonds with the 5' guanine (G, light blue), while the arginine in position -1 (red) forms a similar bond with the 3' guanine (orange). Hydrophobic interactions mediate the contacts between the glutamic acid (E) in position 3 (green) and the cytosine (C, light green). Unlike the other residues which contact the primary DNA strand, the aspartic acid in position 2 (D, bright yellow) hydrogen bonds with the 5' adenine (A, yellow) on the secondary strand. Zinc ion shown as yellow sphere.

Examination of the protein-DNA interactions by the zinc fingers of Specific Protein 1 (Sp1) and Early Growth Response Factor 2 (Krox 20) also revealed the importance of amino acids in the N-terminal portion of the α -helix. Krox 20 is a transcription factor involved in peripheral nervous system development and early myelination, as well as a mediator of phosphatase and tensin homolog (PTEN) signaling (Rebhan et al., 1997). Like Zif268, Krox 20 contains three C2H2 zinc finger domains that bind the consensus sequence 5'-GCGGGGCG-3' (Milbrandt, 1987; Nardelli et al., 1991).

Sp1 is ubiquitously expressed in human tissues serving mainly as an activator of transcription for housekeeping genes and genes involved in growth regulation, but it can also act as a repressor in certain circumstances (Armstrong et al., 1997; Briggs et al., 1986; Dennig et al., 1995; Hagen et al., 1994; Kadonaga et al., 1987; Kwon et al., 1999; Lee et al., 2005; Li et al., 1998; Murata et al., 1994; Ogra et al., 2001; Shou et al., 1998; Suzuki et al., 2000; Zaid et al., 2001). Containing three C2H2 domains, Sp1 binds to the consensus sequence 5'-(G/t)GGGCGG(G/A)(G/A)(C/T)-3' in GC-rich promoters found in many genes (Armstrong et al., 1997; Berg, 1992; Bucher, 1990; Kadonaga and Tjian, 1986).

These early studies with Zif268, Krox20, and Sp1 revealed a simple pattern of C2H2-DNA interactions in which a single amino acid interacted with a single nucleotide in a consistent way. However, as structures for additional proteins became available, a more complex set of interactions was revealed.

1.3.2 α -helix DNA Binding Surface: Non-Canonical

A small group of “non-canonical” zinc fingers appear to participate in more diverse sets of DNA interactions; often displaying interactions mediated by alternate or additional amino acids than those in positions -1, 2, 3, 6 (Table 1.2; (Wolfe et al., 2000)). The first example of this was seen in fingers 4 and 5 of **Glioma-associated protein 1 (Gli1)**, a protein containing five C2H2 domains, separated by a combination conserved and non-conserved linkers (Table 1.2). The crystal structure of Gli1 bound to its consensus sequence, 5'-TGGGTGGTC-3' (Kinzler and Vogelstein, 1990), revealed both similarities and difference to the canonical fingers (Kinzler and Vogelstein, 1990; Pavletich and Pabo, 1993).

First, like the three-fingered proteins mentioned above Gli1 wraps through the major groove, in an “anti-parallel” fashion, and contacts specific nucleotides using residues in the α -helix of its fingers (Pavletich and Pabo, 1993). However, unlike Zif268, Sp1 and Krox 20, which use all fingers equally to contact DNA, only fingers 2-5 of Gli1 contact the DNA, and only fingers 4-5 make sequence specific contacts. Finger 1 does not contact DNA in this structure (Pavletich and Pabo, 1993).

Gli1 also displays a much more complex pattern of amino acid-nucleotide contacts than that of canonical fingers which rely on amino acids in positions -1, 2, 3, and 6 of the α -helix. As shown in figure 1.4, DNA contact residues in finger 4 include amino acids in the canonical positions of 2, 3, and 6 of the α -helix, as well as an alanine in position 1. Similarly, finger 5 contacts DNA using the arginine in position 5, as well as the residues in positions -1, 2, 3 and 6 (Figure 1.4; (Pavletich and Pabo, 1993)).

Finally, as summarized in figure 1.5, residues in both fingers also make non-standard contacts, either by contacting atypical nucleotides, cross-strand binding, or by contacting two nucleotides with a single amino acid (Pavletich and Pabo, 1993; Wolfe et al., 2000).

Like Zif268, Gli1 also interacts with the phosphate backbone using residues in the α -helix (Pavletich and Pabo, 1993). Interestingly, finger 3, which does not contact the nucleotides, instead uses residues in positions -1 and 5 to contact the DNA backbone (Pavletich and Pabo, 1993). Like Zif268, fingers 2, 4-5 make contact to the backbone using various combinations of residues located in α -helix (Figure 1.4; (Pavletich and Pabo, 1991, 1993)).

Despite early results suggesting that ZFPs are comprised of C2H2 domains that are either all canonical fingers or all non-canonical fingers, more recent structures have revealed several examples of proteins that contain both canonical and non-canonical fingers. As mentioned above, TFIIIA was the first transcription factor in which the C2H2 domain was described (Miller et al., 1985). Containing a total of nine C2H2 domains, TFIIIA uses two clusters of three ZFs for DNA-binding, three or four ZFs for RNA-binding, and some fingers that appear to be involved in both DNA- and protein-binding (Del Rio et al., 1993; Del Rio and Setzer, 1993; Friesen and Darby, 1997; Pelham and Brown, 1980; Theunissen et al., 1998). TFIIIA contains a single canonical domain, finger 3, while fingers 1, 2 and 5 are non-canonical (Table 1.2, Figure 1.5; (Wolfe et al., 2000). Table 1.2 and figure 1.5 summarize DNA binding types for all DNA binding C2H2 domains with solved structures.

Figure 1.4: DNA binding by Gli1 C2H2 domains.

Aligned amino acid sequence for the five ZF domains in Gli1. Amino acids involved in nucleotide specific contacts with DNA are shown in red. Phosphate backbone contacting amino acids are shown in bold. Location of the β -sheets and α -helix are indicated by arrows and tube, respectively. Numbers above indicate positions relative to the start of the α -helix.

Figure 1.5: Non-canonical base contacts.

Summary of non-canonical amino acid-nucleotide contacts mediated by C2H2 zinc finger domains. Arrows indicate interactions between amino acid positions (large boxes) and nucleotides (small boxes). Amino acid positions relative to the start of the α -helix are indicated by the large numbers. From (Wolfe et al., 2000).

1.4 Protein Binding by C2H2 Zinc Finger Domains

Unlike ZF-DNA interactions, the interacting residues responsible for specific ZF PPIs have been significantly less studied. Due to a lack of structural and mutagenesis studies, in most cases, little is known about the interaction surface of protein binding C2H2 domains. Only a handful of the C2H2 domains known to be involved in protein-protein interactions have solved structures, and only a few scanning mutagenesis studies have been conducted to examine the location of critical residues.

1.4.1 α -Helix Binding Using DNA Binding Residues

The best characterized ZF PPIs are those between **Friend of Gata1** (FOG1) and its binding partners, the globin transcription factor GATA1 and Transforming Acidic Coiled-coil 3 (TACC3). FOG1 contains a total of nine zinc finger domains (Table 1.3); however, only four are of the classic C2H2 type; the remaining fingers substitute a conserved cysteine for the final histidine (Tsang et al., 1997). FOG1 is known to interact with GATA1 using fingers 1, 5, 6, and 9, which are variant C2HC fingers (Fox et al., 1999). Recently, Liew et al. (2005) determined the structure of the interaction between the C2HC finger 1 of FOG from *Drosophila melanogaster* and a segment of murine GATA-1. The variant C2HC was found to be structurally identical to C2H2 ZFs, except for subtle differences in the C-terminal end of the α -helix (Figure 1.6a; (Liew et al., 2000)). (The reader should note that these C2HC variants are not to be confused with the C-X2-C-X4-H-X4-C class of ZF. This latter class, which includes the well-studied NCp7 nucleocapsid ZFs of HIV, is also referred to as C2HC but has a completely different

protein structure (Morellet et al., 1992)). As shown in Figure 1.6b, the residues in the α -helix of the FOG C2HC domain contact GATA, primarily through polar and hydrophobic interactions. Interestingly, despite similarities in both sequence and structure, C2HC and C2H2 domains are not interchangeable. Matthews et al. (2000) demonstrated that mutation of the C2HC domains of FOG fingers 1 and 9 to C2H2 domains inhibited their ability to interact with GATA without disrupting their ability to fold.

The function of the N-terminal cluster of FOG1 was unknown until recently. This cluster of fingers is comprised of one variant C2HC finger, followed by three classic C2H2 fingers. DNA binding studies failed to demonstrate DNA binding by the cluster, suggesting that these fingers do not participate in DNA interactions (Kwan et al., 2003). After conducting yeast-two hybrid and immunoprecipitation experiments to find and map the PPI between the third (classic) zinc finger of FOG1 and TACC3, Simpson et al. (Simpson et al., 2004) combined NMR and alanine mutagenesis to pinpoint critical amino acid involved in the interaction. Interestingly, they demonstrated that amino acids in the α -helix of FOG1 formed the binding surface for the interaction and that residues in positions normally involved in contacting DNA, positions -1, 2, 3, and 6, were also utilized for protein interactions (Figure 1.7). Additionally, they showed that the interacting surface was longer than that required for DNA binding. It included residues along the entire length of the α -helix, as well as residues located before and after the helix that were oriented into the binding face (Simpson et al., 2004).

Table 1.3: Binding Surfaces for Protein-Protein Interactions Involving C2H2 Domains.

Protein ¹	Binding Surface	Partner	Architecture ²	Ref
FOG1	α helix of F1 of F3	Gata1 TACC3		(Simpson et al., 2004)
Eos	α helix of Fs 5-6	Eos Ikaros Pegasus Trps1		(Westman et al., 2004)
Gli	α helix, Fs 1-2	Intra-protein		(Pavletich and Pabo, 1993)
Zap1	α helix, Fs 1-2	Intra-protein		(Wang et al., 2006)
EEA1	β 1 of F1	Rab5		(Merithew et al., 2003)
Zac1	β 1 of F2	Zac1		(Hoffmann et al., 2003)
MBP-1	Linker	Intra-protein		(Omichinski et al., 1992)

¹All proteins from human except Zac1 (*Mus musculus*) and Zap1 (*Saccharomyces cerevisiae*).

²Linear representation of the domain structure of proteins. Black boxes, C2H2 zinc finger domain; gray boxes, variant C2HC zinc finger domains; F, FYVE zinc finger domain.

Figure 1.6: Comparison of C2H2 and C2HC zinc finger domains.

(A) Superimposed ribbon diagrams of the third C2H2 domain from TFIIIA in *Xenopus laevis* (PDB accession 1TF3, shown in blue) and the first C2HC domain from FOG in *Drosophila melanogaster* (PDB accession 1Y0J, shown in grey) demonstrate that C2HC has the same structure as the classic C2H2 domains. (B) Ribbon diagram of the protein-protein interaction between the first C2HC domain of *Drosophila* FOG (gray) and the N-terminal treble-cleft zinc finger of murine GATA-1 (blue; (Liew et al., 2005). FOG amino acids critical to the interaction are displayed in red. Zinc ions shown as grey (A) or yellow (B) spheres.

Figure 1.7: Protein binding surface of the third C2H2 domain in FOG1.

A ribbon diagram shows the third finger of FOG1 (PDB accession 1SRK). Amino acids critical to the interaction with TACC3 as determined by NMR titration (red), mutation analysis (green), or both (orange) are shown. Zinc ion shown in yellow.

Protein interactions between members of **the Ikaros family** of proteins have also been examined and partially mapped to specific amino acid positions. Ikaros is the founding member of a family of proteins composed of five proteins - Ikaros, Aiolo, Helios, Eos and Pegasus - all of which maintain a conserved, characteristic domain architecture containing two clusters of C2H2 domains. The N-terminal cluster is comprised of 3-4 domains while the C-terminal cluster contains two fingers (Table 1.3). Forming either self-associations or associations with other members of the family, all members of this family use their C-terminal fingers to mediate PPIs, and use their N-terminal fingers to bind DNA (Hahm et al., 1998; Molnar and Georgopoulos, 1994; Morgan et al., 1997; Perdomo et al., 2000). **Trichophinophalageal syndrome 1** (Trps1) contains two C-terminal ZF domains that are homologous to the C-terminal fingers in the Ikaros family. Interestingly, Trps1, which is not a member of the Ikaros family, forms a PPI with the family member Eos, but not the others (Momeni et al., 2000; Westman et al., 2003).

Westman et al. (2004) solved the structure of the C-terminal cluster of two domains in Eos and used alanine mutagenesis to determine amino acids critical for interactions with Ikaros, Pegasus, Trps1, and itself. As observed in FOG1 ZF3, the binding surface for the PPI ran along the α -helix region of finger 5 (Westman et al., 2004). Scanning mutagenesis again highlighted the importance of amino acids in positions 2, 3, 5, and 6 of the α -helix for maintaining the interaction, particularly in finger 5. Additionally, as shown in figures 1.8a and 1.8b, amino acids located in the turn between the β -sheets in both fingers also participated in the interaction, and these

residues are also likely oriented towards the binding surface (Westman et al., 2004).

Interestingly, the alanine substitutions had dissimilar effects on the various PPIs, indicating an underlying selectivity in these interactions (Figure 1.8c).

Figure 1.8: Critical amino acids in Eos-mediated protein interactions.

Homology models of (A) finger 5 and (B) finger 6 of Eos (based on PDB accession 1SRK and 1PAA, respectively) showing amino acids critical to PPIs with Eos, Ikaros, Pegasus, or Trps1. Amino acid positions in which alanine substitution disrupted homodimerization and/or interaction with Ikaros and Pegasus are shown in red. Since most alanine substitutions disrupted the interaction between Eos and Trps1, the positions for which alanine mutation still allowed interactions are shown in green. Zinc ions are shown as yellow spheres. (C) Scanning alanine mutagenesis revealed amino acids critical to the Eos with either Eos (E), Ikaros (I), Pegasus (P), or Trps1 (T) – as indicated on the left. Domain structural elements are indicated above with arrows representing β -sheets and tubes representing α -helices. Alanine mutations with a strongly (red) or moderately/weakly (blue) negative effect on yeast growth in a yeast two hybrid are shown. Alanine mutations that still allowed interactions are shown in green. Positions not affected by alanine substitution shown in black, positions that were not tested shown in light gray.

1.4.2 α -Helix Binding, Non-DNA Binding Residues

One of the earliest examples of a PPI between C2H2 domains was observed between finger 1 and finger 2 in the crystal structure of **glioma-associated protein 1 (Gli1)** (Pavletich and Pabo, 1993). The five-finger Gli1 protein forms several complex interactions, with most fingers able to interact with both protein and DNA (Tables 1.3 and A.1; (Pavletich and Pabo, 1993)). As discussed in greater detail in section 1.3.2, fingers 2-5 bind DNA (Pavletich and Pabo, 1993). Fingers 3-5 mediate a PPI with multiple members of the Zic family of proteins; however, neither the contribution of individual fingers nor the binding surface has been elucidated (Koyabu et al., 2001). Remarkably, while finger 2 is binding DNA, it also forms numerous intra-protein contacts with finger 1 (Figure 1.9a, (Pavletich and Pabo, 1993)).

Wang et al. (2006) reported a similar finding in the *Saccharomyces cerevisiae* protein **Zap1** after solving the structure using NMR. Zap1 contains seven C2H2 domains arrayed in two clusters (Table 1.3). The C-terminal cluster of Zap1 contains five domains and is known to bind DNA (Wang et al., 2006). The N-terminal cluster contains two domains that interact with each other.

A comparison of the Gli1 and Zap1 structures revealed striking similarities in their interaction surfaces (Wang et al., 2006). In each case the interaction occurs in the α -helical region; however, amino acid positions typical to DNA-protein interactions are not used. Instead the interacting face is located about a quarter-turn counterclockwise to the DNA binding surface (when looking down the axis of the alpha helix; Figures 1.9a and b). Despite similarities in the interaction face used by both Gli1 and Zap1, there is

little overlap in the positional location of amino acids required for the interactions, with the exception of α -helix position 8, located in the C-terminal end of the helix (Pavletich and Pabo, 1993; Wang et al., 2006). In both Gli1 and Zap1, amino acids at position 8 in both finger 1 and finger 2 form inter-finger contacts (Figure 1.9 a and b; (Wang et al., 2006)).

Figure 1.9: α -helix inter-finger protein contacts.

Ribbon diagrams depict fingers 1 (blue) and 2 (orange) of (A) Gli1 (PDB accession number 2GLI) and (B) Zap1 (PDB accession number 1ZW8). Interacting amino acids are shown as sticks. Positions typically involved in DNA binding by C2H2 domains are shown in yellow. Zinc ions shown as gray spheres.

1.4.3 β -Sheet Binding

Not all PPIs depend on α -helical amino acids; nor do all C2H2 PPIs preclude DNA binding. One example of this is seen in **early endosome antigen 1 (EEA1)**. EEA1 contains only one C2H2 domain located in the N-terminal region of the protein. It also contains a FYVE domain in the C-terminal region (Table 1.3). While both the C2H2 and FYVE domains are important for PPI with Rab5, the C2H2 domain is sufficient for interaction with Rab5 (Merithew et al., 2003). By combining scanning alanine mutagenesis with surface plasmon resonance (SPR) to measure PPI affinity, Merithew et al. (2003) was able to localize the PPI to the first β -sheet (Figure 1.10). In particular, mutation of either the phenylalanine or the isoleucine residues located at the beginning of the β sheet, immediately before the first cysteine, caused a 100-fold decrease in affinity. Other substitutions, located on the same binding face but different regions of the C2H2 fold, had less of an effect on affinity, decreasing it by 10-fold (Merithew et al., 2003).

Figure 1.10: Amino acids in EEA1 important for dimerization with Rab5.

A homology model of EEA1 (based on 1PAA) shows the amino acids contributing to Rab5 binding. Critical residues determined by alanine mutation and surface plasmon resonance are colored according to fold decrease in binding affinity following mutation to alanine. Red ~ 100 fold decrease, green ~ 30 to 40 fold decrease, blue ~ 5 fold. Zinc ion shown as yellow sphere.

Zac1, a protein involved in regulating apoptosis and cell-cycle arrest, contains seven C2H2 domains, some of which mediate PPIs (Table 1.3; (Hoffmann et al., 2003; Spengler et al., 1997; Varrault et al., 1998)). By combining gel-shift experiments with scanning mutagenesis, Hoffmann et al. (2003) demonstrated that fingers 2-4 and 6-7 bound DNA using residues in positions -1, 2, 3, and 6 of the α -helix (Hoffmann et al., 2003). Using co-immunoprecipitation of *in vitro* translated proteins, they further demonstrated that Zac1 formed homodimers. Homodimerization was dependent on fingers 1 and 2 (Hoffmann et al., 2003), indicating a role in both protein and DNA binding for finger 2. As summarized in figure 1.11, a series of protein deletions enabled Hoffmann et al. (2003) to narrow the site of the PPI to a span of amino acids starting in the linker region between fingers 1 and 2 through the first β -strand of finger 2. The importance of specific amino acids was not determined.

Figure 1.11: Location of PPI binding surfaces in Zac1.

Homology model of Zac1 (based on PDB accession number 1UBD) indicating homodimerization surface. Residues colored red and green delineate finger regions that are strongly or moderately required, respectively. Yellow spheres represent zinc ions.

Similar to GLI and Zap1, intra-protein PPIs can also be mediated by amino acids in the β -sheets. **Major histocompatibility complex enhancer binding protein 1 (MBP1)** contains five C2H2 domains arranged into a cluster of two fingers, a single finger, and a second cluster of two fingers. This protein also forms intra-protein bonds (Table 1.3, Figure 1.12). However, the intra-protein interaction observed in MBP1 is very different than that seen in GLI and Zap1 (Figure 1.8A and B). Omichinski et al. (Omichinski et al., 1992) solved the structure of a synthetic peptide corresponding to the C-terminal pair of C2H2 domains of MBP1. As can be seen in figure 1.12, inter-finger contacts are made between a threonine located in the C-terminal end of the helix of finger 5, a valine in linker region, and a lysine located in the loop between the β -sheet and α -helix of finger 6 (Omichinski et al., 1992). Although these fingers are known to interact with DNA (Baldwin et al., 1990; Clark et al., 1990; Fan and Maniatis, 1990; Maekawa et al., 1989) and the intra-protein contacts do not involve residues typically involved in DNA interaction, the effect of the interaction on DNA binding is not known.

Figure 1.12: Intra-protein binding surface of MBP1.

Inter-finger contacts by MBP1 finger 5 (blue) and finger 6 (orange; PDB accession number 1BBO). Critical amino acids are shown as sticks. Residues that typically contact DNA are colored yellow. Zinc ions shown as gray circles.

1.4.4 Incompletely Annotated Domains

The examples above (Table 1.3) describe PPIs mediated by C2H2 domains using a wide variety of binding surfaces. Unfortunately, the binding surface for the majority of ZF-mediated PPIs has not been determined. Table A.1 presents a survey of over 100 PPIs in which C2H2 domains are ostensibly involved in the interface. In many cases, the interactions were not localized to a specific finger since clusters of multiple fingers were frequently treated as a single binding unit. In other cases, the interaction was found in a high throughput assay and not verified, or the protein fragment tested contained one or more C2H2 domains in conjunction with a large flanking sequence and the binding interface was not assigned to one region. These examples provide strong evidence that C2H2-protein interactions are more widespread than currently appreciated.

1.5 Assessing the Binding Potential of C2H2 Domains

Increasingly, biological experiments are being conducted in a high-throughput fashion. This is particularly true of genomic, proteomic, and informatics based studies attributing function to proteins. Regardless of the goals or outcomes of these studies, functions of otherwise uncharacterized genes and proteins are frequently assigned based on their sequence or structural properties (Lee et al., 2007).

Although C2H2 domains are still primarily considered the DNA binding domain of many transcription factors, both recognition and understanding of their protein binding capabilities is growing. Despite this, the many C2H2 ZFPs that have not yet been functionally characterized are generally assumed to have DNA binding capabilities.

Studies aimed at determining the biological roles of specific ZF proteins frequently reinforce this bias because only DNA binding is tested. On the other hand, protein-interacting C2H2 zinc fingers are frequently discovered fortuitously in the course of studying an individual protein's function. In this context, the domains are generally not tested for DNA binding capability. Even when C2H2 domain(s) are known to bind both protein and DNA, it is only rarely known whether these functions occur separately or simultaneously. As a result of the biased nature of these types of studies, it is difficult to determine how common protein-binding by C2H2 domains is, or even whether protein binding represents an additional or an alternative function for the domain.

To fully understand the diverse functions mediated by the enormous family of ZF proteins, a more inclusive annotation of their component C2H2 domain functions is required. By combining a bioinformatics based approach with a less biased experimental approach, this study attempts to more accurately annotate zinc finger function. This study tests the hypothesis that a less biased approach would produce a more complete description of ZF function. In addition, a bioinformatics based approach was used in an attempted to identify sequence features that distinguish DNA-binding or protein binding function for individual C2H2 domains, for future development of a predictive algorithm for determining binding type based on C2H2 sequence features.

CHAPTER II: DISTINGUISHING PROTEIN BINDING AND DNA BINDING C2H2 DOMAINS

2.1 Introduction

Solved structures of C2H2 domains have revealed that all folded C2H2 domains form the same simple $\beta\beta\alpha$ structure, regardless of variation in the amino acid sequence. Within the bounds of the degenerate consensus sequence: (F/Y)-X-C-X₂₋₅-C-X₃-(F/Y)-X₅- Ψ -X₂-H-X₃₋₅-H, where X is any amino acid and Ψ is any hydrophobic residue, there is enormous variation between individual C2H2 domains. Variation occurs in two ways. First, since there are only four conserved and 3 semi-conserved residues, there is significant variation in the amino acid composition of individual domains. Second, variation also occurs in the number of amino acids between the conserved cysteines and histidines. Finally, variation also occurs in the length and composition of amino acids linking domains together in multi-domained proteins.

This section represents an effort to determine characteristics to distinguish protein binding C2H2 domains from DNA binding C2H2 domains so that an algorithm could be developed to predict binding type. As originally envisioned, it was hoped that the likelihood of protein versus DNA binding by a particular zinc finger could be determined with nothing more than the amino acid sequence. Common computational tools were used to evaluate sequence features for their potential use in developing predictive methods. Although only moderately successful the results here indicate that there are likely differences in sequence features for the two binding types.

2.2 Linker Characterization

It has been previously observed that certain amino acid linker sequences connecting tandem repeats of C2H2 domains were highly correlated with ZF-DNA binding (Laity et al., 2000; Ryan and Darby, 1998). In mammals, tandem ZFs are frequently separated by linkers of 5 amino acids, with roughly 50% of these having the sequence TGEKP (Wolfe et al., 2000). The conserved linker plays an important role in DNA binding, with each residue playing an identifiable role in stabilizing the interaction. Although some known DNA-binding ZFs don't have this linker (for example, Tramtrack ZF1-2; (Fairall et al., 1992)), and some that do not bind DNA do have it (for example, Gli ZF2-3 (Pavletich and Pabo, 1993)), the presence of a TGEKP-like linker is currently the best predictor of DNA-binding. For this analysis, the linker region for 41 protein-binding C2H2 domains, obtained from the protein interaction databases BOND (Alfarano et al., 2005); HPRD (Peri et al., 2003); and MINT (Chatr-aryamontri et al., 2007), as well as additional sources, were analyzed for similarities.

As shown in table 2.1, many protein-binding C2H2 domains also have TGEKP-like linkers. Specifically, linkers between Gli ZF2-3, Bcl6 ZF2-4, PLZF ZF5-8, Sp1 ZF1-3, Zac1 ZF6-7, and YY1 ZF2-4 contain close variations of this motif. Interestingly, with the possible exception of Gli ZF2-3, all of these ZFs are involved in DNA binding as well (Hoffmann et al., 2006; Houbaviy et al., 1996; Kadonaga et al., 1987; Mascle et al., 2003; Sitterlin et al., 1997). Since there does not appear to be an alternate, protein-binding linker (Table 2.1), it is therefore unlikely that a "protein-binding linker signature" can be identified that could be used to predict protein binding. Finally, while the TGEKP

linker may be a useful predictor of DNA-binding function in mammals, its inclusion or exclusion does not appear to have prognostic value for protein binding.

Unfortunately, the usefulness of the TGEKP linker as a predictor of DNA binding is limited. The amino acid sequence connecting tandem repeats of C2H2 domains varies, in both sequence and length, in different species (Bohm et al., 1997; Clarke and Berg, 1998; Tsuzuki and Enver, 2002; Wolfe et al., 2000). C2H2 domains in multiple species were analyzed for the presence or absence of the TGEKP linker. For this analysis, the strict TGEKP linker was used, and “TGEKP-like” linkers were not included. The strict interpretation was chosen because the roles of these amino acids in the formation and stabilization of the domain is known, while the role of various substitutions is unknown. As demonstrated in Table 2.2, outside of mammalian species, less than 10% of C2H2 domains end in a TGEKP linker. However, this analysis did not account for TGEKP-like linkers, and it is possible that these linkers comprise a greater percentage in other species. Finally, this analysis did not include determining whether or not other species have a common linker sequence that is associated with a particular binding type.

