Characterization of the *E. coli* SOS response protein YbfE

by Caitlin S. Kramer

B.S. in Chemistry, Northeastern University

A thesis submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Master of Science

April 22, 2016

Thesis directed by

Penny J. Beuning
Professor of Chemistry and Chemical Biology
Acknowledgements

Thank you Penny, for sharing time, resources, advice, and feedback with me since 2012. There is no doubt that you have greatly influenced my life, lab skills, scientific thinking, and surely my grammar. I am so grateful for the chance you have given me to freely explore my research interests.

I also thank my other committee members, Mary Ondrechen and Carla Mattos, for sharing your time with me and reading this thesis. I greatly appreciate the extra effort you put forth to connect with students in the department.

Much of the modeling work in this thesis was done in my Molecular Modeling class, CHEM 5638 Spring 2015. Additional thanks for Mary for guiding this class, and to its members with whom I worked.

I have had the chance to work with undergraduates on the work done here. I would like to thank Richard Judge, Katleen Klaassen, and Nicole Taylor for helping me make additional progress. I enjoyed working with you and you are all great. Some other people who were absolutely vital to initiation of the project and whose work ethic I admire: Becky Leifer, Meghan Travers, Samir Baig, Sirine Bellou, and Mark Muenter.

Many thanks to my lab mates past and present for being in subgroups and group meetings with me and offering feedback and guidance. Special thanks to Nicholas DeLateur, who trained me and started me on my first independent project with the S. meli genes, and Paul Ippoliti, who asked Penny if I could work with him in the lab in 2012. Thank you Bilyana for being a great friend and coworker, and always listening to my hypotheses and making me think more critically.

Pasha, thank you for always being encouraging and enthusiastic about my work. You have helped me focus on what is important to me.

Finally, I am thankful for my parents for providing everything I ever needed to come as far as I have. I am an extension of you, and I am proud.

Thank you to Northeastern University, the College of Science, and the Department of Chemistry and Chemical Biology for 5 years of excellent opportunities.
Abstract of Thesis

DNA damage occurs constantly in cells under both normal and stress conditions. Sites of damage can be repaired or tolerated as the result of specialized mechanisms. Elucidating these conserved mechanisms is vital to our understanding of the maintenance of genetic material and the development of mutations, which are implicated in a wide variety of diseases including cancer. The molecular properties of the ensemble of proteins involved in DNA damage repair and tolerance define how damage is processed in the cell.

The global response to DNA damage in prokaryotes is controlled through the SOS response, and results in the expression of more than 50 LexA-regulated genes involved in DNA damage repair and tolerance. Recombination, nucleotide excision repair, translesion synthesis, and changes to a variety of cellular processes are invoked during SOS induction. Not all of the LexA-regulated genes have assigned functions, although they are likely to contribute to a robust network of responses to damage. An overview of DNA damage repair and tolerance mechanisms is presented in Chapter 1.

Prior work by our lab has shown that deletion of the uncharacterized LexA-regulated gene \textit{ybfE} is associated with poor survival in \textit{E. coli} exposed to alkylating agents. In order to elucidate the role of \textit{ybfE} in enhancing survival, the function of the \textit{ybfE} gene product was investigated. Chapter 2 examines the genomic context of \textit{ybfE}, and discusses the work done to elaborate its structure and function. YbfE was examined computationally, \textit{in vitro}, and \textit{in vivo} in \textit{E. coli}. The work presented shows that YbfE is a helical, dimeric protein with DNA binding activity. Overexpression of YbfE is lethal to cells, with rescue of this phenotype by deletion of \textit{uvrC}. This could suggest a role for YbfE in nucleotide excision repair, a pathway involved in the repair of bulky DNA adducts.
# Table of Contents

Copyright page.................................................................................................................. ii  
Acknowledgements........................................................................................................... iii  
Abstract of Thesis............................................................................................................. iv  
Table of Contents............................................................................................................... v  
List of Figures.................................................................................................................... vi  
List of Tables..................................................................................................................... vii  
List of Abbreviations.......................................................................................................... viii  

Chapter 1. DNA Damage, the SOS response, and YbfE..................................................... 1  
  1.1 A perspective on DNA damage.................................................................................. 2  
  1.2 An overview of DNA damage responses................................................................. 3  
  1.3 The SOS response..................................................................................................... 4  
  1.4 ybfE is an uncharacterized LexA-regulated gene....................................................... 5  

Chapter 2. Characterization of the SOS response protein YbfE......................................... 11  
  2.1 Introduction............................................................................................................... 12  
  2.2 Materials and Methods............................................................................................ 12  
  2.3 Results....................................................................................................................... 17  
  2.4 Discussion.................................................................................................................. 25  

Chapter 3. Conclusions and Future Work.......................................................................... 32  
  3.1 Conclusions............................................................................................................... 33  
  3.2 Ongoing and Future Work......................................................................................... 34  

References......................................................................................................................... 36
List of Figures

