In vitro and in vivo high-throughput analysis of protein:DNA interactions

by

Seyed Hesam Shahravan

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Department of Chemistry
University of Toronto

© Copyright by S. Hesam Shahravan 2011
In vitro and in vivo high-throughput analysis of protein:DNA interactions

S. Hesam Shahravan

Doctor of Philosophy

Department of Chemistry
University of Toronto

2011

Abstract

In this thesis, emphasis has been placed on development of new approaches for high-throughput analysis of protein:DNA interactions in vitro and in vivo.

In vitro strategies for detection of protein:DNA interaction require isolation of active and soluble protein. However, current methodologies for purification of proteins often fail to provide high yield of pure and tag-free protein mainly because enzymatic cleavage reactions for tag removal do not exhibit stringent sequence specificity. Solving this problem is an important step towards high-throughput in vitro analysis of protein:DNA interactions. As a result, parts of this thesis are devoted to developing new approaches to enhance the specificity of a proteolysis reaction.

The first approach was through manipulation of experimental conditions to maximize the yield of the desired protein products from enterokinase proteolysis reactions of two His-tagged proteins. Because it was suspected that accessibility of the EK site was impeded, that is, a structural problem due to multimerization of proteins, focus was based on use of denaturants as a way to open the structure, thereby essentially increasing the stoichiometry of the canonical recognition site over noncanonical, adventitious sites.
Promoting accessibility of the canonical EK target site can increase proteolytic specificity and cleavage yield, and general strategies promoting a more open structure should be useful for preparation of proteins requiring endoprotease treatment. One such strategy for efficient EK proteolysis is proposed: by heterodimerizing with a separate leucine zipper, the bZIP basic region and amino-terminus can become more open and potentially more accessible to enterokinase.

*In vivo* strategies have the advantage over their *in vitro* counterparts of providing a native-like environment for assessing protein:DNA interactions, yet the most frequently used techniques often suffer from high false-positive and false-negative rates. In this thesis, a new bioprobe system for high-throughput detection of protein:DNA interactions *in vivo* is presented. This system offers higher levels of accuracy and sensitivity as well as accessibility and ease of manipulation in comparison with existing technologies.
In the name of God, the compassionate, the merciful
Acknowledgments

First and foremost, I thank Allah for helping me through every moment of my life. Next, I would like to thank my supervisor, Dr. Jumi A. Shin, for her support and enthusiasm for research, for inspiring me to keep learning and her constant encouragement to achieve higher goals. I also would like to express my special gratitude to I-San Chan for being a great mentor and for showing me the path when I first started, and to Lisa DenBoer and Gang Chen for teaching me invaluable lessons to succeed in graduate school. Thanks to Antonia De Jong, Dr. Sarmitha Sathiamoorthy, and Dr. Jack Iwanczyk for being great teammates and patient listeners to my questions and concerns. I also thank the previous members of Shin group: Melody Qu, Cindy Xu, Joanna Chow, Christopher Damaso, and Cassie Ho for their support and kindness.

Special thanks to the members of my Ph.D. supervisory committee, Dr. Ulrich Krull and Dr. Voula Kanelis, for providing priceless advice that helped me complete my degree, as well as Dr. Andrew Woolley for being a member of my examination committee and offering useful suggestions. I am very thankful to my external appraiser, Dr. Robert Campbell from University of Alberta, for his critical reading of this thesis and providing his comments and suggestions.

I dedicate this work to my family: Mehdi, Fatemeh, Hamed, and Hanieh for their truly outstanding love and thank them for their unconditional support through every stage of my life, and especially to Fatemeh, my best friend and wife, for always believing in me and for her continued love, patience and understanding through rough days.

S. Hesam Shahravan

December 2010
Preface

Parts of this thesis have been organized as a series of manuscripts that are either previously published in peer-reviewed journals (Chapter 2) or submitted for publication (Chapter 4). Therefore, each publication or submitted manuscript has been maintained with only minor modifications (e.g. re-numbering figures). As a result, repetition of some introductory and experimental information was unavoidable. Due to copyright constraints, additional material relevant to a chapter is provided in an Appendix as noted. For each publication, the authors’ contribution is provided on the first page of each corresponding chapter. All bibliographic information is provided in the References section.

Chapter One – In vitro and in vivo high-throughput analysis of protein:DNA interactions – An introduction

Chapter Two – Enhancing the specificity of the enterokinase cleavage reaction to promote efficient cleavage of a fusion tag.

Published in Protein Expression and Purification 59 (2008) 314–319
Authors: S. Hesam Shahravan, Xuanlu Qu, I-San Chan, and Jumi A. Shin

Chapter Three – Efficient cleavage of a fusion tag: an alternative approach using a designed protein-protein interaction to promote a structure more accessible to proteolysis

Authors: S. Hesam Shahravan, I-San Chan, Anna V. Fedorova, and Jumi A. Shin

Chapter Four – A fluorescent protein-based bioprobe for detection of protein:DNA recognition in vivo.

Authors: S. Hesam Shahravan, Isaac T.S. Li, Kevin Truong, and Jumi A. Shin

Chapter Five – Summary and future directions
Other publications submitted and/or accepted and published during M.Sc. and Ph.D.:


Table of Contents

Abstract ....................................................................................................................................... ii
Acknowledgments ....................................................................................................................... v
Preface ........................................................................................................................................ vi
Table of Contents ..................................................................................................................... viii
List of Tables ............................................................................................................................. xii
List of Figures .......................................................................................................................... xiii
List of Abbreviations ................................................................................................................ xv
List of Appendices .................................................................................................................. xvii

1. Chapter One ........................................................................................................................ 1
   1.1 Transcription Factors ...................................................................................................... 2
       1.1.1 Basic region/leucine zipper transcription factor family ........................................... 3
       1.1.2 Basic-region/helix-loop-helix/leucine zipper transcription factor family ............... 5
   1.2 Analysis of protein:DNA interactions .......................................................................... 8
       1.2.1 In vitro strategies for detection of protein:DNA interactions ................................... 8
       1.2.2 In vivo strategies for detection of protein:DNA interactions .................................. 14
   1.3 Thesis objectives ........................................................................................................... 20
   1.4 Thesis organization ......................................................................................................... 21

2. Chapter Two ...................................................................................................................... 22
   2.1 Abstract .......................................................................................................................... 23
   2.2 Introduction .................................................................................................................... 24
   2.3 Materials and Methods ................................................................................................ 26
2.3.1 Preparation of AhR6–C/EBP protein ................................................................. 26
2.3.2 Enterokinase cleavage of AhR6–C/EBP (typical EK reaction) ....................... 27
2.3.3 Isolation and identification of cleavage products by HPLC and ESI-MS .... 28

2.4 Results and Discussion ....................................................................................... 29
2.4.1 Changing pH, temperature, or the amount of EK had no effect on cleavage pattern . ......................................................................................................................... 30
2.4.2 Nonspecific cleavage due to inaccessibility of the target site ..................... 32
2.4.3 Addition of denaturant reduces adventitious enterokinase cleavage of AhR6–
C/EBP ......................................................................................................................... 34
2.4.4 Relative affinities of EK for its cleavage sites in AhR6–C/EBP ....................... 35
2.4.5 Modification of the EK cleavage site as an alternative approach to enhance
 cleavage efficiency and specificity is protein-specific ......................................... 37

3. Chapter Three ...................................................................................................... 41
3.1 Abstract ............................................................................................................. 42
3.2 Introduction ....................................................................................................... 43
3.3 Materials and Methods .................................................................................... 45
3.3.1 Preparation of AhR6–Jun protein ................................................................. 45
3.3.2 Preparation of Fos zipper protein ................................................................ 46
3.3.3 Enterokinase cleavage of AhR6–Jun (typical EK reaction) ....................... 46
3.4 Results and Discussion ....................................................................................... 46
3.4.1 Inefficient proteolysis of AhR6-Jun ............................................................... 48
3.4.2 Addition of denaturant reduces adventitious enterokinase cleavage of AhR6–Jun 49
3.4.3 Addition of Fos zipper inhibits EK activity .................................................. 51

4. Chapter Four ...................................................................................................... 54
4.1 Abstract ............................................................................................................. 55
5.2.2 Use of bioprobe to make a new biological discovery ................................................. 90

References .................................................................................................................................. 93

Appendix A ................................................................................................................................ 108

Appendix B ................................................................................................................................ 115

B.1 Enterokinase cleavage of AhR6–C/EBP (typical EK reaction) ................................. 116

B.2 Isolation and identification of cleavage products by HPLC and ESI-MS ................. 116

Appendix C ................................................................................................................................ 117

C.1 Preparation of 6×DNATarget bioprobes ................................................................. 118

C.2 Preparation of 1×DNATarget/NS bioprobes ......................................................... 119

C.3 Expression of the bioprobe ..................................................................................... 119

C.4 Fluorescence measurement ....................................................................................... 121

C.5 SDS-PAGE analysis of the bioprobe ................................................................. 121

C.6 Screening of DNA binders ...................................................................................... 122

C.7 Fluorescence resonance energy transfer ............................................................. 124

C.7.1 FRET in studies of biological molecules ........................................................ 126
List of Tables

Table 1.1 | Commonly used affinity tags for purification of recombinant proteins ....................... 13

Table 1.2 | Various cleavage agents for affinity tag removal .............................................................. 14

Table 4.1 | Summary of FRET ratio changes for various protein:DNA complexes .................. 65

Supplementary Table 4.1 | Reverse primers used for preparation of base cassette plasmids of 6× bioprobes ..................................................................................................................................... 82

Supplementary Table 4.2 | Reverse primers used for preparation of base cassette plasmids of 1× bioprobes ..................................................................................................................................... 82

Table C.1 | Experimental procedure for expression of the bioprobe ............................................. 120
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Crystal structure of the GCN4 bZIP homodimer bound to the AP-1 DNA site</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>Crystal structures of bHLH protein MyoD and bHLHZ protein Max</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Schematic representation of yeast reporter assays</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>Amino acid sequence of 6×His-tagged AhR6–C/EBP</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Time-course analysis of cleavage products after incubation of AhR6–C/EBP with EK</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>EK cleavage reactions of AhR6–C/EBP with different reagents and under different experimental conditions</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>EK cleavage reactions of AhR6–C/EBP under varying urea concentrations</td>
<td>33</td>
</tr>
<tr>
<td>2.5</td>
<td>HPLC analysis of a typical EK reaction of AhR6–C/EBP</td>
<td>35</td>
</tr>
<tr>
<td>2.6</td>
<td>Time-course analysis of cleavage products after incubation of AhR6–C/EBP with EK with 3 M urea</td>
<td>36</td>
</tr>
<tr>
<td>3.1</td>
<td>An alternative strategy to promote a more open structure for efficient proteolysis of the His tag</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>Amino acid sequence of 6×His-tagged AhR6–Jun</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Time-course analysis of cleavage products after incubation of AhR6–Jun with EK</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>EK cleavage reactions of AhR6–Jun under varying urea concentrations</td>
<td>50</td>
</tr>
<tr>
<td>3.5</td>
<td>Time-course analysis of cleavage products after incubation of AhR6–Jun with EK with 2 M urea</td>
<td>50</td>
</tr>
<tr>
<td>3.6</td>
<td>EK cleavage reactions of AhR6–Jun with varying concentrations of Fos zipper and under different experimental conditions</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>The bioprobe system for detection of protein:DNA recognition <em>in vivo</em></td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Specificity of the bioprobe</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>SDS-PAGE analysis of transcriptional blockage by protein:DNA recognition</td>
<td>67</td>
</tr>
</tbody>
</table>
Figure 4.4 | Mock library screening of two transcription factors .................................................. 72

Supplementary Figure 4.1 | The effect of linker length between CFP and YFP on transcription of the bioprobe reporter gene .............................................................................................................. 83

Supplementary Figure 4.2 | The effect of protein expression in the cell and the expressed bioprobe linker on the bioprobe’s range of detection ........................................................................................................... 84

Supplementary Figure 4.3 | Generation of the CFP-6×AP1-YFP bioprobe ........................................ 85

Figure C.1 | Identification of various colonies after DNA binder screening ................................. 123
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AhR</td>
<td>aryl hydrocarbon receptor</td>
</tr>
<tr>
<td>bHLH</td>
<td>basic-region/helix–loop–helix</td>
</tr>
<tr>
<td>bHLHZ</td>
<td>basic-region/helix-loop-helix/leucine zipper</td>
</tr>
<tr>
<td>bZIP</td>
<td>basic region/leucine zipper</td>
</tr>
<tr>
<td>C/EBP</td>
<td>CCAAT/enhancer binding protein</td>
</tr>
<tr>
<td>CFP</td>
<td>cyan fluorescent protein</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>Dam</td>
<td>DNA adenine methyltransferase</td>
</tr>
<tr>
<td>DamID</td>
<td>DNA adenine methyltransferase identification</td>
</tr>
<tr>
<td>DMS</td>
<td>dimethylsulfate</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>E-box</td>
<td>enhancer box</td>
</tr>
<tr>
<td>EK</td>
<td>enterokinase</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>ESI-MS</td>
<td>electrospray ionization mass spectrometry</td>
</tr>
<tr>
<td>FACS</td>
<td>fluorescence activated cell sorting</td>
</tr>
<tr>
<td>FP</td>
<td>fluorescent protein</td>
</tr>
<tr>
<td>FRET</td>
<td>fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>GFP</td>
<td>green fluorescent protein</td>
</tr>
<tr>
<td>GST</td>
<td>glutathione-S-transferase</td>
</tr>
<tr>
<td>GST–CaM</td>
<td>glutathione S-transferase-calmodulin</td>
</tr>
<tr>
<td>hIL-2</td>
<td>human interleukin-2</td>
</tr>
<tr>
<td>HPLC</td>
<td>high-performance liquid chromatography</td>
</tr>
<tr>
<td>HTH</td>
<td>helix–turn–helix</td>
</tr>
<tr>
<td>IMAC</td>
<td>Immobilized metal ion affinity chromatography</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropylthiogalactoside</td>
</tr>
<tr>
<td>$K_d$</td>
<td>dissociation constant</td>
</tr>
<tr>
<td>LB</td>
<td>luria-bertani</td>
</tr>
<tr>
<td>LM-PCR</td>
<td>ligation-mediated polymerase chain reaction</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Max</td>
<td>myelocytomatosis oncogene (Myc)-associated factor X</td>
</tr>
<tr>
<td>MBP</td>
<td>maltose-binding protein</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MW</td>
<td>molecular-weight marker</td>
</tr>
<tr>
<td>Myc</td>
<td>myelocytomatosis oncogene</td>
</tr>
<tr>
<td>NS</td>
<td>nonspecific</td>
</tr>
<tr>
<td>NT-proCNP</td>
<td>N-terminal proCNP</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PMSF</td>
<td>phenylmethanesulfonyl fluoride</td>
</tr>
<tr>
<td>RP-HPLC</td>
<td>reversed-phase high performance liquid chromatography</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>Y1H</td>
<td>yeast one-hybrid</td>
</tr>
<tr>
<td>Y2H</td>
<td>yeast two-hybrid</td>
</tr>
<tr>
<td>YFP</td>
<td>yellow fluorescent protein</td>
</tr>
</tbody>
</table>
List of Appendices

Appendix A – Re-print license from Elsevier ................................................................. 108

Appendix B – Detailed description of the Methods section for project presented in chapter 2 .... 115

Appendix C – Detailed description of the Methods section for the project presented in chapter 4 and some background information on the theory of fluorescence resonance energy transfer .................. 117
Chapter One

*In vitro* and *in vivo* high-throughput analysis of protein:DNA interactions – An introduction

S. Hesam Shahravan
1.1 Transcription Factors

Transcription factors (TFs) are DNA-binding proteins responsible for transcriptional regulation of gene expression. It has been reported that the 20,000-25,000 protein-coding genes of the human genome are believed to be regulated by transcription factors [1, 2]. TFs regulate many cellular processes, such as replication, development, and DNA repair [3]. Deregulation of the transcriptional activity of numerous genes has been associated with various human diseases. For instance, in cancer, abnormal patterns of gene expression are often caused by anomalous expression of a specific transcriptional activator or repressor [4].

Transcription factors typically comprise a DNA-binding domain that is responsible for directing the protein to a specific DNA target site, and an effector domain that regulates activation or repression of targeted genes. TFs function by binding to specific DNA sequences, typically positioned upstream of the gene promoter region, and regulate gene expression through their protein interactions with global regulatory complexes such as RNA polymerase II and other cofactors present near the gene promoter region [5].

TFs are grouped into families based on their sequences and structural similarities in their DNA-binding domains. Hence, members of each family use homologous structural motifs for DNA recognition [6, 7]. Such families include the basic region/leucine zipper (bZIP) [8-10], helix–turn–helix (HTH) [11], zinc-finger [12], and basic-region/helix–loop–helix (bHLH) [13] classes of proteins.

Two classes of transcription factors are the focus of this thesis: the bZIP family and the basic-region/helix-loop-helix/leucine zipper (bHLHZ) family, which is a member of the bHLH superfamily. Both families utilize a dimer of α-helices to bind specific DNA sequences in the
DNA major groove [14]. The following is a brief introduction to these two classes of transcription factors.

1.1.1 Basic region/leucine zipper transcription factor family

The bZIP motif was first discovered as a conserved sequence pattern in several eukaryotic transcription factors [9], and it is now evident that bZIP proteins comprise a large family of DNA-binding proteins in a variety of species. They are responsible for regulation of a wide range of cellular processes including cell development, differentiation, proliferation, and apoptosis [6, 15].

The bZIP proteins, which can form hetero- and/or homodimers, typically contain 60-80 residues [9, 16, 17] and have two separate subdomains within their DNA-binding domain: a basic region and a leucine zipper region. The basic region (~30 residues) is rich in arginines and lysines, but also contains other residues that are conserved throughout the bZIP family [6]. Leucine zipper sequences are characterized by a heptad repeat of leucine residues (and may include other hydrophobic residues like Met, Val or Ile) over a region of ~35 amino acids within the DNA-binding domain of the bZIP [9]. Swapping basic regions and zipper regions from bZIP proteins has confirmed that the basic region is primarily responsible for sequence-specific DNA-binding activity, and that the leucine zipper is responsible for dimerization specificity [18].

The bZIP motif comprises a small and simple protein structure that recognizes DNA sites with high affinity and specificity [17]. The crystal structures of several bZIP proteins bound to their cognate DNA sequences have been obtained and provide a high-resolution structure of the protein:DNA recognition complex [19-22].
The bZIP proteins bind specifically to DNA as dimers [23]. The crystal structure of the GCN4 bZIP bound to its cognate binding site AP-1 (5′-TGACTCA, Fig. 1.1) shows that the entire structure is a dimer of uninterrupted bipartite α-helices. The N-terminal basic region in each monomer interacts with the DNA site in a sequence-specific manner with each basic region contacting a half-site in the DNA major groove [9, 14, 23]. The basic region of bZIP proteins is partially disordered and undergoes a coil-to-helix transition to form an α-helical extension of the leucine zipper upon sequence-specific DNA binding [24-26]. The C-terminal leucine zipper region dimerizes via an amphipathic coiled-coil structure comprising two parallel α-helices that position the basic regions for targeting the DNA major groove [27].

Figure 1.1 | Crystal structure of the GCN4 bZIP homodimer bound to the AP-1 DNA site. The double-stranded DNA (gray) is at the bottom of the structure. The basic regions of the bZIP α-helices (light gray) and leucine zippers (dark gray) are labeled. The basic region contacts the DNA major groove and is
primarily responsible for the sequence preferences of the bZIP. The leucine zipper mediates dimerization
stability and specificity. The leucine zippers from the two monomers pack together in a coiled-coil
structure (PDB ID: 1YSA [19]).

In addition to homodimerization, bZIP proteins can form heterodimers as well, and
specific "master" bZIP regulators control the cellular activities of their dimerization partners [6].
Heterodimer formation allows for different combination of activation and/or repression domains
and hence can alter the regulatory properties of a molecule bound at a fixed DNA target. For
example, heterodimers can limit DNA-binding activity [28], and may also acquire new DNA-
binding specificities and hence be targeted to sites different from homodimers [29].

1.1.2 Basic-region/helix-loop-helix/leucine zipper transcription factor family

The bHLHZ motif is present in a large number of biologically important proteins. Members of
the bHLHZ family comprise multiprotein complexes by forming a variety of homo- and
heterodimers with diverse biological functions [30]. The bHLHZ family belongs to the larger
bHLH superfamily of transcription factors due to their homologous structures.

First observed by Murre and colleagues in two murine transcription factors [31], bHLH
proteins are found in various organisms from yeast to human and are involved in critical
developmental processes, including sex determination, development of the nervous system and
muscles, and retinal neurogenesis [32-35]. Similar to bZIP proteins, members of the bHLH
superfamily have two highly conserved and functionally different domains, which together
comprise a region of ~60 amino acid residues: a highly conserved DNA-binding basic region and
a dimerization domain. The basic region domain binds the transcription factor to DNA at a
consensus hexanucleotide sequence known as Enhancer box (E-box, 5'-CANNTG), while the
The dimerization domain facilitates interactions with other protein subunits to form homo- and heterodimeric complexes [13, 32]. Homo- and heterodimer formation expands the number of DNA sequences that bHLH proteins can recognize and bind.