Table 2.1: Linker Regions for Protein Interacting C2H2 Domains

Protein/Finger	Finger Seq.	Linker Seq.
Bcl6/F1	FFCNECDCRFSEEASLKRHTLQTH	SDKP
Bcl6/F2	YKCDRCQASFRYKGNLASHKTVH	TGEKP
Bcl6/F3	YRCNICGAQFNRPANLKTHTRIH	SGEKP
Bcl6/F4	YKCETCGARFVQVAHLRAHVLIH	TGEKP
EEA1	FICPQCMKSLGSADELFKHYEAVH	DAGND*
Eos/F5	FKCEHCRILFLDHVMFTIHMGC	GFRDP
Eos/F6	FECNICGYHSQDRYEFSSHIVRGEH	KVG**
Fog1/F3	FVCLICLSAFTTKANCERHLKVH	TDTLS
Gli/F1	TDCRWDGCSQEFDSQEQLVHHINSEH	IHGERKE
Gli/F2	FVCHWGGCSRELRPFKAQYMLVVHMRRH	TGEKP
MBP1/F4	YICEECGIRCKKPSMLKKHIRTH	TDVRP
MBP1/F5	YHCTYCNFSFKTKGNLTKHMKSKAH	SKKCV
OAZ/F14	YPCNQCDLKFSNFESFQTHLKLH	LELLLRK
OAZ/F15	QACPQCKEDFDSQESLLQHLTVH	YMTTSTH
OAZ/F16	YVCECDKQFSSVDDLQKHLDMH	TFVL
OAZ/F17	YHCTLCQEVFDSKVSIVHLAVKH	SNEKKM
OAZ/F18	YRCTACNWDFRKEADLQVHVKSH	LGNPAKA
OAZ/F19	HKCIFCGETFSTEVELQCHITTH	SKK
PLZF/F1	EQCSVCGVELPDNEAVEQHRKLH	SGMKT
PLZF/F2	YGCELCGKRFLDSLRLRMHLLAH	SAGAKA
PLZF/F3	FVCDQCGAQFSKEDALETHRQTH	TGTDMA
PLZF/F4	VFCLLCGKRFAQQSALQQHMEVH	AGVRS
PLZF/F5	YICSECNRTFPSHTALKRHLRSH	TGDHP
PLZF/F6	YECEFCGSCFRDESTLKSHKRIH	TGEKP
Rag1/FA	VKCPAKECNEEVSLKYNHHISSH	KESKEIFVHI***
Rag1/FB	YICTLCDATRLEASQNLVFSITRSH	AENLE*
rOAZ/F29	YDCSQCPQKFFFQTELQNHTMSQH	AQ**
Sp1/F1	HICHIQCGKVYGKTSHLRAHLRWH	TGERP
Sp1/F2	FMCTWSYCGKRFTSRDELQRHKRTH	TGEKK
Sp1/F3	FACPECPKRFRMSDHLKSHIKTH	QNKKG*
YY1/F1	IACPHKGCTKMFRDNSAMRKHLHTH	GPRV
YY1/F2	HVCAECGKAFVSSKLKRHLVH	TGEKP
YY1/F3	FQCTFEGCGKRFSLDFNLRTHVRIH	TGDRP
YY1/F4	YVCPFDGCNKKFAQSTNLKSHILTH	AKAKN*
Zac1/F2	YKCVQPDGKAFVSRYKLMRHMATH	SPQKS
Zac1/F6	HQCDHCERCIFYTRKDVRRLVH	TGCKD
Zac1/F7	FLCQFCAQRFRGRKDHLTRHTKKTH	SQELM*
Zap1/F1	LKCKWKECPESCSSLFDLQRHLLKDH	VSQDFKHMPMEP
Zap1/F2	LACNWEDCDFLGDDTCSIVNHINCQH	GINFDIQFAN***

* Truncated at 5 residues, last C2H2 domain in a string.

** Protein ends after residue.

*** Linker is longer than 10 residues shown here.

Table 2.2: TGEKP Linker in Multiple Proteomes

Species	Domains Ending in TGEKP	Domains NOT Ending TGEKP
<i>Anabaena variables</i>	0%	100%
<i>Anopheles gambiae</i>	8%	92%
<i>Apis mellifera</i>	9%	91%
<i>Arabidopsis thaliana</i>	0%	100%
<i>Bradyrhizobium japonicum</i>	0%	100%
<i>Caenorhabditis briggsae</i>	6%	94%
<i>Caenorhabditis elegans</i>	4%	96%
<i>Canis familiaris</i>	39%	61%
<i>Chlamydomonas reinhardtii</i>	0%	100%
<i>Drosophila melanogaster</i>	8%	92%
<i>Escherichia coli</i>	0%	100%
<i>Gallus gallus</i>	10%	90%
<i>Homo sapiens</i>	41%	59%
<i>Mesorhizobium loti</i>	0%	100%
<i>Mus musculus</i>	39%	61%
<i>Neurospora crassa</i>	7%	93%
<i>Nicotiana tabacum</i>	0%	100%
<i>Oryza sativa (cv. japonica)</i>	0%	100%
<i>Pan troglodytes</i>	43%	57%
<i>Pyrobaculum aerophilum</i>	0%	100%
<i>Pyrococcus abyssi</i>	0%	100%
<i>Rattus norvegicus</i>	27%	73%
<i>Saccharomyces cerevisiae</i>	10%	90%
<i>Schizosaccharomyces pombe</i>	2%	98%
<i>Sinorhizobium meliloti</i>	0%	100%
<i>Solanum tuberosum</i>	0%	100%
<i>Synechococcus sp.</i>	0%	100%
<i>Synechocystis sp.</i>	0%	100%
<i>Zea mays</i>	0%	100%

2.3 Regional Spacing Variation

2.3.1 Background

As mentioned briefly above, C2H2 zinc fingers can vary in the number of amino acids between the conserved cysteines and between the conserved histidine residues. In fact, this variation is so prevalent that several different consensus sequences exist to describe the C2H2 domain. Among these consensus sequences, the most common allows for 2 to 5 amino acids between the paired cysteines, and 3 to 5 amino acids between the paired histidines (Hulo et al., 2008). This analysis considers the variation in spacing to determine if it is related to binding type.

2.3.2 Results

A total of 44 DNA binding ZFs were compiled from the public transcription factor database TRANSFAC (release 7.0, (Matys et al., 2003)) and additional sources. Also, 61 protein-binding ZFs were obtained from the protein interaction databases BOND (Alfarano et al., 2005); HPRD (Peri et al., 2003); and MINT (Chatr-aryamontri et al., 2007), as well as additional sources. After initial selection, a thorough literature review was conducted for each protein to ensure that the contribution of individual C2H2 domains to the interaction had been tested. Known cases of binding type overlap were treated as exceptions and excluded from this analysis. Because each list represents a very small subset of C2H2 domains, any variation detected could be due to random chance. Therefore, a random file of C2H2 domain sequences containing 51 sequences was compiled, with no prior knowledge of function, from a list of C2H2 domains generated by scanning the UniProtKB/Swiss-Prot (release 54.7) using ScanProsite and the pattern

X(2)-C-X(2,4)-C-X(3)-[LIVMFYWC]-X(8)-H-X(3,5)-H-X(5) (Hulo et al., 2008; Sigrist et al., 2002). Each list contained C2H2 domains from *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Xenopus laevis*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*.

Spacing categories were based on the number of amino acids between both of the conserved amino acid pairs (C2C + H3H, C2C + H4H, C2C + H5H, C4C + H3H, C4C + H4H, C4C + H5H). There were no sequences with the C3C spacing. As shown in figure 2.1, the most common spacing arrangement was two amino acids between the cysteines and three amino acids between the histidines (C2C + H3H type) which described the spacing for ~50% of the fingers regardless of binding type.

Almost 30% of the protein binding fingers had C2C + H4H spacing, while only ~17% of the DNA binding fingers contained this type of spacing (Figure 2.1). Similarly, a greater percentage of protein binding fingers had C2C + H5H and C4C + H4H spacing than DNA binding fingers. Conversely, ~23% of the DNA binding fingers had C4C + H3H spacing, but only 6% of the protein binding fingers had the same spacing (Figure 2.1). Unlike the protein and DNA binding fingers, the distribution of the random fingers, outside of a preference for the C2C + H3H spacing, was more even (Figure 2.1).

Distribution of sequences among the categories was assessed by a two-tailed fisher's exact test for significance (Kirkman, 1996; Weisstein, 2008), and the percent difference for each category was calculated. Analysis of the variation in the number of amino acids between conserved residues indicates statistically significant associations between binding type and spacing type (Table 2.3, fisher's exact test, $p = 1.2E-05$, p_{cutoff}

= 0.011). Percent deviation from expectation was analyzed to determine which spacing types were associated with each binding types. Expectation values, as determine by the fisher's exact test, are listed in Table 2.2. DNA binding fingers contained more C4C + H3H fingers than expected and less of the C2C + H5H (Figure 2.2). Protein binding fingers, on the other hand, contained more C2C + H4H fingers, and less C4C + H3H fingers than expected, while the other spacing types were similar to expectation (Figure 2.2).

Table 2.3: Distribution of C2H2 ZF Domains in Each Category				
Spacing Type¹		Protein²	DNA³	Random⁴
C2C + H3H	Observed	29	25	23
	Expected	30.1	21.7	25.2
	% Deviation	-3.7%	15.2%	-8.7%
C2C + H4H	Observed	18	7	6
	Expected	12.1	8.74	10.1
	% Deviation	48.8%	-19.9%	-40.6%
C2C + H5H	Observed	7	1	9
	Expected	6.65	4.79	5.56
	% Deviation	5.3%	-79.1%	61.9%
C4C + H3H	Observed	3	10	4
	Expected	6.65	4.79	5.56
	% Deviation	-54.9%	108.8%	-28.1%
C4C + H4H	Observed	4	1	5
	Expected	3.91	2.82	3.27
	% Deviation	2.3%	-64.5%	52.9%
C4C + H5H	Observed	0	0	4
	Expected	1.56	1.13	1.31
	% Deviation	-100.0%	-100.0%	205.3%

¹Spacing Type defined by the number of amino acids between the conserved Cysteines (C) and Histidines (H). C2 = two amino acids between conserved cysteines, C4 = 4 amino acids between conserved histidine residues, H3 = 3 amino acids between conserved histidine residues, H4 = 4 amino acids between conserved histidine residues, H5 = 5 amino acids between conserved histidine residues. Observed = number of domains with the spacing type for each functional category. Expected = as determined by two-tailed Fisher's exact test. % deviation = [(observed – expected)/expected] x 100.

²Protein binding C2H2 domains.

³DNA binding C2H2 domains.

⁴Domains with unknown binding type.

Figure 2.1: Analysis of variation in the number of amino acids between the conserved cysteine pairs and histidine pairs in C2H2 domains.

156 C2H2 zinc finger sequences were analyzed for the number of amino acids between each of the cysteines and each of the histidines. Sequences were further categorized by functional type (protein binding, DNA binding or random). C2C = two amino acids between conserved cysteines, C4C = 4 amino acids between conserved histidine residues, H3H = 3 amino acids between conserved histidine residues, H4H = 4 amino acids between conserved histidine residues, H5H = 5 amino acids between conserved histidine residues. Bars represent the percent of each binding type.

Figure 2.2: Deviation from expected values for each category.

156 C2H2 zinc finger sequences were analyzed for the number of amino acids between each of the cysteines and each of the histidines. Sequences were further categorized by functional type (protein binding, DNA binding or random). C2 = two amino acids between conserved cysteines, C4 = 4 amino acids between conserved histidine residues, H3 = 3 amino acids between conserved histidine residues, H4 = 4 amino acids between conserved histidine residues, H5 = 5 amino acids between conserved histidine residues. Percent deviation was calculated by the formula: $[(\text{observed} - \text{expected}) / \text{expected}] \times 100$. See Table 2.3 for observed and expected values.

2.3.3 Summary

The results of this analysis demonstrate that there are structural differences between some protein and DNA binding C2H2 domains. The best predictor of binding type appears to be the C4C + H3H spacing group which is strongly associated with DNA binding. Interestingly, the TGEKP linker is also associated with both DNA binding and with H3H spacing (data not shown). As mentioned previously, ~50% of DNA binding fingers in mammals have TGEKP linkers. When the analysis was re-done on the DNA binding domains and non-mammalian domains were excluded, the impact of the H3H spacing becomes even clearer. In this case, ~62% of the domains had the C2C + H3H spacing; furthermore, 100% of the TGEKP linkers were attached to a H3H helix (data not shown).

Studies with protein interacting C2H2 domains that use the α -helix have demonstrated that, in some cases, the binding surface exists along the entire length of the helix (Liew et al., 2005; Simpson et al., 2004; Westman et al., 2004). It is therefore possible that the extra amino acid in the H4H helix would be of more use to a protein binding finger than a DNA binding finger. This might explain the enrichment in protein binding C2C + H4H domains observed in this study.

Other studies have demonstrated the PPI binding surface involves the β -sheets (Merithew et al., 2003). In this case, it might be assumed that the C4C spacing would be more prevalent among the protein binding type, since it allows for additional amino acids that can mediate the interactions. Surprisingly, this study found the opposite to be true; overall, the C4C domains tend to be associated with DNA binding domains. One

explanation for this might be that the function annotation was incomplete. Although special effort was made to exclude domains for which functional overlap was known, many domains have been experimentally tested for only one binding type. In general, this seemed to be particularly true for DNA binding domains (personal observation), probably because the C2H2 domain is primarily known as a DNA binding domain. Therefore, the incomplete annotation is more likely to affect the functional annotation of DNA binding than of protein binding.

Although there were indications that variations in the number of amino acids between the conserved cysteine and histidine residues is related to the finger binding type, it is not entirely predicative of binding types. There are DNA binding fingers that contain C2C + H4H spacing that have been shown to not bind protein (e.g., finger 4 in the Ikaros family of protein (Honma et al., 1999; Kelley et al., 1998; Molnar and Georgopoulos, 1994; Morgan et al., 1997; Perdomo et al., 2000; Sun et al., 1996)), just as there are fingers with C4C + H3H spacing that bind protein but do not participate in sequence specific DNA binding (e.g., finger 1 from WT1 (Stoll et al., 2007)). Finally, because there are so few domains that have been functionally annotated, the actual number of domains within certain categories was low, making it difficult to determine if these results represent real correlations between function and spacing.

2.4 Amino Acid Characterization

2.4.1 Background

Interaction interfaces have previously been computationally probed to identify amino acids involved in the interaction (Skrabanek et al., 2008; Zhou and Qin, 2007). In this approach, amino acids located within and outside the interface are compared in order to identify characteristics that are unique to the interface, or at least sufficiently different from non-interface traits that they are predictive (Chothia and Janin, 1975; Jones and Thornton, 1996, 1997a, b). While meeting with varied success as a predictive indicator, this approach has led to a greater understanding of the characteristics of a binding interface (Zhou and Qin, 2007). Generally, PPI interfaces are enriched in hydrophobic residues, which tend to cluster together, and depleted in other charged residues (Janin et al., 1988; Korn and Burnett, 1991; Lijnzaad and Argos, 1997; Lo Conte et al., 1999; Mueller and Feigon, 2002; Skrabanek et al., 2008; Young et al., 1994; Zhou and Qin, 2007; Zhou and Shan, 2001). For this analysis, the same DNA binding finger and protein binding finger amino acid sequences analyzed above for spacing variation were examined for differences in hydrophobicity and charge.

2.4.2 Results

The same C2H2 fingers analyzed in the section 2.3 were also analyzed here. Initially, amino acid sequences were aligned using publicly available multiple sequence alignment programs (e.g., ClustalW; (Chenna et al., 2003)). However, these programs all had trouble dealing with the variation in the spacing between the pairs of conserved Cys

and His residues. As expected the variable spacing introduced gaps in the alignment; however, the position of gaps and “extra” amino acids varied from finger to finger in the alignment. This was particularly problematic between the His residues, where variation resulted in misaligning some of the C-terminal histidines (data not shown).

Analysis of the solved structures of several C2H2 domains reveals that the “extra” amino acids were always inserted C-terminally, as were the histidine residues, regardless of the number of amino acids or C2H2 domain function (Figure 2.3). Therefore, the alignment of these regions should be: Cxx**C or CxxxxC; and, Hxxx**H, Hxxxx*H, or HxxxxxH, where x represents an amino acid, and * indicates a gap. Because maintaining structure related alignments seemed critical, alignments generated by ClustalW were “corrected” so that the conserved Cys and His residues always aligned with each other, and the gaps between them were located C-terminally. The complete alignments for DNA binding, protein binding, and random fingers can be seen in figures B.1, B.2 and B.3, respectively.

After alignments were generated, sequence logos were created for each alignment using Weblogos (Crooks et al., 2004; Schneider and Stephens, 1990). Sequence logos are superficially similar to consensus sequences; however, there are differences in the way the two are generated and subtle differences can be detected with logos that are lost in generating a consensus sequence. Sequence logos calculate frequency using a method that includes an assessment of the value of the information (Schneider and Stephens, 1990). Relative frequency information for amino acids at particular positions in the sequence is conveyed by the height of the letters in the logo. Conservation is conveyed

by the overall height of the stack of letters for each position (Schneider and Stephens, 1990). As a result, logos highlight predominate residues, provide visual representations of the relative frequency of amino acids at each position, and provide an assessment of the value of degree of conservation as well as the value of the information (Crooks et al., 2004; Mount, 2001; Schneider and Stephens, 1990).

Only minor differences in the logos for DNA and protein binding fingers were observed. Within the first β -sheet there is very little difference in the logos (Figure 2.4, positions 1 to 5). One point of interest is the ranking of phenylalanine (F) and tyrosine (Y) at position 1 which appear to be reversed in DNA and protein binding fingers. However, this association is very weak. DNA binding fingers have a tyrosine 34% of the time and a phenylalanine 31% of the time, while protein binding fingers have phenylalanine 36% of the time and a tyrosine 32% of the time.

Figure 2.3: Location of variable amino acids in C2H2 domains.

Finger 1 and finger 3 of TFIIIA (DNA binding, blue ribbons), finger 3 from FOG1 and finger 6 of Eos (protein binding, yellow ribbons) and finger 1 from Zif268 (binds both DNA and protein, green ribbon) were compared for differences in the location of the “extra” amino acids in C2H2 domains with variable spacing. Zinc ions shown as gray spheres. (A) Regardless of domain function, the first two amino acids (red and blue sticks, respectively) in between conserved Cys residues (black sticks) align and extra amino acids are located C-terminally. There is only slight variation in the orientation of the amino acids at each position, except for in Eos 6, a modeled structure. (B) The asparagine (yellow stick) immediately following the first cysteine in Eos 6 is oriented $\sim 180^\circ$ away from the core, whereas other residues in this position are oriented $\sim 90^\circ$ from the core. (C) Regardless of domain function, extra amino acids between the two His residues (black sticks) are also inserted C-terminally, although the actual orientation of the third residue is variable. (D) The orientation of the amino acids in the third position between the conserved His residues is not related to the binding type.

Figure 2.4: Sequence logos for DNA binding, protein binding and random C2H2 domains.

Sequence logos for DNA and protein binding, as well as random, C2H2 domains were created using Weblogos (Crooks et al., 2004). X-axis indicates position in the sequences; y-axis indicates bit score for the total stack at each position. Amino acids were grouped according to the biochemical characteristics of their R-groups as follows: hydrophobic (FWILMVA), neutral (GPSTY), hydrophilic (QN), basic (KR), or acidic (DE; (Pommie et al., 2004)). Green = hydrophobic, Blue = neutral, Red = basic, Yellow = acidic, Purple = hydrophilic. Conserved cysteine and histidines are shown in black. The protein binding fingers has a gap in the alignment due to an insertion in one sequence (see alignment figure B.2), the gap is indicated by an (*) in the x-axis of the protein binding sequence logo.

Positions 6-7 in figure 2.4 correspond to the location of “extra” amino acids in the C2C versus C4C spacing. The “extra” residues in this position serve to lengthen the loop between the two β -sheets; and, amino acids in this location are oriented away from the core of the fold and away from the DNA (Figure 2.3). DNA binding fingers have a stronger preference for negatively charged amino acids at these positions, particularly position 6, that is not observed in protein binding fingers.

In figure 2.4 positions 8-13 correspond to the second β -sheet. In certain DNA binding fingers, amino acids in this region can make contact with the phosphate backbone (Pavletich and Pabo, 1991, 1993; Wolfe et al., 2000). Because of the distance, this interaction would require a large, positively charged amino acid. This likely explains the slightly enriched lysine (K) at position 10 in the DNA binding fingers versus protein binding fingers (Figure 2.4).

The α -helix is described by positions 14-27 in figure 2.4. Overall, one striking feature of this region is the difference in the level of conservation observed in the DNA binding fingers versus the protein binding finger and the random sequences. In general, there was more variety in the amino acids at every position in the α -helix in the protein binding fingers.

The N-terminal portion of the α -helix, positions 14-20 had only minor differences. Among them, the most interesting is the preference for a negatively charged amino acid, typically an aspartate (D), in position 14 of the protein binding fingers that is not observed in either the DNA binding fingers or the random fingers (Figure 2.4). This position corresponds to the amino acid immediately N-terminal to the start of the α -helix,

and typically contacts nucleotides in DNA binding fingers (Wolfe et al., 2000). At the same position in the DNA binding fingers, there is a preference for a polar or positively charged amino acid (Figure 2.4). Since the overall charge of DNA is negative, it is unlikely that an aspartate in this position would contact DNA. Finally, a leucine (L) is more conserved at position 18 in DNA binding fingers. As a large polar residue, this amino acid side chain contributes to maintaining the domain fold. Protein binding domains seem to have frequent substitutions for other amino acids at this point, however since the other small polar amino acids are most frequently substituted, it is difficult to interpret this result.

The second half of the α -helix, positions 21-27, contained some of the more interesting variation. In DNA binding fingers, the C-terminal portion of the α -helix is not involved in contacting the DNA. In contrast, protein-protein interaction contacts between FOG and EOS and their respective partners are made with amino acids in this region (Liew et al., 2005; Westman et al., 2004). In the protein binding fingers, position 22 had a higher frequency of large hydrophobic residues (LMI), frequently found in protein interfaces (Janin et al., 1988; Lijnzaad and Argos, 1997), than the DNA binding fingers. There is also a difference in the frequency of arginine in position 23 in DNA binding finger as compared to protein. Positions 25 and 26 correspond to the insertion points for extra residues in H3H versus H4-5H fingers.

As expected there was a strong signal for threonine, glycine, glutamate, lysine and proline (TGEKP, respectively) at positions 28-32. This region is the linker region connecting two domains in multi-fingered proteins. Not surprisingly, given the findings

in section 2.2, there is not a corresponding linker pattern observed in protein-binding fingers (Figure 2.4).

2.4.3 Summary

The amino acid composition at specific positions did vary between DNA and protein binding fingers. However, this variation was very small and unlikely to be significant. Based on this analysis, it would be very difficult to predict the binding type of an unknown C2H2 domain. One possible reason for this may be that a more complex analysis, such as one that looks at relationships between the positions, is required. For example, in the sequence logos presented here, there were minor differences in the composition of amino acids located at positions 1 and 14 when DNA binding and protein binding fingers were compared. In the folded structure, amino acids in these two positions interact, with the amino acid in position 1 stacking on the amino acid in position 14, helping to orient the residue (Wolfe et al., 2000). It is possible that a more complex analysis would be able to detect whether or not the composition of amino acids at related positions varies with binding type.

2.5 Discussion

In order to find characteristics that can be used to distinguish between DNA and protein binding types of C2H2 domains, three traits were examined: the linker region, the spacing of conserved residues, and the amino acid composition for each position. Individually, none of these traits is particularly prognostic of binding type. Each of these traits seems to be; at best, only weakly associated with a particular binding type. Of these, a difference in the number of amino acids between the paired Cys and His residues is perhaps the best indicator of binding type. There is also a strong association between

the TGEKP linker and DNA binding; however, this association is limited to mammalian species. Finally, while there was a preference for a negatively charged amino acid in the N-terminal portion of the α -helix in protein binding fingers, a feature that is likely to inhibit DNA binding, this feature is not found in all sequences.

One explanation for the failure of this analysis to find distinguishing characteristics is that the complexity of the task is very high, while the quality of the input data is only marginal. Although this analysis utilized the best characterized C2H2 domains available, the majority have only been tested for one type of binding, and nothing is known about their ability to interact with other biomolecules. At the same time, it is very likely that the sequences analyzed here contain a mixture of binding surfaces. Evidence for this comes from the fact that for each of the traits examined here, the DNA binding fingers had a stronger “signal” than the protein binding fingers did.

Unlike the binding surface for DNA binding fingers, which is both well defined and localized to a small region on the C2H2 finger, protein binding fingers can utilize many different regions of the fold (Brayer and Segal, 2008). Given the enormous effort defining the binding surface requires; and, if potential binding surface areas are as diverse as suggested by the few domains for which it has been described, then complete understanding of C2H2 protein binding is likely years away. In the absence of such knowledge, future informatics based studies should focus on more complex types of analysis such as structure based modeling and simultaneous consideration of multiple traits, each of which contributes to an overall likelihood of binding type. Machine

learning techniques, such as hidden markov modeling or neural networks, may prove more successful.

CHAPTER III: PREDICTING BINDING FUNCTION

3.1 Introduction

Since the completion of the human genome project, as well as the completion of other genome projects, focus has shifted to understanding the function of the genes embedded within these genomes. Currently, elucidating protein function represents a large portion of this effort. One way in which function is being examined is by determining protein interaction capabilities, which provides context for the protein within cellular networks and provides information regarding the function of unknown proteins (Lee et al., 2007; Nariai et al., 2007). Understanding interactions also allows the manipulation of these interactions, potentially leading to the development of new pharmaceuticals.

Historically, transcription factors have been poor drug targets because the DNA-protein interface is often large, and neither the DNA nor the interacting surface of the protein presents a well-defined binding pocket for small molecule binding. On the other hand, transcription factors also participate in PPIs, which are more likely candidates for drug targeting (Arndt, 2006; Jung et al., 2006). As the largest class of protein domains found in transcription factors, the C2H2 zinc finger domain represents an attractive target for the development of drugs that can alter or disrupt gene function.

Until recently, understanding protein interactions has been limited to labor intensive experiments with individual proteins. High-throughput methods such as microarray and mass spectrometry have been developed which increase the number of

proteins that can be examined at a time; however, these methods are expensive, prone to error and not universally applicable (Skrabanek et al., 2008). Increasingly, computational methods are being developed for large-scale prediction of protein interactions and are being used either in conjunction with, or in lieu of wet lab work (Skrabanek et al., 2008).

Predicting PPI by C2H2 domains presents several challenges. Of these, the greatest is that the C2H2 domain is capable of binding many different types of biomolecules, including DNA, protein, RNA and possibly lipids (Brown, 2005; Hall, 2005; Lee et al., 2006; Lu et al., 2003; Merithew et al., 2003). As exemplified by the Ikaros family of proteins, in which 4 of the 6 C2H2 domains bind DNA and the remaining 2 bind protein, individual or clusters of C2H2 domains within a single protein may interact with multiple types of biomolecules (Hahm et al., 1998; Molnar and Georgopoulos, 1994; Morgan et al., 1997; Perdomo et al., 2000). In addition, individual fingers can separately interact with multiple types of biomolecules, as in the case of Gli proteins which use F3-5 to interact with both DNA and protein, but probably not at the same time (Koyabu et al., 2001). Finally, as seen in Zac1 finger 2, individual fingers can also interact with both protein and DNA at the same time (Hoffmann et al., 2006). This means that prior to predicting any PPIs by C2H2 domains, it is necessary to predict the biomolecule(s) with which individual domains are capable of interacting.

This chapter represents an effort to predict the binding type for C2H2 domains. Profile Hidden Markov Modeling, a robust technique frequently used to analyze complex data sets, was used to generate two predictive models, one for protein binding, and one for DNA binding. Predicting protein binding proved more difficult than predicting DNA

binding, possibly because DNA binding relies on a well-defined region of the domain, while the protein binding surface is less well defined. In an effort to separate protein binding domains by their interaction surface, a cluster analysis was also attempted.

3.2 Hidden Markov Modeling

3.1.1 Background

Hidden Markov Modeling (HMM) is one of the most common methods used to analyze complex data sets for the purpose of generating models that can be applied to pattern recognition. Although there are many types and uses for HMMs, profile HMMs are specifically used to analyze sequence composition and patterns (Mount, 2001).

Profile HMMs are implemented by databases such as PFAM and PROSITE to define protein domains, and once developed can be used to locate those same protein domains within other sequences (Bairoch and Apweiler, 1997; Sonnhammer et al., 1998).

Profile HMMs are statistical models that describe the distribution probabilities observed in complex sequences (Eddy, 1996). Probabilities are calculated based on the frequency of each amino acid at a given position in the sequence (Eddy, 1996, 1998).

Unlike other profile methods, HMMs calculate the true frequencies for each position, and therefore, require a smaller number of sequences to derive the model than other profile methods (Eddy, 1998; Mount, 2001). Typically, about 20-100 sequences, referred to as the “training set,” are required to develop a profile HMM (Mount, 2001). Finally, because HMMs describe probabilities over an infinite number of possible sequences, they are uniquely suited to prediction of the likelihood of unknown sequences matching the

model (Eddy, 1998). This means that after training, the resulting profile HMM can be used to search a database for sequences that match the model.