**Figure 1.1** The SOS Response 5
**Figure 1.2** Deletion strain sensitivity to alkylating agents 7
**Figure 1.3** YbfE gene sequence with key features highlighted 8
**Figure 1.4** DELTA-BLAST results for YbfE 9
**Figure 1.5** Hydrophobicity, disorder, and secondary structure character of the YbfE protein sequence 10
**Figure 2.1** Sequence alignment of YbfE, Yiif, and GPS 18
**Figure 2.2** Homology models of YbfE 19
**Figure 2.3** Sequence alignment of YbfE and NikR 20
**Figure 2.4** Induction of 120aa-YbfE over 4 hours 21
**Figure 2.5** EMSA of DNA bound by 97aa-YbfE 21
**Figure 2.6** Formaldehyde crosslinking of 97aa-YbfE dimers 22
**Figure 2.7** CD of 97aa-YbfE 23
**Figure 2.8** Cell sensitivity to YbfE overexpression 24
**Figure 2.9** Rescue of YbfE overexpression lethality by *uvrC* deletion 25


List of Tables

Table 1.1  LexA binding sequences and HI  6
Table 2.1  Homology modeling parameters  12
Table 2.2  Strains and plasmids  13
Table 2.3  POOL residues for monomeric homology model  18
Table 2.4  POOL residues for dimeric homology model  20
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>°C</td>
<td>Degrees Celsius</td>
</tr>
<tr>
<td>Å</td>
<td>Angströms</td>
</tr>
<tr>
<td>BER</td>
<td>Base-excision repair</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>βME</td>
<td>B-mercaptoethanol</td>
</tr>
<tr>
<td>bp</td>
<td>Base pairs</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism</td>
</tr>
<tr>
<td>CV</td>
<td>Column Volumes</td>
</tr>
<tr>
<td>DIG</td>
<td>Digoxigenin</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double stranded DNA</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EMSA</td>
<td>Electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GPS</td>
<td>Geranylgeranyl pyrophosphate synthetase</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-Hydroxyethyl)-1-Piperazinethanesulfonic Acid</td>
</tr>
<tr>
<td>HI</td>
<td>Heterology index</td>
</tr>
<tr>
<td>h</td>
<td>hour</td>
</tr>
<tr>
<td>kD</td>
<td>kilodalton</td>
</tr>
<tr>
<td>min</td>
<td>minutes</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>mL</td>
<td>milliliters</td>
</tr>
<tr>
<td>mM</td>
<td>millimolar</td>
</tr>
<tr>
<td>MMR</td>
<td>Mismatch repair</td>
</tr>
<tr>
<td>mol</td>
<td>mole</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NER</td>
<td>Nucleotide-excision repair</td>
</tr>
<tr>
<td>nM</td>
<td>nanomolar</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Databank</td>
</tr>
<tr>
<td>pI</td>
<td>Isoelectric point</td>
</tr>
<tr>
<td>POOL</td>
<td>Partial-order optimum likelihood</td>
</tr>
<tr>
<td>PVC</td>
<td>Polyvinyl chloride</td>
</tr>
<tr>
<td>RHH</td>
<td>Ribbon-helix-helix</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single-stranded DNA</td>
</tr>
<tr>
<td>TLS</td>
<td>Translesion synthesis</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>V</td>
<td>Volts</td>
</tr>
<tr>
<td>WT</td>
<td>Wild-type</td>
</tr>
<tr>
<td>βME</td>
<td>β-mercaptoethanol</td>
</tr>
<tr>
<td>µL</td>
<td>microliters</td>
</tr>
<tr>
<td>µM</td>
<td>micromolar</td>
</tr>
</tbody>
</table>
CHAPTER 1 – DNA damage, the SOS response, and YbfE
1.1 A perspective on DNA damage

The role of DNA as the genetic material of life is made possible by its chemical properties. The stability of DNA molecules lends to their integrity, while their reactivity allows for the variety of enzymatic reactions that occur to replicate, repair, and diversify the sequence. This reactivity can frequently lead to damage when cellular metabolites, environmental chemicals, or radiation interact with DNA. The immediate structural results take many forms; some well-studied examples include alkylation, pyrimidine dimers, interstrand/intrastrand crosslinks, intercalation, deamination, depurination, depyrimidination, single-stand breaks, and double-strand breaks. Much research has gone into understanding the mechanisms that interconnect specific damage sources, their mechanism of interaction with DNA, and the consequences [1, 2]. The findings enhance our understanding of links between inflammation and oxidative stress [3], sun exposure and skin cancer [4, 5], even antibiotics and the emerging crisis of widespread bacterial antibiotic resistance [6, 7].

Although exogenous genotoxic agents account for a relatively small percentage of the overall genetic insult that most cells experience on a daily basis [8], they are linked to cancer in humans and animals [9]. While humans cannot act to limit our exposure to our own metabolites, we can act to prevent widespread exposure to identified carcinogens. Vinyl chloride, a chemical used most prevalently in the manufacture of polyvinyl chloride (PVC), is processed by cytochrome P-450 in the liver to form reactive species including chloroacetaldehyde. This and other genotoxic metabolites are capable of reacting with the bases of DNA to form a variety of etheno-DNA adducts implicated in liver angiosarcoma [10]. In the pharmaceutical industry, the persistence of synthesis reagents, intermediates, or even degradation products can result in genotoxic impurities in the finished pharmaceutical product. Such a case occurred in 2007 with
Roche’s Viracept, when consumers were estimated to have been exposed to 2.75 mg of ethyl methanesulfonate per day over a period of 3 months [11]. While no product is completely free of impurities, increased manufacturer and consumer awareness of wanton use of genotoxic chemicals has been necessitated by the lessons of history.