The bHLH proteins employ a similar mode of DNA-binding as do bZIP proteins; the only major difference lies in their dimerization domain. In contrast to the leucine zipper domain in bZIP proteins, the dimerization domain in bHLH proteins comprises an antiparallel four-helix bundle; the two basic regions extend from the N-terminus of the HLH subdomain, which not only serves a dimerization function but also optimally positions the DNA-binding basic regions over their target site (Fig. 1.2a). Therefore, both the basic region and HLH dimerization subdomain are essential for formation of functional DNA binding complexes [30, 35]. The loop between the two helices of the four-helix bundle is 5-12 amino acids in length, and the sequence identity of the loop is not conserved among bHLH proteins. Thus, the loop appears to give bHLH proteins flexibility in structure and function.

The bHLHZ family differs from the bHLH superfamily of transcription factors in that in addition to the HLH, it contains a leucine zipper coiled-coil dimerization domain adjacent to the C-terminus of Helix 2, forming a long uninterrupted α-helix in each monomer (Fig. 1.2b). The leucine zippers form a parallel coiled-coil similar to the leucine zipper in the bZIP motif and govern dimerization specificity [30, 36, 37]. The leucine zippers add considerable buried surface area to the dimerization interface, giving stability and specificity to bHLHZ dimer interactions [38]. This is in contrast to bHLH proteins where the HLH domain determines dimerization specificity, also signifying a structural difference between the hydrophobic interface of the HLH domains of bHLH and bHLHZ proteins [39].
Figure 1.2 | Crystal structures of bHLH protein MyoD (a) and bHLHZ protein Max (myelocytomatosis oncogene (Myc)-associated factor X, b) homodimers bound to their DNA target sites. MyoD is a protein that induces skeletal muscle differentiation in cells [40, 41], whereas Max is a member of the Myc/Max/Mad network of transcription factors involved in cell proliferation and differentiation [42]. Both MyoD and Max form a four-helix bundle with Helix 1 and Helix 2 from each monomer. This bundle allows the basic region adjacent with Helix 1 to contact both sides of the DNA major groove. In addition, Max contains an additional dimerization subdomain, the leucine zipper (dark gray), which forms a long uninterrupted α-helix with Helix 2 in each monomer. Similar to bZIP proteins, the partially disordered basic regions of both bHLH and bHLHZ proteins undergo a coil-to-helix transition upon DNA binding (a: PDB ID: 1MDY [43] and b: PDB ID: 1AN2 [30]).
1.2 Analysis of protein:DNA interactions

Proteins control virtually every function of the cell, including signal transduction, genomic information storage and processing, and cellular trafficking. Knowledge of protein science is critical for understanding protein:DNA recognition, for example, transcription and gene regulation [44], and therefore, many research groups have focused their efforts on transcription factors and design of transcription factor-like proteins.

Understanding the interactions between transcription factors and their DNA response elements contributes toward learning how various genes can be regulated. However, only a small percentage of transcription factors have been characterized to date [45]. Hence, there is a great demand for identifying and studying new transcription factors as well as their newly designed synthetic partners.

Protein:DNA interactions can be explored by various in vitro and in vivo strategies, which present different advantages and disadvantages [1, 3, 45]. The following provides a general review of some commonly used techniques.

1.2.1 In vitro strategies for detection of protein:DNA interactions

Numerous methods have been used to characterize protein:DNA interactions in vitro, including electrophoretic mobility shift assay (EMSA) [46, 47], southwestern blotting [48], DNase I and exonuclease III footprinting [49-51], spectroscopic assays such as fluorescence anisotropy [52, 53], and ultraviolet (UV) cross-linking [54] (see the corresponding reference for description of each assay).
In a typical EMSA experiment, the mobility of a protein:DNA complex during electrophoresis on a native polyacrylamide gel is retarded compared to that of free DNA, which moves through the gel via reptation, like a snake. This shift in mobility of the protein:DNA complex is primarily due to the rigidification of DNA where bound by protein, and also the increase in size and mass of the complex moving through the gel. This technique allows determination of dissociation constants ($K_d$) of protein:DNA complexes. In southwestern blotting, a mixture of proteins is first electrophoretically separated on a denaturing polyacrylamide gel based on differences in their molecular weights. Resolved proteins are transferred/blotted onto a nitrocellulose membrane where they are renatured and incubated with a labeled DNA probe.

Footprinting assays allow identification of DNA sites bound by a specific protein. In DNase I footprinting, a radiolabeled DNA fragment is incubated with protein, followed by partial digestion of the DNA fragment with DNase I, an endonuclease that targets the DNA minor groove and cleaves its phosphate backbone nonspecifically. Bound proteins protect their DNA target site from cleavage and create a footprint region in the cleavage ladder that is separated by electrophoresis and compared with a cleavage ladder control of the same labeled DNA fragment in the absence of protein. Exonuclease III footprinting follows a similar concept, although the digestive enzyme binds to a free 3’ hydroxyl on double-stranded DNA and cleaves toward the 5’ direction in a semi-processive manner [1]. In a typical experiment, the enzyme is added to a mixture of protein and DNA, and starts removing mononucleotides from the 3’ termini of DNA fragments before it is blocked and released by a DNA-bound protein. The resulting DNA fragments, which are shortened by the enzyme, are then compared by
electrophoresis with the same DNA fragment digested by exonuclease III in the absence of protein.

Fluorescence anisotropy measurements are based on photoselective excitation of fluorophores by polarized light, and anisotropy is a measure of the degree to which a fluorophore moves/rotates during its excited-state lifetime. Upon binding to protein, the tumbling motion of fluorophore-labeled DNA changes and becomes more anisotropic. Similar to EMSA, this technique allows for determination of dissociation constants of protein:DNA complexes. Ultraviolet cross-linking involves UV irradiation of a protein:DNA complex; this causes formation of covalent linkages between nucleic acids and protein residues that are in close proximity. Molecular weights of complexes are revealed by denaturing polyacrylamide gel analysis. Therefore, both southwestern blotting and UV cross-linking can be used to determine the molecular weight of a specific protein:DNA complex.

1.2.1.1 Protein purification for analysis of protein:DNA interaction in vitro

In vitro methods for detection of protein:DNA recognition require isolation of active, soluble protein, which can be challenging. Reconstitution of individual proteins is also not feasible for large library work, and despite the great advances in recombinant DNA technology, an increasing challenge has always been to develop new and reliable methods to purify recombinant proteins [55]. In light of the scope of this thesis, more details regarding purification of proteins for use in in vitro studies of protein:DNA recognition follows.

In the late 1980’s, the most widely used method for characterization or purification of proteins, either natural or synthetic, was reversed-phase high performance liquid chromatography (RP-HPLC). In addition but to a lesser extent, ion exchange and gel permeation chromatography
were used either separately or along with RP-HPLC [56]. However, it was later realized that none of these separation methods could serve as a basis for purification of an ever-increasing number of biological proteins or their synthetic counterparts and their different analogs.

Immobilized metal ion affinity chromatography (IMAC), which was introduced as an alternative method for protein purification [56-60], is a cost-effective and high-throughput procedure. In recent years, it has been developed as one of the most powerful tools for single-step purification of recombinant proteins [57, 59]. IMAC was primarily developed as a group-specific affinity technique for purification of proteins by Porath and co-workers in 1975 [61]. The basic principle is that various amino acids (e.g. histidine, cysteine, and tryptophan) available on the surface of proteins can act as affinity tags by donating their electrons through their side chains to form stable complexes with certain transition metal ions (i.e. \( \text{Ni}^{2+}, \text{Cu}^{2+}, \text{Zn}^{2+}, \text{Co}^{2+} \)), resulting in separation of tagged proteins from a mixture [62-64].

Protein purification using IMAC offers various valuable features that have driven its rapid growth. Transition metal ions, which act as ligands for the affinity separations, offer more advantages than their biological affinity agents (i.e. antibodies and inhibitors) [65]. This is because small and relatively inexpensive metal ions are more stable under a wider range of conditions for protein purification. Not only can these metal ions be recycled many times without loss of activity, but they also can be incorporated into high-capacity chromatographic support. Elution can be done under relatively mild conditions, and cleaning and regenerating of the column are done easily without reducing protein binding capacity [66]. Additionally, selectivity in separation can be customized by modifying the target protein and through the choice of metal ions and solvent conditions [57, 66].
1.2.1.2 The Importance of removing affinity tags

The use of genetically engineered affinity tags is the most common means of achieving high-throughput protein purification with IMAC (see Table 1.1 for a list of common affinity tags). However, all affinity tags, regardless of size, have the potential to interfere with the biological activity of proteins into which they are incorporated [67]. Tags attached to purified proteins can hinder their crystallization by changing the protein conformation, or else can influence their behavior by altering protein’s biological activity [68]. As a result, it is usually desirable and sometimes essential to remove these tags [67, 68].

Affinity tags can be removed from expressed fusion partners by two different methods (Table 1.2). An affinity tag can be removed by harsh chemical treatment with cyanogen bromide or hydroxylamine [68, 69]. For this purpose, a specific residue must be placed at the junction between the fusion partner and the protein of interest. However, chemical methods are rather nonspecific and may lead to protein denaturation and side chain alteration of residues in the target protein [70, 71].

Alternatively, affinity tags can be cleaved from their fusion partners under milder conditions by enzymes called endopeptidases or endoproteases [68]. Several endoproteases have been developed for tag removal, which present different advantages and disadvantages (Table 1.2). Enterokinase, thrombin and factor Xa are among the most widely used enzymes for this purpose [68]. However, there are drawbacks in removal of tags using endoproteases. One drawback is the need for high ratios of endoprotease to protein, and another is occasionally the long incubation time necessary for tag removal. Under these conditions, endoprotease cleavage often leads to nonspecific cleavage of the protein at secondary sites, different from the
introduced canonical cleavage site [72]. Moreover, endoprotease cleavage can sometimes be a low-yield reaction, often due to inhibition of cleavage by properly folded proteins rendering the cleavage site inaccessible [73].

An alternative approach to cleavage by endoproteases is the use of exopeptidases, whether aminopeptidases or carboxypeptidases [68, 72, 73]. Exopeptidases are natural and derive

---

**Table 1.1** | Commonly used affinity tags for purification of recombinant proteins

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Location</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>His tag</td>
<td>6, 8 or 10 His</td>
<td>N, C, internal</td>
<td>• Low metabolic burden on cell</td>
<td>• Low specificity in IMAC purification</td>
</tr>
<tr>
<td>S-tag (KETAAKFERQHMDS)</td>
<td>15 aa</td>
<td>N, C, internal</td>
<td>• Low metabolic burden on cell</td>
<td>• Harsh elution conditions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High specificity</td>
<td>• Expensive affinity resin</td>
</tr>
<tr>
<td>FLAG™ (DYKDDDDK)</td>
<td>8 aa</td>
<td>N, C</td>
<td>• Low metabolic burden on cell</td>
<td>• Harsh elution conditions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• High specificity</td>
<td>• Expensive affinity resin</td>
</tr>
<tr>
<td>Glutathione-S-transferase (GST)</td>
<td>26 kDa</td>
<td>N</td>
<td>• Efficient translation initiation</td>
<td>• Increased metabolic burden on cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Inexpensive</td>
<td>• Does not enhance solubility</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Mild elution conditions</td>
<td></td>
</tr>
<tr>
<td>Maltose-binding protein (MBP)</td>
<td>40 kDa</td>
<td>N, C</td>
<td>• Efficient translation initiation</td>
<td>• Increased metabolic burden on cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Enhances solubility</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Inexpensive</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Mild elution conditions</td>
<td></td>
</tr>
</tbody>
</table>

a Derived from refs. [67, 73].
from animal sources, but with increasing concern about contamination from such sources, their utility for tag removal in pharmaceutical production of proteins is limited. Moreover, because exopeptidases lack an affinity tag, their own removal poses an additional challenge for purification of the target protein after tag separation [68].

<table>
<thead>
<tr>
<th>Type</th>
<th>Cleavage agent</th>
<th>Cleavage site</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chemical</strong></td>
<td>Cyanogen bromide</td>
<td>Met-↓-Xaa</td>
<td>[74]</td>
</tr>
<tr>
<td></td>
<td>Hydroxylamine</td>
<td>Asn-↓-Gly</td>
<td>[75]</td>
</tr>
<tr>
<td></td>
<td>Formic acid</td>
<td>Asp-↓-Pro</td>
<td>[76]</td>
</tr>
<tr>
<td><strong>Endopeptidase</strong></td>
<td>Enterokinase</td>
<td>Asp-Asp-Asp-Asp-Lys-↓-Xaa</td>
<td>[77]</td>
</tr>
<tr>
<td></td>
<td>Trypsin</td>
<td>Arg/Lys-↓-Xaa</td>
<td>[78]</td>
</tr>
<tr>
<td></td>
<td>Factor Xa</td>
<td>Ile-Glu/Asp-Gly-Arg-↓-Xaa</td>
<td>[79]</td>
</tr>
<tr>
<td></td>
<td>V8-protease</td>
<td>Glu-↓-Xaa</td>
<td>[80]</td>
</tr>
<tr>
<td></td>
<td>Thrombin</td>
<td>Leu-Val-Pro-Arg-↓-Gly-Ser</td>
<td>[81]</td>
</tr>
<tr>
<td><strong>Exopeptidase</strong></td>
<td>Carboxypeptidase A</td>
<td>Poly His-↓-Xaa</td>
<td>[82]</td>
</tr>
<tr>
<td></td>
<td>Carboxypeptidase B</td>
<td>Poly Arg-/PolyLys-↓-Xaa</td>
<td>[83]</td>
</tr>
<tr>
<td></td>
<td>Aminopeptidase I</td>
<td>Glu-Ala-Glu-↓-Xaa</td>
<td>[84]</td>
</tr>
</tbody>
</table>

↓ indicates the cleavage site, and Xaa indicates any nonspecific amino acid.

1.2.2 *In vivo* strategies for detection of protein:DNA interactions

*In vitro* methods for detection of protein:DNA recognition provide a means for quantitative characterization of biomolecular interactions. However, protein function may depend strongly on assay conditions, and a non-native *in vitro* environment can sometimes give results contradictory
to those performed in an \textit{in vivo} assay [85]. Hence, there is a common consensus that detecting a protein:DNA interaction \textit{in vitro} is not necessarily an indication of its relevance \textit{in vivo} [86].

Numerous strategies have been implemented for \textit{in vivo} characterization of protein:DNA interactions, including chromatin immunoprecipitation (ChIP) [87, 88], ligation-mediated PCR (LM-PCR) [89], DNA adenine methyltransferase identification (DamID) [90, 91], and yeast reporter assays [92, 93] (see the corresponding reference for description of each assay). However, ChIP and yeast-hybrid assays are by far the most widely used.

ChIP is one of the most powerful methods to identify and characterize protein:DNA interactions. It has been used widely for studies within the yeast genome, although protocols for performing ChIP in mammalian cells have also emerged [94-97]. Briefly, protein:DNA complexes formed inside living cells are temporarily fixed by a cross-linking agent (e.g. formaldehyde). The chromatin is then isolated and fragmented by sonication. Following incubation with a specific antibody directed against a protein of interest, DNA fragments associated with the protein are selectively immunoprecipitated. Finally, the crosslinks are reversed, and the released DNA fragments are purified for PCR analysis or sequencing.

The main strength of ChIP lies in the fact that it is currently the only method available to directly observe an \textit{in vivo} protein:DNA interaction. However, ChIP suffers from various limitations: it is time-consuming and certain antibodies are inefficient for immunoprecipitation, it regularly generates low signals with respect to negative controls and due to its limited resolution, it is difficult to pinpoint the exact binding site of a protein, which can be problematic if a DNA region of interest contains multiple potential target elements. In addition, cross-linking may fix
interactions of low functional significance, and lack of cross-linking may occur due to inaccessibility of proteins in large complexes [45, 98].

The main goal in ligation-mediated PCR, which is considered an *in vivo* DNA footprinting assay, is to assess the local activity of DNA inside living cells toward modifying agents, such as dimethylsulfate (DMS), UV light, and DNase I, compared with that of purified DNA *ex vivo* [99]. In a typical experiment, two separate samples of the same DNA (one sample inside living cells and in the presence of proteins, and the other purified DNA in the absence of proteins) are first treated with a modifying agent. Following DNA amplification by PCR, the samples are separated on a denaturing polyacrylamide gel, blotted onto a nitrocellulose membrane, and revealed upon hybridization with a radiolabeled probe. DNA regions protected from modification by bound proteins appear as missing bands.

DamID is primarily used to pinpoint the binding sites of transcription factors in eukaryotes. In this technique, identification of binding sites is achieved by expressing a proposed protein of interest as a hybrid fusion with DNA adenine methyltransferase (Dam), which catalyses adenine methylation exclusively in the GATC sequence. Binding of the protein of interest to DNA localizes Dam around its binding site, resulting in methylation of neighboring GATC sequence(s). Since adenine methylation does not naturally occur in eukaryotes, adenine methylation in any region of DNA can be presumed to be located close to the protein’s DNA target site. Adenine methylation can be revealed by a methylation-specific PCR reaction [91].

Yeast reporter assays are the most widely used techniques for *in vivo* detection and characterization of both protein:DNA and protein/protein interactions [92, 93, 100, 101]. Their relative ease of manipulation and high efficiency have made their use widespread. In light of the
scope of this thesis, more detailed information regarding these assays is provided in the following section.

1.2.2.1 Yeast reporter assays: Yeast two- and one-hybrid systems

Yeast two-hybrid (Y2H) and its variant yeast one-hybrid (Y1H) systems are genetic assays used for identification and characterization of protein/protein and protein:DNA interactions. They both exploit the modular nature of eukaryotic transcription factors, which comprise a DNA-binding domain for contacting DNA and an effector domain that regulates activation or repression of targeted genes. In each system, the interaction between either a protein/protein or a protein:DNA pair is detected via \textit{in vivo} reconstitution of a transcriptional activator that turns on expression of a reporter gene (Fig. 1.3).

The power of yeast reporter assays is built on the fact that the “prey”, which in both Y2H and Y1H systems is a protein of interest fused to an activation domain (Fig. 1.3), can be replaced by a cDNA or genomic library that allows screening of all the proteins present in the library. Consequently, over the past two decades, classical yeast reporter assays have been modified and used in a wide range of applications; notably, large-scale screening (e.g. interaction mapping) and drug discovery (e.g. small molecule detection) [102, 103]. Yeast reporter assays do not provide quantitative measurement of binding affinities, although the transcriptional readout from reporter activation can reflect protein:DNA binding activity [104, 105].
Figure 1.3 | Schematic representation of yeast reporter assays. a, In the Y2H system, protein of interest $X$ is fused to a DNA-binding domain that binds a specific DNA target site (i.e. the bait), whereas another protein, $Y$, is fused to a transcription activation domain that is capable of activation of a downstream reporter gene (i.e. the prey). Upon coexpression of both proteins, specific interaction between the bait and the prey brings the DNA-binding domain and the activation domain into close proximity, resulting in reconstitution of a functional transcription factor that activates transcription of the downstream reporter gene. b, In the Y1H system, detection is based on the specific interaction of a protein of interest fused to an activation domain (i.e. prey) with the bait being the cis-acting element, E, upstream of the reporter gene. Two of the most commonly used reporters in yeast reporter assays are HIS3 and lacZ.

Yeast reporter assays offer some advantages over classical biochemical and genetic approaches [106]. First, they provide a native-like environment for in vivo analysis of protein/protein and protein:DNA interactions. They are rather inexpensive and need minimal requirements to initiate screening (i.e. only the cDNA of the gene of interest is needed). In
contrast, biochemical approaches often require large quantities of purified proteins or good quality antibodies. In addition, weak and short-lived interactions, which are often the most interesting in signaling pathways, are more easily detected in yeast reporter assays since the genetic reporter gene strategy results in a considerable amplification. In addition to their ability to screen libraries, these systems also allow for analysis of known interactions, where by pinpointing crucial residues for successful binding, useful knowledge about the interaction, such as identifying residues and/or nucleotides involved in binding, can be obtained.

Yeast reporter assays include some limitations as well. Detection of a protein interaction by Y2H methods relies on the nuclear localization of the proteins under investigation; hence interactions between membrane proteins cannot be detected in a Y2H system, although this problem has been circumvented by introduction of split-ubiquitin or green fluorescent protein (GFP) systems [107, 108]. Additionally, yeast reporter assays suffer from high false-positive and false-negative rates, especially in high-throughput experiments [109-116]. False positives are interactions that are detected in a yeast genetic assay but are not generated by bona fide interactors; whereas, false negatives are real interactions that are not detected in a yeast-based experiment.

Yeast reporter assays are inherently prone to false positives and false negatives primarily because they rely on indirect readout of transcriptional activation of reporter genes [114]. As a result, various factors intrinsic to yeast genetic assays can contribute to the prevalence of false observations. One such factor is the nature of the bait molecule. For instance, baits that are harmful to the viability of the host yeast almost consistently yield a high frequency of false data in yeast library assays. In addition, out-of-frame prey fusions encoding short peptide fragments as well as transformation of yeast cells with multiple prey plasmids are two major sources of
false-positives in the Y2H systems [111]. In Y1H assay, some bait DNA sequences display unusually high leaky expression, which poses the most significant limitation on the applications of the Y1H [117].