3.1.2 Results

Using HMMER (Eddy, 1998), a profile HMM was built for finding protein binding C2H2 domains. Although HMMs can be built using either aligned or unaligned sequences (Eddy, 1998), aligned sequences were used for this analysis. This decision was made because a previous structure based analysis of the location of gaps and insertions revealed that “extra” amino acids were always inserted C-terminally (Figure 2.3; section 2.4.2). The model was trained using 37 well-characterized C2H2 domain sequences (Figure B.4). The two best characterized fingers, Eos fingers 5 and 6, were left out of the training set so they could be used later for evaluating the model.

After profile construction HMMsearch was used to search the UniProt Knowledgebase, release 12.7, for matches to each profile ((Consortium, 2007). As an output, HMMsearch returns a list of sequences that fit the HMM profile along with two sets of scoring information. First, it provides a bit score which evaluates how well the new sequence matches the HMM profile (Eddy, 2003). A bit score above the $\log_2$ of the number of sequences in the database is a recommended cut-off when using HMMER (Eddy, 2003). There are 5473336 sequences in release 12.7 of the UniProt Knowledgebase, so a lower limit bit score cut-off of 23 was used [$\log_2(5473336) \sim 22.4$]. HMMsearch also provides an E-value for evaluating the statistical significance of the bit score. E-values report the probability that the matching sequence was found

randomly. Smaller E-values are best and $E=0.01$ is the recommended upper limit in HMMsearching (Eddy, 2003).

The constructed profile, henceforth referred to as the protein HMM, was 29 positions in length, and represented the entire C2H2 domain. After construction the profile was used by HMMsearch to find matches within the UniProt knowledgebase. Within the matches, the HMM model found all 6 of the fingers in Eos, despite the fact that only 2 of the 6 are protein binding fingers. However, as can be seen in Table 2.3, the scores indicate that only Eos fingers 2 and 3 should be accepted as fitting the model, while finger 1, 4 and 6 should be rejected (do not fit the model). These findings are in conflict with the known function of the domains 2 and 3, which are not known to bind protein, and with finger 6 which is known to bind protein. In addition, the scores for finger 5 of Eos are marginal, with a good bit score but a bad E-value, making it difficult to accept or reject the model for this finger. Therefore, this model is, at best 50% accurate, and is more likely ~30% accurate.

Table 3.1: HMM Profile Scores

Protein	Finger	HMM Profile	Bit Score¹	E-value²
Eos	1	Protein	9.9	29
Eos	2	Protein	32.9*	3.50E-06*
Eos	3	Protein	27.3*	1.80E-04*
Eos	4	Protein	13.5	2.4
Eos	5	Protein	26.3*	6.70E-02
Eos	6	Protein	1.5	9.4E+03
Fog	3	Protein	9.3	8.70E+03
Eos	1	DNA	33.5*	4.30E-04*
Eos	2	DNA	41.6*	1.60E-06*
Eos	3	DNA	36.6*	5.30E-05*
Eos	4	DNA	30.7*	3.10E-03*
Eos	5	DNA	N/A ³	N/A ³
Eos	6	DNA	N/A ³	N/A ³
Fog	3	DNA	10.3	4.20E+03

¹Significant cutoff of 22 was used based on log2 (5473336 entries in Uniprot) ~22.3

²Significant cutoff of E=0.01 (Eddy, 2003)

³Not Detected

*Indicates significant score.

Because the protein HMM profile was not discriminatory, and because the early sequence trait analyses had hinted that DNA binding fingers have a more distinct signature, a second profile HMM was also built using DNA binding domain sequences (Figure B.5). Again, some sequences were withheld to evaluate the effectiveness of the model. In this case, Eos fingers 1-4, known to be DNA binding fingers, were held back. Using DNA binding as the criteria for building the model proved much more accurate. Fingers 5 and 6 of Eos were not detected when using the DNA profile to search Uniprot. Although finger 3 of Fog was detected, both the bit score and the E-value were not significant. Finally, both the bit score and the E-values were significant for fingers 1-4 of Eos. These findings agree with the known function for the domains, suggesting that the DNA profile HMM is highly discriminatory.

3.1.3 Summary

Despite the success of the DNA HMM in discriminating between DNA and non-DNA binding C2H2 domains, the overall goal of developing a method for predicting protein binding by C2H2 domains was not met. While hidden markov modeling is one of the most powerful tools for developing predictive models, profile HMMs are only as good as the training set. Two shortcomings of the dataset likely contributed to the disappointing results of the project. First, although the domains used in the training set are the best characterized protein binding C2H2 domains, very little is actually known about the full capabilities of individual domain on the list. For example, while it is known that all the domains bind proteins, most were not tested for their ability to bind DNA. Secondly, the training set contained domains with at least two different binding

surfaces. As discussed in greater detail in section 3.2, some of the domains are known to use their α -helix as a protein binding surface, while others use their β -sheets. Thus, as with the previously described attempts, the heterogeneous nature of the C2H2 domain's protein binding surface likely contributes the difficulty in predicting this binding type.

3.1.4 Methods

HMMbuild at the Mobyle Portal (<http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py#>) was used to construct the profile HMMs (Eddy, 1996, 1998; Letondal, 2001). HMMbuild variables were set as follows: -swentry $x = 0.75$; -swexit $x = 0.75$. The models were trained on hand aligned sequence files (Figure B.4 and B.5) because previous structure based analysis of the location of gaps and insertions revealed that “extra” amino acids were always inserted C-terminally (section 2.2.3). In total, the protein binding model was trained on 37 sequences, and the DNA binding model was trained on 40 sequences. After building, HMMsearch was used to search the UniProt Knowledgebase, release 12.7, for matches to each profile (Consortium, 2007).

3.2 *Binding Surface Discrimination by Cluster Analysis*

3.2.1 Background

The location of the PPI binding surface in C2H2 domains represents a second challenge to predicting this function. While the DNA-binding surface of C2H2 domains is invariably located in the α -helix, the protein-binding surface may be located in other regions of the fold (Brayer and Segal, 2008). This adds to the burden of PPI prediction, because there is no clear-cut, predictable binding surface used by all C2H2 domains.

One weakness of the approaches to functional assignments utilized in Chapter II, as well as with the prediction of binding type by profile HMM is the assumption of a homogeneous binding face such as that observed in DNA binding fingers. Virtually all contacts to DNA bases extend from the N-terminus of the α -helix. However, the protein binding surfaces appears to be more heterogeneous (Brayer and Segal, 2008). As detailed in section 1.4, C2H2 ZF-protein interaction surfaces can involve the “DNA face” of the helix, the side of the helix, the β -strands, or regions involving both the α -helix and the β -strands. Therefore, a simple alignment of known protein-binding ZF sequences might not be expected to reveal a protein-binding signature. This study examined whether cluster analysis, combined with amino acid characterization, would enable the detection of multiple protein binding signatures and if those could be correlated with binding surface.

3.2.2 Results

The same protein binding C2H2 domains analyzed in section 3.1 were clustered using PhyML, an algorithm used to estimate phylogeny based on maximum likelihood (Guindon and Gascuel, 2003; Guindon et al., 2005). Although clusters can be constructed using a variety of methods, maximum likelihood was chosen because this method is best when there is no clearly recognizable pattern (Mount, 2001). After clustering, groups of sequences were analyzed for differences in amino acid composition using sequence logos (Crooks et al., 2004; Schneider and Stephens, 1990).

As shown in figure 3.1, PhyML created clearly defined clusters. In addition, the sequence logos revealed differences in the amino acid composition for each of the three

groups analyzed. Group 1, containing 13 domains, is enriched in negatively charged amino acids in the α -helix, particularly in position 17 which corresponds to position 2 of the α -helix (Figure 3.1). Group 2, comprised of 12 sequences, also contains negatively charged amino acids in the α -helix located in position -1 of the α -helix (Figure 3.1). Interestingly, Group 3 which contains 16 domains, displays a clearly defined TGEKP linker. This feature, as well as enrichment in positively charged amino acids in the α -helix and positions 10 and 11 is reminiscent of the DNA binding sequence logos observed in section 2.4 (Figures 2.4, 3.1).

In order to determine whether or not these groups corresponded to known binding surfaces, the placement of domains with known binding surfaces was examined. The most interesting association is the clustering of the C2H2 domains from PolH, fingers 5-6 from Eos and the CCHC-type finger from Fog in Group 2 (Figure 3.1). All of these fingers are known use the “DNA binding face” of the α -helix. However, the second finger of Zap1 is also in this group (Figure 3.1). Unlike PolH, Eos fingers 5 and 6, and the CCHC finger of Fog which all use the “DNA binding” face of the helix, Zap1 finger 2’s binding face is located about a quarter-turn counterclockwise (Figures 1.6, 1.8 and Figure 1.9; (Bomar et al., 2007; Wang et al., 2006; Westman et al., 2004; Westman et al., 2003).

The placement of finger 1 of WT1 and finger 7 of Zac1 in Group 3 is also interesting. While both of these fingers are protein binding fingers, they are also part of zinc finger clusters that bind DNA. Structure analysis has shown that finger 1 of WT only makes limited DNA contacts, none of which are sequence specific; however, the

extent of DNA binding by Zac1 finger 7 is currently unknown (Hoffmann et al., 2003).

As with Group 2, the assignment of certain domains to this group is also difficult to explain. For instance, OAZ finger 29 is in this group; and, despite extensive testing, DNA binding has never been demonstrated for this finger (Brayer et al., 2008; Hata et al., 2000; Tsai and Reed, 1998).

Figure 3.1: C2H2 domains clusters reveal amino acid traits related to binding type.

Fifty one (51) protein binding C2H2 domains (Figure B.6) were clustered using PhyML, an algorithm used to estimate phylogeny based on maximum likelihood (Guindon and Gascuel, 2003; Guindon et al., 2005). After clustering, groups of sequences were analyzed for differences in amino acid composition using sequence logos (Crooks et al., 2004; Schneider and Stephens, 1990). X-axis indicates position in the sequences; y-axis indicates bit score for the total stack at each position.

Perhaps the most problematic assignment is the clustering of FOG's third finger and the single finger of EEA1 together in Group 1. Fog's protein binding surface has been clearly demonstrated to be in the α -helix (Fig 1.7, (Simpson et al., 2004)). Although there is no structural information, mutation studies have demonstrated the PPI binding surface for EEA1 is likely located in the β -sheets (Fig 1.10, (Merithew et al., 2003)). Arguably, Group 1 could be split into two groups, which would separate Fog finger 3 and EEA1. Interestingly, separation of Group 1 into two smaller groups reveals a high frequency of negatively charged amino acids in positions 15, 17, and 20 for the cluster containing Fog F3 (Figure 3.2). These positions correspond to positions -1, 2, and 5, of the α -helix, and have been shown to be critical positions in DNA binding, and in the interaction between FOG and TACC3 (Simpson et al., 2004).

Figure 3.2: Splitting Group 1 into two sub-groups reveals distribution of negatively charged amino acids along “DNA binding face” of the α -helix.

Fifty one (51) protein binding C2H2 domains (Figure B.6) were clustered using PhyML, an algorithm used to estimate phylogeny based on maximum likelihood (Guindon and Gascuel, 2003; Guindon et al., 2005). After clustering, groups of sequences were analyzed for differences in amino acid composition using sequence logos (Crooks et al., 2004; Schneider and Stephens, 1990). X-axis indicates position in the sequences; y-axis indicates bit score for the total stack at each position.

3.2.3 Summary

Although limited in scope, this study was comparatively successful. Primarily, the results of this analysis support the idea that no single “signature” can define protein binding C2H2 domains. Unlike C2H2-DNA interactions, the interacting residues responsible for specific C2H2 PPIs have been significantly less studied. Although only four C2H2 domains known to be involved in protein-protein interactions have had their structures solved and/or have been examined by scanning mutagenesis studies for the location of critical residues, those studies have revealed the presence of at least four binding surfaces, two regions of the α -helix, the β -sheets, and the linker region (section 1.4 and (Brayer and Segal, 2008)).

Furthermore, there is compelling evidence to expect an even more diverse assortment of protein binding surfaces. This evidence comes quite unexpectedly from a class of scorpion toxins that bind potassium channels. Many of these toxins have a nearly identical structure to C2H2 ZFs, and some bind their target channel using amino acids similar to those used by C2H2 ZFs to bind DNA (reviewed in (Rodriguez de la Vega et al., 2003)). However, others contact their cognate channel using amino acids projecting from the β -strands, and still others use residues from the loop region between the α -helix and β -strands.

In this respect, this study was overly simplistic. The PhyML generated tree was initially divided into only three clusters. It is likely that this was too few clusters, and each cluster contains domains with different binding surfaces which obscured the binding

surface “signal”. Some evidence to support this comes from the result of splitting Group 1 into 1a and 1b. After splitting Group 1, the increased frequency of negatively charged amino acids distributed along the α -helix in Group 1a, containing domains known to use the α -helix for PPIs, is much more evident.

Finally, as with the spacing variation analysis described in section 2.3, incomplete annotation is likely confounding the results obtained here. Special effort was made to exclude domains for which functional overlap was known; however, many domains have only been experimentally tested for one binding type. Evidence for incomplete annotation can be seen in Group 3 which strongly resembles the logo observed for DNA binding domains in section 2.4, suggesting at least some of the domains in Group 3 are likely to bind DNA in addition to protein (Figures 2.4, 3.1).

3.2.4 Methods

A hand aligned file containing 51 protein binding C2H2 domains (Figure B.6) was clustered using the PhyML, an algorithm used to estimate phylogeny based on maximum likelihood (Guindon and Gascuel, 2003; Guindon et al., 2005). The default substitution model based on Whelan and Goldman (Whelan and Goldman, 2001) was selected, as was the default substitution rate of 4 (Guindon and Gascuel, 2003; Guindon et al., 2005). After clustering, groups of sequences were analyzed for differences in amino acid composition using sequence logos (Crooks et al., 2004; Schneider and Stephens, 1990).

3.3 Discussion

Predicting protein binding for individual C2H2 domains proved to be a much more difficult task than anticipated; however, predicting DNA binding was more successful. There are several possible explanations for the inability to predict protein binding domains. One concern is that the interface surfaces for protein binding are considerably more heterogeneous than those for DNA binding. Virtually all contacts to DNA bases extend from the N-terminus of the α -helix. However, as detailed in Chapter I, C2H2-protein interaction surfaces can involve the “DNA face” of the helix, the side of the helix, the β -strands, or regions involving both the α -helix and the β -strands.

Since one challenge to this process is the variable distribution of the protein binding surfaces on the C2H2 domain, attempts were also made to predict the location of the binding surface of known PPI C2H2 domains. This supported the hypothesis that the list of protein-binding C2H2 domains used to build the protein HMM contained domains with multiple binding surfaces. Due to the limited number of sequences in each cluster it is not possible to generate reliable profile HMMs for each cluster; however, this would be an interesting avenue to pursue in the future.

Finally, despite the robust nature of profile HMMs for analysis of complex data sets, the profiles generated are only as good as the input data. As discussed in Chapter II, one explanation for the failure to find distinguishing sequence characteristics between the binding types was that most of the domains used came from studies that did not attempt a rigorous identification of the full potential functionality of the ZFs; and, it is therefore likely that some of the domains are incompletely annotated. The DNA profile HMM was

able to predict DNA binding with reasonable success. And the binding surface cluster analysis revealed sequence features that “make sense” when compared to the distribution of C2H2 domains with known binding surfaces. However, concerns about the limited number of sequences used in the analysis, as well as concerns about the functional annotation undermine confidence in the results.

CHAPTER IV: DNA BINDING BY ZINC FINGERS

4.1 Introduction

Long recognized as the DNA binding domain present in many transcription factors, the C2H2 domain's DNA binding abilities are well characterized. Unfortunately, DNA binding is frequently assumed when newly discovered proteins containing C2H2 domains are described, especially those discovered in high-throughput assays. Because of this assumption, previous studies aimed at determining the biological roles of specific ZF proteins have introduced bias because only a subset of all possible interactions are investigated. Typically, a C2H2-protein interaction study would identify a zinc finger protein in a yeast two-hybrid assay, and then implicate the role of particular fingers by deletion analysis. However, such studies rarely investigated, for example, if the protein-binding fingers might support DNA-binding in a different context.

There are several domains that are known to interact with both protein and DNA, including the 3-finger proteins Zif268/EGR1 and SP1, in which all three ZFs are known to bind DNA, and have been shown to mediate protein interactions with their zinc fingers (Chapman and Perkins, 2000; Lee et al., 1993; Seto et al., 1993). When combined with concerns raised while attempting to distinguish and predict binding type based on amino acid sequences, this suggests that many domains have sequence features that allow them to perform multiple functional roles. It is therefore possible that the common functional annotation of a ZF-containing protein as DNA-binding, while not inaccurate, is not sufficient to describe the full functionality provided by the ZF domains. To fully

understand the diverse functions mediated by the enormous family of C2H2 ZF proteins, a more accurate annotation of their component ZF functions is required. Therefore, the hypothesis that the functional roles of C2H2 domains can be more accurately revealed by using a less biased approach was tested.

The unbiased determination of potential C2H2 functions necessitates exposing C2H2 domains to all possible DNA or protein sequences and selecting for interacting pairs. In the case of DNA, this objective can be easily accomplished using a target site selection assay such as cyclical amplification and selection of targets (CAST, (Segal et al., 2003)). In this assay, purified ZF protein is exposed to a library of short double-stranded DNA molecules. Bound ZF-DNA complexes are separated in an electromobility shift assay (EMSA). The bound DNA is then recovered and re-amplified for additional rounds of selection or sequenced to determine preferred ZF binding sites (Figure 4.1). This study uses a randomized 21 bp region of DNA, producing a library of 4^{21} or 4.4×10^{12} DNA sequences. Since each ZF recognizes a 3 bp site (or a 4 bp overlapping site), a DNA library of this size presents every possible DNA binding site to a protein containing up to seven ZFs. Although some proteins have been shown to require only two ZFs for DNA binding (Fairall et al., 1992), typically three or more fingers are required for high affinity binding. Cluster sizes therefore needed to contain no less than three fingers and no more than seven.

For this analysis, the human O/E-1-associated zinc finger protein (OAZ) was used as a model protein. OAZ is a large zinc finger protein containing 30 C2H2 domains arranged into 6 clusters (Figure 4.2). The interaction capabilities of OAZ have been

studied previously, and it is known that OAZ binds two different DNA targets, and it does so by using different sets of zinc finger domains (Hata et al., 2000; Tsai and Reed, 1997, 1998). Working with the rat ortholog, rOAZ, Tsai and Reed (Tsai and Reed, 1997, 1998) determined that a fragment equivalent to the combination of clusters 1 and 2 bind a GC-rich, Sp1-like target. More recently, Hata et al. (Hata et al., 2000) reported human OAZ binds a bone morphogenetic response element (BRE) using the fingers in cluster 3. Importantly, these sites represent two separate transcriptional targets for OAZ (Hata et al., 2000; Tsai and Reed, 1997, 1998).

Six fragments of OAZ, corresponding to naturally occurring clusters of zinc fingers (Figure 4.2), were tested for DNA binding by CAST assay. Results demonstrate that only one fragment was capable of binding DNA by CAST assay, suggesting that DNA binding capabilities are not equal among C2H2 domains.

Step 1. PCR:
Amplify library

Step 4. Elute:
For Re-amplification
or Cloning

Step 2. Binding:
Form Complexes

Step 3. Gel Shift:
Separate Complexes

Figure 4.1: Overview of CAST Assay Steps.

The DNA library is amplified by a PCR reaction. After verifying amplification, the library is mixed with purified protein and bound complexes are separated from unbound library on a native PAGE gel. Shifted DNA is eluted from the gel, purified and re-amplified for the next CAST round, or cloned and sequenced to determine target DNA sequence. See methods for details at each step.

Figure 4.2: Schematic representation of human OAZ protein.

The 30 C2H2 zinc finger domains are indicated by rectangles. Brackets above indicate previously described functions of zinc fingers (Hata et al., 2000; Tsai and Reed, 1998). The ZF clusters used in this study are indicated. The aligned amino acid sequence of hOAZ is shown below the schematic, with clusters and ZFs labeled. Positions -1 to 6 of the α -helix, which are important for DNA binding, and the linker regions, are indicated. Conserved cysteines and histidines are highlighted by gray boxes.

MEDESI

Cluster 1 **-1123456** **linker**
ZF1 YTC**DH**CQQDFE SLADLTD **HRAH** RCPGDGDDDPQLSWVASSPS

SKDVASPTQMIGDGCDDLGLGEEEGGTGLP

Cluster 1 **-1123456** **linker**
ZF2 YPC**Q**FCDKSFI RLSYLKR **HEQIH** SDKLP
ZF3 FKCTYCSRLFK HKRSRDR **HIKLIH** TGDKK
ZF4 YHCECEAAFS RSDHLKI **HLKTH** SSSKP
ZF5 FKCTVCKRGFS STSSLQS **HMQAH** KKNKEHLAKSE KEAKKDD

Cluster 2 **-1123456** **linker**
ZF6 FMC**DY**CEDTFS QTEELEK **HVLTRH** PQLSEKAD
ZF7 LQCIHCPEVVFV DENTLLA **HIHQAH** ANQK
ZF8 HKCPMCPEQFS SVEGVYC **HLDSH** RQPDSSNHVSVPDPV

LGSVASMSSATPDSSASVERGSTPDSTLKPLRGQKKMRDDGQGWTKVV

Cluster 3 **-1123456** **linker**
ZF9 YSC**PY**CSKRDFN SLAVLEI **HLKTIH** ADKPQQS
ZF10 HTC**Q**ICLDSMP TLYNLNE **HVRKLIH** KNHAYPVMQFGNISA
ZF11 FHCNYCPEMFV DINSLQE **HIRVSH** CGPNANPSDGNNA
ZF12 FFCNQCSMGFL TESSLTE **HIQQA**H CSVGSAKLESFVVQPTQSFMEV
ZF13 YSC**PY**CTNSPIFG SILKLTG **HIKENH** KNIPLAHSKKSK

AEQSPVSSDVEVSSPKRQRLSASANSISNGE

Cluster 4 **-1123456** **linker**
ZF14 YPC**NQ**CDLKFS NFESFQT **HLKLIH** LELLLRK
ZF15 QACPQCKEDFD SQESLLQ **HLTVH** YMTTSTH
ZF16 YVCESCDKQFS SVDDLQK **HLDDMH** TFVL
ZF17 YHCTLCQEVFD SKVSIQV **HLAVKH** SNEKKM
ZF18 YRCTACNWDFR KEADLQV **HVKHSH** LGNPAKA
ZF19 HKCIFCGETFS TEVELQC **HITTH** SKK
ZF20 YNC**KFC**SKAFH AIILLEK **HLREKH** CVFDAATENG

TANGVPPMATKKAEPADLQGMLLKNPEAPNSHEASEDDVDASEPM

Cluster 5 **-1123456** **linker**
ZF21 YG**CD**ICGAAYT MEVLLQN **HRLRDH** NIRPGEDDGSRKAEFIKGS
ZF22 HKCNVCSRTFF SENGLRE **HLQTH** RGPAAH
ZF23 YMC**PIC**GERFP SLLTLTE **HKVTH** SKSLDT
ZF24 GTCRICKMPLQ SEEEFIE **HCQMH** PDLRNSLTG
ZF25 FR**CVVCM**QTVT STLELKI **HGTFH** MQKLAGSSAAS SPNGQGLQKL

YKCALCLKEFRSKQDLVKLDVNGLPYGLCAGCMARSANGQVGGLAPPEPADRPCAG

Cluster 6 **-1123456** **linker**
ZF26 LR**CP**ECSVKFE SAEDLES **HMQVDH** RDLTPETSGPRKGTQTSPVPRKKT
ZF27 YQ**CI**KCQMTFE NEREIQI **HVANH** MIEEGIN
ZF28 HECKLCNQMFV SPAKLLC **HLIEH** SFEGMGGT
ZF29 FKCPVCFTVFV QANKLQQ **HIFAVH** GQEDKI
ZF30 YDC**SCQ**CPQKFF FQTELQN **HTMSQH** AQ

4.2Results

The six clusters of hOAZ, corresponding to zinc fingers 2-5, 6-8, 9-13, 14-20, 21-25, and 26-30, were subcloned, purified from bacteria as fusions with maltose binding protein (MBP), and applied to CAST analysis. As a positive control for these experiments, a well-characterized 6 finger protein, Aart, was also purified as an MBP fusion and used in the CAST assay (Dreier et al., 2001). This protein not only controls for the technical aspects of the experiment, but also produces a strong enrichment of its preferred target in the early rounds of selection due to its unusually high affinity (70 pM). The visible band of shifted DNA serves to indicate the position of the nearly invisible bands of shifted DNA in samples with proteins of lower affinity.

After four rounds of selection, a substantial portion of the probe DNA was shifted by the Aart positive control protein (Figure 4.3). Only faint bands of shifted DNA were visible in the samples containing Clusters 1-6, which could have been due to either specific or non-specific ZF-DNA interactions. However, it was no longer possible to amplify the excised shifted DNA for Clusters 2-6 in subsequent rounds (Figure 4.4), suggesting the observed bands did not represent specific interactions and were not selected subsequently. These results indicated that Clusters 2-6 were unable to bind specific DNA sequences under these experimental conditions.

In contrast, the shifted DNA in the sample containing Cluster 1 continued to intensify in subsequent rounds, indicating positive selection of specific targets (Figure 4.3). After seven rounds of selection, 15 random clones of the Cluster 1 output were sequenced and revealed a consensus sequence of 5'-TGGGTGTCA-3' (Figure 4.5)

Figure 4.3: CAST analysis of specific DNA-binding activity.

TOP: After 4 rounds of CAST, the autoradiogram shows that some portion of the labeled DNA library is still present in the assay (lower band), but very little is bound by the six clusters of zinc fingers (C1-C6). In contrast, a large proportion of labeled DNA is clearly bound (upper band) by the positive control protein Aart (+).
 BOTTOM: After 7 rounds of CAST, a large portion of the DNA in the Cluster 1 sample and nearly all the DNA in the positive control sample are bound to protein, indicating a strong enrichment for binding sites in these samples. None of the other clusters contained bound or unbound library DNA (not shown), indicating the interactions were too weak to allow for selection and subsequent amplification.

Figure 4.4: Results of PCR amplification of recovered DNA to produce the library for the next round of CAST.

TOP: 15 cycles of amplification produced sufficient DNA for rounds 2-4, although the quality was diminished for experimental samples at Round 4. BOTTOM: For all experiment samples, it was not possible to cleanly amplify DNA for Round 5; however, it was still possible to amplify recovered DNA for Aart (A_L and A_C). Label below each panel refers to CAST round for which DNA is being produced. Labels to the side of each panel refer to number of PCR cycles. 2 = Cluster 2, 3 = Cluster 3, 4 = Cluster 4, 5 = Cluster 5, 6 = Cluster 6, A_L = Aart tested against randomized library, A_C = Aart tested with its target oligo.

Figure 4.5: The binding site selected by Cluster 1 in the CAST assay.

Sequencing of 15 random clones (top) from the CAST output for Cluster 1 revealed a consensus sequence of 5'-TGGGTGTCA-3' (bottom).

4.3 Discussion

C2H2 ZF domains are well known as the DNA binding domains in many transcription factors. A DNA-binding function is frequently assumed for uncharacterized domains; however, some ZFs are also known to mediate interactions with RNA and proteins. Even when uncharacterized C2H2 domains are tested in functional assay, they are frequently only tested against one type of biomolecule, typically DNA.

As the first step in assessing the full binding potential of the human OAZ protein, this study used a library-based method to expose the C2H2 domains in OAZ to a wide variety of potential DNA sequences. Although the study was designed to avoid any pre-conceived expectation of binding type, there are several limitations to this approach. For example, some interactions may require specific post-translational modifications, or might only occur in certain cell types in response to particular stimuli. This could explain our failure to detect DNA binding by Clusters 2 and 3.

Hata et al. (2000) demonstrated binding to a bone morphogenetic protein response element (BRE) by an OAZ fragment equivalent to Cluster 3. However, this interaction was not independent. Specifically, Hata et al. (2000) demonstrated that a protein fragment comprised of Clusters 2 and 3 was able to bind the BRE. Additionally, a fragment comprised of Clusters 3 and 4 was able to bind the BRE with the Cluster 3 portion of the fragment contacting the DNA, but only while the Cluster 4 portion interacted with SMADs under bone morphogenetic protein (BMP) activation (Hata et al., 2000). The separation of Hata et al. (2000)'s two overlapping fragments into the three

clusters used in this study may explain why none were observed to bind DNA in this study.