The angles from which a genotoxic chemical can be characterized are many. Analytical methods such as mass spectrometry and nuclear magnetic resonance can elucidate adduct structure. The Ames test [12] can be used to assay chemicals for mutagenicity and by extension carcinogenicity. The study of what happens in between – the biological processing of DNA adducts - unites these two realms by bridging the gap between adducts and their phenotypic and genotypic consequences. This gap is filled by proteins that interact with DNA, each other, and other components of the cell to coordinate a response to damage.

1.2 An overview of DNA damage responses

The discovery and subsequent elucidation of the different DNA damage response pathways has been of paramount importance to science, as embodied by the 2015 Nobel Prize in Chemistry being awarded to three scientists whose work revolved around three major DNA repair mechanisms. The mechanisms they investigated are nucleotide excision repair (NER), base excision repair (BER), and mismatch repair (MMR), which repair damaged or mismatched DNA bases by excising damage and initiating re-synthesis of DNA based on the undamaged template strand. At their best, these mechanisms have the capability of restoring damaged DNA to its original undamaged structure and sequence. When flawed, they contribute to a wide variety of diseases [8]. Redundancy is a key theme in DNA repair. When these repair pathways fail to fix damage, a cell is not necessarily faced with death - DNA damage can also be tolerated using strategies such as translesion synthesis (TLS) and mutagenesis. Due to their potent ability to
effect structural change of DNA, many DNA damage response genes are tightly regulated. In *E. coli*, 57 genes have been identified whose expression is upregulated following a DNA damage signal [13]. This response is known as the SOS response.

### 1.3 The SOS response

The SOS response (Figure 1.1) is a global response to DNA damage in prokaryotes [8, 14] that is repressed in cells under normal conditions [15]. Activation of the system is regulated by the proteins RecA and LexA. The induction of the response is triggered by the accumulation of single-stranded DNA near stalled replication forks or excised patches. RecA acts as the positive signal in the response when it binds to accumulated single-stranded DNA to form a nucleoprotein filament [16, 17] which acts as a coprotease to facilitate LexA autodigestion [18, 19]. In the uncleaved state, LexA acts as a repressor [20-22] by blocking transcription of over 50 genes [13] involved in cell cycle checkpoints, DNA repair, TLS, and mutagenesis. When LexA is cleaved and vacates its binding sites, the SOS boxes, gene expression is enhanced as RNA polymerase gains access to the gene promoters [23]. The SOS box contains a LexA sequence motif with the consensus sequence TACTG(TA)₅CAGTA [24]. Deviation of this motif from the consensus sequence results in decreased binding affinity between LexA and the SOS box [25], and results in differential basal levels of expression among SOS genes. This results in the execution of different pathways at different times following SOS induction.
1.4 *ybfE* is an uncharacterized LexA-regulated gene

Several of the SOS genes remain uncharacterized [13, 26, 27]. The *ybfE* gene was identified as an SOS gene by computational analysis of the *E. coli* genome [26]. As discussed, the transcription of SOS response genes is repressed by LexA, which binds sequence-specifically to so-called “LexA (or SOS) boxes” in the promoter region of regulated genes. The degree of variation a given sequence has from the consensus sequence is represented by a Heterology Index (HI) which increases with increasing difference compared to the consensus sequence [25]. Levels of constitutive expression for a gene increase with increasing HI, and these genes are less tightly regulated by LexA. The *ybfE* gene has one of the highest HI values of 14.07 (Table 1.1). This relaxed repression was shown by Northern Blot analysis, which showed appreciable...
expression in the absence of DNA damaging treatment compared with more tightly regulated 
SOS genes [26]. *E. coli* cells lacking the uncharacterized LexA-regulated *ybfE* gene are more 
sensitive to the alkylating agents chloroacetaldehyde, benzyl bromide, and styrene oxide 
(Muenter, 2016. Unpublished data. Figure 1.2).

<table>
<thead>
<tr>
<th>Heterology Index</th>
<th>LexA box</th>
<th>Gene promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>TACTGTATATATACAGTA</td>
<td>Consensus sequence</td>
</tr>
<tr>
<td>2.77</td>
<td>TACTGTATAAAACAGTA</td>
<td><em>umuDC</em></td>
</tr>
<tr>
<td>9.60</td>
<td>AGCTGTATTTGTCTAGTA</td>
<td><em>dinS</em></td>
</tr>
<tr>
<td>14.07</td>
<td>AACTGATTAAAAACCCAGCG</td>
<td><em>ybfE</em></td>
</tr>
</tbody>
</table>

**Table 1.1** SOS gene LexA boxes and associated HI values. Adapted from [25].
The 5′ start codon of the *ybfE* gene and thus the translated protein length is inconsistent in databases such as GenBank that house information from genome sequencing projects. The misannotation of the 5′ boundary of genes is an acknowledged limitation to methods used in the annotation of genome sequences—imperfect algorithms take into account possible ORFs, codon usage of the chosen organism, ribosome binding sites, and gene overlap to make the best possible predictions [28]. The selection of different and evolving annotation methods or manipulation of variable input parameters can result in different annotations for the same gene.

At time of writing, the NCBI Protein database contained 1,311 records for the YbfE protein. Of these, 524 (40%) are 97 residues in length; 616 (47%) are 120 residues; the remaining 13% of sequences fall in a range between 41 and 151 residues. The nucleotide coding sequence of 97 residue YbfE (97aa-YbfE) is shown in Figure 1.3 with the start codon beginning at 1. The 150

---

**Figure 1.2** Sensitivity of deletion strains to alkylation damage.

*Inconsistent annotation in databases*
upstream nucleotides appear above, with the LexA binding sequence and alternate 5′ start codon that results in 120-residue YbfE (120aa-YbfE) highlighted. In order to address this potential misannotation and better predict the 5′ end of the gene for further characterization, prior literature that investigated ybfE RNA transcript levels is considered below.