1.3 Thesis objectives

As discussed above, understanding the structural and functional activities of transcription factors is essential to comprehend how various cellular processes can be regulated at a molecular level. Numerous strategies have been introduced in this chapter. *In vitro* methods for detection of protein:DNA interaction require isolation of active and soluble protein, and numerous affinity tags and cleavage agents are available and have been used for this purpose (Tables 1.1 and 1.2). However, current methodologies for purification of proteins often fail to provide high yield of pure and tag-free protein mainly because enzymatic cleavage reactions for tag removal do not exhibit stringent sequence specificity and frequently cleave the target protein at other nonspecific sites. Solving this issue is an important step towards high-throughput *in vitro* analysis of protein:DNA interactions. Therefore, one goal of this thesis is to present and discuss new approaches to enhance the specificity of a proteolysis reaction.

*In vivo* strategies have the advantage over their *in vitro* counterparts of providing a native-like environment for assessing protein:DNA interactions, yet the most frequently used techniques often suffer from high false-positive and -negative rates. Hence, unless new methodologies emerge, high-throughput *in vivo* analysis of protein:DNA interactions will be hampered by the high prevalence of spurious observations. Therefore, the other goal of this thesis is to present a new technique to detect protein:DNA interactions *in vivo*, one that offers higher levels of accuracy and sensitivity in comparison with existing technologies.
1.4 Thesis organization

This thesis comprises five chapters. Chapter one provides a review of various *in vivo* and *in vitro* strategies for studying the interactions of DNA-binding proteins with their target sites along with some background information on transcription factors.

Chapter two presents an *in vitro* study in which the authors describe how the specificity of a proteolytic cleavage reaction for removing a fusion tag is enhanced by simple manipulations of experimental conditions. This chapter was published previously in *Protein Expression and Purification*. The re-print license is provided in Appendix A. A more detailed description of the Methods section is provided in Appendix B.

Chapter three follows the project in described in chapter two by presenting an alternative strategy to augment efficient tag removal using a designed protein-protein interaction. This project is ongoing and work remains to confirm the success of the new strategy.

Chapter four focuses on an *in vivo* study in which we describe a novel fluorescent protein-based bioprobe for detection of protein:DNA recognition *in vivo*. This chapter will be submitted for publication in future. A more detailed description of the Methods section and some background information on the theory of fluorescence resonance energy transfer are provided in Appendix C.

Chapter five provides an overall summary of the findings presented in this thesis and discusses the impact of these findings on high-throughput analysis of protein:DNA interactions. In addition, some proposals for further improvements as well as new research directions will be presented.
Chapter Two

Enhancing the specificity of the enterokinase cleavage reaction to promote efficient cleavage of a fusion tag

S. Hesam Shahravan, Xuanlu Qu, I-San Chan, and Jumi A. Shin

Contributions: I.C. conceived of the idea of enhancing enterokinase proteolysis yields by using denaturants, and J.A.S. suggested that the more "open" protein structure would then be more accessible to cleavage and compete more effectively against adventitious cleavage sites. S.H.S. devised experiments, interpretation, and next-generation experiments, along with X.Q. J.A.S. provided troubleshooting and suggested experiments. S.H.S. and J.A.S. co-wrote and -edited the paper.

Published in: Protein Expression and Purification 59 (2008) 314–319

Note: Elsevier’s permission to re-print is provided in Appendix A. A more detailed description of the Methods section in this chapter is provided in Appendix B.

© 2008 Elsevier Inc. All rights reserved. doi:10.1016/j.pep.2008.02.015
Reprinted by permission from Elsevier
2.1 Abstract

In our work with designed minimalist proteins based on the bZIP motif, we have found our His-tagged proteins to be prone to inclusion body formation and aggregation; we suspect this problem is largely due to the His tag, known to promote aggregation. Using AhR6–C/EBP, a hybrid of the AhR basic region and C/EBP leucine zipper, as representative of our bZIP-like protein family, we attempted removal of the His tag with enterokinase (EK) but obtained the desired cleavage product in very small yield. EK is known for proteolysis at noncanonical sites, and most cleavage occurred at unintended sites. We manipulated experimental conditions to improve specificity of proteolysis and analyzed the cleavage products; no effect was observed after changing pH, temperature, or the amount of EK. We then suspected the accessibility of the EK site was impeded due to protein aggregation. We found that the easily implemented strategy of addition of urea (1–4 M) greatly improved EK cleavage specificity at the canonical site and reduced adventitious cleavage. We believe that this enhancement in specificity is due to a more “open” protein structure, in which the now accessible canonical target can compete effectively with adventitious cleavage sites of related sequence.
2.2 Introduction

Protein purification is often facilitated by use of protein tags: hence, recombinant fusion proteins. The polyHis sequence is commonly used for chelation during immobilized metal–ion affinity chromatography. The His tag can be left on the protein, and often it is; however, the His tag may interfere with certain experiments, including those involving metals or in vivo administration, and it can lead to protein aggregation and inclusion body formation during bacterial expression [118]. Removal of the His tag may therefore be advantageous, and endoprotease sites are typically engineered between the desired protein and tag.

Efficient site-specific proteolysis of fusion proteins is critical, yet the most commonly used endoproteases—enterokinase, factor Xa, and thrombin—do not exhibit stringent sequence specificity and often cleave at nontarget sites [81]. Enterokinase (EK, also known as enteropeptidase) is a membrane-bound serine protease found in the duodenum and initiates activation of pancreatic hydrolases by cleaving and activating trypsinogen. Although the canonical target site for enterokinase is DDDDK, it is known that EK does not exhibit high stringency in its specificity for this sequence [77, 119-121]. For instance, Liew et al. showed that EK preferentially cleaved at an unexpected LKGDR site that was near the carboxyl terminus of N-terminal proCNP and more accessible than the internal canonical DDDDK sequence [120].

In our research on design of minimalist proteins, we express short proteins in *Escherichia coli* that are modeled after the basic region/leucine zipper (bZIP) family of transcription factors, a dimeric, α-helical motif comprising ~60 amino acids [9, 17]. We express our bZIP proteins as fusions comprising an N-terminal 6×His tag with an intervening EK recognition site to facilitate tag removal. We encountered two significant problems during production of AhR6–C/EBP, a
hybrid of the aryl hydrocarbon receptor (AhR) basic region and CCAAT/enhancer binding protein (C/EBP) leucine zipper that we use in this work as representative of our bZIP family of proteins (Fig. 2.1): limited protein solubility and inefficient EK proteolysis of the His tag. We suspected that solubility problems stemmed from the His tag, known to promote aggregation [118]. Previously, we found that by manipulation of temperature during expression and experimental handling (the temperature leap tactic), we could maintain solubility of our His-tagged bZIP derivatives [122, 123]. However, in order to minimize such experimental manipulations, we sought to remove the His tag by use of EK, but we could only obtain the desired cleavage product in vanishingly small yield.

Figure 2.1 | Amino acid sequence of 6×His-tagged AhR6–C/EBP. The enterokinase recognition site is underlined. The AhR6–C/EBP basic region starts at Asp32, and the leucine zipper ends at Leu106. The His tag and EK cleavage site are at the N-terminus. At the C-terminus is the GGCGGYYYY sequence useful for spectroscopic evaluation and diazotization to solid support [124].

We therefore manipulated experimental conditions to maximize the yield of the desired protein products from EK proteolysis. Because we suspected that the accessibility of the EK site was impeded, that is, a structural problem, we focused on use of denaturants as a way to open the
structure, thereby essentially increasing the stoichiometry of the canonical recognition site over noncanonical, adventitious sites.

2.3 Materials and Methods

DNA oligonucleotides were purchased from Operon Biotechnologies (Huntsville, AL). Enzymes were supplied by New England Biolabs (Pickering, ON). Native enterokinase (MW 150 kDa, obtained from calf intestine) was purchased from Roche Applied Science (Laval, QC). Reagents were supplied by EMD Chemicals (Gibbstown, NJ), Fisher Scientific Canada (Ottawa, ON), or Bioshop Canada (Burlington, ON). DNA sequencing was performed on an ABI 3730XL 96 capillary sequencer (Applied Biosystems) at the DNA Sequencing Facility in the Centre for Applied Genomics, Hospital for Sick Children (Toronto, ON). Electrospray ionization mass spectrometry (ESI-MS) was performed on a Micromass ZQ Model MM1 quadruple mass spectrometer (Waters) at the University of Toronto at Mississauga.

2.3.1 Preparation of AhR6–C/EBP protein

The gene for expression of AhR6–C/EBP was constructed by mutually primed synthesis, followed by polymerase chain reaction with terminal primers for gene amplification and purification by nondenaturing polyacrylamide gel electrophoresis. After restriction with BamHI and EcoRI, duplex DNA was cloned into protein expression vector pTrcHis B (Invitrogen, Carlsbad, CA) and sequenced; this vector expresses proteins with an N-terminal 6×His tag. The recombinant plasmid was transformed into *E. coli* strain BL21(DE3) (Stratagene, La Jolla, CA) by electroporation (Bio-Rad *E. coli* Gene Pulser).
Bacterial expression of AhR6–C/EBP was performed in LB medium containing 100 µg/mL ampicillin at 37 °C; induction was initiated at mid-log phase (OD600 ~0.6) with IPTG added to 1 mM final concentration. Cells were harvested after 3 h by centrifugation and lysed by sonication. The 6×His-tagged protein was purified on TALON cobalt metal–ion affinity resin (Clontech, Mountain View, CA) with the wash and elution buffers containing 8 M urea and 6 M guanidine–HCl, respectively, to promote protein solubility. The eluted protein was concentrated by centrifugation with a Centriplus centrifugal filter (Millipore, Bedford, MA). The protein was further purified by HPLC (Beckman System Gold) on a semipreparative reversed-phase C4 column (Vydac, Hesperia, CA) with a gradient of acetonitrile–water plus 0.05% trifluoroacetic acid (v/v) at flow rate 4 mL/min; the gradient started at 10–25% acetonitrile over 15 min, followed by 25–55% acetonitrile over 60 min. Purity of the final product was confirmed by analytical HPLC on a C4 column (Vydac) using the same gradient as above but flow rate 1 mL/min. Identity was verified by ESI-MS; calculated mass: 13034.6 g/mol; found: 13033.0 g/mol with the N-terminal Met cleaved during post-translational modification [125]. HPLC purified protein was lyophilized and stored at -80 °C. The overall yield of purified protein was 2.5 mg/L LB culture.

2.3.2 Enterokinase cleavage of AhR6–C/EBP (typical EK reaction)

The enterokinase reactions (10 µL) were prepared according to the manufacturer’s recommendations. We used the native form of EK supplied by Roche; in earlier experiments, we also used the light chain recombinant version of EK (New England Biolabs, data not shown); however, only with the native version could we obtain reproducible cleavage at the canonical EK target site, so we continued our studies with only the native form. 6.0 µg purified AhR6–C/EBP was incubated with 0.14 µg EK (EK: substrate ratio = 1:42) at 37 °C in 50 mM Tris, pH 7.6. The
reactions were stopped by addition of an equal volume of Tricine Sample Buffer (Bio-Rad, Hercules, CA), which contains 125 mM dithiothreitol (DTT), followed by heating for 10 min at 95 °C, and analysis by 16.5% Tris–Tricine SDS–PAGE with molecular weight markers (Polypeptide SDS–PAGE Molecular Weight Standards, Bio-Rad, shown in Figs. 2.2–2.4 and 2.6). Coomassie stained gels were analyzed with LabImage (version 3.3.3) to calculate cleavage efficiencies. Purity of the native enzyme was assayed by analytical HPLC as above.

2.3.3 Isolation and identification of cleavage products by HPLC and ESI-MS

The enterokinase reaction (100 µL) was prepared as above and stopped by addition of phenylmethanesulfonyl fluoride (PMSF) to a final concentration of 1 mM, followed by heating for 10 min at 95 °C. The stopped reaction was incubated with 30 mM DTT for ≥30 min at room temperature, followed by analytical HPLC as above. The collected fractions were analyzed with ESI-MS at flow rate 40 µL/min with the capillary charged at +3200 V, source temperature 100 °C, and cone voltage 20 V. The mass range 600–1100 m/z was scanned every 1 s for 40 s. Uncalibrated data were processed using MassLynx (version 4.00.00). Molecular masses were analyzed by massXpert (version 1.6.0) to identify cleavage sites. Due to the observed discrepancy between the migration rates of the Bio-Rad molecular weight markers (in particular, the MW marker at 16.9 kDa) and the uncleaved AhR6–C/EBP in Figs. 2.2–2.4 and 2.6, the same typical enterokinase reaction with AhR6–C/EBP (MW 13,033 Da) as above was prepared at large scale to generate our own molecular weight standard (not shown in Figs. 2.2–2.4 and 2.6). All fragments were purified by HPLC, followed by confirmation of identity by ESI-MS before loading onto SDS–PAGE. Our MW standard was therefore used to assign bands A–F in Figs. 2.3, 2.4 and 2.6.
2.4 Results and Discussion

The reaction of AhR6–C/EBP with EK under the manufacturer’s standard conditions (see Materials and methods) is rapid and somewhat nonspecific. In addition to the barely detectable intended cleavage after the canonical EK target sequence DDDDK yielding fragments of 9737 Da (band B in Figs. 2.3, 2.4, and 2.6, Fig. 2.5) and 3315 Da (too small to be visualized in SDS–PAGE, but shown in HPLC, Fig. 2.5), undesired adventitious cleavage occurs (Fig. 2.2). SDS–PAGE analysis of the reaction of AhR6–C/EBP with EK revealed that after just 15 min incubation, 82% of the protein had been cleaved at multiple sites. Prolonged incubation caused further degradation of these cleavage products, as indicated by their gradual disappearance during 2 h incubation. Thus, the prompt formation of these cleavage products, as well as their subsequent disappearance, indicates that, indeed, EK is a very active protease and cleaves our protein at sites other than the DDDDK target site.

![Figure 2.2](image)

**Figure 2.2** | Time-course analysis of cleavage products after incubation of AhR6–C/EBP with EK for the times indicated (top, x-axis, in minutes). Molecular-weight marker lane is at the left (MW, see Materials and methods), and apparent molecular weights are indicated in the y-axis legend. EK proteolysis of AhR6–C/EBP at the DDDDK recognition site should yield fragments of 9737 Da (band B in all SDS–PAGE figures) and 3315 Da (not visualized by SDS–PAGE).
We examined the purity of the enterokinase (the native form) by analytical HPLC to assess whether the adventitious cleavage of AhR6–C/EBP was due to proteases that might be co-purified with the native enzyme (data not shown). We found that the EK used in our study was indeed pure and contained no detectable proteases. Therefore, as similarly observed by Choi et al. and Liew et al., we believe that degradation of AhR6–C/EBP by native enterokinase is the result of an intrinsic broad specificity of EK for various target sites rather than the presence of contaminating proteases [120, 126].

2.4.1 Changing pH, temperature, or the amount of EK had no effect on cleavage pattern

We hypothesized that we could improve the specificity of the EK cleavage reaction of AhR6–C/EBP by manipulating the standard experimental conditions. We first tried changing the pH of the EK reaction buffer (Fig. 2.3). Lowering the pH from 7.6 to 7.0 had no appreciable influence on the reaction rate or cleavage pattern; however, increasing the pH to 8.0 had an inhibitory effect on EK activity, and after 15 min incubation with EK, 42% of our protein remained intact. We then examined whether lowering the reaction temperature would affect the cleavage specificity by decreasing EK activity. Lowering the temperature from 37 °C to 25 °C had no noticeable effect on the reaction rate or cleavage pattern (data not shown), similar to that observed by Menart and coworkers in their study on removal of affinity tags from oligomeric proteins [72].
Figure 2.3 | EK cleavage reactions of AhR6–C/EBP with different reagents and under different experimental conditions. All reactions were stopped after 15 min incubation with EK. Molecular-weight marker lane (MW, see Materials and methods) and apparent molecular weights (y-axis legend) at left. Lane 1, EK reaction according to the manufacturer’s recommendations (typical EK reaction); lane 2, typical EK reaction with 1 M urea; lane 3, typical EK reaction at pH 7.0; lane 4, typical EK reaction at pH 8.0; lane 5, typical EK reaction with one quarter of original EK amount; lane 6, typical EK reaction with half of original EK amount. The masses of the major bands as determined by ESI-MS: (A) 13,033 Da (intact AhR6–C/EBP); (B) 9737 Da; (C) 9210 Da; (D) 8072 Da; (E) 7055 Da; (F) 4978 Da. Note that band B (9737 Da) is very faint; low yields of this cleavage product are obtained when the EK reaction contains only 1 M urea.

We also tried reducing the amount of EK present in the reaction by adding one half or one quarter of the original EK amount: the rate of cleavage was reduced, but specificity was unaffected (Fig. 2.3). If we reduced the amount of EK and increased the incubation time from 2 h to 3 h, we observed the same additional cleavage products and their subsequent degradation, as above (data not shown). Contrary to the results of Liu et al., who reported that reducing the amount of EK could decrease nonspecific cleavage of human parathyroid hormone [127], we found that reducing the concentration of EK only affects the rate of proteolysis of AhR6–C/EBP, not the specificity.
2.4.2 Nonspecific cleavage due to inaccessibility of the target site

These results prompted us to adjust our hypothesis: we suspected that nonspecific proteolysis was due to inaccessibility of the EK target site, and hence, EK cleaved our protein at more accessible sequences, as it is known for cleaving at unexpected sites [77, 119-121]. In our case, the probable reasons for the lack of accessibility to the target site are structural: protein dimerization and aggregation. The bZIP-like AhR6–C/EBP can homodimerize via the C/EBP leucine zipper coiled coil. Additionally, an unintended problem may stem from the observation that His-tagged proteins are prone to aggregation [118, 128, 129]. Thus, protein multimerization, whether by the normal dimerization mechanism or by misfolding and aggregation, may adversely affect EK’s ability to bind to its target site.

We hypothesized that the presence of a denaturant would discourage secondary and tertiary, as well as quaternary, protein structure that shields EK from its target. Thus, we added urea to our cleavage reactions to reduce adventitious cleavage. Addition of urea caused a profound difference in our cleavage reactions (Fig. 2.3). After 15 min incubation of AhR6–C/EBP with EK in the presence of 1 M urea, we observed two major differences in the reaction pattern in comparison to the previously examined standard conditions: (i) a higher proportion of cleavage product at 9210 Da resulted (band C, Fig. 2.3); (ii) the cleavage product at 9737 Da, still only faintly detectable by SDS-PAGE, was produced in higher yield (band B, Fig. 2.3). We performed more cleavage reactions with even higher concentrations of urea; addition of over 1 M urea increased the amount of the 9737 Da product, as indicated by increased density at band B in Fig. 2.4. At 5 M urea, the denaturant began to inhibit EK activity, as indicated by the overall reduced amounts of cleavage products (Fig. 2.4, bands B–F).
Figure 2.4 | EK cleavage reactions of AhR6–C/EBP under varying urea concentrations (top, x-axis). All reactions were stopped after 30 min incubation with EK. MW lane (see Materials and methods), y-axis legend, and masses of bands as determined by ESI-MS same as in Fig. 2.3. Note that the intensity of band B (9737 Da) increases as the concentration of urea is increased during EK cleavage.

We also analyzed EK cleavage of AhR6–C/EBP by use of other denaturants, including guanidine and SDS; both primarily inhibited EK proteolysis at the canonical site rather than improving specificity of cleavage (data not shown). In addition, we performed an unfolding experiment to test whether we could achieve comparable results as observed upon addition of urea to the EK reaction by just manipulating temperature. We prepared a typical EK reaction (as specified in Materials and methods) but without EK and urea. This reaction was heated for 5 min at 90 °C followed by slow cooling to room temperature over 3 h. We then added EK and found that this temperature tactic had no effect on the specificity of the EK reaction (data not shown). Had this temperature tactic worked, we still believe that using urea is a more favorable method for achieving specificity of cleavage at higher yields, as it requires significantly less time.
2.4.3 Addition of denaturant reduces adventitious enterokinase cleavage of AhR6–C/EBP

In order to identify the cleavage products, we used the manufacturer’s standard conditions for reaction of EK with AhR6–C/EBP, isolated the adventitious cleavage products by reversed-phase HPLC (Fig. 2.5), and identified their masses by ESI-MS, which allowed determination of the precise locations of EK cleavage sites in AhR6–C/EBP. EK proteolyzes our protein after Lys31 (its canonical DDDDK site), Arg36, Lys55, Arg72, and Arg91, which explains the total degradation of our protein observed after extended incubation with enterokinase (Fig. 2.2).