The CAST assay did detect DNA binding by Cluster 1, which selected a consensus binding site of 5'-TGGGTGTCA-3'. These results are consistent with the findings of Tsai and Reed (Tsai and Reed, 1998), who detected a palindromic consensus sequence of 5'-GCACCC(A/T)(A/T)GGGTG-3' by CAST assay using the full length rat homolog of OAZ (similarities underlined). However, Tsai and Reed (Tsai and Reed, 1998) also demonstrated homodimerization mediated by the equivalent of Cluster 6 and speculated that selection of a palindromic DNA binding sequence was dependent on homodimerization. In this study, the DNA binding domain (Cluster 1) was separated from the homodimerization domain (Cluster 6), and thus likely selected for only half of the palindrome.

4.4 Materials and Methods

4.4.1 Plasmid Construction

Human full-length OAZ cDNA was provided as a kind gift from Dr. Joan Massagué at the Memorial Sloan-Kettering Cancer Center. The cDNA was subcloned into clusters corresponding to zinc fingers 2-5 (amino acids 78-189), 6-8 (aa 203-289), 9-13 (aa 349-533), 14-20 (aa 572-775), 21-25 (aa 826-987) and 26-30 (aa 1060-1227) by PCR amplification using primers containing appropriate restriction sites. Briefly, OAZ ZF clusters were inserted into the *Bam*HI and *Hind*III sites of the prokaryotic expression vector pMAL-c2X (New England Biolabs) to create fusion proteins consisting of an N-terminal maltose-binding protein (MBP) and C-terminal zinc finger domain.

4.4.2 Protein Purification

Proteins were over-expressed in BL21 (DE3) *E. coli* (Invitrogen) after IPTG induction (0.3 mM). Proteins were purified on an amylose resin column (New England Biolabs), washed with Zinc Buffer A (ZBA; 10 mM Tris pH 7.5, 90 mM KCl, 1 mM MgCl₂, 90 μ M ZnCl₂, 5mM DTT), and eluted with ZBA + 10 mM maltose. Purified protein was aliquoted and stored at –80 °C until use.

4.4.3 Cyclical Amplification and Selection of Targets (CAST) assay

Randomized libraries of double-stranded DNA were created by PCR amplification of 150 pmole of a library oligonucleotide, 5'-GAGCTCATGGAAGTACCATAG-(N₂₁)-GAACGTCGATCACTCGAG-3', with the primers 5'-GAGCTCATGGAAGTACCATAG-3' and 5'-CTCGAGTGATCGACGTTTC-3' (10 cycles; 15 s at 94 °C, 15 s at 70 °C, and 60 s at 72 °C). Library oligos were trace labeled by a single round of PCR with 10 μ Ci [³²P]-dATP. Labeled DNA library was incubated with protein in 1x Binding buffer (ZBA, 1% BSA, 5 mM DTT, 0.17 μ g/ μ l herring sperm DNA, 10% glycerol) for 30 minutes at room temperature and then separated on a 5% TBE polyacrylamide gel in 0.5xTBE buffer. A well-characterized DNA binding zinc finger (Aart; (Dreier et al., 2001)) was also screened against the random N-21 library as positive controls. Dried gels were imaged on a Storm 860 (Molecular Dynamics). Shifted bands were excised from the gel and the DNA was eluted overnight at room temperature in elution buffer (0.1% SDS, 0.5 M NH₃OAc, 10 mM MgOAc).

Eluted DNA was precipitated, washed with ethanol, and re-suspended in 20 μ L DEPC water (Ambion). A non-radioactive PCR reaction was performed to determine the optimal number of amplification rounds, prior to a radioactive reaction to generate the next labeled library for CAST selection. Selection rounds were repeated until 50% of the input was shifted or DNA could no longer be amplified for the subsequent round (4-7 rounds). Target sequence was determined by TOPO-TA cloning amplified DNA into pCR2.1-TOPO (Invitrogen), and submitting samples to the UC Davis DNA Sequencing Facility.

CHAPTER V: PROTEIN BINDING BY ZINC FINGERS

5.1 Introduction

Protein binding by C2H2 domains has only been recently established. Because of their well-established reputation as DNA binding domains, newly discovered proteins are frequently assumed to bind DNA and rarely tested for other functions. Despite this assumption, there are numerous examples of domains which form PPIs and for which the C2H2 domains are implicated in that interaction (Brayer and Segal, 2008). The number of interactions presented by Brayer and Segal (2008) suggests that protein binding is likely under annotated, perhaps due to the biased functional annotation methods used in the past. As presented in Chapter IV, hOAZ contains fingers for which DNA binding could not be detected, some of which are known to be involved in PPIs (Figure 4.3; (Hata et al., 2000; Tsai and Reed, 1998)).

Like Chapter IV, this chapter investigates the hypothesis that a less biased approach would produce a more complete description of ZF function. In comparison to DNA, it is not possible to expose a set of ZFs to all possible protein sequences. The combinatorial permutations of using 20 amino acids are vast. Unlike the largely 2-dimensional major groove of DNA, a protein-protein interaction surface is typically a large ($\approx 800 \text{ \AA}^2$) 3-dimensional structure composed of any number of secondary structural elements (Jones and Thornton, 1996; Stites, 1997). As a practical approach, we used the yeast two-hybrid method to test the protein binding capabilities of 32 C2H2 domains

from two proteins (Figure 5.1). As model proteins, the human 30-finger protein OLF-1/EBF associated zinc finger (OAZ), and human 3-fingered Zif268 were used.

As discussed above, OAZ is a large zinc finger protein containing 30 C2H2 domains arranged into 6 clusters (Figure 4.2). The interaction capabilities of OAZ have been studied previously, and it is known that OAZ can interact with several proteins, and it does so by using different sets of zinc finger domains (Hata et al., 2000; Tsai and Reed, 1997, 1998). Working with the rat ortholog, rOAZ, Tsai and Reed (Tsai and Reed, 1997, 1998) determined that fingers equivalent to cluster 6 mediate homodimerization and also interact with Olf-1/Early B-cell Factor (EBF). Hata et al. (2000) reported human OAZ interacting with SMAD 1 and SMAD 4 using fingers in the fourth cluster. More recently, Finally, an interaction between Parp1 and OAZ has been described, although the significance of this interaction was not determined, nor was it mapped to a specific region of hOAZ (Ku et al., 2003).

Zif268 is a prototypical C2H2 transcription factor. A relatively small protein, Zif268 contains only three C2H2 domains (Figure 5.2), and is a well-known, ubiquitously expressed DNA binding protein. Although Zif268 is known to interact with the RelA subunit of NF- κ B, (Chapman and Perkins (2000), the extent of its protein binding capabilities has never been examined. Conversely its DNA binding abilities are well established, and it is frequently used as the scaffold for custom-engineered ZF proteins that are used to bind a variety of DNA sequences (Carroll et al., 2006; Jamieson et al., 2003; Pavletich and Pabo, 1991; Wright et al., 2006).

Figure 5.1: Overview of the yeast two-hybrid method.

Zinc finger clusters were individually cloned into a vector to create GAL4 DNA-binding domain-ZF fusion proteins (bait), and transformed into yeast. A pre-transformed library of human fetal brain cDNA, fused to the GAL4 activation domain, was purchased from Clontech (prey). Bait and prey yeast strains were grown, allowed to mate, and plated on selective media. When a protein-protein interaction occurs between the bait and prey proteins, the activation domain and the DNA binding domain are brought together, which activates transcription at reporter genes, resulting in growth on selective media.

Construct and test
DNA-BD/bait vector

BD MATCHMAKER™ Pretransformed Library

Reporter Genes: ADE2, HIS3, MEL1, LacZ
Plate on: -Ade/-His/-Leu/-Trp/+X-a-gal

Figure 5.2: Schematic representation of human Zif268.

The 3 C2H2 zinc finger domains are indicated by rectangles. The ZF cluster used in this study are indicated. The aligned amino acid sequence of hOAZ is shown below the schematic, with clusters and ZFs labeled. Positions -1 to 6 of the α -helix, which are important for DNA binding, and the linker regions, are indicated. Conserved cysteines and histidines are highlighted by gray boxes.

5.2 Results

5.2.1 Protein Interactions Detected by Yeast Two-Hybrid Assay

Because of the logistical impossibility of exposing protein fragments to every possible amino acid combination, a yeast two-hybrid method was used to screen a commercial library from human fetal brain (Clontech). This tissue was chosen because it is among the most highly complex in terms of the number of genes expressed within it (Velculescu et al., 1999). The library contained approximately 3.5×10^6 independent clones, which represents the upper limit of the yeast two-hybrid assay. Each of the six hOAZ clusters, and the ZF cluster from Zif268, were subcloned into the Matchmaker Two-Hybrid 3 bait vector (Clontech) and checked for protein expression and autoactivation. The clusters were then examined in the yeast two-hybrid assay. All clusters tested positive for protein interactions, after multiple rounds of re-streaking to retest phenotypes, as indicated by blue colony growth on media that was adenine-, histidine-, leucine-, tryptophan-, and X- α -gal+ (Table 5.1). Approximately 10 random positive colonies from each cluster were sequenced to sample the identities of the putative interacting proteins (Table 5.2).

Several results from the yeast two-hybrid assay seemed striking. Tsai and Reed (Tsai and Reed, 1997, 1998) showed that Cluster 6 of the rat OAZ homolog was responsible for protein-interactions with itself as well as OLF/EBF. Therefore, it is surprising that neither hOAZ nor EBF were among the 14 positive clones for Cluster 6. Similarly, RelA was not among the positive clones for Zif268. Also surprising was that

Cluster 1, which was shown both here and by Tsai and Reed (Tsai and Reed, 1998) to interact with DNA, produced more than 80 positive yeast colonies. Finally, given that protein binding is generally viewed to be an alternative or unusual function for C2H2 domains, it was surprising that each cluster had so many potential partners. However, the yeast two-hybrid assay is associated with a high rate of false-positives. Therefore, samples of the putative protein-interactions for each cluster were verified by the independent method of co-immunoprecipitation.

Table 5.1: Yeast Two-Hybrid Results

Zinc Fingers	No. Clones Screened	No. Colonies ¹
Cluster 1 (ZF2-5)	7.4×10^5	82
Cluster 2 (ZF6-8)	1.4×10^5	10
Cluster 3 (ZF9-13)	4.0×10^4	N/A ²
Cluster 4 (ZF14-20)	3.0×10^5	11
Cluster 5 (ZF21-25)	1.2×10^6	288
Cluster 6 (ZF26-30)	2.8×10^6	14
Zif268	2.8×10^5	15

¹As determined after two rounds of re-streaking (see methods)

²Not all colonies were re-struck. Original plates had 508 colonies; ~50 were re-struck; only a subset of the original colonies, roughly 38% grew on third plate.

Table 5.2: Potential Interaction Partners for OAZ Zinc Finger Clusters

Cluster	Frequency	Protein Name	Amino Acid Start
1	1	GAPDH	97
	1	DBNDD2	45
	1	FADH1	136
	1	XRCC6	311
	1	TUBB2A	1
	1	Weakly similar to PTPRC	72
	1	Copine 6	397
	1	FTL	103
	2	GPBP1L1	18
	1	PIK4CB	620
2	1	VDAC2	108
	1	DHX15	86
	1	JARID1D	31
	1	C2orf16	1719
	2	XPO6	616
	2	DGKI	1060
	2	ADNP	532
3	1	AKAP9	3748
	1	OR2Z1	130
	1	STY13 mRNA	
	1	Similar to VDAC2	212
	1	Alu subfamily J	
	1	CCD57	1
	1	OLIG1	1
	1	ALDOA	1
	1	PGAM1	143
	1	NDUFB8	1
4	1	PGM1	456
	1	PCBP1	21
	4	Similar to p40	7

Table 5.2 cont.

Cluster	Frequency	Protein Name	Amino Acid Start
5	2	PSMA6	85, 113
	1	PPIA	1
	1	TNNI3K	4
	1	PRKAG2	453
	1	RBM4	1
	1	GAPDH	136
	1	CLU	187
	2	PSAP	420, 252
	1	NGFRAP1	1
6	1	RNF157	2
	1	DUS1L	353
	1	MAP1B	2350
	1	Genomic contig	
	1	PRR6	STOP
	1	SYTL1	11
	1	ATP1B1	70
	2	Genomic contig	
Zif268	4	Similar to Loss of heterozygosity, 11	807
	1	Genomic contig	
	1	Saposin-A	420
	1	Notch4	STOP
	1	KRT222P	57
	1	hypothetical protein BC009862	843

5.2.2 Verification of Select Protein Interactions by MBP Pull-down Assay.

Two different methods were used to verify the yeast two-hybrid (Y2H) results. Initially, a “pull-down assay” was used. Here the ZF clusters were subcloned into the prokaryotic expression vector pMAL-c2X for expression of N-terminal MBP-fusion proteins in BL21 *E. coli* cells. For the partner protein, the original pGADT7-Rec plasmids, which appends an N-terminal HA-tag and which were recovered from the Y2H screening, were transformed into BL21 (DE3) *E. coli* cells. Although this plasmid was originally designed for expression of protein in yeast, it contains a pUC replication origin, an ampicillin resistance gene, and a T7 promoter region; and is therefore well suited for the plasmid maintenance and protein expression in *E. coli*. For the assay, ZF clusters and partner proteins were grown separately, and then whole cell lysates were mixed. After an incubation period, protein complexes were passed over an amylose column which retained the MBP-ZF fusion protein, and any binding partner. After washing, complexes were eluted from the column separated on a 12% SDS-PAGE gel, and visualized by probing with anti-MBP and anti-HA antibodies.

The results from the pull-down assays were often difficult to interpret, and not many interactions could be verified (Table 5.3). Results from the MBP pull-down assay suggest the interaction between Cluster 6 and MAP1B and UNC13A are real (Figure 5.3). Likewise, the interaction between Cluster 5 and MAGI2 also appears to be real (data not shown). However, although several attempts were made, with occasional success, there was on-going concern about the number of bands produced within each samples, as well as with elevated background levels (Figure 5.3, Table 5.3). Therefore,

this method was dropped in favor of a mammalian co-immunoprecipitation, and none of these interactions were replicated.

Figure 5.3: Representative MBP pull-down experiment results.

Whole cell lysates of individual cultures of BL21 *E. coli* cultures carrying either the MBP-Zinc finger fusion proteins or the HA-tagged binding partner were run over an amylose column to retain MBP tagged ZFs and their partners. Complexes were eluted, resolved on a 12% SDS-PAGE gel and visualized by anti-MBP and anti-HA antibodies. (A) Cluster 6 forms a protein-protein interaction with MAP1B as evident by the band indicated (upper arrow). There may also be an interaction between cluster 6 and UNC13A as indicated by the slight enrichment in the band indicated by the lower arrow; however, since there are many bands in each lane, it is difficult to determine whether or not this band is an artifact. (B) Cluster 5 may interact with MAGI2 as evident by the slight enrichment of the protein band in the Cluster 5 + MAGI2 lane, versus the protein band in the MAGI2 lane. However, the high background (MAGI2 alone lane) makes it difficult to determine if this is a real PPI, or if MAGI2 is just a “sticky” protein.

Table 5.3: MBP Pull-Down Assay Results

Cluster	Partner	Results ¹
1	ASPP1	Background
1	MADHIP	Background
5	GAPDH	Background
5	FTLL1	Background
5	MAGI2	Positive ²
6	MAP1B	Positive ²
6	UNC13A	Positive ²

¹"Background" indicates that the amount of protein, as indicated by the darkness of the band, was equal regardless of the presence/absence of the ZF protein. "Positive" indicates clear enrichment of the partner protein when the ZF protein was present. See Figure 5.3 for examples.

²Only tested once, therefore not confirmed.

5.2.3 Verification of Select Protein Interactions by Mammalian Co-IP.

Mammalian co-immunoprecipitation (co-IP) proved to be a more successful method for verifying the Y2H hits. Here, the ZF clusters were subcloned into the mammalian expression vector pcDNA 3.1(-) containing an N-terminal FLAG-epitope tag. The putative binding partners were subcloned into pcDNA 3.1(-) containing N-terminal HA-epitope tagged fusion protein. Plasmids expressing FLAG-tagged ZF clusters were co-transfected into the cell line HEK263T with an equal amount of either HA-tagged putative partner or empty vector. After harvesting cells, protein complexes were immunoprecipitated with anti-FLAG beads, resolved on a 12% SDS-PAGE gel, and transferred to a PVDF membrane for visualization using anti-HA antibodies. In addition, amino acids 221-570 of the protein EBF, which had been previously shown to interact with Cluster 6 in the rat homolog rOAZ, were used as a positive control (Tsai and Reed, 1997, 1998).

The co-immunoprecipitation assay in HEK293T cells verified interactions for Cluster 1 with XRCC6, but not Ferritin or GPB1L1 (Figure 5.4 a and c). XRCC6, also known as Ku70, is well known for its involvement in DNA repair, and is also known to be involved in V(D)J recombination. Also verified were the interactions between Cluster 6 and human EBF and RNF157 (Table 5.4, Figure 5.4a). EBF is involved in B-cell development and the rat homolog has previously been shown to interact with rOAZ (Tsai and Reed 1998). The function of RNF157 is unknown (Rebhan et al., 1997). Finally, the interaction between Zif268 and BAF57, a protein involved in chromatin remodeling, was verified (Figure 5.5b).

Table 5.4: Mammalian Co-IP Results

Cluster	Partner	Result ¹
1	XRCC6	Positive
1	Ferritin	Negative
1	TUBB2A	Background
1	GPBP1L1	Not Detected
1	GPBP1L1	Not Detected
2	ADNP	Not Detected
2	JARID1D	Not Detected
3	VDAC2	Not Detected
3	OLIG1	Positive? ²
4	PCBP1	Positive? ²
4	PGM1	Background
5	RBM4B	Not Detected
5	ASB3	Not Detected
6	RNF157	Positive
6	EBF	Positive
Z	BAF57	Positive

¹Each partner was tested a minimum of 2 times. "Positive" indicates clear enrichment of the partner protein when the ZF protein was present (Figure 5.4a). "Background" indicates the amount of protein, as indicated by the darkness of the band, was equal regardless of the presence/absence of the ZF protein (Figure 5.4b). "Negative" indicates an absence of the partner protein after co-IP and the presence of the partner protein in the input (Figure 5.4c). "Not Detected" indicates that partner protein was not detected in either the co-immunoprecipitation or the input sample.

²Inconsistent results, tested multiple times. See text for details.

Cluster 3's interaction with Olig1, a protein involved in the formation and maturation of oligodendrocytes, was tested three times with an inconclusive outcome. The first attempt yielded positive results confirming the interaction; however, this result could not be replicated and the ensuing attempts show high background levels of Olig1 in the absence of Cluster 3 (data not shown). Similarly, Cluster 4's interaction with PCPB1 was tested two times: the first time verified the interaction, but replication resulted in elevated background levels of PCPB1 in the absence of Cluster 4 (data not shown).

Expressed protein for the two binding partners tested for each of Clusters 2 and 5 could not be detected nor could GPBP1L1, a putative partner for Cluster 1. Although there are many possible reasons for this, including no expression, proteosomal degradation of the partner protein, folding such that the epitope tag is unavailable for antibody binding, or loss of the protein during sample preparation, the reason was not determined (Table 5.4). Finally, like the pull-down assays, there were several tested interactions that could not be proved or disproved due to high background (Figure 5.4b, Table 5.4).

Lastly, all of the putative binding partners for Cluster 6 that were tested in both the MBP pull-down assay and the co-IP assay were confirmed. Since this could be an indication that Cluster 6 was a "sticky" protein fragment and that these interactions were not, in fact, specific, the reciprocal binding capabilities for Cluster 6 and Cluster 1 was tested. To test this Cluster 6 was screened against the confirmed binding partner for Cluster 1, XRCC6. Similarly, Cluster 1 was screened against two of Cluster 6's

confirmed binding partners, EBF and RNF157. No significant interactions for Cluster 6 with XRCC6 nor Cluster 1 with RNF157 or EBF were detected (data not shown).

Figure 5.4: Representative mammalian co-immunoprecipitation results.

HEK293 cells were co-transfected with FLAG-tagged Cluster 1 and (A) HA-tagged XRCC6, (B) HA-tagged TUBB2A, (C) HA-tagged Ferritin DNA, followed by immunoprecipitation using anti-FLAG agarose beads, and probing with anti-HA (top panels) or anti-FLAG (lower panels) antibodies. Results were repeated a minimum of 3 times (except Ferritin which was only tested 2 times). Panel (B) also exemplifies typical “background” results, where the band for TUBB2A is equal for the lanes with and without Cluster 1.

A

B

Figure 5.5: Protein-protein interactions mediated by Cluster 6 of hOAZ and Cluster Z of Zif268.

HEK293T cells were co-transfected with (A) FLAG-tagged Cluster 6 and HA-tagged EBF and RNF157, or (B) FLAG-tagged Cluster Z and HA-tagged BAF57 DNA, followed by co-immunoprecipitation using anti-FLAG agarose beads, and probing with anti-HA (top panels) or anti-FLAG (bottom panels) antibodies.

5.2.4 Observed Protein Interactions are Dependent on Zinc

To further examine if the proper folding of the ZFs was important to the interaction, some of the co-immunoprecipitations were repeated in the presence of 5 mM EDTA. This method had been shown previously to chelate the zinc ions and disrupt the zinc fingers' activity (Tsai and Reed, 1998). For these experiments, Cluster 1 and its partner, XRCC6, as well as Cluster 6 with both RNF157 and EBF, were tested. Addition of EDTA abolished the interaction between RNF157 and Cluster 6 demonstrating the importance of an intact C2H2 domain to this interaction (Figure 5.6b). Although the Cluster 6-EBF and Cluster 1-XRCC5 interactions were not completely abolished by the addition of EDTA, the interactions were reduced to background levels. This result suggested that proper folding of the zinc fingers was important to these interactions (Figure 5.6 a and b).

A

B

Figure 5.6: Protein-protein interactions mediated by Clusters 1 and 6.

HEK293T cells were co-transfected with (A) FLAG-tagged Cluster 1 and HA-tagged XRCC6, (B) FLAG-tagged Cluster 6 and HA-tagged RNF157 or HA-tagged EBF DNA, followed by immunoprecipitation using anti-FLAG agarose beads and probing with anti-HA (top panels) or anti-FLAG (lower panels) antibodies. 5mM EDTA or 20U DNase (+/-) was added to the whole cell lysates prior to immunoprecipitation, as indicated. Experiments were repeated a minimum of 3 times and results presented here are representative.

5.2.5 Observed Protein Interactions are Not Dependent on DNA

Both Cluster 1 and XRCC6 are known to bind DNA. It was therefore possible that the observed protein interaction was actually mediated by the simultaneous binding to an unexpected DNA. EBF also has a DNA binding domain. The DNA-binding capability of RNF157 is unknown, and Cluster 6 was found to not interact with DNA (see Chapter IV). Based on similar previous experiments (Nieto et al., 2002), the DNA-dependence of these four protein interactions was tested by adding 20U of DNase to the co-immunoprecipitation reaction. Although somewhat reduced overall, XRCC6 clearly co-precipitated with Cluster 1 compared to background (Figure 5.6a). These results indicate that the interaction between Cluster 1 and XRCC6 is likely not DNA-dependent. As expected, the addition of 20U of DNase had no effect on the interaction between Cluster 6 and either EBF or RNF157 (Figure 5.6b).

5.3 Discussion

By using the yeast two-hybrid method to expose the domains to a wide variety of potential DNA and protein interaction partners, these experiments were designed to avoid any pre-conceived expectations of the function of the ZFs. However, there are several limitations to this approach. An interaction partner might be poorly represented or absent from the fetal brain library. This might be a reasonable explanation for not detecting the expected Cluster 6-EBF interaction in the yeast two-hybrid library screening. Some interactions may require specific post-translational modifications or might only occur in certain cell types in response to particular stimuli. This could explain the failure to detect

the expected Cluster 4-SMAD protein interaction, which had been shown to be dependant on the activation of bone morphogenetic protein (BMP) (Hata et al., 2000).

Despite missing some expected interactions, the yeast two-hybrid library screen resulted in several potential protein interaction partners for all clusters in our yeast two-hybrid assay. Because the yeast two-hybrid assay has a high rate of false positives, hits identified in a yeast two-hybrid assay are frequently verified using a different method such as pull-down assays or mammalian cell co-immunoprecipitation (van Criel et al., 1999). Although time limitations precluded testing all 65 hits listed in Table 5.2, at least two potential partners were tested for all clusters. Furthermore, novel PPIs were verified for each cluster, except Cluster 2, albeit inconclusively for Clusters 3 and 4.

The interaction between Cluster 1 and XRCC6 was confirmed by the co-IP results. Although the biological significance of the interaction was not examined here, XRCC6 has been observed to interact with C2H2 zinc fingers in other transcription factors (Ishiguro et al., 2007). Addition of DNase did not disrupt any of the interactions, but all were found to be sensitive to disruption by EDTA. These results indicate that the PPIs were not facilitated by the presence of DNA, and that a folded C2H2 domain was required for the interaction. Similarly, the interaction between the zinc fingers of Zif268 and BAF57 was also confirmed. BAF57 has been previously demonstrated to interact with the DNA binding domains of other transcription factors, including the C2H2 domains of Sp1, a protein closely related to Zif268 (Kadam et al., 2000).

Several new interactions for Cluster 6 were confirmed. Among these were MAP1B and UNC13A, two proteins involved in neurogenesis. MAP1B is critical to

axonogenesis and dendrite formation (Gonzalez-Billault et al., 2002; Rebhan et al., 1997). UNC13A is a diacylglycerol and phorbol ester receptor with demonstrated roles in synaptic vesicle maturation and neurotransmitter release (Rebhan et al., 1997). These findings are in agreement with earlier findings by Tsai and Reed (Tsai and Reed, 1997, 1998) that the rat homolog of OAZ was involved in regulating the switch between neural tissue proliferation and neural differentiation. Furthermore, Tsai and Reed (1997, 1998) speculated that the switch was due, at least in part, to a change from an interaction with EBF to homodimerization, both of which were mediated by Cluster 6. This study also confirmed that like their rat homologs, hOAZ interacts with hEBF via Cluster 6. Finally, the confirmed interaction between Cluster 6 and RNF157 suggests the possibility of additional functions for hOAZ. However, since RNF157 is functionally uncharacterized, the biological significance of the interaction remains elusive.

Cluster 5 of OAZ was previously functionally uncharacterized. The tentatively confirmed interaction with MAGI2 therefore represents a new example of C2H2 domains which are capable of interacting with protein, but do not bind DNA. It may also represent a novel function for hOAZ. Although MAGI2's function is not well understood, it seems to be involved in regulating Notch signaling in neuronal cells, possibly having a role in neuron migration (Wright et al., 2004; Xu et al., 2001). In addition, MAGI2 seems to be a scaffold protein that plays a role in assembling neurotransmitter receptors (Rebhan et al., 1997). Finally, MAGI2 is known to interact with both TGF β and SMAD proteins (Takamune et al., 2000; Xu et al., 2001).

Unfortunately, the interactions between Clusters 3 and 4 and their partners could not be conclusively verified here. It is interesting that previously described functions for the putative binding partners implicate them in neurogenesis, which provides additional support for these interactions being real. However, each of these interactions was found by screening a brain library. Therefore, additional testing is required to verify the interactions between Clusters 3 and 4 and Olig1 and PCBP1, respectively.

Cluster 3 appears to interact with Olig1, a protein expressed exclusively in oligodendrocytes and oligodendrocyte precursors with an essential role in differentiation and maturation of oligodendrocytes (Arnett et al., 2004; Lu et al., 2002; Lu et al., 2000). Neuronal differentiation in Olig1⁺ cells has also been demonstrated to be regulated through BMP receptor 1a (Samanta et al., 2007). Furthermore, Hata et al. (2000) demonstrated that interaction of an OAZ fragment that included Cluster 3 lead to the activation of transcription from a BRE site in the Xvent2 gene, previously shown to repress neural differentiation, under BMP stimulation, and that this interaction was dependent on BMP activation of SMAD proteins. It is likely that OAZ plays a role in this process, however additional study will be required to determine its placement within the pathway(s).

Cluster 4 appears to interact with PCBP1. PCBP1 is generally known as an RNA binding protein, involved in post-transcriptional regulation; however, it has recently been shown to act as a transcription factor at a variety of genes, including mu-opioid receptor in mice (Kim et al., 2005). In addition, Janssen and Marynen (Janssen and Marynen, 2006) verified an interaction between ZNF384 and PCBP1, although the interacting

region for ZNF384 was not mapped and there is only ~40% homology between the domains in ZNF384 and Cluster 4.