Evidence in favor of 120aa-YbfE

In [26], the ybfE gene was shown to be LexA-regulated by Northern Blot which employed the use of a complementary DIG-labeled oligo. The complementary sequence (highlighted in Figure 1.3) for this oligo is located entirely upstream of the 5′ start coding for 97aa-YbfE, experimentally indicating that the transcript of the gene begins prior to this site. Additional evidence comes from genome microarray experiments used for transcriptional response profiling. There are 49 array experiments listed in the Gene Expression Omnibus (GEO) database [29, 30] that include gene expression data for ybfE. These arrays all use
oligonucleotide probes designed specifically for each gene in the species of interest. One such commercially available array is the Affymetrix GeneChip® E. coli Genome 2.0 Array [31], which uses 11 perfect match probes per target gene. Of the 11 probes for ybfE, three contain sequences that would only be present in the earlier 5’ start. These probe recognition sequences are highlighted in yellow in Figure 1.3. The successful experimental use of ybfE probes complementary to sequence that would only be present in the longer transcript of the gene is indicative of misannotation in databases for E. coli ybfE genes. These findings also support a putative 120aa-YbfE gene product of 13.9 kDa. The modeling and experimental characterization that follow in Chapter 2 consider both 97aa-YbfE and 120aa-Ybfe, and highlight functional differences that are likely attributable to 120aa-YbfE being the product expressed in vivo.

Protein sequence analysis

There are a variety of methods that can be used to make predictions about the properties of a protein including molecular weight, pI, hydrophobicity, order/disorder, secondary structure, and conservation across species. A Domain-Enhanced BLAST of the NCBI protein database yielded the results show in Figure 1.4. The amino acid sequence of YbfE is conserved in Gammaproteobacteria, and its different occurrences across species comprise the PRK11675 superfamily. The C-terminal region of PRK11675 superfamily proteins contains a ribbon-helix-helix motif implicated in dimerization and DNA-binding. Secondary structure prediction and
homology modeling suggest a predominantly α-helical secondary structure for YbfE (Figure 1.4b and 1.5c).

Hydropathy [32, 33] and disorder [34] plots (Figures 1.5a and 1.5b) predict a disordered, hydrophilic region from residues 42-75 connecting an ordered α-helical N-terminus and the RHH domain. Secondary structure prediction [35] shows largely α-helical secondary structure, with some β-sheet patches in the N-terminus corresponding to regions of increased order and hydrophobicity.

**Figure 1.5 a.)** Hydropathic character of amino acid residues plotted against the primary sequence of YbfE. **b.)** Prediction of disorder plotted against the YbfE sequence. **c.)** Secondary structure prediction for YbfE
CHAPTER 2- Characterization of the SOS response protein YbfE
2.1 Introduction

The general mechanisms of DNA damage repair and tolerance are conserved across all kingdoms of life. DNA damage repair and tolerance in bacterial cells is facilitated by the SOS response, which is triggered by the relatively poor ability of the replicative polymerase to process bulky adducts or strand breaks. Single-stranded DNA, which accumulates near the replication fork, is bound by the protein RecA to create a nucleoprotein-filament that facilitates the autocatalytic cleavage of the LexA repressor bound to SOS-boxes upstream of LexA-regulated genes. In order to characterize the role of the uncharacterized LexA-regulated gene ybfE and its product in enhancing the cellular response to alkylation damage, a three-pronged approach comprising computational modeling, \textit{in vitro} assays using purified protein, and phenotypic characterization was undertaken.

2.2 Materials and Methods

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling speed</td>
<td>Slow</td>
</tr>
<tr>
<td>PsiBLASTs</td>
<td>3</td>
</tr>
<tr>
<td>EValue Max</td>
<td>0.5</td>
</tr>
<tr>
<td>Max no. templates</td>
<td>5</td>
</tr>
<tr>
<td>Templates SameSeq</td>
<td>1</td>
</tr>
<tr>
<td>OligoState</td>
<td>4</td>
</tr>
<tr>
<td>Max no. alignments</td>
<td>5</td>
</tr>
<tr>
<td>LoopSamples</td>
<td>50</td>
</tr>
<tr>
<td>TermExtension</td>
<td>10</td>
</tr>
</tbody>
</table>

\textbf{Table 2.1.} Homology modeling parameters input to YASARA.

\textbf{Homology modeling}  The amino acid sequence of \textit{E. coli} YbfE was obtained from NCBI (CDU39555.1) in the form of a FASTA file. This file was loaded in YASARA and models were built using the standard homology building macro, \texttt{hm\_build.mcr}. Two different homology models were built, one from a PDB search for templates and another using an assigned template, which was done using the template NikR structure (PDB 2HZV). The parameters are shown in Table 2.1. The resulting .yob files were saved as .pdb files and edited for presentation using Pymol. The PDB files were further analyzed by submitting them to POOL (Partial-Order Optimum
Likelihood) [36, 37] and CombFunc [38] servers for analysis. In cases where the resulting homology model was multimeric, only the monomer was submitted for analysis.