The largest cleavage product with mass 9737 Da, which appeared in EK reactions containing more than 1 M urea (Fig. 2.4, band B), is the result of cleavage after the DDDDK target site. Hence, addition of urea to our cleavage reactions indeed facilitated cleavage of AhR6–C/EBP at the intended EK target site. We then incubated AhR6–C/EBP with EK in reaction buffer containing 3 M urea for 5 h (Fig. 2.6). Comparison of these results with those obtained earlier without urea (Fig. 2.2) demonstrates the much slower disappearance of the cleavage fragment at 9210 Da, formed after EK cleavage at Arg36, as indicated by band C in Fig. 2.6, thereby indicating inhibition of further digestion after Lys55, Arg72, and Arg91. We note that under all tested conditions, the adventitious 9210 Da product appears to be in higher yield than the desired 9737 Da product, as qualitatively visualized by SDS–PAGE; under HPLC analysis, we were unable to separate fully the 9210 Da and 9737 Da peaks (Fig. 2.5), and therefore, we were unable to assess quantitatively the amounts of these species.
2.4.4 Relative affinities of EK for its cleavage sites in AhR6–C/EBP

Analysis of the SDS–PAGE of cleavage reactions containing 3 M urea (Fig. 2.6) further increases our understanding of the relative affinities of EK for its different cleavage sites in AhR6–C/EBP. Enterokinase cleaves our protein at sites preceded by either Lys or Arg. Light et
[al. have reported that the minimum sequence for EK cleavage requires a basic amino acid (Lys or Arg) at the P1 position, which directly precedes the cleavage site, and an acidic amino acid (Asp or Glu) at the P2 site, which precedes P1; they also reported that the rate of hydrolysis of a cleavage site increases when more acidic residues occupy the P3–P5 positions [130]. Although all of the cleavage sites observed in AhR6–C/EBP (–DK31, –SR36, –IK55, –DR72, and –DR91) have either Lys or Arg in the P1 position (Fig. 2.5), not all of the cleavage sites have an acidic amino acid in the P2 position (–SR36, –IK55); hence, the requirement for an acidic residue at P2 is not as stringent as that for a basic residue at P1 in AhR6–C/EBP.

Figure 2.6 | Time-course analysis of cleavage products after incubation of AhR6–C/EBP with EK with 3 M urea for the times indicated (top, x-axis, in hours). MW lane (see Materials and methods), y-axis legend, and masses of bands as determined by ESI-MS same as in Fig. 2.3.

Moreover, SDS–PAGE analysis of EK reactions containing 3 M urea shows that within the first 30 min of incubation of AhR6–C/EBP with EK (Fig. 2.6), the cleavage product present in the largest amount is that resulting from cleavage at –SR36 (Fig. 2.6, band C, 9210 Da). Serine is a polar, uncharged amino acid, yet cleavage at –SR36 is the fastest of all the sites (–IK55, –DR72, and –DR91). Regarding AhR6–C/EBP, EK proteolysis does not require an acidic residue in
the P2 position; additionally, any site containing a basic residue at P1 and an acidic residue at P2 is not necessarily an actual cleavage site for EK. For example, we were unable to see any cleavage of our protein at –DR<sup>65</sup> (Fig. 2.1), although we did observe cleavage at –DR<sup>72</sup> and –DR<sup>91</sup> cleavage sites. A probable explanation is that in AhR6–C/EBP, –DR<sup>65</sup> is preceded by four basic residues that may inhibit proteolysis [130].

2.4.5 Modification of the EK cleavage site as an alternative approach to enhance cleavage efficiency and specificity is protein-specific

Others have examined the effects of residues downstream from the canonical EK site in their efforts to enhance the specificity of enterokinase cleavage. Hosfield and Lu investigated the influence of the first downstream residue after the canonical site on the EK cleavage efficiency of their glutathione S-transferase-calmodulin fusion protein (GST–CaM) [131]. They found that EK is permissive regarding the amino acid immediately downstream of the EK recognition site in their protein. Of the residues occupying this position, smaller amino acids allow for more efficient cleavage, whereas those that create steric hindrance in the binding pocket result in lower cleavage efficiency. Similarly, Liew et al. studied the effect of addition of an SRLLR motif immediately downstream of the EK target site on cleavage efficiency of their thioredoxinfused N-terminal proCNP (NT-proCNP) [132]. They reported steady increase in the rate of hydrolysis of their protein at the DDDDK site upon incremental additions of SRLLR motifs downstream from the DDDDK site. In our case, we took a different approach to address the problem of proteolytic specificity. By addition of denaturant, we were able to achieve increased EK cleavage specificity of AhR6–C/EBP without any modifications to the protein construct. This experimental modification is easier to implement and may be more generally applicable.
We believe that the problem of inefficient cleavage in AhR6–C/EBP at the canonical EK site stems from its inaccessibility, and thus, adventitious sites of related sequence can outcompete the canonical site for EK proteolysis. His-tagged AhR6–C/EBP can homodimerize via the C/EBP leucine zipper and is prone to inclusion body formation, both of which can adversely affect EK’s ability to bind to its target site. Others have also observed similar phenomena of adventitious cleavage and have speculated that this may result from inaccessibility of the EK target site during proteolysis [120, 133]. For instance, Liew et al. observed less efficient cleavage at the adventitious LKGDR sequence when they moved it from the C-terminus of NT-proCNP to a more internal location [120].

Enhancement of cleavage specificity at the DDDDK site achieved by methods involving replacing residues upstream or downstream of the canonical site is protein-specific, and hence, may not be universally applied to all fusion proteins. As an example, Hosfield and Lu reported that in their GST-CaM fusion protein, 84% cleavage at the DDDDK site resulted when Asp occupied the position immediately after the canonical site [131]. In our case, the EK target site in AhR6–C/EBP is also immediately followed by Asp (Fig. 2.1). Yet, until standard experimental conditions were altered by addition of denaturant, little to no cleavage was observed after the canonical EK target site (Figs. 2.2 and 2.3).

Thus, we suspect that such methods involving enhancement of EK cleavage by modifying the target site and flanking regions depend on protein structure. AhR6–C/EBP may be dimerized, even aggregated potentially, thereby restricting access to the EK target; opening this structure by use of denaturant can alleviate this problem. In contrast, Hosfield and Lu’s GST–CaM fusion may not have a tendency to dimerize or form aggregates, and hence, this structural issue may not be the reason for their EK cleavage inefficiency.
Others have also observed that protein structure can be responsible for successful proteolysis. Kim et al. studied protease cleavage efficiency of glucagon-fused human interleukin-2 by modifying the sequence upstream from the canonical EK site [133]. Human Interleukin-2 (hIL-2) contains one disulfide linkage and is folded into a bundle-shaped protein comprising four α-helices [134]. Heterologous expression of hIL-2 involves formation of heterogeneous aggregates by nonspecific intermolecular disulfide linkages [135]. Kim et al. carried out expression of hIL-2 by dissolving insoluble aggregates by a pH-shift and performed subsequent EK cleavage without any denaturing agent in the reaction buffer [133]. They reported low cleavage efficiency, which they speculated to be due to the EK recognition site being sterically hindered. In order to enhance the cleavage efficiency, they inserted the negatively charged DDDD sequence upstream from the EK target site. Contrary to their expectations, this modification to the target site did not enhance EK cleavage specificity; rather, they observed increased adventitious cleavage of hIL-2, and they concluded that hIL-2 was inherently an inefficient substrate for EK [133]. Zhang et al. prepared recombinant 6×His-tagged hepcidin, a highly disulfide-bonded peptide containing eight Cys residues and four intermolecular disulfide bonds [129]. They used 8 M urea to promote solubility and prevent precipitation of hepcidin resulting from disulfide-linked multimerization during protein preparation, and their EK reaction buffer contained 2 M urea. After a 16 h EK cleavage reaction, they obtained intact, tag-free hepcidin in ~20% yield [129].

With hIL-2, Kim et al. did not use denaturant; rather, they tried to increase EK recognition by modifying the target site [133]. For hIL-2, this approach was not successful, although for GST–CaM and NT-proCNP, this type of strategy was effective [131, 132]. Zhang et al. used high concentrations of urea with hepcidin and successfully obtained desired EK cleavage
product [129]. We surmise that an open, accessible recognition site for EK is critical for efficient proteolysis, and that the success of different strategies for promoting open protein structure can be case-dependent.

Addition of urea to our EK reactions of bZIP-like AhR6–C/EBP not only promotes the desired cleavage after the canonical EK site, but also inhibits additional proteolytic cleavage at adventitious sites: we suspect that this is achieved when the canonical site is as accessible as any other potential cleavage site, and is therefore better able to compete with these adventitious sites for recognition by EK. Addition of denaturant to the EK reaction buffer may therefore be a general strategy for discouraging protein structure that shields the intended cleavage site from EK recognition. However, use of higher amounts of urea (~5 M) had an inhibitory effect on the activity of EK during cleavage of AhR6–C/EBP (Fig. 2.4). Therefore, promoting accessibility of the canonical EK target site can increase proteolytic specificity and cleavage yield, and general strategies that promote a more open structure should be useful for preparation of proteins requiring endoprotease treatment.
Chapter Three

Efficient cleavage of a fusion tag: an alternative approach using a designed protein-protein interaction to promote a structure more accessible to proteolysis

S. Hesam Shahravan, I-San Chan, Anna V. Fedorova and Jumi A. Shin

Contributions: I.C. and A.V.F. conceived the idea of enhancing enterokinase proteolysis yields by using a designed protein-protein interaction. I.C. performed some preliminary experiments. S.H.S. devised the next-generation experiments, and performed interpretation and analysis of data. J.A.S. provided troubleshooting and direction.

Note: This chapter describes an ongoing project.
3.1 Abstract

We previously showed that addition of denaturant to enterokinase (EK) proteolysis reaction of bZIP-like AhR6-C/EBP greatly improved EK cleavage specificity at the canonical site and inhibited adventitious cleavage. We believe that this enhancement in specificity was due to a more “open” protein structure, in which the accessible canonical target could compete effectively with adventitious cleavage sites of related sequence. However, use of higher amounts of urea had an inhibitory effect on the activity of EK, and we were unable to increase the yield of the desired protein product significantly. In this work, we are proposing an alternative strategy for more efficient EK proteolysis: by addition of a separate leucine zipper to heterodimerize with the bZIP, we surmise that the basic region and amino-terminus becomes more open and more accessible to enterokinase. We used a synthetic peptide corresponding to the minimal leucine zipper region from Fos to heterodimerize with AhR6-Jun, a hybrid of the AhR basic region and Jun leucine zipper. However, addition of the Fos zipper to EK reactions inhibited EK activity, and we were unable to test our new strategy using this zipper. Therefore, work remains to verify the success of the new approach using other combinations of substrate and zipper proteins.
3.2 Introduction

The main advantage for using affinity tags to facilitate protein purification is that they can be generalized to wide classes of proteins that can be purified using a common method, as opposed to developing highly customized procedures for specific proteins [68]. However, a tag can interfere with protein conformation and biological activity and can lead to undesired flexibility in structural studies and toxicity [136, 137].

Removal of a tag may therefore be advantageous, and endoprotease sites are commonly engineered between the desired protein and tag. However, many endoproteases do not exhibit stringent sequence specificity and often cleave at noncanonical sites [81]. We previously showed that removal of a His tag by enterokinase (EK) resulted in inefficient proteolysis of the tag during production of AhR6–C/EBP, a hybrid of the aryl hydrocarbon receptor (AhR) basic region and CCAAT/enhancer binding protein (C/EBP) leucine zipper, modeled after the basic region/leucine zipper (bZIP) family of transcription factors [9, 17], and that we could only obtain the desired cleavage product in vanishingly small yield [138].

In order to maximize yield of the tag-free intact protein product from EK proteolysis, we manipulated experimental conditions. Because we suspected that accessibility of the EK site was impeded due to protein aggregation, we focused on use of denaturants as a way to open the structure, thereby essentially increasing the stoichiometry of the canonical recognition site over noncanonical, adventitious sites [138]. We showed that addition of urea to our EK reactions of bZIP-like AhR6–C/EBP not only promotes the desired cleavage after the canonical EK site, but also inhibits additional proteolytic cleavage at adventitious sites. However, use of higher amounts of urea (~5 M) had an inhibitory effect on the activity of EK during cleavage of AhR6–
C/EBP, and hence we were unable to increase the yield of the desired protein product considerably.

We are therefore developing an alternative strategy to promote a more open structure. Because the bZIP is a dimer, we propose to use a separate leucine zipper to heterodimerize with the bZIP, thereby leaving the basic region and amino-terminus open and potentially more accessible to enterokinase (Fig. 3.1). This strategy of using a designed protein that noncovalently interacts with the recombinant protein and promotes proper structure may allow us to obtain good yields of purified bZIP.

**Figure 3.1** | An alternative strategy to promote a more open structure for efficient proteolysis of the His tag. (A) Protein aggregation, whether by normal dimerization of the bZIP-like protein via the leucine zipper coiled coil or by misfolding and aggregation of His-tagged proteins, can adversely affect EK’s ability to bind to its target site. (B) By using a short leucine zipper to heterodimerize with the bZIP, the basic region and amino-terminus become more open and potentially more accessible to enterokinase.
3.3 Materials and Methods

The experimental procedures have been described in detail previously [138]; the following is a brief summary specific for the AhR6-Jun protein.

3.3.1 Preparation of AhR6–Jun protein

The protein expression vector for AhR6–Jun was created by replacing the gene encoding C/EBP zipper with that of the Jun zipper in pTrcHis B AhR6-C/EBP expression vector. pTrcHis B AhR6-C/EBP expression vector was first digested with *Xho*I and *EcoR*I restriction endonucleases, purified by gel electrophoresis, and ligated to a previously designed gene encoding Jun zipper also flanked by *Xho*I and *EcoR*I restriction sites. The recombinant plasmid was sequenced before transformation into *E. coli* strain BL21(DE3) (Stratagene) by electroporation.

Bacterial expression of AhR6–Jun was performed in Luria-Bertani (LB) medium containing 50 µg/mL ampicillin at 37 °C, with induction at mid-log phase (OD600 ~0.6) with final concentration of 1 mM IPTG. Cells were collected by centrifugation and lysed by sonication. The N-terminal 6×His-tagged protein was initially purified on TALON cobalt metal–ion affinity resin (Clontech) with the wash and elution buffers containing 8 M urea and 6 M guanidine–HCl, respectively, to promote protein solubility. The eluted protein was further purified by HPLC (Beckman System Gold) on a semipreparative reversed-phase C4 column (Vydac). Purity of the final product was confirmed by analytical HPLC. Identity was verified by electrospray ionization mass spectrometric analysis (ESI-MS, Waters); calculated mass: 13860.6 g/mol; found: 13861.0 g/mol with the N-terminal Met cleaved during post-translational modification [125].
3.3.2 Preparation of Fos zipper protein

Fos zipper protein (38-residues, ELDRLEAEETDQLEDEKSALQTEIANLLKEKEKLEFILY) was obtained from Biomer Technology (Concord, CA) as a crude product synthesized using standard Fmoc [N-(9-fluorenyl)methoxycarbonyl] chemistry followed by cleavage and deprotection. The peptide was purified by HPLC as described above. Identity was verified by ESI-MS: calculated mass, 4509.1 g/mol; found, 4511.9 g/mol.

3.3.3 Enterokinase cleavage of AhR6–Jun (typical EK reaction)

The enterokinase reactions (10 µL) were prepared according to the manufacturer’s recommendations with the native form of EK supplied by Roche. 6.0 µg purified AhR6–Jun was incubated with 0.14 µg EK at 37 °C in 50 mM Tris, pH 7.6. The reactions were stopped by addition of an equal volume of Tricine Sample Buffer (Bio-Rad), containing 125 mM dithiothreitol (DTT), followed by heating for 10 min at 95 °C, and analysis by 12% Tris–Tricine SDS–PAGE with molecular weight markers (Polypeptide SDS–PAGE Molecular Weight Standards, Bio-Rad).

3.4 Results and Discussion

In order to promote a proper structure for efficient EK proteolysis using a designed protein-protein interaction, we based our design on the well-characterized Fos-Jun heterodimer [139, 140]. Fos and Jun, which are the protein products of the nuclear proto-oncogenes c-fos and c-jun, bind to DNA preferentially as a heterodimer, with dimerization partner specificity dictated by the leucine zipper regions [141]. In addition, while Jun can homodimerize weakly, Fos does not homodimerize. We surmised that by using Fos-Jun interaction, we can nucleate a more properly folded, nonaggregated structure where the basic region of the bZIP would be more open, and the
enterokinase cleavage site, which is at the N-terminus of the basic region, may be more accessible to protease cleavage.

In order to test our strategy, we replaced the gene for the C/EBP leucine zipper in AhR6-C/EBP, which we studied previously [138], with a gene encoding the leucine zipper region of Jun: hence, AhR6-Jun protein (Fig. 3.2). We used a synthetic peptide corresponding to the minimal leucine zipper region from Fos to heterodimerize with AhR6-Jun. AhR6-Jun is the ideal protein with which we can assess our strategy, for although it can homodimerize via the Jun-Jun leucine zipper coiled coil, it preferentially forms heterodimers in the presence of Fos. Therefore, an excess of Fos zipper may outcompete any AhR6-Jun homodimers or aggregates that obstruct enterokinase target sites.

![Amino acid sequence of 6×His-tagged AhR6–Jun. The enterokinase recognition site is underlined. The AhR6–Jun basic region starts at Asp32. The leucine zipper starts at Glu80 and ends at Asn115. The His tag and EK cleavage site are at the N-terminus. At the C-terminus is the GGCGGYYYY sequence useful for spectroscopic evaluation and diazotization to solid support [124].](image-url)
3.4.1 Inefficient proteolysis of AhR6-Jun

Similar to our observations with AhR6-C/EBP, the reaction of AhR6–Jun with EK under the manufacturer’s standard conditions (see Materials and methods) was rapid and nonspecific (Fig. 3.3). SDS–PAGE analysis of the reaction of AhR6–Jun with EK revealed that after just 15 min incubation, the protein had been cleaved at multiple undesired adventitious sites, and extended incubation caused further degradation of these cleavage products. Thus, the prompt formation of these cleavage products, as well as their subsequent disappearance, indicates that EK consistently cleaves AhR6-Jun at sites other than the DDDDK target site.

![Figure 3.3](image)

**Figure 3.3** | Time-course analysis of cleavage products after incubation of AhR6–Jun with EK for the times indicated (top, x-axis, in minutes). Molecular-weight marker lane is at the left (MW, see ref. [138]), and apparent molecular weights are indicated in the y-axis legend. EK proteolysis of AhR6–Jun (13.8 kDa) at the DDDDK recognition site should yield fragments of 10564 Da and 3315 Da. This SDS-PAGE analysis shows that after just 15 min incubation, all of the protein has been cleaved at multiple sites. Prolonged incubation causes further degradation of these cleavage products, as indicated by their gradual disappearance during 2 h incubation.

We re-examined the purity of the enterokinase by analytical HPLC to assess whether the adventitious cleavage of AhR6–Jun was due to proteases that might be co-purified with the
native enzyme. We confirmed once more that the EK used in our study was pure and contained no detectable proteases (data not shown). Therefore, as observed before, we believe that degradation of AhR6–Jun by native enterokinase is indeed the result of an intrinsic broad specificity of EK for various target sites rather than the presence of contaminating proteases [120, 126, 138].

3.4.2 Addition of denaturant reduces adventitious enterokinase cleavage of AhR6–Jun

Similar to our observations with AhR6-C/EBP, we suspected that nonspecific proteolysis of AhR6-Jun could also be due to inaccessibility of the EK target site, and hence, EK cleaved the protein at more accessible sequences. Similar to AhR6-C/EBP, the His-tagged AhR6–Jun can homodimerize via the Jun leucine zipper coiled coil, and is prone to protein aggregation [118]. In addition, another possible explanation for poor proteolysis could be that the EK active site is blocked due to interaction with the 6×His tag from the dimer partner.

Therefore, we added urea to the cleavage reactions to test if the presence of denaturant would once again enhance the specificity of the EK proteolysis reaction and reduce adventitious cleavage. Indeed, addition of urea produced a similar outcome in the cleavage pattern of our AhR6-Jun reactions (Fig. 3.4), and additional proteolytic cleavage at adventitious sites was inhibited. With AhR6-Jun, the denaturant began to inhibit EK activity at 3 M urea, as indicated by the overall reduced amounts of cleavage products.

Similarly, we then incubated AhR6–Jun with EK in reaction buffer containing 2 M urea for 4 h (Fig. 3.5). Comparison of these results with those obtained earlier without urea (Fig. 3.3) demonstrates the much slower disappearance of the cleavage fragments, indicating that addition
of denaturant to the EK reaction buffer can be a general strategy for discouraging protein structure that shields the intended cleavage site from EK recognition.

**Figure 3.4** | EK cleavage reactions of AhR6–Jun under varying urea concentrations (top, x-axis). All reactions were stopped after 30 min incubation with EK. Arrow A indicates the intact His-tagged AhR6-Jun, and arrow B indicates a cleavage product (molecular mass not identified). Molecular-weight marker lane is at the left (MW, see ref. [138]), and apparent molecular weights are indicated in the y-axis legend. This SDS-PAGE analysis shows that addition of urea to EK cleavage reactions of AhR6-Jun inhibits additional proteolytic cleavage at adventitious sites, as shown by the greater presence of the cleavage fragment indicated by arrow B in comparison with other smaller cleavage products.