Finally, both Clusters 1 and 3 had previously been described as DNA binding fingers. Here, both Clusters 1 and 3 were observed to be capable of protein binding, although largely unverified for Cluster 3. Therefore, the data support the hypothesis that unanticipated functional information can be obtained by using this less biased approach. In addition, the data indicate that some ZFs that were determined to have a particular function in one context may in fact be capable of supporting other types of interactions.

5.4 Materials and Methods

5.4.1 Plasmid construction

Bait constructs were created by cloning each OAZ zinc finger cluster into the *NcoI* and *XmaI* sites of the bait vector, pGBTK7 DNA BD (Clontech). This vector appends an N-terminal GAL4 DNA-binding domain (GAL4-BD) and a C-MYC epitope tag. Since Cluster 3 contained an internal *NcoI* site, *NdeI* was used instead. Following sequence verification of the insert, each bait vector was transformed into the AH109 yeast strain and grown on SD/-Trp/X- α -gal agar plates to verify transformation and reporter gene activation. Expression of GAL4-BD-zinc finger fusion proteins was verified by western blot using antibodies to the C-MYC epitope tag (Sigma).

5.4.2 Yeast 2 Hybrid

OAZ zinc finger clusters were individually screened against a commercial human fetal brain cDNA library (Clontech #638804) that had been pre-cloned into pGADT7-Rec

and pre-transformed into yeast strain Y187. Screening was performed using Clontech's Matchmaker Two-Hybrid System 3 according to the manufacturer's protocol for high stringency screening. Individual colonies displaying a positive phenotype were re-struck twice on SD/-Ade/-His/-Leu/-Trp/+X- α -gal for phenotype verification, prior to DNA isolation for sequencing and co-immunoprecipitation. Yeast plasmid DNA was isolated using the Zymoprep yeast plasmid miniprep kit (Zymo Research) per the manufacturer's instruction for sequencing and for transformation into BL21 (DE3) *E. coli* (Invitrogen) for the MBP pull-down assay.

5.4.3 MBP Pull-Down Assay

Zinc finger clusters from hOAZ and Zif268 were cloned into the *BamHI* and *HindIII* sites of the prokaryotic expression vector pMAL-c2X (New England Biolabs) to create fusion proteins consisting of an N-terminal maltose-binding protein (MBP) and C-terminal zinc finger domain and transformed into BL21 (DE3) *E. coli* (Invitrogen). MBP-tagged and HA-tagged proteins were over-expressed in BL21 (DE3) *E. coli* (Invitrogen) after IPTG induction (0.3 mM). Centrifuged cell pellets were resuspended in Zinc Buffer A (ZBA; 10 mM Tris pH 7.5, 90 mM KCl, 1 mM MgCl₂, 90 μ M ZnCl₂, 5mM DTT) and sonicated on ice to lyse the cells. Whole cell lysate was cleared by centrifugation at 14,000 rpm for 15 minutes at 4 °C, and supernatant was transferred to a clean tube. Total protein concentration was determined by Bradford analysis.

Columns were prepared with 1 mL packed amylose resin (New England BioLabs) and washed with 10 column volumes of ZBA. Equilibrated volumes of cleared whole cell lysate for MBP-ZF proteins and putative HA-tagged partner proteins were mixed

(1:3) and loaded onto the amylose column. Amylose column were incubated, rotating, for 2 hours at 4 °C, followed by washing of the column with 20 column volumes of ZBA. Protein complexes were eluted with ZBA + 10 mM maltose and an aliquot was mixed with SDS-loading buffer (50 mM Tris pH 6.8, 100 mM DTT, 2% SDS, 0.1% bromophenol blue, 10% glycerol) and resolved on a 12% SDS-PAGE gel at 25 mA.

Proteins were transferred to a PVDF membrane (75 mA for 1.5 hours). After transfer, membrane was blocked with 5% MPT (100 mM NaH₂PO₄, 27 mM KCl, 1.37 mM NaCl, 0.05% Tween, 5% Nonfat dry milk), then incubated with anti-HA-HRP antibody (1:500; Sigma) for ~4 hours and anti-FLAG-HRP antibody (1:10,000) for ~30 minutes. Bands were visualized using the ECL chemiluminescence detection kit (GE Healthcare).

5.4.4 In Vivo Co-Immunoprecipitation Assay

Each OAZ ZF cluster, as well as the ZF cluster for Zif268, was cloned into the *Bam*HI and *Hind*III sites of pcDNA3.1(-) (Invitrogen), in frame with an engineered N-terminal FLAG epitope tag. Putative binding partners selected from the yeast two-hybrid library were PCR amplified using the Matchmaker 5' and 3' AD LD-insert screening amplimers (Clontech), and cloned into a pcDNA3.1(-) vector containing engineered *Sfi*I sites and an N-terminal HA epitope tag. Full-length human Early B-cell factor (EBF) cDNA (ATCC #9891743) was also subcloned by PCR amplified (aa 221-570) and cloned into the *Bam*HI and *Hind*III sites of a pcDNA3.1(-) vector containing an engineered N-terminal HA epitope tag.

HEK293T cells were cultured in Dulbecco's modification of eagle's medium (DMEM; Cellgro) with 4.5 g/L glucose and L-glutamine and without sodium pyruvate, plus 5% new born calf serum (JR Scientific) and 1x penicillin/streptomycin/neomycin (Sigma). Prior to transfection, cells were plated at a density of 4.4×10^6 cells/mL on 10 cm plates. The next day, FLAG-zinc finger constructs were transfected into HEK293T cells with either empty pcDNA3.1(-) vector, or HA-partner pcDNA3.1(-) vector using Lipofectamine 2000 per the manufacturer's protocol. A total of 12 μ g of DNA was mixed with Lipofectamine (1:3) in 3 mL OptiMEM 1 reduced serum media (Gibco) and incubated for 4 hours at 37 °C, before addition of 7 mL DMEM. After 16 hours the transfection media was removed and replaced with fresh DMEM, and the cells were incubated an additional 24 hours before harvesting.

After removing DMEM and washing cells twice with PBS, 1 mL ice cold lysis buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 0.5% Triton X-100, 1x Complete EDTA-free protease inhibitor cocktail (Roche)) was added to each plate and cells were harvested by scraping with a rubber spatula. Lysed cells were transferred to a 1.5 mL microfuge tube and incubated on ice for 10 minutes prior to the addition of 3 M NaCl to a final concentration of 150 mM, followed by an additional incubation on ice for 5 minutes. Whole cell lysate was cleared by centrifugation at 14,000 rpm for 15 minutes at 4 °C, and supernatant was transferred to a clean microfuge tube. Total protein concentration was determined by Bradford analysis. After equilibrating protein concentrations, 1 mL of sample was mixed with 40 μ L of M2 anti-FLAG agarose beads (Sigma) pre-washed with TNT buffer (50 mM Tris-HCl, pH7.5, 150 mM NaCl, 0.05% Triton X-100). 100 μ M

ZnCl₂, 5 mM EDTA, or 20U TURBO DNase (Ambion) were added, where indicated, and rotated overnight at 4 °C.

Beads were washed 8x with 1 mL of either 150 mM wash (50 mM Tris 7.5, 150 mM NaCl, 0.5% triton X-100, 20 μ M ZnCl₂), 300 mM wash (50 mM Tris 7.5, 300 mM NaCl, 0.5% triton X-100, 20 μ M ZnCl₂), or EDTA- wash (50 mM Tris 7.5, 150/300 mM NaCl, 0.5% triton X-100). Between each wash beads were collected by centrifugation at 1000 rpm for 1 min at 4 °C and supernatant was removed. After final wash, beads were re-suspended in SDS-loading buffer (50 mM Tris pH 6.8, 100 mM DTT, 2% SDS, 0.1% bromophenol blue, 10% glycerol) and resolved on a 12% SDS-PAGE gel at 25 mA. Proteins were transferred to a PVDF membrane (75 mA for 1.5 hours). After transfer, membrane was blocked with 5% MPT (100 mM NaH₂PO₄, 27 mM KCl, 1.37 mM NaCl, 0.05% Tween, 5% Nonfat dry milk), then incubated with anti-HA-HRP antibody (1:500; Sigma) for ~4 hours. Bands were visualized using the ECL chemiluminescence detection kit (GE Healthcare). Membranes were stripped (100 mM 2-mercaptoethanol, 2% SDS, 62.5 mM Tris-HCl pH 6.7) and re-probed with anti-Flag-HRP antibody (1:10,000; Sigma).

CHAPTER VI: PREDICTING BINDING TYPE FOR INDIVIDUAL C2H2 DOMAINS

6.1Background

Having determined, with non-biased functional assays, the binding types of clusters of zinc fingers for both OAZ and Zif268, the next feasible step was to try to determine what individual fingers within each cluster were likely doing. Typically, DNA binding requires a string of at least three C2H2 domains. In contrast, numerous studies have indicated that a single C2H2 can interact with protein.

This study attempts to predict the contribution of individual fingers within each zinc finger cluster to the interactions observed in this study and in previous studies. Individual C2H2 domains from hOAZ and Zif268 zinc finger clusters were examined for their potential protein and DNA binding capabilities (Figures 6.1, 6.2). For this analysis, the best distinguishing characteristics from Chapter II were used. The predictions made by both the DNA and protein HMMs are also presented.

MEDESI			
		-1123456	linker
ZF1	YTC [■] DHC [■] QQDFE	SLADLTD H [■] RAH [■]	RCPGDGDDDPQLSWVASSPS
SKDVASPTQMIGDGC [■] DLGLGEEEGGTGLP			
Cluster 1			
		-1123456	linker
ZF2	YPC [■] QFC [■] DKSFI	RLSYLKR HE [■] QIH [■]	SDKLP
ZF3	FKCTYCSRLFK	HKRSRDR HIKLH	TGDKK
ZF4	YHCHECEAAFS	RSDHLKI HLKTH	SSSKP
ZF5	FKCTVCKRGFS	STSSLQS HMQAH	KKNKEHLAKSEKEAKKDD
Cluster 2			
		-1123456	linker
ZF6	FMCDYCEDTFS	QTEELEK HVLTRH	PQLSEKAD
ZF7	LQCIHCPEV [■] FV	DENTLLA HIHQAH	ANQK
ZF8	HKCPMCPEQFS	SVEGVYC HLDSH	RQPDSSNHSVSPDPV
LGSVASMSSATPDSSASVERGSTPDSTLKPLRGQKKMRDDGQGWTKVV			
Cluster 3			
		-1123456	linker
ZF9	YSCPYCSKRDFN	SLAVLEI HLKTIH	ADKPQOS
ZF10	HTCQICLDSMP	TLYNLNE HVRKLH	KNHAYPVMQFGNISA
ZF11	FHCNYCPEMFA	DINSLQE HIRVSH	CGPNANPSDGNNNA
ZF12	FFCNQCSMGFL	TESSLTE HIQQA [■] H	CSVGSAKLESPVVQPTQSFMEV
ZF13	YSCPYCTNSPIFG	SILKLT [■] K HIKENH	KNIPLAHSKKSK
AEQSPVSSDVEVSSPKRQRLSASANSISNGE			
Cluster 4			
		-1123456	linker
ZF14	YPCNQCDLKFS	NFESFQT HLKLH	LELLLRK
ZF15	QACPQCKEDFD	SQESLLQ HLT [■] VH	YMTTSTH
ZF16	YVCESCDKQFS	SVDDLQK HLLDMH	TFVL
ZF17	YHCTLQC [■] EVFD	SKVSIQV HLAVKH	SNEKKM
ZF18	YRCTACNWD [■] FR	KEADLQV HVKHS [■] H	LGNPAKA
ZF19	HKCIFCGETFS	TEVELQC HITTH	SKK
ZF20	YNCKFC [■] SKAFH	AIILLEK HLREKH	CVFDAATENG
TANGVPPMATKKAEPADLQGMLLNPEAPNSHEASEDDVDASEPM			
Cluster 5			
		-1123456	linker
ZF21	YGC [■] DICGAAYT	MEVLLQN HRLRDH	NIRPGEDDGSRKKA [■] EFIKGS
ZF22	HKCNVCSRTFF	SENGLRE HLQTH	RGPAKH
ZF23	YMCPI [■] CGERFP	SLLT [■] LTE HKVTH	SKSLDT
ZF24	GTCRICKMPLQ	SEEEFIE HCQMH	PDLRNSLTG
ZF25	FRCVVC [■] MQTVT	STLELKI HGTFH	MQKLAGSSAASSPNGQGLQKL
YKCALCLKEFRSKQDLVKLDVNGLPYGLCAGCMARSANGQVGGLAPPEPADRPCAG			
Cluster 6			
		-1123456	linker
ZF26	LRCPECSVKFE	SAEDLES HMQVDH	RDLTPETSGPRKGTQTSPVPRKKT
ZF27	YQCIKCQMTFE	NEREIQI HVANH	MIEEGIN
ZF28	HECKLCNQMF [■] D	SPAKLLC HLIEH	SFEGMGGT
ZF29	FKCPVCFT [■] VFV	QANKLQQ HIFAVH	GQEDKI
ZF30	YDCSQCPQKFF	FQTELQN HTMSQH	AQ

Figure 6.1: C2H2 Domains in human OAZ

The aligned amino acid sequence of hOAZ is shown with clusters and ZFs labeled. Positions -1 to 6 of the α -helix, which are important for DNA binding, and the linker regions, are indicated. Conserved cysteines and histidines are highlighted by gray boxes.

			-1123456		linker		
ZF1	YACP	VE	SCDRRF	S	RSDELTR	HIRIH	TGQKP
ZF2	FQCRI	CMRNFS			RSDHLTT	HIRTH	TGEKP
ZF3	FACDI	CGRKFA			RSDEKRR	HTKIH	LRQKD

Figure 6.2: C2H2 Domains in human Zif268.

The aligned amino acid sequence of Zif268 is shown with ZFs labeled. Conserved cysteines and histidines are highlighted by gray boxes. Positions -1 to 6 of the α -helix, which are important for DNA binding, and the linker regions, are indicated.

6.2Results

As shown in Table 6.1, three traits were used in this analysis. The presence/absence of a TGEKP(like) linker, regional spacing between conserved cysteines and histidine, and the overall charge in the α -helix. Scoring was based on a point system, and points were independently awarded for DNA and protein binding characteristics. For example if the trait under consideration was a DNA binding trait, domains with that trait were awarded a “DNA point,” while domains without the trait received nothing.

Prediction was based on the total number of points.

All domains were examined for the presence/absence of a TGEKP(like) linker, which indicates DNA binding (section 2.2). Scoring was based on similarity to the TGEKP linker. Domains at the ends of clusters were not considered. Not surprisingly, both fingers of Zif268 received DNA points, with finger 1 getting 2 points, and finger 2 getting 3 points. Fingers 3 and 4 of OAZ also received 2 DNA points, while Finger 2 of OAZ with a weakly similar linker received only 1 point (Figure 6.1, Table 6.2).

Table 6.1: Single C2H2 Domain Prediction

	LINKER	SPACING	α-HELIX CHARGE		BINDING TYPE		HMM
	DNA Only		Protein	DNA	Predicted	Tested	
1	-	-	2	-	Protein	N/A	-
2	1	1 DNA	-	3	DNA	Both	Both
3	2	1 DNA	2	5	DNA/Both	Both	-
4	2	1 DNA	2	3	DNA/Both?	Both	Both
5	-	1 DNA	-	-	Weak DNA	Both	Both
6	0	2 Protein	3	1	Protein	DNA	-
7	0	2 Protein	2	-	Protein	DNA	-
8	-	1 DNA	1	-	DNA/Both	DNA	-
9	0	2 Protein	1	-	Protein	Both?	-
10	0	2 Protein	1	-	Protein	Both?	-
11	0	2 Protein	2	-	Protein	Both?	-
12	0	2 Protein	2	-	Protein	Both?	-
13	0	2 Protein	-	2	Protein/Both	Both?	-
14	0	1 DNA	1	-	Protein	Protein	-
15	0	1 DNA	1	-	Protein	Protein	-
16	0	2 Protein	2	1	Protein	Protein	-
17	0	2 Protein	-	1	Protein	Protein	-
18	0	2 Protein	2	1	Protein	Protein	-
19	0	1 DNA	2	-	Both	Protein	-
20	-	2 Protein	1	1	Protein	Protein	-
21	0	2 Protein	1	-	Protein	Unknown	-
22	0	1 DNA	2	1	Protein	Unknown	Protein
23	0	1 DNA	1	-	Protein	Unknown	Protein
24	0	1 DNA	4	-	Protein	Unknown	
25	-	1 DNA	1	1	DNA/Both	Unknown	
26	0	2 Protein	3	-	Protein	Protein	Protein
27	0	1 DNA	2	1	Protein	Protein	
28	0	1 DNA	-	1	DNA	Protein	
29	0	2 Protein	-	1	Protein	Protein	
30	-	2 Protein	1	-	Protein	Protein	Protein
Z1	2	1 DNA	2	2	Both	Both	DNA
Z2	3	1 DNA	1	2	Both	Both	Both
Z3	-	1 DNA	2	4	Both	Both	Protein

The regional spacing analysis in section 2.3 indicated that the C4C + H3H type was associated with DNA binding, while the C2C + H4H tended to be associated with protein binding. Since only finger 1 of Zif268 contained the C4C spacing type, only the spacing between the histidine residues was considered. One DNA point was assigned for the C2C + H3H spacing. However, since there is also an association between the TGEKP linker and the H3H spacing, and because there is not an equivalent linker for protein binding, two protein points were awarded to domains with C2C + H4H spacing. The spacing score for each domain can be seen in Table 6.1.

The analysis of amino acid characteristics in section 2.4 was not particularly successful; however there were indications that protein binding fingers have negatively charged amino acids in the N-terminal portion of the α -helix. DNA binding fingers use amino acids in the same region when contacting DNA. Since negatively charged amino acids are likely to preclude DNA binding, one protein point was awarded for each negatively charged amino acid (DE) in positions -1 through 6 of the α -helix (Figure 6.2, Table 6.1). Similarly, one DNA point was awarded for every positively charged amino acid (HKR) in the α -helix.

6.3 Discussion

Overall, the predicted contribution of individual zinc finger domains agrees with the known binding types for each cluster. Both this study and Tsai and Reed (1998) demonstrated DNA binding for fingers 2-5 of OAZ (Cluster 1). However, protein binding was also demonstrated Cluster 1 by this study. The sequence based prediction

suggests that fingers 3 and 4 are able to bind both DNA and protein. Each contains the characteristic TGEKP(like) linker and H3H spacing of DNA binding, but also have negatively charged amino acids in the α -helix. Finger 2 on the other hand, is more characteristic of DNA binding. For the most part, these results agree with the predictions made by each profile HMM, which scored fingers 2, 4 and 5 above the cutoff for both the protein HMM and the DNA HMM. Finger 3 scored just below the threshold (bit score = 19.9, E-value 0.04) for the DNA profile, but was not detected with the protein HMM. Together these results agree with the finding that this domain is capable of supporting interactions with both DNA and protein. In addition, it is likely that the interaction with XRCC6 is mediated through fingers 2 and 3. Interestingly, Finger 5 did not appear to have characteristics of either binding type.

Cluster 2 was demonstrated to bind DNA in two independent studies (Hata et al., 2000; Tsai and Reed, 1998). Those studies, however, examined Cluster 2 only as part of a larger fragment containing additional domains (Figure 4.1; (Hata et al., 2000; Tsai and Reed, 1998)). This study did not detect DNA binding by Cluster 2. Fingers 6 and 7 were predicted to bind protein, while finger 8 may be capable of binding DNA. However, the only DNA characteristic displayed by finger 8 is the C2C + H3H helix, so this prediction could easily be incorrect.

The results for Cluster 3 are the most in conflict between this study and other studies. Hata et al. (2000) found Cluster 3 to be a DNA binding cluster. Like Cluster 2, DNA binding by Cluster 3 was only demonstrated within the framework of larger fragments containing additional domains (Figure 4.1: (Hata et al., 2000; Tsai and Reed,

1998)). This study split those fragments and did not detect DNA binding with the isolated cluster. The results of the yeast two-hybrid suggest that the cluster also mediates protein-protein interactions; however, the co-immunoprecipitation results were inconclusive. The predictions presented here indicated that most of the fingers are protein binding fingers. It is interesting however, that finger 13, like finger 8 is weakly DNA-like. Furthermore, fingers 14 and 15, of Cluster 4, are also weakly DNA-like. Hata et al. (2000) demonstrated DNA binding by a protein fragment containing these fingers 8 and 13 fingers (as well as others) that was independent of protein interactions, while a fragment containing fingers 13, 14, and 15 bound DNA in conjunction with SMAD binding. Although it is unusual for DNA interactions to depend on only two fingers, it is not unheard of (Fairall et al., 1992; Fairall et al., 1993; Hata et al., 2000; Tsai and Reed, 1998).

The rest of the predictions agree with known function for OAZ, with the exception of fingers 19, 25, and 28. Each of which is also weakly DNA-like despite being in clusters involved in DNA binding. However, it is unlikely that individual fingers would be capable of mediating sequence specific DNA binding, and, unlike with fingers 8 and 13, the distance between each of these fingers is large.

Zif268 contains three zinc fingers and is well known for its DNA binding capabilities. With a solved structure demonstrating not only the contribution of individual fingers to DNA binding, but also indicating critical amino acids, there is no doubt all three fingers are able to bind DNA. However, this study determined these fingers were also capable of binding protein, although the contribution of individual

fingers is not known. Not surprisingly, Zif268 had clear DNA-binding characteristics.

All fingers of contained negatively charged amino acids in the α -helix, suggesting that it is capable of binding protein.

It is the location of the negatively charged amino acids in all three fingers of Zif268 that most demonstrate that the predictive value of the amino acid characteristics is very low. In all three fingers of Zif268, there is a D in position 2 of the α -helix, and this amino acid has been shown, in a solved structure, to interact with DNA. As a result, it is clear that the prediction of protein binding by C2H2 domains is still far from accurate. Despite this, as has been demonstrated consistently throughout this thesis, it appears to be much easier to accurately predict DNA binding capabilities, versus protein binding capability, of C2H2 domains.

CHAPTER VII: SUMMARY

The most significant conclusion of this study is that DNA binding is likely a more restricted function of zinc fingers while their potential for mediating protein binding is likely greater. The finding that putative protein interactions were observed for all ZF clusters tested, but DNA interactions were only observed for Cluster 1 of OAZ (and is known for Zif268), strongly supports this conclusion. Also supporting this conclusion was the comparative success for determining DNA binding sequence characteristics, as well as a reasonably accurate predictive algorithm.

In some respects, this conclusion is obvious. There are a wide variety of protein-interaction surfaces on the C2H2 domain (Brayer and Segal, 2008), while the DNA-interaction surface is restricted to the N-terminus of the α -helix (Wolfe et al., 2000). Similarly, the potential binding site within the partner biomolecule is more restricted, being limited to the size and shape of the major groove for DNA, versus an almost infinite variety of sizes and shapes in proteins. This conclusion is also supported by the observations of others, such as Mackay (Gamsjaeger et al., 2007), who recently pointed out that in some cases “a single classical ZF is capable of mediating protein–protein interactions, and that an array of such domains is necessary for high affinity DNA binding.” However, if the potential of C2H2 ZFs for mediating protein interactions is

indeed greater than their potential for DNA binding, it would suggest a need for a dramatic reevaluation of zinc finger function.

The data presented here also supports the hypothesis that unanticipated functional information can be obtained by using a less biased approach, and that some C2H2 domains that were determined to have a particular function in one context may in fact be capable of supporting other types of interactions. Combined with recent observations that protein binding by C2H2 domains may be much more common than previously believed (Brayer and Segal, 2008); the common functional annotation of a ZF-containing protein as DNA-binding, while not inaccurate, is not sufficient to describe the full functionality provided by the ZF domains. Furthermore, if further studies were to find these conclusions to be generally true for C2H2 ZFs in other proteins, it would suggest that many previously described C2H2 domains have been mischaracterized. This is particularly true of domains that have been functionally annotated based on descriptive computational methods such as Gene Ontology, or by studies that have only examined DNA binding.

Finally, although the C2H2 protein domain is one of the simplest folds found in nature, it is proving to be profoundly intricate in its protein-protein interactions. With recent advances, including structural information and more complete mutagenesis studies, the characterization of C2H2 ZF-mediated PPIs is starting to approach our understanding of ZF-DNA interactions. However, the complexity and variety of PPIs add further challenges to this task. DNA-binding C2H2 domains rely on a binding surface comprised of a small number of amino acids invariably located in the N-terminal

region of the α -helix. In contrast, protein-binding C2H2 domains utilize many different regions of the fold, including the β -sheets, the linker regions, as well as residues in the α -helix. Different surfaces of the α -helix can be used, and the overall protein binding surface is frequently larger than that observed for DNA binding.

Aside from C2H2 ZF-protein binding being underestimated and under-annotated, it has probably also been underappreciated. The possibility of developing ZF PPIs as a new class of drug targets, or designing custom ZFs protein-binding domains, has lagged behind the manipulation of C2H2 ZF-DNA interactions. Today, C2H2 domains can be engineered to bind almost any DNA target desired. This has led to the generation of artificial proteins that can be used as research tools and therapeutics (Ghosh et al., 2006; Jamieson et al., 2003; Porteus and Carroll, 2005).

Typically, synthetic proteins are generated by mixing and matching naturally occurring C2H2 domains to create a protein with novel DNA binding properties (Bae et al., 2003), or by altering specific residues within the framework of a standard C2H2 domain, typically that of either Zif268 or Sp1. Target specificity is achieved by selecting amino acids in the DNA contacting N-terminal portion of the α -helix that are appropriate for the DNA target. Today, there are websites that facilitate the design of target specific DNA binding C2H2 domains.

McCarty et al. (2003) reported the first successful attempt to create novel C2H2 ZF-protein interactions by mixing together C2H2 domains from separate proteins. Going one step further, Giesecke et al. (2006) created an artificial gene network able to activate transcription from an endogenous gene in human cells. Currently, there is neither a way

to predict protein binding by C2H2 zinc finger domains nor any tools to facilitate the design of target specific protein binding C2H2 domains.

The manipulation of DNA-binding ZFs to create diverse sets of custom DNA-binding proteins, research tools and drug therapies has resulted in a much greater understanding of how these domains interact with DNA. The recent works of Westman (Westman et al., 2004), McCarty (McCarty et al., 2003), and Giesecke (Giesecke et al., 2006) have demonstrated that protein-binding ZFs can also be manipulated to create modified and selective protein-binding surfaces. It is hoped that continued engineering efforts will eventually produce custom protein-interaction tools as well as greater insights into the structural features underlying C2H2 ZF PPI function and specificity and, therefore better insights into the biology of one of the largest protein superfamilies of found in nature.