**Bacteriological techniques** All strains were grown at 37 °C in Luria broth (LB) supplemented with ampicillin (100 μg/mL) or chloramphenicol (25 μg/mL). Competent cells were prepared with the CaCl₂ method [39]. Cloning constructs were confirmed by sequencing (Eton Bioscience, MA). Transformation was carried out by incubating cells and plasmid on ice for 20 min, followed by a 5-min incubation at 37 °C, followed by 10 min on ice. A 1 h recovery at 37 °C with LB preceded plating on LB-agar plates with selective antibiotic.

**Strains** The full list of strains used is in Table 2.2. AB1157 derivatives were created by P₁vir transduction with lysate prepared from Keio collection single-gene knockouts [40] procured from the *E. coli* Genetics Stock Center (Yale University). For protein expression, the Rosetta (DE3) pLysS (Novagen) strain was used. DH5α was used for plasmid maintenance.

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Genotype or characteristics</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pCA24N</td>
<td>N-terminal His tag, Cm&lt;sup&gt;R&lt;/sup&gt;</td>
<td>[41]</td>
</tr>
<tr>
<td>pET15b</td>
<td>N-terminal His tag, cloning vector, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Novagen</td>
</tr>
<tr>
<td>pGYCXNKN</td>
<td>pGB2 derivative, pSC101 ori, LexA consensus sequence with o₁&lt;sup&gt;C&lt;/sup&gt; mutation, Spec&lt;sup&gt;R&lt;/sup&gt;</td>
<td>[42]</td>
</tr>
<tr>
<td>pShYbfE15b</td>
<td>Short <em>ybfE</em> cloned in the pET15b vector at the NdeI and XhoI site</td>
<td>This work</td>
</tr>
<tr>
<td>pLYbfE15b</td>
<td>Long <em>ybfE</em> cloned in the pET15b vector at the NdeI and XhoI site</td>
<td>This work</td>
</tr>
<tr>
<td>pShYbfE24N</td>
<td>Short <em>ybfE</em> cloned in the pCA24N vector</td>
<td>[41]</td>
</tr>
<tr>
<td>pLYbfE24N</td>
<td>Long <em>ybfE</em> cloned in the pCA24N vector</td>
<td>This work</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strain</th>
<th></th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH5α</td>
<td><em>fhuA</em>2 <em>lac(del)U169 phoA glnV44 Φ80' lacZ(del)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17</em></td>
<td>Lab stock</td>
</tr>
<tr>
<td>Rosetta (DE3) pLysS</td>
<td>F&lt;sup&gt;−&lt;/sup&gt; ompT hsdS&lt;sub&gt;B&lt;/sub&gt;(r&lt;sub&gt;B&lt;/sub&gt; m&lt;sub&gt;B&lt;/sub&gt;) gal dcm (DE3) pLysSRARE (Cam&lt;sup&gt;R&lt;/sup&gt;)</td>
<td>Novagen</td>
</tr>
<tr>
<td>AB1157</td>
<td><em>thr-1, araC14, leuB6</em>(Am), Δ(gpt-proA)62, lacY1, tsx-33, qsr'-0, glnX44(AS), galK2(Oc), λ, Rac-0,*</td>
<td>Lab stock</td>
</tr>
</tbody>
</table>
Conjugation The two putative ORFs for the *ybfE* gene were amplified by colony PCR from the AB1157 chromosome. The 5’ and 3’ primers contained NdeI and XhoI restriction sites, respectively, for cloning into the pET15b plasmid (Novagen). For the long ORF, the primer pair 5’-GCCATAGCATATGTATTATGGCGCGCTATCCATCCG-3’ and 5’-GGTCTCGAGATATTAAACGATGCCCTGACTACGCAGCGC-3’ was used to generate a 386-bp amplicon. For the short ORF, the 5’ primer 5’-GCCATAGCATATGGCCAAAGAACAAACGGACCGTACG-3’ and the same 3’ primer were used to generate a 313 bp amplicon. Amplicons and pET15b were incubated at 37 °C for 3 h with NdeI and XhoI enzymes (New England Biolabs, MA). SAP and ligation were carried out using a kit (Roche). DH5α cells were transformed with ligated plasmid and selected on ampicillin. Cloning of the long *ybfE* ORF into pCA24N was done as previously described [41].

**Table 2.2.** Plasmids and strains used in this work. * - Becky Leifer is a former member of the Beuning Lab.
Phenotypic characterization  Strains were incubated at 37 °C for 16 h with aeration. The saturated overnight cultures were serially diluted in a 96-well plate from $10^{-1}$ to $10^{-7}$ in 0.85% saline in triplicate for each strain. A 10-µL aliquot from each well was plated on LB-agar plates containing 25 µg/mL chloramphenicol and 0 mM, 0.15 mM, 0.25 mM, 0.5 mM, 0.75 mM, or 1 mM IPTG. The plates were incubated for 16 h at 30 °C, 37 °C, or 42 °C. Colony growth for each serial dilution was ascertained by counting and multiplied by the dilution factor to give CFUs, for a total of 21 replicates for each strain at a given temperature and concentration of IPTG. (Lawn growth was not included in calculations.) CFUs were normalized to 0 mM IPTG values for each strain and temperature.