**Figure 3.5** | Time-course analysis of cleavage products after incubation of AhR6–Jun with EK with 2 M urea for the times indicated (top, x-axis, in hours). Molecular-weight marker lane is at the left (MW, see...
ref. [138]), and apparent molecular weights are indicated in the y-axis legend. This SDS-PAGE analysis demonstrates the much slower disappearance of cleavage fragments in comparison with results obtained earlier (Fig. 3.3) upon addition of urea to the reaction buffer.

### 3.4.3 Addition of Fos zipper inhibits EK activity

We next investigated whether we could alternatively improve the specificity of the EK reaction by using the synthetic Fos zipper to heterodimerize with AhR6-Jun protein (Fig. 3.1). We first tried finding the optimal concentration of Fos zipper necessary to promote an accessible structure for more efficient cleavage. Hence, we prepared EK reactions containing varying molar ratios of AhR6-Jun to Fos zipper (Fig. 3.6).

Preliminary experiments demonstrated that the activity of EK was strongly inhibited, as indicated by the absence of cleavage products. In fact, with 1:5 molar ratio of AhR6-Jun:Fos zipper, no cleavage fragments were observed even after extended incubation (~24 hrs) or when twice the original EK amount was added. Similarly, EK activity was also abolished when we included an equal amount of Fos zipper (i.e. 1:5 molar ratio of protein to zipper) in the typical EK reaction of AhR6-C/EBP protein that we studied earlier (data not shown). Thus, this effect was not substrate-dependent, but appeared to affect EK activity only.

One possible explanation for this observation includes the inhibitory effect of Fos zipper on the catalytic pocket of EK. As reported by Lu et al. in their structural studies of enterokinase, the major determinant of target recognition for EK is Lys99 [142]. Sadler and coworkers previously demonstrated that substitution of this positively charged residue by alanine abolished the ability of EK to recognize and cleave at its target site. In light of the fact that the Fos zipper we use in this study bears a high concentration of net negative charge at the pH used in the EK reaction (see Materials and methods), we hypothesize that addition of high concentrations of Fos
zipper to our EK reactions, causes shielding of Lys99 via electrostatic interactions, resulting in inhibition of EK activity.

Figure 3.6 | EK cleavage reactions of AhR6–Jun with varying concentrations of Fos zipper and under different experimental conditions. Arrow A indicates the intact His-tagged AhR6-Jun, and arrow B indicates the Fos zipper peptide. Reactions 1-3 and 5 were stopped after 2 hrs incubation with EK, whereas reaction 4 was stopped after 24 hrs incubation with EK. Lane 1, typical EK reaction (see Materials and methods) with 1:3 molar ratio of AhR6-Jun to Fos zipper; lane 2, typical EK reaction with 1:4 molar ratio of AhR6-Jun to Fos zipper; lanes 3-4, typical EK reaction with 1:5 molar ratio of AhR6-Jun to Fos zipper; lane 5, typical EK reaction with 1:5 molar ratio of AhR6-Jun to Fos zipper and with twice the original EK amount. This SDS-PAGE analysis demonstrates the strong inhibition of EK activity upon addition of Fos zipper to the reaction buffer, as indicated by the absence of cleavage products.

Therefore, work remains to verify if an accessible structure for a more efficient cleavage reaction can be promoted using alternative combinations of substrates and zipper proteins. As an example, a C/EBP leucine zipper corresponding to the leucine zipper region from AhR6-C/EBP (30 residues, LEQKVLELTSNDRLRKRVEQSLRSLTY) can be a suitable candidate, and we previously reported that our AhR6-C/EBP can also homodimerize via the C/EBP leucine zipper
The C/EBP zipper does not carry a high concentration of net negative charge at the pH used in the EK reaction; thus, it is not expected to have an inhibitory effect on EK activity similar to that of the Fos zipper. So, by introducing an excess of a C/EBP zipper to dimerize with AhR6-C/EBP, we might be able to nucleate a more properly folded, nonaggregated structure where the enterokinase cleavage site would be more accessible to protease cleavage.
Chapter Four

A fluorescent protein-based bioprobe for detection of protein:DNA recognition \textit{in vivo}

S. Hesam Shahravan, Isaac T.S. Li, Kevin Truong, and Jumi A. Shin

\textbf{Contributions:} S.H.S. and J.A.S. designed and interpreted experiments, and S.H.S. performed experiments and analyzed data. I.T.S.L. conceived the idea of the bioprobe reporter. K.T. contributed to the design of the project. S.H.S. and J.A.S. prepared the manuscript.

\textbf{Note:} A more detailed description of Methods section in this chapter and some background information on the theory of fluorescence resonance energy transfer are provided in Appendix C.
4.1 Abstract

We describe a bacterial bioprobe system for rapid and facile detection of protein:DNA recognition \textit{in vivo}. The bioprobe reporter comprises the genes of two fluorescent proteins (FPs) separated by a potential DNA target site. Upon translation of the bioprobe gene, fluorescence resonance energy transfer (FRET) can occur between the two FPs. However, if a co-expressed transcription factor binds to the DNA target, transcription of the second FP is impeded, resulting in loss of FRET partner. Using ratiometric FRET, we show that qualitative evaluation of protein:DNA recognition can be reliably made on transcription factors from the bZIP and bHLHZ families and their DNA target sites. The bioprobe is capable of detecting a wide range of protein:DNA binding affinities comparable to well-established electrophoretic assays and yeast genetic assays. Our system can be applied to high-throughput searches of both protein and DNA libraries; binding and nonbinding complexes can be distinguished by visual inspection of bacterial colonies on culture plates. Our bioprobe presents a number of advantages over existing techniques when applied to understanding protein:DNA structure-function relationships, assaying protein/protein interactions, and screening small molecules that influence protein:DNA or protein/protein recognition.
4.2 Introduction

Protein:DNA interactions can be explored by various in vitro and in vivo strategies, which present different advantages and disadvantages. In vitro methods provide better quantitative characterization but require isolation of active, soluble protein, which can be challenging and impractical in high-throughput assays. Additionally, protein function may depend strongly on assay conditions. Hence, a non-native in vitro environment can give rise to results contradictory to those performed in an in vivo assay [85]. Alternatively, in vivo assays provide a native-like environment for studying the protein:DNA interaction. For instance, yeast reporter assays [92, 93, 100, 101] are widely used and offer the capability to screen libraries, but suffer from elevated rates of false positives and false negatives [110, 111, 114]—as high as 45% was estimated in a yeast two-hybrid (Y2H) mapping of protein interactions in C. elegans [143]. Chromatin immunoprecipitation (ChIP) is also useful, but requires significant time and efficient antibodies for immunoprecipitation [45]. These drawbacks can hinder performing large-scale assays, such as mapping entire protein:DNA interactomes of large-genome organisms and identification of transcriptional regulatory elements in the human genome [1, 144, 145].

We therefore sought to develop a system for high-throughput detection of in vivo protein:DNA recognition capable of searching both protein and DNA libraries. Our bacterial system contains two separate vectors (Fig. 4.1): a transcription-factor (TF) expression vector, and a bioprobe reporter vector that contains a potential TF target site flanked by the genes encoding Cerulean (a cyan fluorescent protein, CFP) and Venus (a yellow fluorescent protein, YFP) [146, 147]. If the TF recognizes the DNA target, transcription of the second FP gene, YFP, is impeded by bound TF, yet the first FP gene, CFP, will always be transcribed. Full-length transcription of the bioprobe ensues when no protein:DNA recognition occurs. Additionally,
detection of signal from expressed CFP serves as an intrinsic reference for cell viability and successful protein expression for every sample.

Figure 4.1 | The bioprobe system for detection of protein:DNA recognition in vivo. a, Our system comprises an *E. coli* host cell that contains the expression vectors for a transcription factor (TF) and a bioprobe reporter (R). The bioprobe contains the genes for two fluorescent proteins (CFP and YFP) flanking a DNA target site, as indicated by the curved arrow. b, TF recognizes the DNA target and impedes transcription of YFP. c, Full transcription of the bioprobe results when no protein:DNA recognition occurs.

For each analysis, we measure whole-cell fluorescence of *E. coli* cells expressing both the TF and bioprobe and calculate a fluorescence resonance energy transfer (FRET) ratio defined as the ratio of YFP maximum emission intensity at 528 nm over that of CFP at 475 nm (see Appendix C for information on the theory of FRET). A change in this ratio, when compared to the FRET ratio of a reference sample expressing the same bioprobe but no TF, is taken as an assessment of protein:DNA recognition in live cells.
We used our bioprobe to evaluate qualitative protein:DNA binding affinities for two cases: (1) a single protein was assayed against different DNA sequences, and (2) different proteins were assayed against a single DNA sequence. We then tested the bioprobe's utility as a means for *in vivo* identification of specific protein:DNA interactions from a library.

### 4.3 Methods

#### 4.3.1 Preparation of bioprobe

In order to prepare the CFP-6×AP1-YFP bioprobe, we first constructed plasmid CFP-3×AP1-YFP that contains the genes for CFP and YFP separated by 3 tandem copies of the AP-1 site in pTriEx-1.1 Hygro (Novagen). This plasmid contains a base cassette that allows for further expansion of number of DNA sites (see ref. [148] for more information on this technique). Site 1 was fixed as the restriction site for *Nco*I, site 2a as *Spe*I, site 2b as *Nhe*I, and site 3 as *Xho*I. The CFP-3×AP1-YFP base cassette was created by C-terminal extension of a CFP gene fragment using the following primers: forward, 5´-TACCATGGGCCCTGACTAGTGGATC-3´, and reverse, 5´-AGGCTAGCTGAGTCATGAGTCACTGAGTCACCACCACTAGTCTTGTACAGCTCGT-3´ to yield the CFP-3xAP1 fragment, which was digested with *Nco*I and *Nhe*I and subcloned into a pTriEx-1.1 Hygro vector containing the gene for YFP (referred to as pCfvtx in ref. [148]).

To create the CFP-6×AP1-YFP bioprobe, we digested the CFP-3×AP1-YFP plasmid twice; first at the *Spe*I and *Xho*I sites to obtain the 3×AP1-YFP insert, and then at the *Nhe*I and *Xho*I sites to obtain the pTriEx-1.1 Hygro CFP-3×AP1 host vector. The two desired DNA fragments were purified by agarose gel electrophoresis and ligated to generate the CFP-6xAP1-YFP plasmid (Supplementary Fig. 4.3). This plasmid was chemically transformed into *E. coli*
strain DH5α (Invitrogen); plasmid DNA was extracted (Wizard® Plus SV Miniprep, Promega) and sequenced (DNA Sequencing Facility, Hospital for Sick Children, Toronto, ON). The CFP-2×NS-YFP bioprobe control was similarly constructed except for the reverse primer used for construction of its CFP-NS-YFP base cassette: 5′-AGGCTAGCCGAACCTCAGGAAATTCCTCCTGCACTAAGTTGTTACAGCTCGT-3′ (see Supplementary Tables 4.1 and 4.2 for the entire list of the reverse primers used in this study).

The wt bZIP, \textit{DPAALKRANTEAARRSRARKLQRMKQLEQKVLELTSNDRLRKRVEQLSRELDL}, and C/EBP bZIP, \textit{DPSEYRVRRENIAVRKSRDKQRMKQLEQKVLELTSNDRLRKRVEQLSRELDL} (note that the basic region sections are italicized and underlined), were prepared by cloning their corresponding genes between the \textit{Bam}HI and \textit{Eco}RI restriction sites of pET28a(+) expression vector (Novagen) and sequenced [122]. The set of Max proteins were similarly prepared in pET28a(+) [149, 150].

### 4.3.2 Expression of bioprobe

Electrocompetent \textit{E. coli} BL21(DE3)pLysS cells (Stratagene) were first transformed with the pET28a(+) transcription-factor expression vector, followed by transformation with the bioprobe reporter vector, and transferred onto a Luria-Bertani (LB) plate containing 30 µg/mL kanamycin and 50 µg/mL ampicillin for overnight incubation at 37 °C. For each transformation, a single colony was chosen for highest intensity by observing YFP fluorescence on the plate using a Lightools Illumatool LT-9900 imaging system equipped with a 488/10 nm excitation filter and 530 nm viewing filter. Each chosen colony was grown in LB media containing the same amounts
of antibiotic above. Protein expression was induced at OD$_{600}$ 0.6-0.7 with addition of 1 mM isopropylthiogalactoside (IPTG); growth was continued overnight at 37 °C (10-16 hr; a time-course experiment showed that 10 hr was sufficient).

4.3.3 Fluorescence measurement

Steady-state fluorescence measurements were performed on a Perkin Elmer LS 50B fluorimeter. Whole-cell fluorescence of each sample (1 mL) in LB medium was directly measured with no pre-measurement treatment (i.e. no washing). Each sample was excited at 440 nm and emission scanned in the range of 460-560 nm, with excitation and emission slits set at 10 and 2.5 nm, respectively. In light of the need to make spectral corrections as discussed by Gordon et al. in their studies using similar steady-state sensitized emission assays [151], three spectral corrections were applied to each raw fluorescence spectrum before FRET ratio calculation: (1) for each sample, a background spectrum, defined as the spectrum obtained from the same sample before protein induction, was subtracted to eliminate any fluorescence contribution from LB and E. coli cells; (2) to remove CFP spectral bleed-through from YFP intensity, a spectrum originally obtained from a sample containing only CFP was normalized with respect to maximum CFP intensity measured for each sample and subtracted from the same sample’s raw fluorescence spectrum; (3) to remove fluorescence contribution from direct-excitation of YFP, a spectrum originally obtained from a sample containing only YFP was normalized with respect to maximum YFP intensity measured for each sample and subtracted from the same sample’s raw fluorescence spectrum.
4.3.4 SDS-PAGE analysis of bioprobe

Four BL21(DE3)pLysS cell cultures were grown overnight with IPTG induction: (1) cells containing both the CFP-6×AP1-YFP bioprobe and pET28a(+) vector for expression of the wt bZIP, (2) cells containing both the CFP-6×AP1-YFP bioprobe and pET28a(+) vector for expression of the C/EBP bZIP, (3) cells containing both the CFP-2×NS-YFP bioprobe and pET-28a(+) vector for expression of the wt bZIP, and (4) cells containing both the CFP-6×AP1-YFP bioprobe and pET-28a(+) vector with no bZIP gene inserted (expresses 58-mer from original pET28a(+))

MGSSHHHHHHSSGLVPRGSHMASMTGGQOMGRGSEFELRRQACGRTRAPPPIPPLRSGC,

MW=6239.1 Da). Each 20 mL cell culture was harvested and re-suspended in 500 μL lysis buffer (20 mM Tris-HCl, 0.3 M NaCl, 5 mM β-mercaptoethanol, 8% glycerol, 6 M urea, pH 8.0). The cells were sonicated for 40 s, and the mixture centrifuged at 13,400 g for 10 min. 10 μL supernatant was mixed with 10 μL SDS buffer (50 mM Tris-HCl, 100 mM DTT, 2% SDS, 0.1% bromophenol blue, 10% glycerol, pH 6.8) without heating and resolved by 12% SDS-PAGE, followed by visualization under illumination by our Lightools fluoroimaging system (once with a 440/20 nm excitation filter and 475 nm viewing filter for CFP emission, and once with a 488/10 nm excitation filter and 530 nm viewing filter for YFP emission) and staining by Coomassie brilliant blue solution containing 50% methanol.

4.3.5 Screening of DNA binders

A 1:1 mixture of CFP-6×AP1-YFP bioprobe cells expressing either the wt bZIP or C/EBP bZIP was prepared by mixing equal amounts of BL21(DE3)pLysS cells containing the CFP-6×AP1-YFP bioprobe and either the wt bZIP protein expression vector or the C/EBP bZIP protein
expression vector. This mixture was transferred onto a LB plate (30 µg/mL kanamycin, 50 µg/mL ampicillin, 1 mM IPTG) for overnight incubation at 37 °C (10-16 hrs). After incubation, each plate was photographed while under illumination (Lightools fluoroimaging system, 488/10 nm excitation filter, 530 nm viewing filter) before analysis using a Nikon Eclipse L150 fluorescence microscope (ND16 filter, 4X objective) equipped with a laser light source (Coherent 406 nm/25 mW, z405/20x excitation filter) and dichroic 450 nm long-pass filter (z405 RDC) and attached to a diode array detector (Ocean Optic CCD QE65000). Colonies displaying appropriate fluorescence were identified and grown in LB media (30 µg/mL kanamycin, 50 µg/mL ampicillin) overnight at 37 °C. Cells were pelleted and plasmid DNA was extracted. The identity of each selected colony was revealed by restriction digestion of plasmids followed by agarose gel electrophoretic separation. Plasmids containing the gene for wt bZIP were cleaved by both PstI and MluI restriction endonucleases, whereas plasmids containing the C/EBP bZIP gene were only linearized under the same conditions, as this plasmid does not contain the PstI restriction site.

4.4 Results

4.4.1 Detecting protein:DNA recognition in vivo

In order to test our system's capability to measure protein:DNA recognition in vivo, we first applied it to the well-studied interaction between the GCN4 bZIP and its cognate binding site AP-1 (5´-TGACTCA) [19-21]. The dimeric bZIP (basic region/leucine zipper) comprises monomers of approximately 60 residues; each monomer uses a basic region for DNA-binding function and a leucine zipper for dimerization via a coiled-coil structure. We used our previously constructed wt bZIP (wild-type), a derivative of the GCN4 bZIP; our wt bZIP comprises the
basic region of GCN4 (residues 226-254) and leucine zipper from C/EBP (residues 312-338), and binds to the AP-1 target with high specificity and affinity (Table 4.1) [152, 153]. The wt bZIP maintains α-helical structure and DNA-binding function comparable to the native GCN4 bZIP [152, 154].

The bioprobe reporter vector was created by insertion of six tandem repeats of the AP-1 site between the genes of CFP and YFP. This vector was transformed into an *E. coli* strain containing a separate TF-expressing vector in which the gene for the wt bZIP was inserted. A reference system was similarly created except that the TF-expressing vector remained the original pET28a(+) vector, which expresses a control 58-residue protein. Therefore, both cell systems expressed the CFP-6×AP1-YFP bioprobe, but only one also expressed the wt bZIP. Hence, any change in the observed FRET ratios between the two samples could be correlated with wt bZIP activity in the cells.

From overnight cultures, we measured whole-cell fluorescence of each sample and calculated a FRET ratio after applying the necessary spectral corrections (see Methods section). We then calculated the change in FRET ratios between the sample in question vs. the reference system: for the wt bZIP:AP-1 pair, the change in FRET ratios was -52±3% (Fig. 4.2a and Table 4.1). This dramatic change in FRET ratios demonstrated that the bioprobe produces strong and reliable signals.

4.4.2 Specificity of the wt bZIP:AP-1 interaction

We performed two control experiments to show that both the wt bZIP transcription factor and AP-1 DNA target are responsible for the 52% drop in FRET ratios. First, we replaced the wt bZIP with the C/EBP bZIP, which does not bind to the AP-1 sequence [18], and would be
expected to give no change in FRET ratios; the wt bZIP and C/EBP bZIP share the same C/EBP leucine zipper, but possess different basic regions, which determine DNA sequence specificity (see Methods section). Thus, the C/EBP bZIP is an ideal control to assess whether DNA-binding activity by the TF is necessary for signal detection. Indeed, co-expression of the C/EBP bZIP in the bioprobe resulted in a clean 0% change in FRET ratios (Fig. 4.2b and Table 4.1). Second, we designed the CFP-2×NS-YFP bioprobe, in which two copies of a nonspecific (NS) DNA sequence replaced the six tandem AP-1 repeats; we previously showed that NS is not targeted by the wt bZIP [153]. Similarly, we observed no drop in FRET ratios (Fig. 4.2c and Table 4.1).

**Figure 4.2** | Specificity of the bioprobe. Representative emission spectra (spectrally corrected, see Methods section) were obtained by measuring whole-cell fluorescence of *E. coli* cells expressing either both TF and bioprobe (blue solid line) or bioprobe only (red dotted line) for the following protein:DNA complexes: a, wt bZIP and 6×AP-1; b, C/EBP bZIP and 6×AP-1; c, wt bZIP and nonspecific DNA.

These two controls indicated that the specific combination of protein and DNA sequence was necessary to achieve the change in detected signal. Furthermore, we tested the C/EBP bZIP with its cognate C/EBP target sequence (CFP-6×C/EBP-YFP). This protein:DNA complex also has a low nM $K_d$ value (~3 nM [155]) and showed a 54±9% drop in FRET ratios, similar to the
wt bZIP:AP-1 complex (Table 4.1). Thus, the bioprobe consistently detected another strong protein:DNA interaction.