APPENDIX A: OTHER PROTEIN BINDING DOMAINS

Table A.1: Proteins Containing C2H2 Domains Implicated in Protein-Protein Interactions

Protein	Alias/Symbol	Fingers	Partner	Ref
Aiolos	ZnFN1A3, IKZF3	5-6	Ikaros, Aiolos, Pegasus, Helios	(Kelley et al., 1998; Morgan et al., 1997; Perdomo et al., 2000)
AT-binding transcription factor 1	ZFHX3, ATBF1	22-23	PIAS3	(Nojiri et al., 2004)
B-cell CLL/lymphoma 11A	Bcl11a (MGI), EVI9, CTIP1	1-2	ARP1	(Avram et al., 2000)
		4	Self	(Avram et al., 2002)
B-cell CLL/lymphoma 11B	ZfphRit1 alpha, CTIP2	2-4	Tat, HP1	(Rohr et al., 2003)
		5-7	Tat, HP1	(Rohr et al., 2003)
B-cell CLL/lymphoma 6	Znf 51, BCL6	1-3	PLZF	(Dhordain et al., 2000)
		3-6	HDAC5/HDAC7	(Lemercier et al., 2002; Mascle et al., 2003)
		4-5	ETO	(Chevallier et al., 2004)
		All	LRF	(Davies et al., 1999)
		All	Miz1	(Phan et al., 2005)
		All	c-Jun, JunD, JunB	(Vasanwala et al., 2002)
Bone marrow zinc finger 2	ZNF224, BMZF2	6-10	WT1 -KTS	(Lee et al., 2002)
CCCTC-binding factor	CTCF	All	YB-1	(Chernukhin et al., 2000)
		All	YY1	(Donohoe et al., 2007)
DAZ interacting protein 1, testis 1	DZIPt1	1	DAZ	(Moore et al., 2004)

Table A.1: continued

Protein	Alias/Symbol	Fingers	Partner	Ref
DAZ interacting protein 1, testis 2	DZIPt2	1	DAZ	(Moore et al., 2004)
Early growth response 1	Zif263, ZNF225, Egr1	All	RELA	(Chapman and Perkins, 2000)
Early hematopoietic zinc finger	ZNF521, Evi3	All	SMAD1, SMAD4	(Bond et al., 2004)
Ecotropic viral integration site 1	Evi1, PRDM3	1, 6	Gata1	(Laricchia-Robbio et al., 2006)
Eos	ZNFN1A4, IKZF4	5-6	Eos, Ikaros, Pegasus, Trps1	(Honma et al., 1999; Perdomo et al., 2000; Westman et al., 2004)
Fanconi anemia zinc finger protein	ZBTB32, FAZF, ZNF538	All	GATA2	(Tsuzuki and Enver, 2002)
FLT3-interacting zinc finger 1	ZNF798, Fiz1	1-4 or 7-11	Flt3	(Wolf and Rohrschneider, 1999)
GLI-Kruppel family member GLI3	Gli3	3-5	Smad1/2	(Liu et al., 1998)
Glioma-associated oncogene 1	Gli1	3-5	Zic-1, -2, -3	(Koyabu et al., 2001)
Growth factor independent 1	ZNF163, GFI1	3-5	PU.1	(Dahl et al., 2007)
		All	MTG8 -human	(McGhee et al., 2003)
Helios	ZNFN1A2, IKZF2	5-6	Helios, Ikaros, Aiolos	(Kelley et al., 1998)
Hunchback	hb	5-6	hunchback	(McCarty et al., 2003)
Ikaros	ZNFN1A1, IKZF1	5-6	Ikaros, Aiolos, Helios, Eos, Pegasus	(McCarty et al., 2003; Sun et al., 1996)
KRAB box containing zinc finger protein	Krim1 (RGD)	2	MYC	(Hennemann et al., 2003)
Kruppel-like factor 1 (erythroid)	KLF1, EKLF	All	FLI1	(Starck et al., 2003)

Table A.1: continued

Protein	Alias/Symbol	Fingers	Partner	Ref
Kruppel-like factor 13	KLF13, FKLF2	All	CBP/p300	(Song et al., 2002)
Lola locus isoform 3D	<i>LOLA3D</i>	1-2	JIL1 Kinase	(Zhang et al., 2003)
Myc-associated zinc finger protein-related factor	ZNF278, PATZ1, MAZR	All	MITF	(Morii et al., 2002)
NRC-interacting factor 1	Zfp335	5-6	NRC	(Mahajan et al., 2002)
Pegasus	ZNFN1A5, IKZF5	4-5	Pegasus, Eos, Ikaros, Aiolos	(Perdomo et al., 2000)
polymerase (DNA directed), eta	POLH	1	Ubiquitin	(Bomar et al., 2007)
Promyelocytic leukemia zinc finger protein	ZNF145, ZBTB16	1-3	RARa, GR, ERa	(Martin et al., 2003)
		1-6	GATA2	(Tsuzuki and Enver, 2002)
		6-7	HB-EGF-C	(Nanba et al., 2003)
		All	BCL6	(Dhordain et al., 2000)
RE1-silencing transcription factor	REST, NRSF	9	Co-REST	(Andres et al., 1999)
Recombination activating gene 1	Rag1	B	RAG2	(Aidinis et al., 2000)
Recombination activating gene 1	Rag1	A	Rag1	(Rodgers et al., 1996)
ROAZ	ZFP423	29	Olf-1	(Tsai and Reed, 1998)
Schnurri	shn	1-2, 4-8	MAD	(Dai et al., 2000)
Senseless	Sens	2-3	Scute	(Acar et al., 2006)
Sequoia	seq	All	dsh	(Formstecher et al., 2005)
Serendipity	SRY	6	Self	(Payre et al., 1997)
Smad- and Olf-interacting zinc finger protein	ZNF423, OAZ	14-19	SMAD-1, -4	(Hata et al., 2000)

Table A.1: continued

Protein	Alias/Symbol	Fingers	Partner	Ref
Sp1 transcription factor	Sp1	1	E2F1	(Rotheneder et al., 1999)
		All	BCoR, NCoR, SMRT	(Lee et al., 2005)
		All	E2F	(Karlseder et al., 1996)
		All	Huntingtin	(Li et al., 2002)
		All	MYC	(Gartel et al., 2001)
		All	TAF1	(Suzuki et al., 2003)
		All	YY1	(Lee et al., 1993; Seto et al., 1993)
Sp2 transcription factor	Sp2	1	E2F1	(Rotheneder et al., 1999)
Sp3 transcription factor	Sp3	1	E2F1	(Rotheneder et al., 1999)
Sp4 transcription factor	Sp4	1	E2F1	(Rotheneder et al., 1999)
Transcription factor 8	Tcf8, ZEB1	5	Oct1	(Smith and Darling, 2003)
Trichorhinophalangeal syndrome 1	Trps1, zfp GC79	8-9	Eos	(Westman et al., 2004)
Uncoordinated protein 98	unc-98	All	UNC-97 (pinch)	(Mercer et al., 2003)
Wilms Tumor 1 (-KTS)	WT1 (-KTS)	1-2	WTAP	(Little et al., 2000)
		1-2	CBP/p300	(Lee and Haber, 2001)
		All	BMZF2, Ciao-1, Par4	(Lee and Haber, 2001; Lee et al., 2002)
		All	NHRPU	(Spraggon et al., 2007)
Wilms Tumor 1 (+KTS)	WT1 (+KTS)	1-2	CBP/p300	(Lee and Haber, 2001)
		2-4	U2AF65	(Davies et al., 1998)

Table A.1: continued

Protein	Alias/Symbol	Fingers	Partner	Ref
Ying Yang 1	YY1	1	ATF/CREB	(Zhou et al., 1995)
		1-2	Sp1	(Lee et al., 1993; Seto et al., 1993)
		1-2	YAF2	(Kalenik et al., 1997)
		2-4	TAFII55	(Austen et al., 1997)
		3-4	Adenovirus E1A	(Lee et al., 1995; Lewis et al., 1995)
		All	CTCF	(Donohoe et al., 2007)
		All	MYC	(Austen et al., 1998)
		All	Smad4	(Kurisaki et al., 2003)
		All	TBP, CBP	(Austen et al., 1997)
		All	TFIIB	(Austen et al., 1997)
Zic family member 1	Zic1	3-5	Gli-1, -2, -3	(Koyabu et al., 2001)
Zic family member 2	Zic2	3	Ku70, Ku80, PARP, RHA	(Ishiguro et al., 2007)
Zinc finger 148	ZNF148, ZBP89	All	p53	(Bai and Merchant, 2001)
Zinc finger and BRCA1-interacting protein with a KRAB domain 1	ZNF350, ZBRK1	7-8	BRCA1	(Tan et al., 2004)
Zinc finger and BTB domain-containing protein 7A	ZBTB7A	All	BCL6	(Davies et al., 1999)
Zinc finger protein 161	Zfp106, zf5	1-5	Self	(Numoto et al., 1999)
Zinc finger protein 219	Zfp219	7-9	mSufu	(Paces-Fessy et al., 2004)
Zinc finger protein 251	Znf251	1-5	Smad1	(Colland et al., 2004)

Table A.1: continued

Protein	Alias/Symbol	Fingers	Partner	Ref
Zinc finger protein 295	znf295, zbtb21	1-9	Zfp161	(Wang et al., 2005)
Zinc finger protein 41	znf41	9-16	Smad2	(Colland et al., 2004)
Zinc finger protein 484	ZNF484	2-5	Smad8	(Colland et al., 2004)
Zinc finger protein 512	ZNF512B	3-5	Many	(Colland et al., 2004)
Zinc finger protein 512	ZNF512	5-6	Many	(Colland et al., 2004)
Zinc finger protein 76	znf76	1-5	Smad1	(Colland et al., 2004)
Zinc finger protein 8	znf8	1-6	Many	(Colland et al., 2004)
Zinc finger protein 8	Znf8	All	Smad8a	(Colland et al., 2004)
Zinc finger protein 83	znf83	8-15	Smad3, Smad8	(Colland et al., 2004)
Zinc finger, X-linked, duplicated A	ZXDA	All	ZXDC	(Al-Kandari et al., 2007)
Zkscan17 (MGI)	zfp496	All	jumonji/jarid2	(Mysliwiec et al., 2007)

APPENDIX B: SUPPLEMENTARY MATERIAL

hHEL_4	HKCNY--CGRSYKQRSSLEEHEKERC--HNYLQN--
hEOS_4	YKCN--CGRSYKQOSTLEEHEKERC--HNYLQS--
hAIO_1	MNCDV--CGLSCISFNVLMLVHKRS--HTGERP--
hEOS_1	LKCDV--CGMVCIGPNVLMVHKRS--HTGERP--
Peg_2	HRCHL--CPFASAYERHLEAHMRS--HTGEKP--
TFIIIA_1	YICSFADCGAAYNKNWKLQAHLC--HTGEKP--
Aiolos_2	FQCNQ--CGASFTQKGNLLRHIL--HTGEKP--
Ikaros_2	FQCNQ--CGASFTQKGNLLRHIL--HSGEKP--
hZAP1_4	FECKD--CGKAFIQKSNLIRHQRT--HTGEKP--
ZBRK1_1	HVCSE--CGKAFIKKSWLTDHQVM--HTGEKP--
Zbrk1_4	YICSE--CGKGFIQKGNLIVHQRI--HTGEKP--
Zbrk1_5	YICNE--CGKGFIQKTCLIAHQRF--HTGKTP--
Tcf8_3	FKCTE--CGKAFKYKHHLEKHLRI--HSGEKP--
ATF2	FLCTAPGCGQRF--TNEHDHLAVHKKH--HEMTLK--
phWT1_4	FSCRWPSCQKKFARSDELVRHHNM--HQRNM---
YY1_3	FQCTFEGCGKRFSLDNLRTHVRI--HTGDRP--
SWI5_1	FECLFPGCTKTFKRRYNIRSHIQ--HLED RP--
YY1_4	YVCPFDGCNKKFAQSTNLKSHILT--HAKAKN--
TFIIIA_2	FPCKEEGCEKGFTSLHHLTRHSLT--HTGEKN--
hGLI_2	FVCHWGGCSRPFKAQYMLVVHMRR--HTGEKP--
phWT1_3	FQCKT--CQRKFSRSDHLKTHTRT--HTGKTS--
Zac1_3	HQCAH--CEKTFNRKDHLKHNHLQT--HDPNKM--
SWI5_2	YSCDHPGCDKAFVRNHDILRHKK--HQEKAY--
TFIIIA_5	YECPHGCDKRFSLPSRLKRHEKV--HAG-----
LolaK_1	YVCRH--CGKKYRWKSTLRRHENVE--CGGKEP--
Zfp219_1	FPCPV--CGKRFRFNSILALHLRA--HPGAQA--
Zfp219_5	FRCQV--CGQSFTQSWFLKGHMRK--HKASFD--
Zac1_4	FGCEE--CGKKYNTMLGYKRHLAL--HAASSG--
Ikaros_3	FKCHL--CNYACRRRDALTGHLRT--HSVGP--
Helios_3	FKCPF--CSYACRRRDALTGHLRT--HSVGP--
Zfp219_2	FQCPH--CGHRAAQRALLRSHLRT--HQPERP--
LolaK_2	HPCPY--CSYKAKQRGNLGVHVRK--HHPEKP--
GAGA	ATCPI--CYAVIRQSRNLRRHELRL--HFAKPG--
Rest_1	FRCKP--CQYEAEESEEQFVHHIRV--HSAKKF--
Zap1_5	YKCKT--CKRCFSSEETLVQHTRT--HSGEKP--
hb_1	YKCKT--CGVVAITKVDFWAHTRT--HMKPDK--
TTKB_1	YRCKV--CSRVTYTHISNFCRHYVTS--HKRNK--
ATBF1_21	RRCPF--CRALFKAKTALEAHIRSR--HWHEAK--
hb_2	LQCPK--CPFVTEFKHHLEYHIRK--HKNQKP--
serenD_3	KQCRY--CPKSF--SRPVNTLRHMRM--HWDK--
hPEG_3	YKCEL--CSFRCSDRSNLSHRRRRK--HKMVPI--
Tcf8_4	YECPN--CKKRF--SHSGSYSSHSSKKCISLIP--
TTKB_2	YPCPF--CFKEFTRKDNMTAHVKII--HKIENP--
TFIIIA_3	FTCDSDGCDLRFTTKANMKKHFNRF--HNIKIC--

* * * *

Figure B.1: Alignment of DNA-binding C2H2 domains used in amino acid characterization.

```

IKAROS_5  YKCEH--CRVL-FLDHVMTIHMGC--HGFRDP-
HELIOS_5  FKCEH--CRVL-FLDHVMTIHMGC--HGYRDP--
aiolos_5  YRCDH--CRVL-FLDYVMFTIHMGC--HGFRDP--
Pegasus_5 HHCQH--CDMY-FADNILYTIHMGC--HGYENP--
hb_5      YECKY--CDIF-FKDAVLYTIHMGY--HSCDDV--
OAZ_19    HKCIF--CGET-FSTEVELQCHITT--HSKK----
OAZ27     YQCIK--CQMT-FENEREQIHVAN--HMIEEG--
IKAROS_6  FECNM--CGYH-SQDRYEFSSHITRGEHRFHMS--
HELIOS_6  LECNI--CGYR-SQDRYEFSSHIVRGEHTFH----
EOS_6     FECNI--CGYH-SQDRYEFSSHIVRGEHKVG----
AIOLOS_6  FECNM--CGDR-SHdryEFSSHIARGEHRSLLK--
Pegasus_6 FQCNI--CGCK-CKNKYDFACHFARG--HITNID--
hb_6      FKCNI--CGEK-CDGPVGLFVHMARNAHS-----
Zfp335_5  FPCPV--CGRV-YPMQKRLTQHMKT--HSTEKP--
PLZF_1    EQCSV--CGVE-LPDNEAVEQHRKL--HSGMKT--
OAZ_15    QACPQ--CKED-FDSQESLLQHLTV--HYMTTS--
OAZ_17    YHCTL--CQEV-FDSKVSIVHLAVK--HSNEKK--
OAZ28     HECKL--CNQM-FDSPAKLLCHLIE--HSFEGM--
FAZF_1    YACSV--CGKR-FSLKHQMETHYRV--HTGEKP--
ZBRK1_7   FECSE--CGKA-FSTKQKLIVHQR--HTGERP--
ZBRK1_8   YGCNE--CGKA-FAYMSCLVKHKRI--HTREKQ--
krim1_2   HTCKQ--CGKA-FSRPSQLHKHERI--HTGEKP--
hGLI_3    HKCTFEGCRKS-YSRLENLKLTHLRS--HTGEKP--
hWT1_1    FMCAYPGCNKR-YFKLSHLQMHSRK--HTGEKP--
ZFP219_7  KDCPF--CGKS-FRSAHHLKVHLRV--HTGERP--
Tcf8_5    YACDL--CDKI-FQKSSSLLRHKYE--HTGKRP--
Zfp335_6  HMCCK--CGKS-FKKRYTFKMHLT--HIQAVA--
serenD_6  HYCSV--CPKS-FTERYTLKMHMKT--HEGDVV--
PLZF_2    YGCEL--CGKR-FLDSLRLRMHLA--HSAGAK--
FAZF_2    FCSL--CPQR-SRDFSAMTKHLRT--HGAAPY--
phYY1_1   IACPHKGCTKM-FRDNSAMRKHLHT--HGPRV---
ZAC1_6    HQCDH--CERC-FYTRKDVRRLV--HTGCKD--
ROAZ_29   YDCSQ--CPQK-FFFQTELQNHTMSQ--HAQ-----
OAZ30     YDCSQ--CPQK-FFFQTELQNHTMSQ--HAQ-----
phEEA1_1  FICPQ--CMKS-LGSADELFKHYEAV--HDAGND--
OAZ29     FKCPV--CFTV-FVQANKLQHHIFAV--HGQEDK--
lolaF_1   FHCAV--CNKS-YLRKRHLQRHMRDE--CIGIPP--
lolaF_2   FNCEF--CSSR-FRRKYHMRHLVSK--HGIPPA--
ZAC1_7    FLCQF--CAQR-FGRKDHLTRHTKKT--HSQELM--
yZAP1_1   LKCKWKECPES-CSSLFDLQRHLLKD--HVSQDF--
Zfp106_3  YRCWWHGCSLI-FGVVDHLKQHLLTD--HTNPNF--
OAZ_16    YVCES--CDKQ-FSSVDDLQKHLLD--HTFVL---
OAZ_18    YRCTA--CNWD-FRKEADLQVHVKHS--HLGNPA--
DZIP1_1   YQCHF--CDKA-FMNQAFLOSHIQR--HTEENS--
DZIP2_1   HTCHL--CDKT-FMNATFLRGHIQR--HAGVAE--
ZPF219_8  YKCPH--CDYA-GTQSGSLKYHLQRH--HREQR---
OAZ_20    YNCKF--CSKA-FHAILLEKHLREK--HCVFDA--
ZPF219_9  YRCLF--CPFA-TGAPELMALHLQV--HHSRRA--
FAZF_3    YRCSL--CGAG-CPSLASMQAHMRG--HSPSQL--
REST_9    FVCIF--CDRS-FRKGKDYSKHLNR--HLNVNY--

```

Figure B.2: Alignment of protein-binding C2H2 domains used in amino acid characterization.

Continued next page.

OAZ26	LRCPE--CSVK-FESAEDLESHMQVD-HRDLTP--
OAZ_14	YPCNQ--CDLK-FSNFESFQTHLKL--HLELLL--
Fog_CCHC	FPCKD--CGIW-YRSENLQAHLLEY-CASRQG--
ATBF1_22	LVCRK--CQAG-FSDEEAARSHLKSL-CFFGQS--
ATBF1_23	YHCLA--CESA-LCGEEALSQHLESALHKHRTI--
PolH_1	VPCEK--CGSL-VP-VWDMPEHMDY--HFALEL--
hGLI_1	TDCRWDGCSQE-FDSQEQLVHHINSE-HIHGER--
yZAP1_2	LACNWEDCDFL-GDDTCSIVNHINCQ-HGINFD--
RAG1_2	YICTL--CDATRLEASQNLVFHSITRSHAENLE--
EOS_5	FKCEH--CRIL-FLDHVMFTIHMGC--HGFRDP--
Fog_3	FVCLI--CLSA-FTTKANCERHLKV--HTDTLS--
	* * * *

Figure B.2 cont.: Alignment of protein-binding C2H2 domains used in amino acid characterization.

```

sp|Q01954|BNC1_HUMAN/928-956/U
sp|Q14872|MTF1_HUMAN/140-169/U
sp|P55201|BRPF1_HUMAN/21-52/US
sp|Q86VZ6|JAZF1_HUMAN/173-203/
sp|P10072|HKR1_HUMAN/282-309/U
sp|Q8TAX0|OSR1_HUMAN/175-202/U
sp|Q96JB3|HIC2_HUMAN/561-588/U
sp|A0PJY2|FEZF1_HUMAN/260-287/
sp|Q9H116|GZF1_HUMAN/575-602/U
sp|Q9NPC7|MYNN_HUMAN/302-329/U
sp|O14753|OVOL1_HUMAN/118-145/
sp|Q8NI51|BORIS_HUMAN/514-542/
sp|Q5JPB2|CT174_HUMAN/172-201/
sp|Q9UHK0|NUFP1_HUMAN/174-201/
sp|Q14526|HIC1_HUMAN/437-464/U
sp|Q96JM3|K1802_HUMAN/738-765/
sp|P29119|FURI1_XENLA/652-683/
sp|Q9BXB1|LGR4_HUMAN/856-884/U
sp|P43243|MATR3_HUMAN/291-320/
sp|Q9H3H1|MOD5_HUMAN/395-424/U
sp|Q96NB3|CCD16_HUMAN/51-80/US
sp|Q96AP4|CF113_HUMAN/2-29/USE
sp|Q5F1R6|DJC21_HUMAN/482-511/
sp|Q86V15|CASZ1_HUMAN/551-580/
sp|Q13133|NR1H3_HUMAN/113-140/
sp|Q9H8Y5|ANKZ1_HUMAN/72-101/U
sp|Q5VVQ6|OTU1_HUMAN/318-347/U
sp|P92189|STIL_DROME/45-74/USE
sp|Q9UPR6|K1086_HUMAN/271-300/
sp|Q13118|KLF10_HUMAN/429-456/
sp|Q9BQA5|MIZF_HUMAN/15-44/USE
sp|Q68CP9|ARID2_HUMAN/1632-166
sp|Q8IVH2|FOXP4_HUMAN/307-337/
sp|O60678|ANM3_HUMAN/48-76/USE
sp|Q6IQ32|ADNP2_HUMAN/770-798/
sp|Q00453|RGM1_YEAST/49-79/USE
sp|P35555|FBN1_HUMAN/1898-1929
sp|Q8N4C8|MINK1_HUMAN/127-156/
sp|Q07407|FGFR1_DROME/460-487/
sp|Q07008|NOTC1_RAT/554-584/US
sp|P97766|CFC1_MOUSE/130-161/U
hMZf1_1
sp|Q5TEC3|ZN697_HUMAN/189-216/
sp|Q14586|ZN267_HUMAN/716-743/
sp|Q14591|ZN271_HUMAN/286-313/
sp|Q9NYT6|ZN226_HUMAN/755-782/
sp|O75437|ZN254_HUMAN/266-293/
sp|Q99592|ZN238_HUMAN/466-494/
sp|Q9P2Y4|ZN219_HUMAN/57-84/US
sp|Q96CS4|ZN689_HUMAN/429-456/
ITCHL--CQKTYSNKGTFRAHYKTV--HLRQLH--
YQCTFEGCPRTYSTAGNLRTHQKT--HRGEYT--
YECPVETCRKVYKSYSGIEYHLYHYDHDNPPP--
FACPVPGCKKRYKNVNGIKYHAKNG--HRTQIR--
YVCRE--CGRGFTWKSNIITHQRT--HSGEKP--
FVCKF--CGRHFTKSYNLLIHERT--HTDERP--
FACDE--CGMRFTQYRLTEHMRV--HSGEKP--
FTCEV--CGKVFNAHYNLTRHMPV--HTGARP--
FMCNA--CGRTFTDKSTLRRHTSI--HDKNTP--
PMCNT--CGKVFSEASSLRRHMRI--HKGVKP--
FTCRV--CQKAFTYQRMNLNRHMKC--HNDVKR--
FTCLS--CNKCFRQKQLLNAHFRKY--HDANFI--
FPCAT--CGIAFKTQSNLYKHRRTQTHLNNSR--
FFCDT--CDRGFKNQEKYDKHMSE--HTKCPE--
YVCIP--CGKGFPSSEQLNAHVEA--HVEEEE--
YKCTI--CGKAFLLESLLKNHVAA--HGQSLL--
TICCTSKCASLLQDQSYRTHHSCPAYHTNDDS--
TVCDC--CESFLLTKPVSCKHLIKS--HSCPAL--
HLCISI--CDLPVHSNKEWSQHINGASHSRRCQ--
HLCDL--CDRIIIGDREWAHAHIKSKSHLNQLK--
LSCAL--CNTPVKSELLWQTHVLGKQHREKVA--
LSCNI--CGETVTSEPDMAHLIV--HMESEI--
ISCTT--CHSEFP SRNKLFDHLKATGHARVPS--
YHCMQVGCNKVYTSTSDVMTHENF--HKKNTQ--
LSCEG--CKGFFRRSVIKGAHYIC--HSGGHC--
LFCST--CDQTFQNHQEQREHYKLDWHRFNLK--
LRCMV--CQKGLTGQAEAREHAKETGHTNFGE--
DSCTV--CRISFYGYPNCSKHVISRKHLMMGE--
HYCDI--CKISCAGPQTYREHLGGQKHKKEA--
FACPM--CDRRFMRSDDLTKHARR--HLSAKK--
LQCEWGS CSFVCSTMEKFFEHTVQ--HLQQHL--
FMCLWQSCKKWFQTPSQVIFYHAATE--HGKDV--
GECKWPGCETLCEDLGQFIKHLNTE--HALDDR--
TPCLF--CNRLFTSAEETFSHCKSE--HQFNID--
LGCLF--CPCTFHDIKGLSEHSRNR--HLGKKK--
IHCPNSFCKGFISSDSNQMH SILD--HVRDRL--
DACGNGTCRNTIGSFNCRNNGFILSHNNDCI--
EDCIAIYICREILRGLAHLHAHKVI--HRDIKG--
EMCKI--CLIIINMVGLENIHGRV--HSVYLD--
GYCELTCTCDGEPPEVIHDCLI--HDGQVH--
ARCLGTRCRQDCLPFCQSYAHIRCCLHSLLS--
GRCDV--CGKVSQRSNLLRHQKI--HTGERP--
TICPD--CGESFSPGAFLQHQR--HRLAEA--
YKCEE--CGKAFNSRSYLIAHQRS--HTREKL--
YKCNE--CGKAFNQSSVLILHQRI--HTGEKP--
YKCEI--CGKSFSWRSNLTVHHRI--HVGDKS--
YKCEE--CGEAFNRSSNLTTHKII--HTGEKP--
HACKW--CERRFTQSGDLRYHRIKF--HCLVNV--
FPCPV--CGKRFRFNSILALHLRA--HPGAQA--
YACDL--CSKRFAQWSHLAQHQLL--HTGEKP--

```

Figure B.3: Alignment of random C2H2 domains used in amino acid characterization.

Zfp335_5	FPCPV--CGRV-YPMQKRLTQHMKT--HSTKEP--
PLZF_1	EQCSV--CGVE-LPDNEAVEQHRKL--HSGMKT--
FAZF_1	YACSV--CGKR-FSLKHQMETHYRV--HTGEKP--
ZBRK1_7	FECSE--CGKA-FSTKQKLIVHQRT--HTGERP--
ZBRK1_8	YGCNE--CGKA-FAYMSCLVKHKRI--HTREKQ--
krim1_2	HTCKQ--CGKA-FSRPSQLHKHERI--HTGEKP--
hGLI_3	HKCTFEGCRKS-YSRLLENLKTHLRS--HTGEKP--
hWT1_1	FMCAYPGCNKR-YFKLSHLQMHRSK--HTGEKP--
ZFP219_7	KDCPF--CGKS-FRSAHHLKVHLRV--HTGERP--
Tcf8_5	YACDL--CDKI-FQKSSSLLRHKYE--HTGKRP--
Zfp335_6	HMC DK--CGKS-FKKRYTFKMHLT--HIQAVA--
serendipityD_6	HYCSV--CPKS-FTERYTLKMHMKT--HEGDVV--
PLZF_2	YGCEL--CGKR-FLDSLRLRMHLLA--HSAGAK--
FAZF_2	FSCSL--CPQR-SRDFSAMTKHLRT--HGAAPY--
phYY1_1	IACPHKGCTKM-FRDNSAMRKHLHT--HGPRV---
Fog_3	FVCLI--CLSA-FTTKANCERHLKV--HTDTLS--
ZAC1_6	HQCDH--CERC-FYTRKDVRRLV--HTGCKD--
ROAZ_29	YDCSQ--CPQK-FFFQTELQNHTMSQ-HAQ-----
phEEA1_1	FICPQ--CMKS-LGSADELFKHYEAV-HDAGND--
lolaF_1	FHCAV--CNKS-YLRKRHLQRHMRDE-CIGIPP--
lolaF_2	FNCEF--CSSR-FRRKYHMRHLVSK-HGIPPA--
ZAC1_7	FLCQF--CAQR-FGRKDHLTRHTKKT-HSQELM--
yZAP1_1	LKCKWKECPES-CSSLFDLQRHLLKD-HVSQDF--
Zfp106_3	YRCWWHGCSLI-FGVVDHLKQHLLTD-HTNPNF--
DZIP1_1	YQCHF--CDKA-FMNQAFQLQSHIQRR-HTEENS--
DZIP2_1	HTCHL--CDKT-FMNATFLRGHIQRR-HAGVAE--
ZPF219_8	YKCPH--CDYA-GTQSGSLKYHLQRH-HREQR---
ZPF219_9	HRCLF--CPFA-TGAPELMALHLQV--HHSRRA--
FAZF_3	YRCSL--CGAG-CPSLASMQAAMRG--HSPSQL--
REST_9	FVCIF--CDRS-FRKGKDYSKHLNR--HLNVNY--
Fog_CCHC	FPCKD--CGIW-YRSERNLQAHLLEY-CASRQG--
ATBF1_22	LVC RK--CQAG-FSDEEAARSHLKSL-CFFGQS--
ATBF1_23	YHCLA--CESA-LCGEEALSQHLESALHKHRTI--
PolH_1	VPCEK--CGSL-VP-VWDMPEHMDY--HFALEL--
hGLI_1	TDCRWDGCSQE-FDSQEQLVHHINSE-HIHGER--
yZAP1_2	LACNWEDCDFL-GDDTCSIVNHINCQ-HGINFD--
RAG1_2	YICTL--CDATRLEASQNLVFHSITRSHAENLE-

Figure B.4: Alignment of domains used to build the protein HMM profile.