Protein expression and purification  97aa-YbfE was expressed with an N-terminal His-tag from the pET15b vector. To induce protein expression, IPTG was added to a final concentration of 0.5 mM in culture between OD$_{600}$ 0.4-0.6. The culture was incubated at 30 °C for 4 h, before cells were harvested by centrifugation at 4 °C for 10 min at 6000 x g. The cell pellet was stored at -80 °C until thawing was carried out overnight at 4 °C in lysis buffer (50 mM HEPES pH 7.5, 100 mM NaCl, 2 mM βME, 100 µM PMSF, Protease Inhibitor Cocktail (Sigma)) using 5 mL/g pellet wet weight. All subsequent steps were carried out at 4 °C. Lysozyme was added to 1 mg/g and incubated for 1 h. The pellet was then further lysed by sonication. The cell lysate was clarified by centrifugation at 11k x g for 1 h followed by sterile filtration with a 0.45 µm filter. Clarified lysate was loaded on a 5 mL equilibrated HiTrap Heparin HP column (GE). The column was washed with 25 CV of A buffer (50 mM HEPES pH 7.5, 100 mM NaCl, 2 mM βME). Bound protein was eluted in 5 mL fractions over a 25 CV gradient to 100% B buffer (50 mM HEPES pH 7.5, 1 M NaCl, 2 mM βME). Fractions were analyzed by 14% SDS-PAGE. Fractions containing 97aa-YbfE were dialyzed overnight against A$_2$ buffer (50 mM HEPES pH
8.0, 300 mM NaCl, 2 mM βME, 10 mM imidazole), then loaded on an equilibrated 5 mL HisTrap HP (GE) column and washed with 15 CV A2 buffer. Bound protein was eluted in 5 mL fractions over a 20 CV gradient to 100% B2 buffer (50 mM HEPES pH 8.0, 300 mM NaCl, 2 mM βME, 500 mM imidazole). Fractions were analyzed by 14% SDS-PAGE and fractions containing pure 97aa-YbfE were concentrated and exchanged into Storage buffer (50 mM HEPES pH 7.5, 200 mM NaCl, 2 mM βME, 10% glycerol). The N-terminal His-tag was removed using the Thrombin CleanCleave Kit (Sigma).

**EMSAs** YbfE was incubated at concentrations of 0-4 μM with 1.7 μM 45 bp linear dsDNA (5'-ACTCCTGGCGACTTGTA TTCAGCTAAGACACTGCACTGGATTAAG-3') in 10 mM Tris pH 8 and 1 mM EDTA for 20 min at ambient temperature. Each entire reaction mixture was loaded on a 15% acrylamide-TAE gel and run for 85 min at 125 V before staining with SYBR Gold (Thermo) and imaging on a STORM 860 Phosphorimager using the Blue (450 nm) channel.

**Formaldehyde Crosslinking** Each protein (1 nmol) was incubated in 50 mM HEPES pH 8, 100 mM KCl, 10 mM Mg(OAc)2, 1 mM DTT, and 1% formaldehyde for 30 min at room temperature. In reactions containing DNA, the DNA concentration was 1.56 nM of 45 bp linear dsDNA. Aliquots (10-μL) were removed at time points of 0, 5, 10, 20, and 30 min after the addition of formaldehyde and immediately quenched with 4X SDS-dye, and stored on ice until analysis by 14% SDS-PAGE at 125 V for 85 min.

**Circular Dichroism** The short YbfE construct 97aa-YbfE was dialyzed in 15 mM sodium phosphate pH 7.5 for 16 h at 4 °C. Concentration was determined by Bradford assay (Bio-Rad) and the protein diluted to 5 μM with 15 mM sodium phosphate pH 7.5. Analysis was carried out
in a Jasco J-715 spectropolarimeter using a 1 mm path length quartz cuvette (Starna). A buffer blank spectrum was obtained and subtracted from the spectrum with protein.

2.3 Results

Homology Modeling

Homology models based on templates found by YASARA

The PDB template search run by YASARA when no template was provided yielded 2 hits:

1.) Geranylgeranyl Pyrophosphate Synthetase from Pyrococcus horikoshii Ot3
   PDB ID: 1WY0-A
   89% coverage

Geranylgeranyl pyrophosphate synthetase (GPS) is an enzyme that performs condensation reactions to join isopentenyl diphosphate and geranylgeranyl-diphosphate. The sequence alignment between this protein and YbfE covered the N-terminal domain as well as some of the RHH domain. The RHH domain was more similar to the second template, which follows.

2.) Yiif from Shigella flexneri (serotype 5b, strain 8401)
   PDB ID: 2K5J-A
   44% coverage

Yiif is an RHH protein conserved in bacteria including E. coli. The protein does not have an assigned function, and the structure was solved as part of a Structural Genomics project. The sequence alignment of Yiif with YbfE covered the RHH domain of the protein. Both of the templates were monomers and thus yielded monomeric homology models. The model built using the Yiif template covered 42/120 of the target residues, mainly in the RHH motif. The aligned residues share 31% sequence identity and 57% sequence similarity (Figure 2.1).
The final model based on Yiif was ranked Optimal by YASARA in terms of dihedrals and packing, but it did not cover the complete target sequence. The model built from the GPS template covered 107/120 target residues with 13% sequence identity and 35% sequence similarity.

The final model based on the GPS template was ranked Good by YASARA, and covered 100% of the target sequence. The hybrid model built from the two templates covered 100% of the target sequence of YbfE and was ranked Good overall. Residues 60-120, which included the RHH motif, were made from the Yiif template, while the N-terminal residues were made from the GPS template.