<table>
<thead>
<tr>
<th>DNA target</th>
<th>DNA sequence</th>
<th>Binding protein</th>
<th>$K_d$ (nM)</th>
<th>Change in FRET ratios (%$^c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$6 \times d$</td>
<td>$1 \times d$</td>
</tr>
<tr>
<td>AP-1</td>
<td>TGACTCA</td>
<td>wt bZIP</td>
<td>9$^{e,f}$, 13$^{g,h}$</td>
<td>- $52 \pm 3$ $^i$</td>
</tr>
<tr>
<td>C/EBP</td>
<td>TTGCGCAA</td>
<td>wt bZIP</td>
<td>120$^{g,h}$</td>
<td>- $48 \pm 3$</td>
</tr>
<tr>
<td>Arnt E-box</td>
<td>TCACGTGA</td>
<td>wt bZIP</td>
<td>570$^{g,l}$</td>
<td>- $12 \pm 5$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>wt bZIP</td>
<td>840$^{g,l}$</td>
<td>- $22 \pm 10$</td>
</tr>
<tr>
<td>HRE</td>
<td>GCACGTAG</td>
<td>wt bZIP</td>
<td>1400$^{g,l}$</td>
<td>+ $3 \pm 5$</td>
</tr>
<tr>
<td>AP-1</td>
<td>TGACTCA</td>
<td>C/EBP bZIP</td>
<td>3$^{g,l}$</td>
<td>- $54 \pm 9$</td>
</tr>
<tr>
<td>NS</td>
<td>ATGCAGAGAAGGAA</td>
<td>TTCCTGAAGGGTTTCG</td>
<td>wt bZIP</td>
<td>Nonbinding$^k$</td>
</tr>
<tr>
<td>C/EBP</td>
<td>TTGCGCAA</td>
<td>C/EBP bZIP</td>
<td>14$^{e,m}$</td>
<td>- $30 \pm 5$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>MaxbHLHZ</td>
<td>536$^{e,n}$</td>
<td>- $41 \pm 6$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>MaxFos</td>
<td>861$^{e,o}$</td>
<td>- $37 \pm 6$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>MMbHLH/ERY</td>
<td>32$^{o,p,q}$</td>
<td>- $32 \pm 16$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>Max1bHLH-C/EBP'</td>
<td>Nonbinding$^{o,p,q}$</td>
<td>- $6 \pm 11$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>MaxΔAS</td>
<td>Nonbinding$^{o,p}$</td>
<td>- $6 \pm 11$</td>
</tr>
<tr>
<td>Max E-box</td>
<td>CCACGTGG</td>
<td>MaxbHLH</td>
<td>Nonbinding$^{m,q}$</td>
<td>- $2 \pm 17$</td>
</tr>
</tbody>
</table>

$^a$ Indicates the DNA target inserted between the genes of CFP and YFP in the bioprobe construct. $^b$ The nucleotide sequence for a single copy of each DNA target. $^c$ Percent changes in FRET ratios are mean±s.d. from ≥ 3 independent samples. $^d$ Indicates the number of copies of each DNA target site in the bioprobe construct. $^e$ Dissociation constants of protein:DNA complexes obtained by fluorescence anisotropy with a single copy of each DNA target site. $^f$ Previously published in ref. [152]. $^g$ Dissociation constants of protein:DNA complexes obtained by quantitative EMSA with a single copy of each DNA target site. $^h$ Previously published in ref. [153], assayed by EMSA. $^i$ Similar drop in FRET ratios was observed with a 3×AP1/NS bioprobe (see Supplementary Information for discussion of CFP-3×AP1/NS-YFP). $^j$ Previously published in ref. [156]. $^k$ Previously published in ref. [18], assayed by DNase I footprinting. $^l$ Previously published in ref. [155]. $^m$ Previously published in ref. [149]. $^n$ Unpublished results. $^o$ Previously published in ref. [150]. $^p$ Assayed by fluorescence anisotropy. $^q$ Assayed by Y1H. $^r$ See Supplementary Information for more discussion of binding activity.
4.4.3 Protein:DNA recognition blocks transcription of YFP

Our hypothesis was that a protein, even a small bZIP domain, bound to a specific DNA target is sufficient to block the *E. coli* RNA polymerase machinery from completing transcription of the full-length bioprobe gene. Specifically, transcription of the YFP gene would be impeded, resulting in a mixture of full-length (CFP-linker-YFP) and truncated (CFP only) transcripts (Fig. 4.1). Complete blockage of transcription will not occur, as TFs are not covalently bound at the DNA target. Because there are numerous studies demonstrating varying ability of RNA polymerases to transcribe through a protein:DNA complex, with only a few studies discussing possible reasons for this ability (reviewed in refs. [157-159]), we attempted to isolate either transcripts or expressed protein fragments corresponding to the truncated and full-length bioprobe proteins that could give rise to the ratiometric FRET data and provide an orthogonal means for validation of the bioprobe design.

We performed SDS-PAGE analysis because FPs retain some fluorescence even after SDS-PAGE (samples cannot be heated). We expected only the full-length bioprobe protein containing both CFP and YFP if no wt bZIP:AP-1 interaction occurred, that is, no expression of the wt bZIP. In the presence of the wt bZIP, there should also exist a second, smaller protein fragment comprising only CFP.

We clearly visualized the two expected protein fragments in the sample where the wt bZIP was expressed (Fig. 4.3). We first illuminated the gel under 440 nm light, which directly excites any protein fragment containing CFP; thus, we visualized all bioprobe protein fragments, and the bioprobe-only control shows the full-length protein (Fig. 4.3, arrow I). Using a 475 nm viewing filter to capture CFP emission, only for cells expressing the wt bZIP did we observe the second,
Figure 4.3 | SDS-PAGE analysis of transcriptional blockage by protein:DNA recognition. Panels a, b, and c are the same SDS-PAGE shown under different conditions. Arrow I indicates the full-length bioprobe construct (theoretical mass: 56.0 kDa), arrow II indicates the CFP-truncated bioprobe (approximately 27 kDa, depends on the exact location of transcription blockage), and arrow III indicates the transcription factors (wt bZIP and C/EBP bZIP are 11.2 and 11.7 kDa, respectively). Lane 1: MW marker (Precision Plus Proteins WesternC™, Bio-Rad) with apparent molecular weights indicated on y-axis. Lane 2: 6×AP1 bioprobe with no TF expressed. Lane 3: 2×NS bioprobe with wt bZIP expressed.
Lane 4: 6×AP1 with C/EBP bZIP expressed. Lane 5: 6×AP1 bioprobe with wt bZIP expressed. Lane 6: Cerulean fluorescent protein control. a, Illumination under 440/20 nm light, visualization through a 475 nm viewing filter for CFP emission. b, Illumination under 488/10 nm light, visualization through a 530 nm viewing filter for YFP emission. c, Coomassie brilliant blue staining exposes all protein fragments. Note the discrepancy between the apparent molecular weights of the MW markers and full-length bioprobe constructs; because samples were not heated before loading, proteins can remain folded in a native state, with SDS coating the folded structure's exterior. Observation of fluorescence from the bioprobe reporter can only occur if CFP and YFP are properly folded.

smaller fragment (Fig. 4.3a, arrow II). When this same gel was then visualized for YFP emission through a 530 nm filter, this smaller fragment disappeared after the excitation light was switched to 488 nm (Fig. 4.3b). Therefore, the lower molecular weight fragment contained only CFP, and this protein's migration corresponded with that of the slightly longer CFP-only control in the neighboring lane. The SDS-PAGE analysis confirmed that the wt bZIP:AP-1 binding event causes a transcriptional blockage of YFP, and that the observed changes in FRET ratios are due to the presence or absence of YFP in the expressed constructs.

4.4.4 The bioprobe can detect a clear signal from one DNA target site

We next considered whether multiple repeats of a DNA target site are necessary for reliable detection of a protein:DNA interaction. If a clear, reproducible ratiometric FRET signal from a bioprobe containing a single copy of a DNA target can still be detected, then the utility of the bioprobe can be extended to screening DNA libraries against a desired protein: i.e. random duplexes from a cDNA library will not contain 6 repeated sites. In the CFP-6×AP1-YFP bioprobe, it is possible for more than one AP-1 site to be occupied simultaneously by bound bZIP, and multiple proteins would increase steric blockage, preventing RNA polymerase from completing its transcription activities. In the context of our system, increasing the number of
DNA targets is useful when screening a protein library against a desired DNA sequence. We emphasize that the reason multiple DNA sites are used in reporter assays, such as the yeast one-hybrid (Y1H), is not because proteins will occupy all sites simultaneously, but to increase the probability of protein binding; one bound protein is sufficient for reporter activation.

We created CFP-1×AP1/NS-YFP, where NS is used to maintain the same overall linker length between CFP and YFP as that for the 6×AP1 construct; we found a clear, reproducible 23±2% decrease in FRET ratios (Table 4.1). Therefore, a single strong protein:DNA interaction can be reliably detected. In every bioprobe, we used the NS sequence to preserve the same number of nucleotides as that for the 6×AP-1 construct (84 base pairs) between the genes of CFP and YFP. An appropriate linker length between the two fluorophores must be achieved for efficient ratiometric FRET measurements (see Supplementary Information and Supplementary Fig. 4.1 for further discussion of linker lengths), and a positive control protein is useful to ensure that the protein:DNA complex can be detected.

4.4.5 Detection of a wide range of protein:DNA binding affinities

The bioprobe is capable of detecting a strong, reproducible signal from a single DNA target site. We therefore considered its range of detection; that is, the wt bZIP:AP-1 interaction is a strong TF:DNA complex, but can weaker interactions be detected as well? Whether screening a protein or DNA library, a tremendous utility would be the ability to detect not just the strong binders, but moderate and even weak binders, which can serve as templates for further rational design and directed evolution. Detection of strong interactions can be done with yeast reporter assays, but signals from weaker interactions can be obscured by background and a challenge to validate. A
real demonstration of the strength of our system would be reliable detection of weaker protein:DNA interactions.

We first tested a series of wt bZIP:DNA complexes for which we had previously measured $K_d$ values ranging from 120 to 1400 nM (Table 4.1) [153, 156]. Each $K_d$ value was measured with a single DNA target site and compared with bioprobe measurements obtained from a single copy of a DNA target, as well as with 6 tandem repeats of DNA target to investigate how target number affects the bioprobe’s limit of detection. Although the $K_d$ for the wt bZIP:C/EBP DNA complex (120 nM) is 10-fold weaker than that with the AP-1 site [153], 6 copies and 1 copy of the more weakly bound C/EBP site gave decreases in FRET ratios identical to those measured with the 6× and 1× versions of the AP-1 bioprobe (Table 4.1). Biopropes containing the very weakly bound Arnt E-box or Max E-box site demonstrate that high nM $K_d$ values lie at the threshold of detection. In all cases, smaller drops in FRET ratios from the 1× vs. 6× versions of each bioprobe were observed, and larger standard deviations were measured from weaker complexes.

We then tested the bioprobe's utility toward screening a protein library, that is, a fixed DNA target site screened against a library of proteins. Similar to the experiments above, we previously measured $K_d$ values of Max derivatives bound to a single Max E-box target site (Table 4.1) [149, 150]. Max belongs to the basic-region/helix-loop-helix/leucine zipper (bHLHZ) family of transcription factors and can homodimerize and bind to the Max E-box [10]. The MaxbHLHZ:Max-Ebox complex has a low nM $K_d$ value (14 nM [149]) and showed a 30±5% drop in FRET ratios with the 6×MaxEbox bioprobe. MMbHLH/ERY and MaxFos show that high nM $K_d$ values are reliably detected with the 6×MaxEbox bioprobe. Interestingly, in previous
studies, we found Max1bHLH-C/EBP to be an extremely weak binder of the Max E-box [150], yet with the 6×MaxEbox bioprobe, we observed a clear decrease in FRET ratios (32±16%).

Both sets of experiments above, whether fixed wt bZIP protein tested against a series of DNA sites (1× or 6× targets) or fixed Max E-box DNA sites (CFP-6×MaxEbox-YFP) tested against a series of Max mutants, confirm that a wide range of protein:DNA binding affinities, from low to high nM $K_d$ values, can be detected (see Supplementary Information for more discussion of range of detection). In particular, as shown with members of the bZIP and bHLHZ families, the threshold for reliable detection in the 6× versions of the bioprobe appears to be around 800 nM $K_d$, and for 1× versions, up to 500 nM $K_d$. We note that similar to other in vivo techniques, the thresholds of the bioprobe system are not absolute and depend on the specific protein:DNA pairs tested. For example, we observed variations in the intensity of fluorescent signals and detection thresholds that may depend on differences between bZIP vs. bHLHZ expression and stability in the cell or the expressed bioprobe linker between the two FPs (see Supplementary Information and Supplementary Fig. 4.2 for more discussion of factors influencing the range of detection).

4.4.6 The bioprobe provides a facile means for screening libraries

We next considered whether our system is capable of high-throughput screening of protein:DNA complexes, in which case, a simple method would be to conduct the assay on plates where individual colonies of cells displaying appropriate fluorescence could be easily distinguished. We designed a mock library screen in which a 1:1 mixture of CFP-6×AP1-YFP bioprobe cells expressing either the wt bZIP or C/EBP bZIP was spread on a culture plate. After overnight
growth, fluoroimaging revealed that some colonies were significantly brighter, indicating the likelihood of more YFP (Fig. 4.4).

By eye, we chose five colonies displaying high YFP intensity and presumed to express the nonbinding C/EBP bZIP, and five weakly fluorescent colonies presumed to express the strong binder wt bZIP. Restriction analysis of plasmids confirmed that colonies containing the wt bZIP could be separated from those containing the C/EBP bZIP with 90% success (data not shown).

Figure 4.4 | Mock library screening of two transcription factors. A culture plate containing a 1:1 mixture of CFP-6×AP1-YFP bioprobe cells expressing either wt bZIP or C/EBP bZIP is shown under illumination by a 488/10 nm light and visualized through a 530 nm viewing filter for YFP emission. For each transcription factor, five colonies were chosen based on YFP emission intensity, as indicated by circles and squares for wt bZIP and C/EBP bZIP, respectively.
To gain quantitative measurements from this plate assay, we used a fluorescence microscope to allow calculation of in-colony FRET ratios. We chose five colonies for each of the two predicted TFs, with comparison to the FRET ratio from a reference colony with no expressed TF. This time, all ten colonies were predicted correctly (see Appendix C Fig. C.1). Therefore, whether by eye or by obtaining FRET ratios, our system presents a facile and accurate means for high-throughput detection of protein:DNA complexes from colony growth on plates. This result points to the possibility that the bioprobe will find utility in screening libraries of protein:DNA complexes, which can be performed by eye, by using a fluorescence microscope, or by fluorescence activated cell sorting (FACS).

4.5 Discussion

We were inspired by the pioneering work of Fields and Song in their development of yeast reporter assays [92]. Likewise, the bioprobe is a genetic assay akin to the successful GFP-based protein-folding assay designed by Waldo et al. [160], but not a true reporter assay where a protein:DNA recognition event is visualized by the product of a downstream reporter gene. Thus, the design of the bioprobe system was intended to facilitate evaluation of in vivo protein:DNA interactions in ways that build on these existing technologies: (1) Internal protein expression control in every sample. Protein expression can be assessed by signal from CFP, and DNA binding from YFP, which can be normalized to that from CFP. In addition, the two FPs allow for ratiometric FRET measurements, which are intrinsically quantitative and less prone to artifacts from concentration differences, as compared with absolute emission intensity measurements [161]. (2) Accessibility and ease of manipulation. Bioprobe results are available within hours or overnight, and no SDS-PAGE or immunoblotting is required. Widely available bacteria strains and plasmids are used; measurement is performed on cell cultures with no special treatment.
Detection is performed on a standard fluorimeter. (3) Library screens can be performed by visual inspection of cell colonies growing on plates by using fluorescence imaging.

Our bioprobe design is different compared to other biosensor applications based on ratiometric FRET. In typical biosensor applications, the donor and acceptor are brought together either by intermolecular interactions or intramolecular conformational changes [162, 163]. Examples include protein/protein interactions or a Ca\(^{2+}\) sensor based on conformational changes in calmodulin upon ion binding [164, 165]. In these cases, both donor and acceptor are always present and depend on achieving a certain distance for efficient FRET; thus, detection is at the "protein stage." One drawback of protein biosensors is their limited FRET range; for example, conformational changes may not always be large enough to produce significant FRET changes. Another drawback is that the fusion FP can negatively affect the folding and activity of the sensor unit in protein biosensors; a major design obstacle encountered in the many failed attempts to create protein biosensors is that the sensor domain simply loses activity once fused to the FP. In contrast, the bioprobe signal arises from the "DNA stage" before transcription occurs: detection depends on just two species in the sample, the full-length and truncated bioprobe proteins. Hence, our system is not a true "biosensor" that is reversible and generates its own signal based on an event (e.g. increased FRET when FPs move closer), so we use the term "bioprobe."

A rough comparison with the Y1H, an assay we use routinely, indicates that the 6× versions of the bioprobe show limits of detection at least as good as that of the Y1H assay using multiple copies of a DNA target site. However, we have never tried a single DNA target in the Y1H. With 1× versions of the bioprobe, protein:DNA complexes up to 500 nM \(K_d\) lie at the threshold of detection, which is comparable to \textit{in vitro} techniques like fluorescence anisotropy,
footprinting, and EMSA. Yeast reporter assays can have high false-positive and negative rates [109, 110, 112, 113, 166] primarily due to their reliance on indirect readout by transcriptional activation of reporter genes [114]; we also find that many "hits" do not validate. Large-scale Y2H screens, such as those used in building interactome maps, can have false-positive rates as high as 64% and false-negative rates of 43-71% [167, 168]. Nonetheless, this assay maintains broad versatility and utility [103, 110, 116], as demonstrated by the observation that over 50% of protein/protein interactions published in Medline were obtained by yeast genetic assays [113, 169].

The nature of the bait molecule, which is a DNA target in the Y1H and a protein in the Y2H, is a major contributor to spurious signals; for example, broadly interactive “sticky” baits often yield high frequencies of false positive clones in yeast library screens, and baits that are harmful to the host can give false negatives. Huang and Bader estimated 10% and 51% bait-specific false-positive and false-negative rates, respectively, in their study of Y2H screens [112], and Serebriiski and Golemis found different bait proteins in the Y2H can produce differences in the rate of false positives observed [114]. In the Y1H assay, some bait DNA sequences display unusually high leaky expression, which are only uncovered by trial and error, and this poses the most significant constraint on the application of the Y1H screening tool [117].

A number of false positives originate from indirect metabolic effects in the Y2H assay, including changes in cell permeability, viability, growth rates and transcription rates [170]. In contrast, the bioprobe measurement derives from comparing FRET ratios from full-length and truncated bioprobe proteins between the sample and reference, so most indirect metabolic effects should not affect bioprobe measurements. False negatives in the Y2H can result from protein misfolding leading to loss of function [112], which can also occur in the bioprobe. Similarly,
testing for leaky *HIS3* expression and 3-AT inhibitor concentrations in yeast assays is comparable to testing linker length between CFP and YFP in the bioprobe assay. Thus, for any system, users must establish the parameters for conducting their own particular experiments to obtain clear signals, followed by validation using orthogonal techniques.

These reasons can explain why in most cases, with the bioprobe, we found good correlation between *in vivo* bioprobe FRET measurements and quantitative *in vitro* binding affinities for the various protein:DNA complexes tested; possibly this strong level of validation stems from the bioprobe's independence from a reporter-driven signal, and its dependence on a physical interaction between the protein and DNA target that blocks transcription. In the bioprobe, a toxic expressed TF would also harm the host, but no false signal would be observed: no colony growth, and hence, no fluorescence signal would result. Moreover, the length of the DNA target inserted as bait between CFP and YFP, rather than identity of its sequence, is the primary determinant for successful FRET ratio measurements, and thus, we recommend that a known positive control be performed (see Supplementary Information for further discussion of linker lengths and comparison of bioprobes as well as Appendix C for more discussion of other factors influencing the efficiency of FRET).

Compared with existing technologies, the bioprobe presents a number of advantages for augmenting our understanding of protein:DNA structure-function relationships including accuracy and sensitivity, as well as being economical and straightforward in implementation. Moreover, the bioprobe's utility can be extended to high-throughput screening of libraries of protein:DNA complexes, as well as assaying protein/protein recognition, with a protein:DNA interaction providing the detected signal (see ref. [171] for a similar design) or *in vivo* screening of small-molecule libraries for ability to disrupt protein:DNA or protein/protein interactions.
4.6 Supplementary Information

4.6.1 Supplementary Discussion

4.6.1.1 DNA target site number between CFP and YFP affects changes in FRET ratios

In addition to the CFP-1×AP1/NS-YFP bioprobe, we created a CFP-3×AP1/NS-YFP bioprobe, where NS is the same nonspecific DNA described in the main text. This construct maintains the same overall linker length between CFP and YFP as that for the 6×AP1 construct (84 base pairs): only the target-site number is decreased. The CFP-3×AP1/NS-YFP construct, with co-expressed wt bZIP, showed the identical drop in FRET ratios as the 6×AP-1 bioprobe (50±4% and 52±3%, respectively, Table 4.1). As discussed in the main text, the CFP-1×AP1/NS-YFP with co-expressed wt bZIP gave a 23±2% decrease in FRET ratios, which is still indicative of a clear, detectable, and reproducible signal.

Therefore, in this high affinity protein:DNA complex (wt bZIP:AP-1), 1 DNA target site gives an easily detected signal, and 3 DNA target sites gives essentially the maximum FRET drop we detect in the bioprobe, i.e. approximately 50% change in FRET ratios. Thus, more than 3 target sites will not provide an improvement in detection in the AP-1 bioprobe system.