TFIIIA_1-3_	YICSFADCGAAYNKNWKLQAHLC--HTGEKP--
hZAP1_4	FECKD--CGKAFIQKSNLIRHQRT--HTGEKP--
ZBRK1_1	HVCSE--CGKAFIKKSWLTDHQVM--HTGEKP--
Zbrk1_4	YICSE--CGKGF IQKGNLIVHQRI--HTGEKP--
Zbrk1_5	YICNE--CGKGF IQKTCLIAHQRF--HTGKTP--
Tcf8_3	FKCTE--CGKAFKYKHHLKEHLRI--HSGEKP--
ATF2	FLCTAPGCGQRFTNEDHLAVHKKH--HEMTLK--
phWT1_4	FSCRWPSCQKKFARSDELVRHHNM--HQRNM--
YY1_3	FQCTFEGCGKRFSLDFNLRTHVRI--HTGDRP--
SWI5_1	FECLFPGCTKTFKRRYNIRSHIQT--HLED RP--
YY1_4	YVCPFDGCNKKFAQSTNLKSHILT--HAKAKN--
TFIIIA_2	FPCKEEGCEKGFTSLHHLTRHSLT--HTGEKP--
hGLI_2	FVCHWGGCSRPFKAQYMLVVHMR--HTGEKP--
phWT1_3	FQCKT--CQRKFSRSDHLKTHTRT--HTGKTS--
Zac1_3	HQCAH--CEKTFNRKDHLKNHLQT--HDPNKM--
SWI5_2	YSCDHPGCDKAFVRNHD LIRHKKS--HQEKAY--
TFIIIA_5	YECPEGCDKRFSLPSRLKRHEKV--HAG-----
LolaK_1	YVCRH--CGKKYRWKSTLRRHENVE--CGGKEP--
Zfp219_1	FPCPV--CGKRFRFNSILALHLRA--HPGAQA--
Zfp219_5	FRCQV--CGQSFTQSWFLKGHMRK--HKASFD--
Zac1_4	FGCEE--CGKKYNTMLGYKRHLAL--HAASSG--
Zfp219_2	FQCPH--CGHRAAQRALLRSHLRT--HQPERP--
LolaK_2	HPCPY--CSYKAKQRGNLGVHVRK--HHPEKP--
GAGA	ATCPI--CYAVIRQSRNLRRHEL R--HFAKPG--
Rest_1	FRCKP--CQYEAEESEQFVHHIRV--HSAKKF--
Zap1_5	YKCKT--CKRCFSSEETLVQHTRT--HSGEKP--
hb_1	YKCKT--CGVVAITKVDFWAHTRT--HMKPDK--
TTKB_1	YRCKV--CSR VYTHISNFCRHYVTS--HKRN VK--
ATBF1_21	RRCPF--CRALFKAKTALEAHIRSR--HWHEAK--
serendipityD_3	KQCRY--CPKSFSRPVNTLRHMRM--HWDKKK--
hPEG_3	YKCEL--CSFRCSDRSNLSHRRRK--HKMVPI--
Tcf8_4	YECPN--CKKRFSHSGSYSSHISKKCISLIP--
TTKB_2	YPCPF--CFKEFTRKDNMTAHVKII--HKIENP--
TFIIIA_3	FTCDSDGCDLRFTTKANMKKHFNRF--HNIKIC--

Figure B.5: Alignment of domains used to build the DNA HMM profile.

```

EOS_5
OAZ_19
OAZ27
EOS_6
Zfp335_5
PLZF_1
OAZ_15
OAZ_17
OAZ28
FAZF_1
ZBRK1_7
ZBRK1_8
krim1_2
hGLI_3
hWT1_1
ZFP219_7
Tcf8_5
Zfp335_6
serendipityD_6
PLZF_2
FAZF_2
phYY1_1
Fog_3
ZAC1_6
ROAZ_29
OAZ30
phEEA1_1
OAZ29
lolaF_1
lolaF_2
ZAC1_7
yZAP1_1
Zfp106_3
OAZ_16
OAZ_18
DZIP1_1
DZIP2_1
ZPF219_8
OAZ_20
ZPF219_9
FAZF_3
REST_9
OAZ26
OAZ_14
Fog_CCHC
ATBF1_22
ATBF1_23
PolH_1
hGLI_1
yZAP1_2
RAG1_2

FKCEH--CRIL-FLDHVMFTIHMGC--HGFRDP--
HKCIF--CGET-FSTEVELQCHITT--HSKK----
YQCIK--CQMT-FENEREIQIHVAN--HMIEEG--
FECNI--CGYH-SQDRYEFSSHIVRGEHKVG----
FPCPV--CGRV-YPMQKRLTQHMKT--HSTEKP--
EQCSV--CGVE-LPDNEAVEQHRKL--HSGMKT--
QACPQ--CKED-FDSQESLLQHLTV--HYMTTS--
YHCTL--CQEV-FDSKVSIVHLAVK-HSNEKK--
HECKL--CNQM-FDSPAKLLCHLIE--HSFEGM--
YACSV--CGKR-FSLKHQMETHYRV--HTGEKP--
FECSE--CGKA-FSTKQKLIVHQRT--HTGERP--
YGCNE--CGKA-FAYMSCLVKHKRI--HTREKQ--
HTCKQ--CGKA-FSRPSQLHKHERI--HTGEKP--
HKCTFEGCRKS-YSRLENLKTHLRS--HTGEKP--
FMCAYPGCNKR-YFKLSHLQMHRSK--HTGEKP--
KDCPF--CGKS-FRSAHHLKVHLRV--HTGERP--
YACDL--CDKI-FQKSSSLLRHKYE--HTGKRP--
HMC DK--CGKS-FKKRYTFKMHLT--HIQAVA--
HYCSV--CPKS-FTERYTLMHMKT--HEGDVV--
YGCEL--CGKR-FLDSLRLRMHLLA--HSAGAK--
FSCSL--CPQR-SRDFSAMTKHLRT--HGAAPY--
IACPHKGCTKM-FRDNSAMRKHLHT--HGPRV--
FVCLI--CLSA-FTTKANCERHLKV--HTDTLS--
HQCDH--CERC-FYTRKDVRRHLVV--HTGCKD--
YDCSQ--CPQK-FFFQTELQNHTMSQ-HAQ-----
YDCSQ--CPQK-FFFQTELQNHTMSQ-HAQ-----
FICPQ--CMKS-LGSADELFKHYEAV-HDAGND--
FKCPV--CFTV-FVQANKLQQHIFAV-HGQEDK--
FHCFAV--CNKS-YLRKRHLQRHMRDE-CIGIPP--
FNCEF--CSSR-FRRKYHMRHLVSK-HGIPPA--
FLCQF--CAQR-FGRKDHLTRHTKKT-HSQELM--
LKCKWKECPES-CSSLFDLQRHLLKD-HVSQDF--
YRCWWHGCSLI-FGVVDHLKQHLLTD-HTNPNF--
YVCES--CDKQ-FSSVDDLQKHLLDM-HTFVL---
YRCTA--CNWD-FRKEADLQVHVKHS-HLGNPA--
YQCHF--CDKA-FMNQAFQLQSHIQRR-HTEENS--
HTCHL--CDKT-FMNATFLRGHIQRR-HAGVAE--
YKCPH--CDYA-GTQSGSLKYHLQRH-HREQR---
YNCKF--CSKA-FHAIILLEKHLREK-HCVFDA--
HRCLF--CPFA-TGAPELMALHLQV--HHSRRA--
YRCSL--CGAG-CPSLASMQAAMRG--HSPSQL--
FVCIF--CDRS-FRKGKDYSKHLNR--HLNVVY--
LRCPE--CSVK-FESAEDLESHMQVD-HRD LTP--
YPCNQ--CDLK-FSNFESFQTHLKL--HLELLL--
FPCKD--CGIW-YRSENLQAHLLEY-CASRQG--
LVCRK--CQAG-FSDEEAARSHLKS-LCFFGQS--
YHCLA--CESA-LCGEEALSQHLESALHKHRTI--
VPCEK--CGSL-VP-VWDMPEHMDY--HFALEL--
TDCRWDGCSQE-FDSQEQLVHHINSE-HIHLGER--
LACNWEDCDFL-GDDTCSIVNHINCQ-HGINFD--
YICTL--CDATRLEASQNLVFSITRSHAENLE--

```

Figure B.6: Alignment of domains used in cluster analysis.

APPENDIX C: PUBLISHED MANUSCRIPTS

Brayer, K.J., Kulshreshtha, S., and Segal, D.J. (2008). The Protein-Binding Potential of C2H2 Zinc Finger Domains. *Cell biochemistry and biophysics*. In press.

Brayer, K.J., and Segal, D.J. (2008). Keep Your Fingers Off My DNA: Protein-Protein Interactions Mediated by C2H2 Zinc Finger Domains. *Cell biochemistry and biophysics* 50, 111-131.

REFERENCES

- Acar, M., Jafar-Nejad, H., Giagtzoglou, N., Yallampalli, S., David, G., He, Y., Delidakis, C., and Bellen, H.J. (2006). Senseless physically interacts with proneural proteins and functions as a transcriptional co-activator. *Development (Cambridge, England)* *133*, 1979-1989.
- Aidinis, V., Dias, D.C., Gomez, C.A., Bhattacharyya, D., Spanopoulou, E., and Santagata, S. (2000). Definition of minimal domains of interaction within the recombination-activating genes 1 and 2 recombinase complex. *J Immunol* *164*, 5826-5832.
- Al-Kandari, W., Koneni, R., Navalgund, V., Aleksandrova, A., Jambunathan, S., and Fontes, J.D. (2007). The zinc finger proteins ZXDA and ZXDC form a complex that binds CIITA and regulates MHC II gene transcription. *Journal of molecular biology* *369*, 1175-1187.
- Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., *et al.* (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic acids research* *33*, D418-424.
- Andres, M.E., Burger, C., Peral-Rubio, M.J., Battaglioli, E., Anderson, M.E., Grimes, J., Dallman, J., Ballas, N., and Mandel, G. (1999). CoREST: a functional corepressor required for regulation of neural-specific gene expression. *Proceedings of the National Academy of Sciences of the United States of America* *96*, 9873-9878.
- Armstrong, S.A., Barry, D.A., Leggett, R.W., and Mueller, C.R. (1997). Casein kinase II-mediated phosphorylation of the C terminus of Sp1 decreases its DNA binding activity. *The Journal of biological chemistry* *272*, 13489-13495.
- Arndt, H.D. (2006). Small molecule modulators of transcription. *Angewandte Chemie (International ed)* *45*, 4552-4560.
- Arnett, H.A., Fancy, S.P., Alberta, J.A., Zhao, C., Plant, S.R., Kaing, S., Raine, C.S., Rowitch, D.H., Franklin, R.J., and Stiles, C.D. (2004). bHLH transcription factor Olig1 is required to repair demyelinated lesions in the CNS. *Science (New York, NY)* *306*, 2111-2115.
- Austen, M., Cerni, C., Luscher-Firzlauff, J.M., and Luscher, B. (1998). YY1 can inhibit c-Myc function through a mechanism requiring DNA binding of YY1 but neither its transactivation domain nor direct interaction with c-Myc. *Oncogene* *17*, 511-520.

REFERENCES - *continued*

- Austen, M., Luscher, B., and Luscher-Firzlaff, J.M. (1997). Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein. *The Journal of biological chemistry* 272, 1709-1717.
- Avram, D., Fields, A., Pretty On Top, K., Nevriy, D.J., Ishmael, J.E., and Leid, M. (2000). Isolation of a novel family of C(2)H(2) zinc finger proteins implicated in transcriptional repression mediated by chicken ovalbumin upstream promoter transcription factor (COUP-TF) orphan nuclear receptors. *The Journal of biological chemistry* 275, 10315-10322.
- Avram, D., Fields, A., Senawong, T., Topark-Ngarm, A., and Leid, M. (2002). COUP-TF (chicken ovalbumin upstream promoter transcription factor)-interacting protein 1 (CTIP1) is a sequence-specific DNA binding protein. *The Biochemical journal* 368, 555-563.
- Bae, K.H., Kwon, Y.D., Shin, H.C., Hwang, M.S., Ryu, E.H., Park, K.S., Yang, H.Y., Lee, D.K., Lee, Y., Park, J., *et al.* (2003). Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat Biotechnol* 21, 275-280.
- Bai, L., and Merchant, J.L. (2001). ZBP-89 promotes growth arrest through stabilization of p53. *Mol Cell Biol* 21, 4670-4683.
- Bairoch, A., and Apweiler, R. (1997). The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *Journal of molecular medicine (Berlin, Germany)* 75, 312-316.
- Baldwin, A.S., Jr., LeClair, K.P., Singh, H., and Sharp, P.A. (1990). A large protein containing zinc finger domains binds to related sequence elements in the enhancers of the class I major histocompatibility complex and kappa immunoglobulin genes. *Mol Cell Biol* 10, 1406-1414.
- Berg, J.M. (1992). Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. *Proceedings of the National Academy of Sciences of the United States of America* 89, 11109-11110.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic acids research* 28, 235-242.

REFERENCES - *continued*

- Bohm, S., Frishman, D., and Mewes, H.W. (1997). Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins. *Nucleic acids research* 25, 2464-2469.
- Bomar, M.G., Pai, M.T., Tzeng, S.R., Li, S.S., and Zhou, P. (2007). Structure of the ubiquitin-binding zinc finger domain of human DNA Y-polymerase eta. *EMBO reports* 8, 247-251.
- Bond, H.M., Mesuraca, M., Carbone, E., Bonelli, P., Agosti, V., Amodio, N., De Rosa, G., Di Nicola, M., Gianni, A.M., Moore, M.A., *et al.* (2004). Early hematopoietic zinc finger protein (EHZF), the human homolog to mouse Evi3, is highly expressed in primitive human hematopoietic cells. *Blood* 103, 2062-2070.
- Brayer, K.J., Kulshreshtha, S., and Segal, D.J. (2008). The Protein-Binding Potential of C2H2 Zinc Finger Domains. *Cell biochemistry and biophysics*.
- Brayer, K.J., and Segal, D.J. (2008). Keep Your Fingers Off My DNA: Protein-Protein Interactions Mediated by C2H2 Zinc Finger Domains. *Cell biochemistry and biophysics* 50, 111-131.
- Briggs, M.R., Kadonaga, J.T., Bell, S.P., and Tjian, R. (1986). Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science* (New York, NY) 234, 47-52.
- Brown, R.S. (2005). Zinc finger proteins: getting a grip on RNA. *Current opinion in structural biology* 15, 94-98.
- Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology* 212, 563-578.
- Carroll, D., Morton, J.J., Beumer, K.J., and Segal, D.J. (2006). Design, construction and in vitro testing of zinc finger nucleases. *Nature protocols* 1, 1329-1341.
- Chapman, N.R., and Perkins, N.D. (2000). Inhibition of the RelA(p65) NF-kappaB subunit by Egr-1. *The Journal of biological chemistry* 275, 4719-4725.
- Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic acids research* 35, D572-574.

REFERENCES - *continued*

- Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research* *31*, 3497-3500.
- Chernukhin, I.V., Shamsuddin, S., Robinson, A.F., Carne, A.F., Paul, A., El-Kady, A.I., Lobanenko, V.V., and Klenova, E.M. (2000). Physical and functional interaction between two pluripotent proteins, the Y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF. *The Journal of biological chemistry* *275*, 29915-29921.
- Chevallier, N., Corcoran, C.M., Lennon, C., Hyjek, E., Chadburn, A., Bardwell, V.J., Licht, J.D., and Melnick, A. (2004). ETO protein of t(8;21) AML is a corepressor for Bcl-6 B-cell lymphoma oncoprotein. *Blood* *103*, 1454-1463.
- Chothia, C., and Janin, J. (1975). Principles of protein-protein recognition. *Nature* *256*, 705-708.
- Christy, B., and Nathans, D. (1989). DNA binding site of the growth factor-inducible protein Zif268. *Proceedings of the National Academy of Sciences of the United States of America* *86*, 8737-8741.
- Clark, L., Matthews, J.R., and Hay, R.T. (1990). Interaction of enhancer-binding protein EBP1 (NF-kappa B) with the human immunodeficiency virus type 1 enhancer. *Journal of virology* *64*, 1335-1344.
- Clarke, N.D., and Berg, J.M. (1998). Zinc fingers in *Caenorhabditis elegans*: finding families and probing pathways. *Science (New York, NY)* *282*, 2018-2022.
- Colland, F., Jacq, X., Trouplin, V., Mouglin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J.M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome research* *14*, 1324-1332.
- Consortium, U. (2007). The Universal Protein Resource (UniProt). *Nucleic acids research* *35*, D193-197.
- Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome research* *14*, 1188-1190.
- Dahl, R., Iyer, S.R., Owens, K.S., Cuylear, D.D., and Simon, M.C. (2007). The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. *The Journal of biological chemistry* *282*, 6473-6483.

REFERENCES - *continued*

- Dai, H., Hogan, C., Gopalakrishnan, B., Torres-Vazquez, J., Nguyen, M., Park, S., Raftery, L.A., Warrior, R., and Arora, K. (2000). The zinc finger protein schnurri acts as a Smad partner in mediating the transcriptional response to decapentaplegic. *Developmental biology* 227, 373-387.
- Davies, M.A., Koul, D., Dhesi, H., Berman, R., McDonnell, T.J., McConkey, D., Yung, W.K., and Steck, P.A. (1999). Regulation of Akt/PKB activity, cellular growth, and apoptosis in prostate carcinoma cells by MMAC/PTEN. *Cancer research* 59, 2551-2556.
- Davies, R.C., Calvio, C., Bratt, E., Larsson, S.H., Lamond, A.I., and Hastie, N.D. (1998). WT1 interacts with the splicing factor U2AF65 in an isoform-dependent manner and can be incorporated into spliceosomes. *Genes & development* 12, 3217-3225.
- Del Rio, S., Menezes, S.R., and Setzer, D.R. (1993). The function of individual zinc fingers in sequence-specific DNA recognition by transcription factor IIIA. *Journal of molecular biology* 233, 567-579.
- Del Rio, S., and Setzer, D.R. (1993). The role of zinc fingers in transcriptional activation by transcription factor IIIA. *Proceedings of the National Academy of Sciences of the United States of America* 90, 168-172.
- Dennig, J., Hagen, G., Beato, M., and Suske, G. (1995). Members of the Sp transcription factor family control transcription from the uteroglobin promoter. *The Journal of biological chemistry* 270, 12737-12744.
- Dhordain, P., Albagli, O., Honore, N., Guidez, F., Lantoine, D., Schmid, M., The, H.D., Zelent, A., and Koken, M.H. (2000). Colocalization and heteromerization between the two human oncogene POZ/zinc finger proteins, LAZ3 (BCL6) and PLZF. *Oncogene* 19, 6240-6250.
- Donohoe, M.E., Zhang, L.F., Xu, N., Shi, Y., and Lee, J.T. (2007). Identification of a Ctfc cofactor, Yy1, for the X chromosome binary switch. *Molecular cell* 25, 43-56.
- Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D., and Barbas, C.F., 3rd (2001). Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *The Journal of biological chemistry* 276, 29466-29478.
- Eddy, S.R. (1996). Hidden Markov models. *Current opinion in structural biology* 6, 361-365.
- Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)* 14, 755-763.

REFERENCES - *continued*

- Eddy, S.R. (2003). HMMER User's Guide, version 2.3.3 (St. Louis, HHMI and Washington University).
- Fairall, L., Harrison, S.D., Travers, A.A., and Rhodes, D. (1992). Sequence-specific DNA binding by a two zinc-finger peptide from the *Drosophila melanogaster* Tramtrack protein. *Journal of molecular biology* 226, 349-366.
- 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.
- Fan, C.M., and Maniatis, T. (1990). A DNA-binding protein containing two widely separated zinc finger motifs that recognize the same DNA sequence. *Genes & development* 4, 29-42.
- Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., *et al.* (2005). Protein interaction mapping: a *Drosophila* case study. *Genome research* 15, 376-384.
- Fox, A.H., Liew, C., Holmes, M., Kowalski, K., Mackay, J., and Crossley, M. (1999). Transcriptional cofactors of the FOG family interact with GATA proteins by means of multiple zinc fingers. *The EMBO journal* 18, 2812-2822.
- 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 of the United States of America* 84, 4841-4845.
- Friesen, W.J., and Darby, M.K. (1997). Phage display of RNA binding zinc fingers from transcription factor IIIA. *The Journal of biological chemistry* 272, 10994-10997.
- 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.
- Gartel, A.L., Ye, X., Goufman, E., Shianov, P., Hay, N., Najmabadi, F., and Tyner, A.L. (2001). Myc represses the p21(WAF1/CIP1) promoter and interacts with Sp1/Sp3. *Proceedings of the National Academy of Sciences of the United States of America* 98, 4510-4515.
- Ghosh, I., Stains, C.I., Ooi, A.T., and Segal, D.J. (2006). Direct detection of double-stranded DNA: molecular methods and applications for DNA diagnostics. *Mol BioSystems* 2, 551-560.

REFERENCES - *continued*

- Giesecke, A.V., Fang, R., and Joung, J.K. (2006). Synthetic protein-protein interaction domains created by shuffling Cys2His2 zinc-fingers. *Molecular systems biology* 2, 2006-2011.
- Gonzalez-Billault, C., Engelke, M., Jimenez-Mateos, E.M., Wandosell, F., Caceres, A., and Avila, J. (2002). Participation of structural microtubule-associated proteins (MAPs) in the development of neuronal polarity. *Journal of neuroscience research* 67, 713-719.
- Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52, 696-704.
- Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids research* 33, W557-559.
- Hagen, G., Muller, S., Beato, M., and Suske, G. (1994). Sp1-mediated transcriptional activation is repressed by Sp3. *The EMBO journal* 13, 3843-3851.
- Hahm, K., Cobb, B.S., McCarty, A.S., Brown, K.E., Klug, C.A., Lee, R., Akashi, K., Weissman, I.L., Fisher, A.G., and Smale, S.T. (1998). Helios, a T cell-restricted Ikaros family member that quantitatively associates with Ikaros at centromeric heterochromatin. *Genes & development* 12, 782-796.
- Hall, T.M. (2005). Multiple modes of RNA recognition by zinc finger proteins. *Current opinion in structural biology* 15, 367-373.
- Hata, A., Seoane, J., Lagna, G., Montalvo, E., Hemmati-Brivanlou, A., and Massague, J. (2000). OAZ uses distinct DNA- and protein-binding zinc fingers in separate BMP-Smad and Olf signaling pathways. *Cell* 100, 229-240.
- Hennemann, H., Vassen, L., Geisen, C., Eilers, M., and Moroy, T. (2003). Identification of a novel Kruppel-associated box domain protein, Krim-1, that interacts with c-Myc and inhibits its oncogenic activity. *The Journal of biological chemistry* 278, 28799-28811.
- Hoffmann, A., Barz, T., and Spengler, D. (2006). Multitasking C2H2 zinc fingers link Zac DNA binding to coordinated regulation of p300-histone acetyltransferase activity. *Mol Cell Biol* 26, 5544-5557.
- Hoffmann, A., Ciani, E., Boeckardt, J., Holsboer, F., Journot, L., and Spengler, D. (2003). Transcriptional activities of the zinc finger protein Zac are differentially controlled by DNA binding. *Mol Cell Biol* 23, 988-1003.

REFERENCES - *continued*

- Honma, Y., Kiyosawa, H., Mori, T., Oguri, A., Nikaido, T., Kanazawa, K., Tojo, M., Takeda, J., Tanno, Y., Yokoya, S., *et al.* (1999). Eos: a novel member of the Ikaros gene family expressed predominantly in the developing nervous system. *FEBS letters* 447, 76-80.
- Houbaviy, H.B., Usheva, A., Shenk, T., and Burley, S.K. (1996). Cocystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13577-13582.
- Huang, R.P., Liu, C., Fan, Y., Mercola, D., and Adamson, E.D. (1995). Egr-1 negatively regulates human tumor cell growth via the DNA-binding domain. *Cancer research* 55, 5054-5062.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., and Sigrist, C.J. (2008). The 20 years of PROSITE. *Nucleic acids research* 36, D245-249.
- Ishiguro, A., Ideta, M., Mikoshiba, K., Chen, D.J., and Aruga, J. (2007). ZIC2-dependent transcriptional regulation is mediated by DNA-dependent protein kinase, poly(ADP-ribose) polymerase, and RNA helicase A. *The Journal of biological chemistry* 282, 9983-9995.
- Iuchi, S. (2001). Three classes of C2H2 zinc finger proteins. *Cell Mol Life Sci* 58, 625-635.
- Jamieson, A.C., Miller, J.C., and Pabo, C.O. (2003). Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov* 2, 361-368.
- Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *Journal of molecular biology* 204, 155-164.
- Janssen, H., and Marynen, P. (2006). Interaction partners for human ZNF384/CIZ/NMP4--zyxin as a mediator for p130CAS signaling? *Experimental cell research* 312, 1194-1204.
- Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13-20.
- Jones, S., and Thornton, J.M. (1997a). Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology* 272, 121-132.

REFERENCES - *continued*

- Jones, S., and Thornton, J.M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *Journal of molecular biology* 272, 133-143.
- Jung, D., Choi, Y., and Uesugi, M. (2006). Small organic molecules that modulate gene transcription. *Drug discovery today* 11, 452-457.
- Kadam, S., McAlpine, G.S., Phelan, M.L., Kingston, R.E., Jones, K.A., and Emerson, B.M. (2000). Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes & development* 14, 2441-2451.
- Kadonaga, J.T., Carner, K.R., Masiarz, F.R., and Tjian, R. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* 51, 1079-1090.
- Kadonaga, J.T., and Tjian, R. (1986). Affinity purification of sequence-specific DNA binding proteins. *Proceedings of the National Academy of Sciences of the United States of America* 83, 5889-5893.
- Kalenik, J.L., Chen, D., Bradley, M.E., Chen, S.J., and Lee, T.C. (1997). Yeast two-hybrid cloning of a novel zinc finger protein that interacts with the multifunctional transcription factor YY1. *Nucleic acids research* 25, 843-849.
- Karlseder, J., Rotheneder, H., and Wintersberger, E. (1996). Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. *Mol Cell Biol* 16, 1659-1667.
- Kelley, C.M., Ikeda, T., Koipally, J., Avitahl, N., Wu, L., Georgopoulos, K., and Morgan, B.A. (1998). Helios, a novel dimerization partner of Ikaros expressed in the earliest hematopoietic progenitors. *Curr Biol* 8, 508-515.
- Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., *et al.* (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic acids research* 33, D297-302.
- Kim, S.S., Pandey, K.K., Choi, H.S., Kim, S.Y., Law, P.Y., Wei, L.N., and Loh, H.H. (2005). Poly(C) binding protein family is a transcription factor in mu-opioid receptor gene expression. *Molecular pharmacology* 68, 729-736.
- Kinzler, K.W., and Vogelstein, B. (1990). The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Mol Cell Biol* 10, 634-642.
- Kirkman, T.W. (1996). Statistics to Use. In <http://www.physics.sjsu.edu/stats/>.

REFERENCES - *continued*

Korn, A.P., and Burnett, R.M. (1991). Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins* 9, 37-55.

Koyabu, Y., Nakata, K., Mizugishi, K., Aruga, J., and Mikoshiba, K. (2001). Physical and functional interactions between Zic and Gli proteins. *The Journal of biological chemistry* 276, 6889-6892.

Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003). Structural classification of zinc fingers: survey and summary. *Nucleic acids research* 31, 532-550.

Ku, M.C., Stewart, S., and Hata, A. (2003). Poly(ADP-ribose) polymerase 1 interacts with OAZ and regulates BMP-target genes. *Biochemical and biophysical research communications* 311, 702-707.

Kurisaki, K., Kurisaki, A., Valcourt, U., Terentiev, A.A., Pardali, K., Ten Dijke, P., Heldin, C.H., Ericsson, J., and Moustakas, A. (2003). Nuclear factor YY1 inhibits transforming growth factor beta- and bone morphogenetic protein-induced cell differentiation. *Mol Cell Biol* 23, 4494-4510.

Kwan, A.H., Czolij, R., Mackay, J.P., and Crossley, M. (2003). Pentaprobe: a comprehensive sequence for the one-step detection of DNA-binding activities. *Nucleic acids research* 31, e124.

Kwon, H.S., Kim, M.S., Edenberg, H.J., and Hur, M.W. (1999). Sp3 and Sp4 can repress transcription by competing with Sp1 for the core cis-elements on the human ADH5/FDH minimal promoter. *The Journal of biological chemistry* 274, 20-28.

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 Cys(2)-His(2) zinc fingers. *Journal of molecular biology* 295, 719-727.

Laity, J.H., Lee, B.M., and Wright, P.E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current opinion in structural biology* 11, 39-46.

Laricchia-Robbio, L., Fazzina, R., Li, D., Rinaldi, C.R., Sinha, K.K., Chakraborty, S., and Nucifora, G. (2006). Point mutations in two EVI1 Zn fingers abolish EVI1-GATA1 interaction and allow erythroid differentiation of murine bone marrow cells. *Mol Cell Biol* 26, 7658-7666.

Lee, B.M., Xu, J., Clarkson, B.K., Martinez-Yamout, M.A., Dyson, H.J., Case, D.A., Gottesfeld, J.M., and Wright, P.E. (2006). Induced fit and "lock and key" recognition of 5S RNA by zinc fingers of transcription factor IIIA. *Journal of molecular biology* 357, 275-291.

REFERENCES - *continued*

- Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature reviews* 8, 995-1005.
- Lee, J.A., Suh, D.C., Kang, J.E., Kim, M.H., Park, H., Lee, M.N., Kim, J.M., Jeon, B.N., Roh, H.E., Yu, M.Y., *et al.* (2005). Transcriptional activity of Sp1 is regulated by molecular interactions between the zinc finger DNA binding domain and the inhibitory domain with corepressors, and this interaction is modulated by MEK. *The Journal of biological chemistry* 280, 28061-28071.
- Lee, J.S., Galvin, K.M., and Shi, Y. (1993). Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1. *Proceedings of the National Academy of Sciences of the United States of America* 90, 6145-6149.
- Lee, J.S., See, R.H., Galvin, K.M., Wang, J., and Shi, Y. (1995). Functional interactions between YY1 and adenovirus E1A. *Nucleic acids research* 23, 925-931.
- 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 (New York, NY)* 245, 635-637.
- Lee, S.B., and Haber, D.A. (2001). Wilms tumor and the WT1 gene. *Experimental cell research* 264, 74-99.
- Lee, T.H., Lwu, S., Kim, J., and Pelletier, J. (2002). Inhibition of Wilms tumor 1 transactivation by bone marrow zinc finger 2, a novel transcriptional repressor. *The Journal of biological chemistry* 277, 44826-44837.
- Lemercier, C., Brocard, M.P., Puvion-Dutilleul, F., Kao, H.Y., Albagli, O., and Khochbin, S. (2002). Class II histone deacetylases are directly recruited by BCL6 transcriptional repressor. *The Journal of biological chemistry* 277, 22045-22052.
- Letondal, C. (2001). A Web interface generator for molecular biology programs in Unix. *Bioinformatics (Oxford, England)* 17, 73-82.
- Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., and Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic acids research* 30, 242-244.
- Lewis, B.A., Tullis, G., Seto, E., Horikoshi, N., Weinmann, R., and Shenk, T. (1995). Adenovirus E1A proteins interact with the cellular YY1 transcription factor. *Journal of virology* 69, 1628-1636.

REFERENCES - *continued*

- Li, N., Seetharam, S., and Seetharam, B. (1998). Characterization of the human transcobalamin II promoter. A proximal GC/GT box is a dominant negative element. *The Journal of biological chemistry* 273, 16104-16111.
- Li, S.H., Cheng, A.L., Zhou, H., Lam, S., Rao, M., Li, H., and Li, X.J. (2002). Interaction of Huntington disease protein with transcriptional activator Sp1. *Mol Cell Biol* 22, 1277-1287.
- Liew, C.K., Kowalski, K., Fox, A.H., Newton, A., Sharpe, B.K., Crossley, M., and Mackay, J.P. (2000). Solution structures of two CCHC zinc fingers from the FOG family protein U-shaped that mediate protein-protein interactions. *Structure* 8, 1157-1166.
- Liew, C.K., Simpson, R.J., Kwan, A.H., Crofts, L.A., Loughlin, F.E., Matthews, J.M., Crossley, M., and Mackay, J.P. (2005). Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. *Proceedings of the National Academy of Sciences of the United States of America* 102, 583-588.
- Lijnzaad, P., and Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins* 28, 333-343.
- Little, N.A., Hastie, N.D., and Davies, R.C. (2000). Identification of WTAP, a novel Wilms' tumour 1-associating protein. *Human molecular genetics* 9, 2231-2239.
- Liu, F., Massague, J., and Ruiz i Altaba, A. (1998). Carboxy-terminally truncated Gli3 proteins associate with Smads. *Nature genetics* 20, 325-326.
- Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of molecular biology* 285, 2177-2198.
- Lu, D., Searles, M.A., and Klug, A. (2003). Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature* 426, 96-100.
- Lu, Q.R., Sun, T., Zhu, Z., Ma, N., Garcia, M., Stiles, C.D., and Rowitch, D.H. (2002). Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. *Cell* 109, 75-86.
- Lu, Q.R., Yuk, D., Alberta, J.A., Zhu, Z., Pawlitzky, I., Chan, J., McMahon, A.P., Stiles, C.D., and Rowitch, D.H. (2000). Sonic hedgehog--regulated oligodendrocyte lineage genes encoding bHLH proteins in the mammalian central nervous system. *Neuron* 25, 317-329.
- Mackay, J.P., and Crossley, M. (1998). Zinc fingers are sticking together. *Trends in biochemical sciences* 23, 1-4.

REFERENCES - *continued*

- Maekawa, T., Sakura, H., Sudo, T., and Ishii, S. (1989). Putative metal finger structure of the human immunodeficiency virus type 1 enhancer binding protein HIV-EP1. *The Journal of biological chemistry* *264*, 14591-14593.
- Mahajan, M.A., Murray, A., and Samuels, H.H. (2002). NRC-interacting factor 1 is a novel cotransducer that interacts with and regulates the activity of the nuclear hormone receptor coactivator NRC. *Mol Cell Biol* *22*, 6883-6894.
- Martin, P.J., Delmotte, M.H., Formstecher, P., and Lefebvre, P. (2003). PLZF is a negative regulator of retinoic acid receptor transcriptional activity. *Nucl Recept* *1*, 6.
- Masclé, X., Albagli, O., and Lemercier, C. (2003). Point mutations in BCL6 DNA-binding domain reveal distinct roles for the six zinc fingers. *Biochemical and biophysical research communications* *300*, 391-396.
- Matthews, J.M., and Sunde, M. (2002). Zinc fingers--folds for many occasions. *IUBMB life* *54*, 351-355.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., *et al.* (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research* *31*, 374-378.
- McCarty, A.S., Kleiger, G., Eisenberg, D., and Smale, S.T. (2003). Selective dimerization of a C2H2 zinc finger subfamily. *Molecular cell* *11*, 459-470.
- McGhee, L., Bryan, J., Elliott, L., Grimes, H.L., Kazanjian, A., Davis, J.N., and Meyers, S. (2003). Gfi-1 attaches to the nuclear matrix, associates with ETO (MTG8) and histone deacetylase proteins, and represses transcription using a TSA-sensitive mechanism. *Journal of cellular biochemistry* *89*, 1005-1018.
- Mercer, K.B., Flaherty, D.B., Miller, R.K., Qadota, H., Tinley, T.L., Moerman, D.G., and Benian, G.M. (2003). *Caenorhabditis elegans* UNC-98, a C2H2 Zn finger protein, is a novel partner of UNC-97/PINCH in muscle adhesion complexes. *Molecular biology of the cell* *14*, 2492-2507.
- Merithew, E., Stone, C., Eathiraj, S., and Lambright, D.G. (2003). Determinants of Rab5 interaction with the N terminus of early endosome antigen 1. *The Journal of biological chemistry* *278*, 8494-8500.
- Michael, S.F., Kilfoil, V.J., Schmidt, M.H., Amann, B.T., and Berg, J.M. (1992). Metal binding and folding properties of a minimalist Cys2His2 zinc finger peptide. *Proceedings of the National Academy of Sciences of the United States of America* *89*, 4796-4800.

REFERENCES - *continued*

- Milbrandt, J. (1987). A nerve growth factor-induced gene encodes a possible transcriptional regulatory factor. *Science (New York, NY)* 238, 797-799.
- Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO journal* 4, 1609-1614.
- Molnar, A., and Georgopoulos, K. (1994). The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. *Mol Cell Biol* 14, 8292-8303.
- Momeni, P., Glockner, G., Schmidt, O., von Holtum, D., Albrecht, B., Gillesen-Kaesbach, G., Hennekam, R., Meinecke, P., Zabel, B., Rosenthal, A., *et al.* (2000). Mutations in a new gene, encoding a zinc-finger protein, cause tricho-rhino-phalangeal syndrome type I. *Nature genetics* 24, 71-74.
- Moore, F.L., Jaruzelska, J., Dorfman, D.M., and Reijo-Pera, R.A. (2004). Identification of a novel gene, DZIP (DAZ-interacting protein), that encodes a protein that interacts with DAZ (deleted in azoospermia) and is expressed in embryonic stem cells and germ cells. *Genomics* 83, 834-843.
- Morellet, N., Jullian, N., De Rocquigny, H., Maigret, B., Darlix, J.L., and Roques, B.P. (1992). Determination of the structure of the nucleocapsid protein NCp7 from the human immunodeficiency virus type 1 by 1H NMR. *The EMBO journal* 11, 3059-3065.
- Morgan, B., Sun, L., Avitahl, N., Andrikopoulos, K., Ikeda, T., Gonzales, E., Wu, P., Neben, S., and Georgopoulos, K. (1997). Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *The EMBO journal* 16, 2004-2013.
- Morii, E., Oboki, K., Kataoka, T.R., Igarashi, K., and Kitamura, Y. (2002). Interaction and cooperation of mi transcription factor (MITF) and myc-associated zinc-finger protein-related factor (MAZR) for transcription of mouse mast cell protease 6 gene. *The Journal of biological chemistry* 277, 8566-8571.
- Mount, D.W. (2001). *Bioinformatics : sequence and genome analysis* (Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press).
- Mueller, T.D., and Feigon, J. (2002). Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions. *Journal of molecular biology* 319, 1243-1255.

REFERENCES - *continued*

- Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., *et al.* (2007). New developments in the InterPro database. *Nucleic acids research* 35, D224-228.
- Murata, Y., Kim, H.G., Rogers, K.T., Udvardi, A.J., and Horowitz, J.M. (1994). Negative regulation of Sp1 trans-activation is correlated with the binding of cellular proteins to the amino terminus of the Sp1 trans-activation domain. *The Journal of biological chemistry* 269, 20674-20681.
- Mysliwiec, M.R., Kim, T.G., and Lee, Y. (2007). Characterization of zinc finger protein 496 that interacts with Jumonji/Jarid2. *FEBS letters* 581, 2633-2640.
- Nanba, D., Mammoto, A., Hashimoto, K., and Higashiyama, S. (2003). Proteolytic release of the carboxy-terminal fragment of proHB-EGF causes nuclear export of PLZF. *The Journal of cell biology* 163, 489-502.
- Nardelli, J., Gibson, T.J., Vesque, C., and Charnay, P. (1991). Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* 349, 175-178.
- Nariai, N., Kolaczyk, E.D., and Kasif, S. (2007). Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE* 2, e337.
- Nieto, J.M., Madrid, C., Miquelay, E., Parra, J.L., Rodriguez, S., and Juarez, A. (2002). Evidence for direct protein-protein interaction between members of the enterobacterial Hha/YmoA and H-NS families of proteins. *Journal of bacteriology* 184, 629-635.
- Nojiri, S., Joh, T., Miura, Y., Sakata, N., Nomura, T., Nakao, H., Sobue, S., Ohara, H., Asai, K., and Ito, M. (2004). ATBF1 enhances the suppression of STAT3 signaling by interaction with PIAS3. *Biochemical and biophysical research communications* 314, 97-103.
- 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. *Proceedings of the National Academy of Sciences of the United States of America* 95, 2938-2943.
- Numoto, M., Yokoro, K., and Koshi, J. (1999). ZF5, which is a Kruppel-type transcriptional repressor, requires the zinc finger domain for self-association. *Biochemical and biophysical research communications* 256, 573-578.
- Ogra, Y., Suzuki, K., Gong, P., Otsuka, F., and Koizumi, S. (2001). Negative regulatory role of Sp1 in metal responsive element-mediated transcriptional activation. *The Journal of biological chemistry* 276, 16534-16539.

REFERENCES - *continued*

- Omichinski, J.G., Clore, G.M., Robien, M., Sakaguchi, K., Appella, E., and Gronenborn, A.M. (1992). High-resolution solution structure of the double Cys2His2 zinc finger from the human enhancer binding protein MBP-1. *Biochemistry* *31*, 3907-3917.
- Omichinski, J.G., Pedone, P.V., Felsenfeld, G., Gronenborn, A.M., and Clore, G.M. (1997). The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nature structural biology* *4*, 122-132.
- Paces-Fessy, M., Boucher, D., Petit, E., Paute-Briand, S., and Blanchet-Tournier, M.F. (2004). The negative regulator of Gli, Suppressor of fused (Sufu), interacts with SAP18, Galectin3 and other nuclear proteins. *The Biochemical journal* *378*, 353-362.
- Page, D.C., Mosher, R., Simpson, E.M., Fisher, E.M., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A., and Brown, L.G. (1987). The sex-determining region of the human Y chromosome encodes a finger protein. *Cell* *51*, 1091-1104.
- Parraga, G., Horvath, S.J., Eisen, A., Taylor, W.E., Hood, L., Young, E.T., and Klevit, R.E. (1988). Zinc-dependent structure of a single-finger domain of yeast ADR1. *Science (New York, NY)* *241*, 1489-1492.
- Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science (New York, NY)* *252*, 809-817.
- Pavletich, N.P., and Pabo, C.O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science (New York, NY)* *261*, 1701-1707.
- Payre, F., Buono, P., Vanzo, N., and Vincent, A. (1997). Two types of zinc fingers are required for dimerization of the serendipity delta transcriptional activator. *Mol Cell Biol* *17*, 3137-3145.
- Pelham, H.R., and Brown, D.D. (1980). A specific transcription factor that can bind either the 5S RNA gene or 5S RNA. *Proceedings of the National Academy of Sciences of the United States of America* *77*, 4170-4174.
- Perdomo, J., Holmes, M., Chong, B., and Crossley, M. (2000). Eos and pegasus, two members of the Ikaros family of proteins with distinct DNA binding activities. *The Journal of biological chemistry* *275*, 38347-38354.
- Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjana, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., *et al.* (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research* *13*, 2363-2371.

REFERENCES - *continued*

- Phan, R.T., Saito, M., Basso, K., Niu, H., and Dalla-Favera, R. (2005). BCL6 interacts with the transcription factor Miz-1 to suppress the cyclin-dependent kinase inhibitor p21 and cell cycle arrest in germinal center B cells. *Nature immunology* 6, 1054-1060.
- Pommie, C., Levadoux, S., Sabatier, R., Lefranc, G., and Lefranc, M.P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* 17, 17-32.
- Porteus, M.H., and Carroll, D. (2005). Gene targeting using zinc finger nucleases. *Nat Biotechnol* 23, 967-973.
- Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* 13, 163.
- Rodgers, K.K., Bu, Z., Fleming, K.G., Schatz, D.G., Engelman, D.M., and Coleman, J.E. (1996). A zinc-binding domain involved in the dimerization of RAG1. *Journal of molecular biology* 260, 70-84.
- Rodriguez de la Vega, R.C., Merino, E., Becerril, B., and Possani, L.D. (2003). Novel interactions between K⁺ channels and scorpion toxins. *Trends in pharmacological sciences* 24, 222-227.
- Rohr, O., Lecestre, D., Chasserot-Golaz, S., Marban, C., Avram, D., Aunis, D., Leid, M., and Schaeffer, E. (2003). Recruitment of Tat to heterochromatin protein HP1 via interaction with CTIP2 inhibits human immunodeficiency virus type 1 replication in microglial cells. *Journal of virology* 77, 5415-5427.
- Rotheneder, H., Geymayer, S., and Haidweger, E. (1999). Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *Journal of molecular biology* 293, 1005-1015.
- Ryan, R.F., and Darby, M.K. (1998). The role of zinc finger linkers in p43 and TFIIIA binding to 5S rRNA and DNA. *Nucleic acids research* 26, 703-709.
- Samanta, J., Burke, G.M., McGuire, T., Pisarek, A.J., Mukhopadhyay, A., Mishina, Y., and Kessler, J.A. (2007). BMPR1a signaling determines numbers of oligodendrocytes and calbindin-expressing interneurons in the cortex. *J Neurosci* 27, 7397-7407.
- Sander, T.L., Stringer, K.F., Maki, J.L., Szauter, P., Stone, J.R., and Collins, T. (2003). The SCAN domain defines a large family of zinc finger transcription factors. *Gene* 310, 29-38.

REFERENCES - *continued*

- Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research* 18, 6097-6100.
- Segal, D.J., Beerli, R.R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksche, B., Lund, C.V., Magnenat, L., Valente, D., *et al.* (2003). Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* 42, 2137-2148.
- Seto, E., Lewis, B., and Shenk, T. (1993). Interaction between transcription factors Sp1 and YY1. *Nature* 365, 462-464.
- Shi, Y., and Berg, J.M. (1995). A direct comparison of the properties of natural and designed zinc-finger proteins. *Chemistry & biology* 2, 83-89.
- Shou, Y., Baron, S., and Poncz, M. (1998). An Sp1-binding silencer element is a critical negative regulator of the megakaryocyte-specific α IIb gene. *The Journal of biological chemistry* 273, 5716-5726.
- Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics* 3, 265-274.
- Simpson, R.J., Yi Lee, S.H., Bartle, N., Sum, E.Y., Visvader, J.E., Matthews, J.M., Mackay, J.P., and Crossley, M. (2004). A classic zinc finger from friend of GATA mediates an interaction with the coiled-coil of transforming acidic coiled-coil 3. *The Journal of biological chemistry* 279, 39789-39797.
- Sitterlin, D., Tiollais, P., and Transy, C. (1997). The RAR α -PLZF chimera associated with Acute Promyelocytic Leukemia has retained a sequence-specific DNA-binding domain. *Oncogene* 14, 1067-1074.
- Skrabanek, L., Saini, H.K., Bader, G.D., and Enright, A.J. (2008). Computational prediction of protein-protein interactions. *Molecular biotechnology* 38, 1-17.
- Smith, G.E., and Darling, D.S. (2003). Combination of a zinc finger and homeodomain required for protein-interaction. *Molecular biology reports* 30, 199-206.
- Song, C.Z., Keller, K., Murata, K., Asano, H., and Stamatoyannopoulos, G. (2002). Functional interaction between coactivators CBP/p300, PCAF, and transcription factor FKL2. *The Journal of biological chemistry* 277, 7029-7036.

REFERENCES - *continued*

- Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic acids research* 26, 320-322.
- Spengler, D., Villalba, M., Hoffmann, A., Pantaloni, C., Houssami, S., Bockaert, J., and Journot, L. (1997). Regulation of apoptosis and cell cycle arrest by Zac1, a novel zinc finger protein expressed in the pituitary gland and the brain. *The EMBO journal* 16, 2814-2825.
- Spraggon, L., Dudnakova, T., Slight, J., Lustig-Yariv, O., Cotterell, J., Hastie, N., and Miles, C. (2007). hnRNP-U directly interacts with WT1 and modulates WT1 transcriptional activation. *Oncogene* 26, 1484-1491.
- Starck, J., Cohet, N., Gonnet, C., Sarrazin, S., Doubeikovskaia, Z., Doubeikovski, A., Verger, A., Duterque-Coquillaud, M., and Morle, F. (2003). Functional cross-antagonism between transcription factors FLI-1 and EKLF. *Mol Cell Biol* 23, 1390-1402.
- Stites, W.E. (1997). Protein-protein interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. *Chem Rev* 97, 1233-1250.
- Stoll, R., Lee, B.M., Debler, E.W., Laity, J.H., Wilson, I.A., Dyson, H.J., and Wright, P.E. (2007). Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. *Journal of molecular biology* 372, 1227-1245.
- Sun, L., Liu, A., and Georgopoulos, K. (1996). Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. *The EMBO journal* 15, 5358-5369.
- Suzuki, T., Kimura, A., Nagai, R., and Horikoshi, M. (2000). Regulation of interaction of the acetyltransferase region of p300 and the DNA-binding domain of Sp1 on and through DNA binding. *Genes Cells* 5, 29-41.
- Suzuki, T., Muto, S., Miyamoto, S., Aizawa, K., Horikoshi, M., and Nagai, R. (2003). Functional interaction of the DNA-binding transcription factor Sp1 through its DNA-binding domain with the histone chaperone TAF-I. *The Journal of biological chemistry* 278, 28758-28764.
- Takamune, N., Misumi, S., and Shoji, S. (2000). Cyclic zinc-dithiocarbamate-S,S'-dioxide blocks CXCR4-mediated HIV-1 infection. *Biochemical and biophysical research communications* 272, 351-356.

REFERENCES - *continued*

- Tan, W., Zheng, L., Lee, W.H., and Boyer, T.G. (2004). Functional dissection of transcription factor ZBRK1 reveals zinc fingers with dual roles in DNA-binding and BRCA1-dependent transcriptional repression. *The Journal of biological chemistry* 279, 6576-6587.
- Theunissen, O., Rudt, F., and Pieler, T. (1998). Structural determinants in 5S RNA and TFIIIA for 7S RNP formation. *European journal of biochemistry / FEBS* 258, 758-767.
- Tsai, R.Y., and Reed, R.R. (1997). Cloning and functional characterization of Roaz, a zinc finger protein that interacts with O/E-1 to regulate gene expression: implications for olfactory neuronal development. *J Neurosci* 17, 4159-4169.
- Tsai, R.Y., and Reed, R.R. (1998). Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz. *Mol Cell Biol* 18, 6447-6456.
- Tsang, A.P., Visvader, J.E., Turner, C.A., Fujiwara, Y., Yu, C., Weiss, M.J., Crossley, M., and Orkin, S.H. (1997). FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* 90, 109-119.
- Tsuzuki, S., and Enver, T. (2002). Interactions of GATA-2 with the promyelocytic leukemia zinc finger (PLZF) protein, its homologue FAZF, and the t(11;17)-generated PLZF-retinoic acid receptor alpha oncoprotein. *Blood* 99, 3404-3410.
- van Crielinge, W., Cornelis, S., Van De Craen, M., Vandenabeele, P., Fiers, W., and Beyaert, R. (1999). GAL4 is a substrate for caspases: implications for two-hybrid screening and other GAL4-based assays. *Mol Cell Biol Res Commun* 1, 158-161.
- Varrault, A., Ciani, E., Apiou, F., Bilanges, B., Hoffmann, A., Pantaloni, C., Bockaert, J., Spengler, D., and Journot, L. (1998). hZAC encodes a zinc finger protein with antiproliferative properties and maps to a chromosomal region frequently lost in cancer. *Proceedings of the National Academy of Sciences of the United States of America* 95, 8835-8840.
- Vasanwala, F.H., Kusam, S., Toney, L.M., and Dent, A.L. (2002). Repression of AP-1 function: a mechanism for the regulation of Blimp-1 expression and B lymphocyte differentiation by the B cell lymphoma-6 protooncogene. *J Immunol* 169, 1922-1929.
- Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., *et al.* (1999). Analysis of human transcriptomes. *Nature genetics* 23, 387-388.

REFERENCES - *continued*

- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. *Science* (New York, NY) *291*, 1304-1351.
- Wang, J., Kudoh, J., Takayanagi, A., and Shimizu, N. (2005). Novel human BTB/POZ domain-containing zinc finger protein ZNF295 is directly associated with ZFP161. *Biochemical and biophysical research communications* *327*, 615-627.
- Wang, Z., Feng, L.S., Matskevich, V., Venkataraman, K., Parasuram, P., and Laity, J.H. (2006). Solution structure of a Zap1 zinc-responsive domain provides insights into metalloregulatory transcriptional repression in *Saccharomyces cerevisiae*. *Journal of molecular biology* *357*, 1167-1183.
- Weiss, M.A., and Keutmann, H.T. (1990). Alternating zinc finger motifs in the male-associated protein ZFY: defining architectural rules by mutagenesis and design of an "aromatic swap" second-site revertant. *Biochemistry* *29*, 9808-9813.
- Weisstein, E.W. (2008). Fisher's Exact Test. In *MathWorld - A Wolfram Web Resource* <http://mathworld.wolfram.com/FishersExactTest.html>, W.R. Inc., ed. (CRC Press LLC).
- Westman, B.J., Perdomo, J., Matthews, J.M., Crossley, M., and Mackay, J.P. (2004). Structural studies on a protein-binding zinc-finger domain of Eos reveal both similarities and differences to classical zinc fingers. *Biochemistry* *43*, 13318-13327.
- Westman, B.J., Perdomo, J., Sunde, M., Crossley, M., and Mackay, J.P. (2003). The C-terminal domain of Eos forms a high order complex in solution. *The Journal of biological chemistry* *278*, 42419-42426.
- Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution* *18*, 691-699.
- Wolf, I., and Rohrschneider, L.R. (1999). Fiz1, a novel zinc finger protein interacting with the receptor tyrosine kinase Flt3. *The Journal of biological chemistry* *274*, 21478-21484.
- Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* *29*, 183-212.
- Wright, D.A., Thibodeau-Beganny, S., Sander, J.D., Winfrey, R.J., Hirsh, A.S., Eichinger, M., Fu, F., Porteus, M.H., Dobbs, D., Voytas, D.F., *et al.* (2006). Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. *Nature protocols* *1*, 1637-1652.

REFERENCES - *continued*

- Wright, G.J., Leslie, J.D., Ariza-McNaughton, L., and Lewis, J. (2004). Delta proteins and MAGI proteins: an interaction of Notch ligands with intracellular scaffolding molecules and its significance for zebrafish development. *Development* (Cambridge, England) *131*, 5659-5669.
- 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.
- Xu, J., Paquet, M., Lau, A.G., Wood, J.D., Ross, C.A., and Hall, R.A. (2001). beta 1-adrenergic receptor association with the synaptic scaffolding protein membrane-associated guanylate kinase inverted-2 (MAGI-2). Differential regulation of receptor internalization by MAGI-2 and PSD-95. *The Journal of biological chemistry* *276*, 41310-41317.
- Young, L., Jernigan, R.L., and Covell, D.G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci* *3*, 717-729.
- Zaid, A., Hodny, Z., Li, R., and Nelson, B.D. (2001). Sp1 acts as a repressor of the human adenine nucleotide translocase-2 (ANT2) promoter. *European journal of biochemistry / FEBS* *268*, 5497-5503.
- Zhang, W., Wang, Y., Long, J., Girton, J., Johansen, J., and Johansen, K.M. (2003). A developmentally regulated splice variant from the complex lola locus encoding multiple different zinc finger domain proteins interacts with the chromosomal kinase JIL-1. *The Journal of biological chemistry* *278*, 11696-11704.
- Zhou, H.X., and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics* (Oxford, England) *23*, 2203-2209.
- Zhou, H.X., and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* *44*, 336-343.
- Zhou, Q., Gedrich, R.W., and Engel, D.A. (1995). Transcriptional repression of the c-fos gene by YY1 is mediated by a direct interaction with ATF/CREB. *Journal of virology* *69*, 4323-4330.