Figure 2.1 Sequence alignment between YbfE and the templates found by YASARA that were used to build the hybrid homology model. The N-terminal domain is shown in blue, while the RHH motif is shown in orange. An additional 184 residues from the N-terminal domain of GPS were not included in the alignment.

The hybrid model from the two templates covered 100% of the target sequence of YbfE and was ranked Good overall. Residues 60-120, which included the RHH motif, were made from the Yiif template, while the N-terminal residues were made from the GPS template. POOL analysis was performed on the hybrid model PDB file and the residues were ranked based on likelihood they are involved in protein function. The top 10 POOL-predicted residues are shown in Table 2.3. CombFunc analysis of the PDB file predicted a biochemical function of DNA binding, and produced results for nests, groups of adjacent amino acids that are frequently involved in catalytic functions. The nest residues were [Pro43, Gly44, Arg45] and [Arg95, Asn96, Met97]. These and the top five POOL ranked residues are shown in the homology model (Figure 2.2a).

<table>
<thead>
<tr>
<th>Rank</th>
<th>POOL score</th>
<th>Residue Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.13E-04</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>3.40E-05</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>1.38E-05</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>9.60E-06</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>6.21E-06</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>3.30E-06</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>2.16E-06</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td>1.21E-06</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>7.68E-07</td>
<td>62</td>
</tr>
<tr>
<td>10</td>
<td>6.08E-07</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 2.3. Top 10 POOL ranked residues for the hybrid model.
Figure 2.2  a.) Monomeric homology model built from templates 1WY0-A and 2K5J-A. The RHH domain is shown in yellow and the N-terminal domain is shown in blue. The top five POOL residues are shown in cyan, and the six nest residues are shown in orange. The three green hydrophobic residues correspond to conserved hydrophobic residues in the RHH motif. b.) The consensus dimeric RHH structure. The three conserved hydrophobic residues in the β-strand are shown as green spheres. The monomer in green is aligned similarly to the yellow RHH motif in the homology model. c.) Dimeric homology model built using NikR as a template. Arginine interactions with the DNA phosphate backbone are shown.
**NikR-template**

The homology model built using the specified template NikR covered 38/120 target residues from YbfE (Figure 2.3). The template consisted of a NikR tetramer, a Mg$^{2+}$ substrate, and bound DNA. The homologous region was the shared RHH motif, which was the only domain included in the model. Sequence identity in this region was 37%, and sequence similarity was 66%. The overall homology model was ranked Good by YASARA. The resulting model is shown in Figure 2.2c. A monomer from the homology model was analyzed using POOL; the top 10 ranked residues are shown in Table 2.4.

<table>
<thead>
<tr>
<th>Rank</th>
<th>POOL score</th>
<th>Residue Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.77E-04</td>
<td>GLU:A 104</td>
</tr>
<tr>
<td>2</td>
<td>1.69E-04</td>
<td>GLU:A 105</td>
</tr>
<tr>
<td>3</td>
<td>1.63E-04</td>
<td>GLU:A 101</td>
</tr>
<tr>
<td>4</td>
<td>7.29E-06</td>
<td>ILE:A 103</td>
</tr>
<tr>
<td>5</td>
<td>1.2E-06</td>
<td>LEU:A 107</td>
</tr>
<tr>
<td>6</td>
<td>1E-06</td>
<td>SER:A 100</td>
</tr>
<tr>
<td>7</td>
<td>8.64E-07</td>
<td>MET:A 106</td>
</tr>
<tr>
<td>8</td>
<td>4.8E-07</td>
<td>MET:A 108</td>
</tr>
<tr>
<td>9</td>
<td>2.43E-07</td>
<td>ARG:A 99</td>
</tr>
<tr>
<td>10</td>
<td>1.44E-07</td>
<td>LEU:A 102</td>
</tr>
</tbody>
</table>

**Table 2.4.** Top 10 POOL ranked residues for the model based on NikR.

**Figure 2.3** Sequence alignment between YbfE and the NikR template. The N-terminal domain of YbfE is not included in the alignment, while the C-terminal of NikR is shown in blue.
Induction and purification of YbfE proteins

The proteins 6His-97aa-YbfE and 6His-120aa-YbfE are strongly overexpressed after addition of IPTG to the culture, as shown in Figure 2.4. However, after lysis under the same conditions, 6His-97aa-YbfE remains soluble while 6His-120aa-YbfE aggregates in the cell debris pellet. The formation of insoluble protein in the case of 120aa-YbfE but not 97aa-YbfE suggests a marked difference in folding or activity between the two proteins.

Purification of 97aa-YbfE resulted in high yields of His-tagged protein detectable by αHis-tag western blot. Using the same purification protocol with 120aa-YbfE, very little protein was recovered from cell lysate, and it did not interact with the αHis-tag antibody. Due to the small yield and questionable quality of the purified 120aa-YbfE protein, 97aa-YbfE with His-tag removed was used for further in vitro characterization.

YbfE binds linear dsDNA as shown by EMSA

To investigate whether YbfE binds DNA as predicted, 97aa-YbfE was incubated at increasing concentrations with a constant amount of DNA. When these...
reactions are resolved by electrophoresis and stained for DNA, a shifted band appears beginning at 1 μM YbfE and increases in intensity up to 4 μM YbfE. All of the DNA appears to be bound at 4 μM YbfE, which is roughly a stoichiometry of 2:1 (protein:DNA). The appearance of a smear in the shifted bands is likely the result of complex destabilization as the sample enters the gel. These reactions show DNA binding beginning to occur between 500 nM and 1 μM concentrations of 97aa-YbfE.