4.6.1.2 An optimal linker length between CFP and YFP must be achieved for efficient ratiometric FRET measurements

For every bioprobe constructed, we always maintained the same number of nucleotides as that for the 6×AP-1 construct (84 base pairs) between the genes of CFP and YFP by using the NS sequence. When we constructed a 3×API bioprobe (51 base pairs, as opposed to 84 base pairs...
for 3×AP1/NS), SDS-PAGE analysis showed that although YFP transcription was impeded in the presence of wt bZIP, the FRET ratio data was inconsistent, indicating that an appropriate DNA linker length, and its corresponding translated protein linker length, between the two fluorophores must be achieved for efficient ratiometric FRET measurements.

Interestingly, we found that for bioprobes containing 9, 12, and 24 copies of the AP-1 target site (corresponding to 117, 150, and 282 base pairs, respectively, between the genes of CFP and YFP), there was no decrease in FRET ratios upon co-expression of the wt bZIP. These results were confirmed by SDS-PAGE, which revealed that YFP transcription was not blocked upon co-expression of wt bZIP; only the full-length bioprobe was produced, and no truncated, CFP-only protein was detected in both experiments (Supplementary Fig. 4.1b). In contrast, when similar bioprobes having the identical numbers of base pairs between the genes of CFP and YFP were made with 9, 12, and 24 copies of the Max E-box target site, YFP transcription was successfully blocked upon co-expression of MaxFos.

Two possible explanations for the unsuccessful blockage of transcription with 9×, 12×, and 24×AP-1 bioprobes, despite the presence of wt bZIP, include plasmid DNA adopting a conformation such that the AP-1 targets were not accessible to wt bZIP, and/or the protein:DNA complex successfully occurred, but the resultant structure was an insufficient block of the transcriptional apparatus.

We therefore conclude that an optimal linker length between the genes expressing CFP and YFP is around 84 base pairs, based on all the DNA target sites examined in this study. The ability of bioprobes, with longer lengths between the genes for CFP and YFP, to provide a sufficient block to YFP transcription appears to be sequence-specific and depend on linker
identity. For example, as discussed above, for the 9×, 12×, and 24× versions of the AP-1 bioprobe with co-expressed wt bZIP, YFP transcription was not blocked, as shown by SDS-PAGE and bioprobe measurements. However, for the 9×, 12×, and 24× versions of the MaxE-box bioprobe with co-expressed MaxFos, SDS-PAGE revealed successful blockage of transcription (Supplementary Fig. 4.1a). Hence, use of a positive control protein to ensure that a desired protein:DNA complex can be detected with a specific bioprobe is recommended.

4.6.1.3 Further discussion about the range of protein:DNA binding affinities that can be detected in the bioprobe

All of the $K_d$ values discussed in the main text were obtained by either quantitative electrophoretic mobility shift assay (EMSA) or fluorescence anisotropy, as measured on a single DNA target site (i.e. not 6 repeats). These binding affinities gained by in vitro methods provided a comparison for the range of detection available from the bioprobe in vivo.

We tested a series of wt bZIP:DNA complexes for which we had previously measured $K_d$ values ranging from 120 to 1400 nM (Table 4.1 and references therein). These $K_d$ values were obtained by EMSA. Both the 6×C/EBP bioprobe and 6×AP1 bioprobe expressed with the wt bZIP exhibited the same decreases in FRET ratios (48±3% and 52±3%, respectively), despite that the $K_d$ for the wt bZIP:C/EBP DNA complex (120 nM) is 10-fold weaker than that for the wt bZIP:AP-1 complex. Likewise, the 1×C/EBP/NS bioprobe gave a comparable drop in FRET ratios as did the 1×AP1/NS bioprobe (26±6% and 23±2%, respectively; Table 4.1). Thus, either 6 copies or 1 copy of the more weakly bound C/EBP site could be readily detected, so the bioprobe did not differentiate the binding affinities in this 10-fold range of $K_d$ values, essentially up to around 100-200 nM $K_d$. 
The 6× and 1× versions of the Arnt E-box expressed with the wt bZIP \( (K_d \ 570 \text{ nM}, \ 50\text{-fold}) \) gave decreases in FRET ratios of 12±5% and 5±4%, respectively (Table 4.1). Similar decreases were observed with the Max E-box \( (K_d \ 840 \text{ nM}, \ \text{flanking base pairs differ from that of the Arnt E-box}) \), and no change in FRET ratios was measured with 6×HRE \( (K_d \ 1400 \text{ nM}) \). Therefore, in the 6× versions of the bioprobe, up to around 500-800 nM \( K_d \) can be detected, and with the 1× versions, up to 500 nM \( K_d \) is the threshold.

In the protein library screening experiments with the CFP-6×MaxEbox-YFP bioprobe binding to a series of Max protein derivatives that we had previously analyzed by fluorescence anisotropy, the threshold for detection was comparable to that of the wt bZIP (Table 4.1). Note that the bZIP (e.g. wt bZIP and C/EBP) and bHLHZ (e.g. MaxbHLHZ and all the Max derivatives) are two different transcription factor families. The MaxbHLHZ:Max-Ebox complex has a 14 nM \( K_d \) value and showed a 30±5% drop in FRET ratios with the 6×MaxEbox bioprobe. MaxFos and MMbHLH/ERY both display weak binding affinities for the Max E-box target (536 and 861 nM, respectively). Yet with the 6×MaxEbox bioprobe, both gave readily detectable decreases in FRET ratios of 41±6% and 37±6%, respectively.

Another interesting observation is that in previous studies, we found Max1bHLH-C/EBP to be an extremely weak binder of the Max E-box: no binding with one Max E-box site was observed in fluorescence anisotropy, and only a weak signal was detected in the Y1H assay with 4 tandem Max E-box sites inserted upstream of the reporter gene. With the 6×MaxEbox bioprobe, however, we observed a clear decrease in FRET ratios (32±16%). We also had previously found that MaxΔAS and MaxbHLH were nonbinders of the Max E-box target; both gave inconsistent drops in FRET ratios of 6±11% and 2±17%, respectively, thereby proving to be nonbinders of the Max E-box, consistent with previous studies.
These measurements on members of the bZIP and bHLHZ families demonstrate that two different transcription factor families give comparable results in the bioprobe, and reinforce the conclusion that the bioprobe is capable of reliable detection among different protein:DNA systems.

The discrepancy in the FRET ratio data we observed between wt bZIP with 6×ArntEbox bioprobe and MaxFos with 6×MaxEbox bioprobe (both complexes have $K_d$ values ~ 550 nM, yet gave decreases in FRET ratios of 12±5% and 41±6%, respectively), prompted us to investigate what other important factors, aside from protein:DNA binding affinity, may have an effect on the bioprobe’s range of detection. We expected that differences between bZIP vs. bHLHZ expression and stability in the cell and/or the expressed bioprobe linker between the two FPs may be the main culprits behind our FRET ratio observations.

We therefore analyzed by SDS-PAGE the whole-cell lysates of the two samples mentioned above (Supplementary Fig. 4.2). Given the equal loading of cell lysate samples in lanes 3 and 5 (i.e. identical OD$_{600}$ values before lysis), not only was the amount of expressed full-length bioprobe construct for the ArntEbox bioprobe significantly greater than that of the MaxEbox bioprobe, as seen by the differences in intensity of the bands indicated by arrow I, but the amount of expressed wt bZIP and MaxFos proteins were also different (Supplementary Fig. 4.2, arrows II and III, respectively) with MaxFos expressed by more than ~55% over wt bZIP, as revealed by densitometric analysis.

Our observations indicated that the differences in translated protein linker and the variations in expression level of wt bZIP and MaxFos proteins are the likely reasons that the wt bZIP:Arnt E-box complex gives a 12% drop in FRET ratios in the 6×ArntEbox bioprobe, but
that the equally strong MaxFos:Max E-box complex gives a larger 41% drop in FRET ratios in the 6×MaxEbox bioprobe (Table 4.1). Similarly, this may be the same reason that weak DNA binders MaxFos and MMBHLH/ERY give a larger drop in FRET ratios than does the strong DNA binder MaxbHLHZ. Possibly, MaxbHLHZ is not expressed as well or as stable in the *E.coli* environment as the other two mutants.

4.6.2 Supplementary Tables

**Supplementary Table 4.1 | Reverse primers used for preparation of base cassette plasmids of 6× bioprobes**

<table>
<thead>
<tr>
<th>DNA target</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP-1</td>
<td>5´-AGGCTAGCTGAGTCATGACTGACACTGACACCACCACCTAGTTGTACAGCTCGT-3´</td>
</tr>
<tr>
<td>C/EBP</td>
<td>5´-AGGCTAGCTTGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>Arnt E-box</td>
<td>5´-AGGCTAGCTTGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>Max E-box</td>
<td>5´-AGGCTAGCAGTACAGGTACCTTTCTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>HRE</td>
<td>5´-AGGCTAGCTGAGTCATGACTGACACTGACACCACCACCTAGTTGTACAGCTCGT-3´</td>
</tr>
<tr>
<td>NS</td>
<td>5´-AGGCTAGCCGAACTTCTCAGGAATTCCTTTTCTCAGCTACAGCTCGT-3´</td>
</tr>
</tbody>
</table>

*a* Indicates the DNA target inserted between the genes of CFP and YFP in the bioprobe construct. *b* The nucleotide sequence for the reverse primer containing 3 copies of each DNA target site (base cassette plasmids contain 3 copies of the DNA target; larger bioprobe plasmids are built from this plasmid).

**Supplementary Table 4.2 | Reverse primers used for preparation of base cassette plasmids of 1× bioprobes**

<table>
<thead>
<tr>
<th>DNA target</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP-1</td>
<td>5´-AGGCTAGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>C/EBP</td>
<td>5´-AGGCTAGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>Arnt E-box</td>
<td>5´-AGGCTAGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
<tr>
<td>Max E-box</td>
<td>5´-AGGCTAGCAGTACAGGTACCTTTCTCTCTCTCTGCTACAGCTCGT-3´</td>
</tr>
</tbody>
</table>

*a* Indicates the DNA target inserted between the genes of CFP and YFP in the bioprobe construct. *b* The nucleotide sequence for the reverse primer containing a single copy of each DNA target site.
4.6.3 Supplementary Figures and Legends

Supplementary Figure 4.1 | The effect of linker length between CFP and YFP on transcription of the bioprobe reporter gene. Panel a represents one SDS-PAGE, and panel b represents another SDS-PAGE: top, Illumination under 440/20 nm light, visualization through a 475 nm viewing filter for CFP emission; bottom, Coomassie brilliant blue staining exposes all protein fragments. Arrow I indicates the full-length bioprobe construct (56.0 kDa), arrow II indicates the CFP-truncated bioprobe (approximately 27 kDa, depends on the exact location of transcription blockage), and arrow III indicates the transcription factors (MaxFos and wt bZIP, 11.6 and 11.2 kDa, respectively). a, Bioproses analyzed in lanes 2-5 were co-expressed with MaxFos protein. Lane 1: MW marker (Precision Plus Proteins WesternC™, Bio-Rad) with apparent molecular weights indicated on y-axis. Lane 2: 6×MaxEbox bioprobe. Lane 3: 9×MaxEbox bioprobe. Lane 4: 12×MaxEbox bioprobe. Lane 5: 24×MaxEbox bioprobe. Note that the CFP-truncated bioprobe band for the 6×MaxEbox sample in lane 2 is relatively faint and hence not clearly visualized on the SDS-PAGE. b, Lanes 1, 3, 5, and 7 show the full-length translated bioprobe construct controls with no TF expressed. Lane 1: 6×AP1 bioprobe. Lane 2: 6×AP1 bioprobe with wt bZIP expressed. Lane 3: 9×AP1 bioprobe. Lane 4: 9×AP1 bioprobe with wt bZIP expressed. Lane 5: 12×AP1 bioprobe. Lane 6: 12×AP1 bioprobe with wt bZIP expressed. Lane 7: 24×AP1 bioprobe. Lane 8: 24×AP1 bioprobe with wt
bZIP expressed. Note the discrepancy between the apparent molecular weights of the MW markers and full-length bioprobe constructs (see caption of Fig. 4.3).

**Supplementary Figure 4.2** | The effect of protein expression in the cell and the expressed bioprobe linker on the bioprobe’s range of detection. SDS-PAGE analysis of total cell lysate is shown. Arrow I indicates the full-length bioprobe construct (56.0 kDa), arrow II indicates the expressed wt bZIP protein (11.2 kDa), and arrow III indicates the expressed MaxFos protein (11.6 kDa). Lane 1: MW marker (SeeBlue® Plus2 Pre-Stained Standard, Invitrogen) with apparent molecular weights indicated on y-axis. Lanes 2-3: 6×ArntEbox bioprobe with wt bZIP protein pre- and post-induction, respectively. Lane 4-5: 6×MaxEbox bioprobe with MaxFos protein pre- and post-induction, respectively. Note the discrepancy between the apparent molecular weights of the MW markers and full-length bioprobe constructs, and between wt bZIP and MaxFos proteins (see caption of Fig. 4.3).
Supplementary Figure 4.3 | Generation of the CFP-6×AP1-YFP bioprobe. Using our cassette-based strategy to create the CFP-6×AP1-YFP bioprobe, we digested the CFP-3×AP1-YFP plasmid twice; first at the SpeI and XhoI sites to obtain the 3×AP1-YFP insert, and then at the NheI and XhoI sites to obtain the pTriEx-1.1 Hygro CFP-3×AP1 host vector. The two desired DNA fragments were purified by agarose gel electrophoresis and ligated to generate the CFP-6xAP1-YFP plasmid. Note the vector map is not to scale.
Chapter Five

Summary and future directions
5.1 Conclusions

Understanding the interactions between transcription factors and their DNA response elements contributes toward learning how various genes can be regulated. This entails identifying and characterizing transcription factors as well as their newly designed synthetic partners, which can be accomplished by various in vitro and in vivo techniques. In this thesis, emphasis has been placed on development of new approaches for high-throughput analysis of protein:DNA interactions in vitro and in vivo.

In vitro strategies for detection of protein:DNA interaction require isolation of active and soluble protein. However, current methodologies for purification of proteins often fail to provide high yield of pure and tag-free protein mainly because many enzymatic cleavage reactions for tag removal do not exhibit stringent sequence specificity. Using members from the bZIP family of transcription factors as cleavage substrates, we attempted removal of the His tag with enterokinase (EK) but obtained the desired cleavage product in very small yield.

We manipulated experimental conditions and showed that addition of denaturant to our EK reactions of bZIP-like AhR6-C/EBP and AhR6-Jun proteins greatly improved EK cleavage specificity at the canonical site and inhibited adventitious cleavage. We believe this enhancement in specificity was due to a more "open" protein structure, in which the accessible canonical target could compete effectively with adventitious cleavage sites. However, use of higher amounts of urea (~3-5 M) inhibited EK activity, and we were unable to increase the yield of the desired protein product significantly. Therefore, we proposed an alternative strategy for efficient EK proteolysis: by using an added leucine zipper to heterodimerize with the bZIP, we surmised that
the basic region and amino-terminus would become more open and potentially more accessible to enterokinase. Work remains to verify the success of this approach in the future.

In vitro methods for detection of protein:DNA recognition provide a means for quantitative characterization of biomolecular interactions. However, protein function may depend strongly on assay conditions, and there is a common consensus that detecting a protein:DNA interaction in vitro is not necessarily an indication of its relevance in vivo. Various strategies have been implemented for in vivo characterization of protein:DNA interactions, yet the most frequently used techniques often suffer from high false-positive and high false-negative rates.

We describe a novel bacterial bioprobe system for rapid and facile detection of protein:DNA recognition in vivo. The bioprobe reporter comprises the genes of two fluorescent proteins separated by a potential DNA target site. Upon translation of the bioprobe gene, FRET can occur between the two FPs. However, if a co-expressed transcription factor binds to the DNA target, transcription of the second FP is impeded, resulting in loss of FRET partner. Full-length transcription of the bioprobe ensues when no protein:DNA recognition occurs. Additionally, detection of signal from the first FP serves as an intrinsic reference for cell viability and successful protein expression for every sample. Therefore, the bioprobe signal arises from the "DNA stage" before transcription occurs: detection depends on just two species in the sample, the full-length and truncated bioprobe proteins. The bioprobe produces results within hours or overnight. Widely available bacteria strains and plasmids are used and no SDS-PAGE or immunoblotting is required. Measurement is performed on cell cultures with no special treatment, and detection is performed on a standard fluorimeter.
Using ratiometric FRET, we show that qualitative evaluation of protein:DNA recognition can be reliably made by our bioprobe on transcription factors from the bZIP and bHLHZ families and their DNA target sites. The bioprobe can produce a clear signal from one DNA target site and is capable of detecting a wide range of protein:DNA binding affinities comparable to well-established electrophoretic assays and yeast genetic assays; as shown with members of the bZIP and bHLHZ families, the threshold for reliable detection in the $6\times$ versions of the bioprobe appears to be around 800 nM $K_d$, and for $1\times$ versions, up to 500 nM $K_d$. Our system is applicable to high-throughput searches of both protein and DNA libraries; binding and nonbinding complexes can be distinguished by visual inspection of bacterial colonies on culture plates.

With the bioprobe, we consistently found correlation between in vivo bioprobe FRET measurements and quantitative in vitro binding affinities for the various protein:DNA complexes tested. We attribute this strong level of validation to bioprobe's independence on indirect readout by transcriptional activation of reporter genes, and its dependence on a physical interaction between the protein and DNA target that blocks transcription. In the bioprobe, the length of the DNA target inserted as bait between CFP and YFP, rather than identity of its sequence, is the primary determinant for successful FRET ratio measurements, and thus, similar to other existing technologies, we recommend that a known positive control be performed to ensure that a desired protein:DNA complex can be detected with a specific bioprobe.

5.2 Future work

The following sections are some proposals for further improvements as well as new research directions.
5.2.1 Improvement of EK proteolysis reaction

We are planning to use an alternative combination of substrate and zipper protein to verify that the inhibition of EK activity described in Chapter 3 is specific to the Fos zipper. We will use instead the C/EBP leucine zipper (30 residues, LEQKVLELTSNDRLRKRVEQLSRELDTLY), which can homodimerize to the leucine zipper region from AhR6-C/EBP presented in Chapter 2. We are aware that the C/EBP zipper does not carry a high concentration of net negative charge at the pH used in the EK reaction (pH 7.6), and hence it is not expected to have an inhibitory effect similar to that of the Fos zipper on the activity of EK, given our hypothesis about how the Fos zipper inhibits EK activity. So, by introducing an excess of a C/EBP zipper to dimerize with AhR6-C/EBP, we might be able to nucleate a more properly folded, nonaggregated structure where the enterokinase cleavage site would be more accessible to protease cleavage. Therefore, we are still exploring the original idea presented in Chapter 3 and shown in Fig. 3.1 about using a mini-helix to dimerize with the bZIP leucine zipper and force a more open structure with a more accessible canonical EK cleavage site.

5.2.2 Use of bioprobe to make a new biological discovery

In order to demonstrate the capability of our bioprobe system to make a new biological discovery, we are designing a new E-box binding protein.

The wt bZIP protein used in Chapter 4 is a yeast GCN4 bZIP derivative that binds with very weak affinity to the core E-box binding site (5′-CACGTG-3′, $K_d$ 570 nM and 840 nM for Arnt E-box (5′-TCACGTGA-3′) and Max E-box (5′-CCACGTGGG-3′) sites, respectively). We are trying to design a new bZIP protein that is a strong binder of the E-box sequence and use our bioprobe system for discovery and validation. We are using EmBP1 protein as a template for our
design; EmBP1 is a plant bZIP transcription factor that binds to the 5'-GCCACGTGCG-3' sequence with affinity similar to that of the wt bZIP:AP-1 complex [172-174].

Two sets of controls are necessary: (1) wt bZIP (DPAALKRARNTAARRSKRLQRMKOLEQKVLELETSDNDRRLKRVEQLSRELDTL) bound to AP-1 or Max E-box DNA site (i.e. the CFP-6×AP1-YFP and CFP-6×MaxEbox-YFP bioprobes, respectively). These measurements are presented in Chapter 4; (2) Native EmBP1 bZIP (ELKRERRKQSNRESARRSRRLKQQECEELAKVSELTAAANGTLRSELDQLKEDKTEVENKQLMGKILG) will also be tested with 6×AP1 and 6×MaxEbox bioprobes. The EmBP1 bZIP is expected to bind strongly to the Max E-box DNA sequence (5'-CCACGTGG-3', note the similarity to the native EmBP1 binding sequence). The EmBP1 bZIP has been shown to bind the AP-1 site approximately 100-fold weaker than to the Max E-box site [172].

To produce new biology, we will construct and analyze EmBP1-C/EBP bZIP, a hybrid of the EmBP1 basic region and C/EBP leucine zipper, bound to the AP-1 or Max E-box DNA site. The EmBP1-C/EBP fusion is expected to be a strong and specific binder of the Max E-box sequence. Additionally, we will try to convert our wt bZIP into a strong E-box binder. To do this, we have used sequence alignment and high-resolution structural information in our rational design of five targeted mutations that appear to be critical for EmBP1 recognition of its DNA target site; these include an insertion of four amino acids and mutation of one amino acid in the GCN4 basic region of wt bZIP. Furthermore, we plan to make a protein library with one randomized amino acid, encoded by the NNK codon that covers all 20 amino acids and removes 2 stop codons. The resulting protein library can be screened by visual inspection of fluorescent colonies on plates, as demonstrated previously in Chapter 4. Selected clones will be sequenced to uncover the identity of the randomized residue. Therefore, these designs based on EmBP1 will
test our ability to re-design the wt bZIP into a strong E-box binder and test the utility of the bioprobe toward discovering "hits" from a screen of a randomized library.
References


Appendix A

This section contains the permission from Elsevier to re-print the research article “Enhancing the specificity of the enterokinase cleavage reaction to promote efficient cleavage of a fusion tag”, published in Journal of Protein Expression and Purification 59 (2008) 314–319, which is presented in Chapter 2 of this thesis.
This is a License Agreement between Hesam Shahravan ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier
Elsevier Limited
The Boulevard, Langford Lane
Kidlington,Oxford,OX5 1GB,UK

Registered Company Number
1982084

Customer name
Hesam Shahravan

Customer address
3359 Mississauga Rd. North
Mississauga, ON M2N7L6

License number
2477260450812

License date
Jul 27, 2010

Licensed content publisher
Elsevier

Licensed content publication
Protein Expression and Purification

Licensed content title
Enhancing the specificity of the enterokinase cleavage reaction to promote efficient cleavage of a fusion tag
INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

   “Reprinted from Publication title, Vol / edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER].” Also Lancet special credit - “Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier.”

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically
revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

**LIMITED LICENSE**

The following terms and conditions apply only to specific license types:

15. **Translation:** This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. **Website:** The following terms and conditions apply to electronic reserve and author websites:
**Electronic reserve:** If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:
This license was made in connection with a course, *This permission is granted for 1 year only.*
You may obtain a license for future website posting. All content posted to the web site must maintain the copyright information line on the bottom of each image, A hyper-text must be included to the Homepage of the journal from which you are licensing at [http://www.sciencedirect.com/science/journal/xxxxx](http://www.sciencedirect.com/science/journal/xxxxx) or the Elsevier homepage for books at [http://www.elsevier.com](http://www.elsevier.com), and Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. **Author website** for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version, A hyper-text must be included to the Homepage of the journal from which you are licensing at [http://www.sciencedirect.com/science/journal/xxxxx](http://www.sciencedirect.com/science/journal/xxxxx), As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier’s online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article’s Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting. Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. **Author website** for books with the following additional clauses:

Authors are permitted to place a brief summary of their work online only. A hyper-text must be included to the Elsevier homepage at [http://www.elsevier.com](http://www.elsevier.com)

All content posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website** (regular and for author): A hyper-text must be included to the Homepage of the journal from which you are licensing at [http://www.sciencedirect.com/science/journal/xxxxx](http://www.sciencedirect.com/science/journal/xxxxx), or for books to the Elsevier homepage at [http://www.elsevier.com](http://www.elsevier.com)

20. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and
include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions:**

v1.6

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10822032. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2477260450812 and noting the reason for cancellation.

Questions? [customercare@copyright.com](mailto:customercare@copyright.com) or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.
Appendix B

This section presents a more detailed and general description of the Methods section originally provided in Chapter 2 (Enhancing the specificity of the enterokinase cleavage reaction to promote efficient cleavage of a fusion tag) of this thesis.
B.1 Enterokinase cleavage of AhR6–C/EBP (typical EK reaction)

The enterokinase reactions (10 µL) were prepared according to the manufacturer’s recommendations. We used the native form of EK (150 kDa) supplied by Roche. 6.0 µg AhR6–C/EBP (purified by TALON affinity chromatography followed by reversed-phase HPLC) was incubated with 0.14 µg EK (EK:substrate ratio = 1:42) at 37 °C in 50 mM Tris, pH 7.6. The reactions were stopped by addition of an equal volume of Tricine Sample Buffer (Bio-Rad), which contains 125 mM dithiothreitol (DTT, added fresh to the sample buffer), followed by heating for 10 min at 95 °C. Samples can be stored at -20 °C until SDS-PAGE is performed. To detect a possible nonspecific cleavage caused by either autolysis or proteolytic contaminations of the recombinant protein, a control incubation without enterokinase is highly recommended (control contains all components above except no enzyme).

B.2 Isolation and identification of cleavage products by HPLC and ESI-MS

The typical enterokinase reaction (100 µL; this preparatory sample is 10-fold more than that described above) was prepared as above and stopped by addition of phenylmethanesulfonyl fluoride (PMSF) to a final concentration of 1 mM PMSF, followed by heating for 10 min at 95 °C. The heating step is important as addition of PMSF alone does not fully quench the reaction. The stopped reaction was incubated with 30 mM DTT for ≥30 min at room temperature, followed by analytical HPLC on a C4 column (Vydac). The collected fractions were analyzed with ESI-MS at flow rate 40 µL/min with the capillary charged at +3200 V, source temperature 100 °C, and cone voltage 20 V. The mass range 600–1100 m/z was scanned every 1 s for 40 s. Uncalibrated data were processed using MassLynx (version 4.00.00). Molecular masses were analyzed by massXpert (version 1.6.0) to identify cleavage sites.
Appendix C

This section presents a more detailed and general description of the Methods section originally provided in Chapter 4 (A fluorescent protein-based bioprobe for detection of protein:DNA recognition \textit{in vivo}) of this thesis as well as some background information on the theory of fluorescence resonance energy transfer.
C.1 Preparation of 6×DNATarget bioprobes

In order to prepare a CFP-6×DNATarget-YFP bioprobe, we first constructed plasmid CFP-3×DNATarget-YFP that contains the genes for CFP and YFP separated by 3 tandem copies of the DNA target in pTriEx-1.1 Hygro (Novagen). This plasmid contains a base cassette that allows for further expansion of number of DNA sites (see ref. [148] for more information on this technique). In our design, site 1 was always fixed as the restriction site for NcoI, site 2a as SpeI, site 2b as Nhel, and site 3 as XhoI. Note that site 2a and site 2b should always produce compatible ends.

For each DNA target examined in this study, the base cassette was created by C-terminal extension of a CFP gene fragment (pTriEx-1.1 Hygro CFP-3×AP1-YFP plasmid can also be used) using these primers: forward, 5´-TACCATGGGCCTGACTAGTGGATC-3´, and the corresponding reverse primer in Supplementary Table 4.1. The resulting PCR fragment was digested with NcoI and Nhel, purified (QIAEX II, Qiagen), and subcloned into a pTriEx-1.1 Hygro vector containing the gene for YFP (referred to as pCfvtx in ref. [148]). The original pCfvtx vector can be obtained by restriction digestion of any bioprobe plasmid already in stock with NcoI and Nhel restriction enzymes.

To create a CFP-6×DNATarget-YFP bioprobe, we digested the corresponding CFP-3×DNATarget-YFP base cassette plasmid twice; first at the SpeI and XhoI sites to obtain the 3×DNATarget-YFP insert, and then at the Nhel and XhoI sites to obtain the pTriEx-1.1 Hygro CFP-3×DNATarget host vector (see Supplementary Fig. 4.3 for a schematic diagram of the cloning procedure). The two desired DNA fragments were purified (QIAquick, Qiagen) after agarose gel electrophoresis and ligated to generate the CFP-6×DNATarget-YFP plasmid. This
plasmid was chemically transformed into *E. coli* strain DH5α; plasmid DNA was extracted (Wizard® Plus SV Miniprep, Promega) and sequenced. Note that a CFP-12×DNA target-YFP bioprobe can be created by repeating the same procedure above with the CFP-6×DNA target-YFP bioprobe.

### C.2 Preparation of 1×DNATarget/NS bioprosbes

All bioprosbes containing a single copy of a DNA target site (CFP-1×DNATarget/NS-YFP, which comprises a linker length of 84 bp between the genes for CFP and YFP) were similarly constructed as above for construction of the 6× bioprosbes except different reverse primers were used (see Supplementary Table 4.2).

To create a CFP-1×DNATarget/NS-YFP bioprobe, we digested the corresponding CFP-1×DNATarget-YFP base cassette plasmid at the *Nhe*I and *Xho*I sites to obtain the pTriEx-1.1 Hygro CFP-1×DNATarget host vector and ligated it to a previously obtained NS-YFP insert (obtained by restriction digestion of CFP-NS-YFP bioprobe at the *Spe*I and *Xho*I sites) to generate the CFP-1×DNATarget/NS-YFP plasmid. A CFP-2×DNATarget-YFP (84 bp) bioprobe can be created by repeating the same procedure above using CFP-1×DNA target-YFP bioprobe.

### C.3 Expression of the bioprobe

Electrocompetent *E. coli* BL21(DE3)pLysS cells were first transformed with the pET28a(+) transcription-factor expression vector, followed by transformation with the bioprobe reporter vector, and plating onto a fresh LB plate containing 30 µg/mL kanamycin and 50 µg/mL ampicillin for overnight incubation at 37 °C. Note that the sequence in which the two expression
vectors are transformed into *E. coli* BL21(DE3)pLysS cells does not affect the overall efficiency of transformation.

For each transformation and after overnight incubation on the plate, a single colony was chosen for highest intensity by visualizing YFP fluorescence using a Lightools Illumatool LT-9900 imaging system equipped with a 488/10 nm excitation filter and 530 nm viewing filter. Note that this is an important step that must be done before proceeding to bioprobe expression as colonies not displaying high fluorescence intensity have been seen to later fail to express substantial amounts of fluorescent proteins. Protein expression was performed according to the protocol provided in Table C.1.

**Table C.1 | Experimental procedure for expression of the bioprobe**

<table>
<thead>
<tr>
<th>Step</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inoculate 5 mL fresh LB media (antibiotic added) with a single colony from transformation plate. Shake this culture overnight at 37 °C (speed 250 rpm).</td>
</tr>
<tr>
<td>2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Inoculate 50 mL fresh LB media (antibiotic added) with 1 mL overnight culture. Shake this culture at 37 °C to OD&lt;sub&gt;600&lt;/sub&gt; 0.6-0.7. This process takes between 2-3 hrs.</td>
</tr>
<tr>
<td>3</td>
<td>Once the desired OD is reached, use 1 mL culture for zero-time reading measurement.</td>
</tr>
<tr>
<td>4</td>
<td>Add IPTG to final concentration 1mM and continue growth overnight at 37 °C (10-16 hrs; a time-course experiment showed that 10 hrs was sufficient).</td>
</tr>
<tr>
<td>5</td>
<td>Use 1 mL aliquot of this sample for measurement.</td>
</tr>
</tbody>
</table>

<sup>a</sup> Can be performed on a smaller scale, if needed, but it is highly recommended that the volume ratios are not altered.
C.4 Fluorescence measurement

Whole-cell fluorescence of each sample (precisely 1 mL) in LB medium was directly measured with no pre-measurement treatment (i.e. no washing) using a conventional fluorimeter (Perkin Elmer LS 50B). Each sample was excited at 440 nm and the emission was recorded in the range of 460-560 nm (scan rate ~250 nm/min), with excitation and emission slits set at 10 and 2.5 nm, respectively. Each sample was measured three times with the final raw spectrum taken as the average of all three measurements.

In light of the need to make spectral corrections as discussed by Gordon et al. in their studies using similar steady-state sensitized emission assays [151], three spectral corrections were applied to each raw fluorescence spectrum before FRET ratio calculation: (1) for each sample, a background spectrum (i.e. sample from step 3 in Table C.1), defined as the spectrum obtained from the same sample before protein induction, was subtracted to eliminate any fluorescence contribution from LB and *E. coli* cells; (2) to remove CFP spectral bleed-through from YFP intensity, a spectrum originally obtained from a sample containing only CFP was normalized with respect to maximum CFP intensity measured for each sample and subtracted from the same sample’s raw fluorescence spectrum; (3) to remove fluorescence contribution from direct-excitation of YFP, a spectrum originally obtained from a sample containing only YFP was normalized with respect to maximum YFP intensity measured for each sample and subtracted from the same sample’s raw fluorescence spectrum.

C.5 SDS-PAGE analysis of the bioprobe

BL21(DE3)pLysS cell cultures containing both TF- and bioprobe-expression vectors were grown overnight with IPTG induction. Each 20 mL cell culture was harvested and re-suspended in 500
μL lysis buffer (20 mM Tris-HCl, 0.3 M NaCl, 5 mM β-mercaptoethanol, 8% glycerol, 6 M urea, pH 8.0). The cells were sonicated for 40 s while chilled on ice and centrifuged at 13,400 g for 10 min at 4 °C. 10 μL supernatant was mixed with 10 μL SDS buffer (50 mM Tris-HCl, 100 mM DTT, 2% SDS, 0.1% bromophenol blue, 10% glycerol, pH 6.8) without heating and resolved by 12% SDS-PAGE at room temperature (SDS-PAGE running buffer: 25 mM Tris, 192 mM Glycine, 0.1% SDS, pH 8.8), followed by visualization under illumination by our Lightools fluoroimaging system (once with a 440/20 nm excitation filter and 475 nm viewing filter for CFP emission, and once with a 488/10 nm excitation filter and 530 nm viewing filter for YFP emission) and staining by Coomassie brilliant blue solution containing 50% methanol to fix the protein samples.

C.6 Screening of DNA binders

Screening of DNA binders was carried out by transferring a mixture of cells containing potential DNA binders (BL21(DE3)pLysS cells containing a specific bioprobe and an expression vector carrying the gene for a potential DNA binder) onto a freshly made LB plate (30 μg/mL kanamycin, 50 μg/mL ampicillin, and 1 mM IPTG added directly to the plate solution) for overnight incubation at 37 °C (10-16 hrs). After incubation, each plate was photographed while under illumination (Lightools fluoroimaging system, 488/10 nm excitation filter, 530 nm viewing filter) before analysis using a Nikon Eclipse L150 fluorescence microscope (ND16 filter, 4X objective) equipped with a laser light source (Coherent 406 nm/25 mW, z405/20x excitation filter) and dichroic 450 nm long-pass filter (z405 RDC) and attached to a diode array detector (Ocean Optic CCD QE65000). The laser light was focused on individual colonies, and in-colony FRET ratios were calculated.
Colonies displaying appropriate fluorescence were identified and grown in LB media (30 μg/mL kanamycin, and 50 μg/mL ampicillin) overnight at 37 °C. Cells were pelleted and plasmid DNA was extracted. The identity of each selected colony was revealed by restriction digestion of plasmids followed by agarose gel electrophoresis separation. In an experiment where wt bZIP and C/EBP bZIP were the only two members of the protein library, plasmids containing the gene for wt bZIP were cleaved by both PstI and MluI restriction endonucleases, whereas plasmids containing the C/EBP bZIP gene were only linearized under the same conditions, as this plasmid does not contain the PstI restriction site (Fig. C.1).

**Figure C.1** | Identification of various colonies after DNA binder screening. The agarose gel electrophoresis separation of extracted and digested plasmids of each of the ten colonies chosen by fluorescence microscopy is shown. Plasmids containing the gene for wt bZIP give two cleavage fragments at 4573 bp and 994 bp after treatment by PstI and MluI restriction endonucleases, whereas plasmids containing the C/EBP bZIP gene are only linearized under the same conditions (one fragment at 5576 bp). Note that each sample also contains the CFP-6×AP1-YFP bioprobe plasmid that also gets digested by PstI and MluI restriction endonucleases (resulting in two fragments at 5368 bp and 2954 bp).
C.7 Fluorescence resonance energy transfer

Fluorescence (or Förster) resonance energy transfer is the distance- and orientation-dependent non-radiative transfer of energy from a donor fluorophore in a higher-energy state to an acceptor fluorophore in a lower-energy state [175, 176]. In a FRET experiment, a donor fluorophore is excited and transfers its energy of excitation to an acceptor fluorophore via a long-range resonant dipole-dipole interaction. FRET does not require the acceptor to be fluorescent; in fact, the acceptor can be a chromophore that is non-fluorescent (i.e. a quencher). However, if the acceptor is fluorescent, it will fluoresce with its characteristic emission profile as if it had been directly excited [162].

The rate of energy transfer is described by equation C.1, the Förster rate equation [175, 176].

\[
 k_{ET} = \frac{1}{\tau_D} \left( \frac{R_0}{r} \right)^6
\]  

(C.1)

\[
 R_0^6 = 8.79 \times 10^{-25} (\kappa^2 n^{-4} Q_D J(\lambda))
\]  

(C.2)

\[
 J(\lambda) = \int_0^\infty F_D(\lambda) \epsilon_A(\lambda) \lambda^4 d\lambda
\]  

(C.3)

where \( \tau_D \) is the donor lifetime in the absence of acceptor, \( r \) is the interfluorophore distance, and \( R_0 \) is the Förster distance (the interchromophore distance at which FRET efficiency is 50%). As seen above, the rate of energy transfer depends on many other factors as well, including the quantum yield of the donor (\( Q_D \)), the relative orientation of the donor and acceptor transition dipoles (\( \kappa \), frequently taken as 2/3 for dynamic random averaging of the donor and acceptor), the
index of refraction of the medium through which transfer occurs \((n, \text{ typically assumed as 1.4 for biomolecules in aqueous solution})\), and the extent of spectral overlap of the emission spectrum of the donor with the absorption spectrum of the acceptor \((J(\lambda), \text{ usually } >30\% \text{ overlap is required for observation of FRET [177]})\) which itself depends on the acceptor molar absorption coefficient \((\varepsilon_A)\), and \(F_D(\lambda)\), the peak-normalized fluorescence intensity of the donor as a function of wavelength. The scaling constant in equation C.2 is set such that if the wavelength is expressed in cm and \(J(\lambda)\) is in units of M\(^{-1}\) cm\(^3\), the Förster distance is given in units of cm\(^6\).

The efficiency of energy transfer is defined as the fraction of photons absorbed by the donor that are transferred to the acceptor. This fraction is the ratio of transfer rate to the total decay rate of the donor in the presence of acceptor. FRET efficiency \((E)\) is given by equation C.4, which when combined with equation C.1, can be rearranged to yield equation C.5.

\[
E = \frac{k_{ET}}{k_{ET} + \tau_D^{-1}} \quad \text{(C.4)}
\]

\[
E = \frac{R_0^6}{R_0^6 + r^6} \quad \text{(C.5)}
\]

\[
E = 1 - \frac{F_{DA}}{F_D} = 1 - \frac{\tau_{DA}}{\tau_D} = 1 - \frac{Q_{DA}}{Q_D} \quad \text{(C.6)}
\]

FRET efficiency can be measured experimentally from changes in fluorescence intensity \((F)\), fluorescence lifetime \((\tau)\), or quantum yield \((Q)\) of donor in the absence and presence of acceptor. In equation C.6, the subscripts D and DA denote isolated donor and donor in the presence of acceptor, respectively. It is important to recognize that most applications of FRET in studies of biological molecules involve ratiometric imaging of emission intensity or imaging of
donor fluorescence lifetime. In ratiometric imaging, the biological sample is illuminated with wavelengths of light that specifically excite the donor molecule while images of the fluorescence emission are collected from the sample through two different emission channels corresponding to the emissions of the donor and acceptor. This allows for calculation of a FRET ratio (defined as the ratio of acceptor to donor fluorescence intensity) which is typically used as a substitute for FRET efficiency in live-cell measurements [162].

**C.7.1 FRET in studies of biological molecules**

Equation C.5 shows that FRET efficiency is strongly dependent on distance when intermolecular distance between donor and acceptor is close to $R_0$; moreover, it shows that through experimental determination of FRET efficiency, the donor-acceptor distance ($r$) can be readily obtained once the Förster distance is calculated. In fact, this has resulted in FRET being described as a “spectroscopic ruler”, because the Förster distance can be reliably predicted from the spectral properties of donor and acceptor [175, 178]. Although many early applications of FRET in structural biology involved its use as a distance indicator between two fluorophores, the development of fluorescent protein technology in recent years has propelled the development of novel FRET-based biosensors useful for performing a wide range of tasks, including imaging of concentrations of ions and small molecules, the activity of enzymes, and changes in the conformation of proteins in living cells (see ref. [179] for a comprehensive overview of FRET-based biosensors).

For most applications of FRET in studies of biomolecules, the biochemical signal is a change in FRET efficiency that is usually caused by changes in relative distance and orientation between donor and acceptor fluorophores. However, one should consider that other factors can
influence the success of a FRET-based strategy for use in a given experiment. For example, in Chapter 4 we demonstrated that for bioprobes containing 9, 12, and 24 copies of the AP-1 target site (corresponding to 117, 150, and 282 base pairs, respectively, between the genes of CFP and YFP), there was no decrease in FRET ratios upon co-expression of the wt bZIP. If only distance were at issue in our experiments, we would have expected our strategy to fail once the linker length was increased past a certain extent (84 base pairs). However, this was not the case: when similar bioprobes having the identical numbers of base pairs between the genes of CFP and YFP were made with 9, 12, and 24 copies of the Max E-box target site, YFP transcription was successfully blocked upon co-expression of MaxFos. This indicated that the success of our FRET-based approach can also depend on the amino acid identity of the linker as well.