*YbfE forms dimers that can be cross-linked by formaldehyde*

The high value of the predicted pI of YbfE prevented us from using native PAGE to characterize oligomerization. 120aa-YbfE had a predicted pI of 9.81, while 97aa-YbfE had a predicted pI of 10.09 [43]. In order to examine the prediction of dimerization on the basis of the RHH domain, formaldehyde crosslinking was chosen because of its short crosslinking distance (2.3-2.7 Å) to covalently link interacting proteins. In Figure 2.6, the 0 min time point was taken prior to the addition of formaldehyde and shows YbfE in a band at ~14 kDa. In the subsequent time points, the formation of dimer at ~28 kDa is observed coinciding with a reduction in the intensity of the monomer band. No higher molecular weight bands were observed, suggesting that higher-order oligomers do not form under the conditions investigated. This experiment was repeated with the inclusion of DNA at 1.56 nM concentration with no observable difference in oligomerization (data not shown), suggesting that DNA is not required for dimerization nor does it induce the formation of additional higher-order oligomers.

**Figure 2.6** Formaldehyde crosslinking of YbfE showing the formation of dimer was carried out in the presence of 1% formaldehyde as described in Materials and Methods. The formation of dimer is apparent after 5 min.
The secondary structure of 97aa-YbfE is α-helical

Difficulty obtaining a melting point for YbfE by differential scanning fluorimetry prompted us to consider that YbfE may be disordered in the absence of DNA. To investigate this, circular dichroism spectroscopy was employed to characterize the native secondary structure of YbfE. The resulting spectrum is shown in Figure 2.7, which displays the characteristics of a prototypical α-helical secondary structure with a positive band at 193 nm and negative bands at 208 nm and 222 nm. The addition of DNA did not bring about a change in this secondary structure (data not shown), but YbfE-DNA binding may have been disrupted under the phosphate buffer conditions. EMSAs carried out in phosphate buffer did not exhibit gel shifts compared to the same experiments done in TE or HEPES in the presence of NaCl. This spectral data agree with the homology model and secondary structure prediction, and indicate that YbfE is not disordered in the absence of DNA.

Overexpression of YbfE is lethal to cells

The proteins 120aa-YbfE (AB1157 pLYbfE24N) and 97aa-YbfE (AB1157 pShYbfE24N) were expressed from the ASKA pCA24N plasmid. Expression was induced using varying concentrations of IPTG in the solid medium during overnight growth.
The strain harboring the empty vector (AB1157 pCA24N) experienced no growth defects at any of the IPTG concentrations tested. As shown in Figure 2.8, AB1157 pShYbfE24N showed decreased survival on the order of $10^4$ compared with wild-type at the same concentrations of IPTG. AB1157 pLYbfE24N showed decreased survival on the order of $10^7$ compared with AB1157 pCA24N.

**Figure 2.8 a.)** Overexpression of 97aa-YbfE and 120aa-YbfE compared to EV at different levels of induction with IPTG. The 120aa-YbfE causes increased cell death compared to 97aa-YbfE. **b.)** Image of serial dilutions representing the survival of different strains upon YbfE overexpression. The most concentrated samples are at the top.

*Deletion of uvrC rescues the overexpression lethal phenotype*

Ribbon-helix-helix proteins can act as transcription factors. To uncover potential functional links between other DNA damage response genes and *ybfE*, 12 knockout strains (PB101 Δ*recJ*, PB108 Δ*aadA*, PB109 Δ*alkB*, PB113 Δ*mutS*, PB114 Δ*nfo*, PB117 Δ*recN*, PB120 Δ*sbmC*, PB122 Δ*symE*, PB124 Δ*uvrC*, PB125 Δ*uvrD*, PB126 Δ*uvrY*, and PB128 Δ*yebG*, Table 2.2. Not all data shown.) were subjected to 120aa-YbfE overexpression from the pLYbfE24N plasmid. The *uvrC* knockout strain PB124 (AB1157 Δ*uvrC*) performed similarly to AB1157
harboring the pCA24N empty vector, rescuing the cells from the lethal overexpression effects of YbfE. (Figure 2.9).

![Graph](image)

**Figure 2.9.** Survival of WT and PB124 strains overexpressing 120aa-YbfE at an IPTG concentration of 0.15 mM.

### 2.4 Discussion

Computational modeling can act as a vital first step in the characterization of new or uncharacterized proteins. In this work, we used free and openly-accessible resources in addition to molecular modeling software to help guide hypotheses and experiments. Because protein structure and function are so intimately linked, a homology model is an important place to start understanding an uncharacterized protein. It allows visualization of the molecule of interest and identification of putative functional residues. In the work presented, YASARA was used to create a homology model which was then analyzed with POOL [36] and CombFunc [38] to predict functional residues. YASARA creates homology models by running PSI-BLAST [44]
against Uniprot [45] and generating a position-specific profile that is used to search the Protein Data Bank [46] for a suitable template. POOL processes three-dimensional coordinate files and ranks residues on the basis of electrostatic properties and local geometry. Highly ranked residues are predicted to be involved in the function of a protein. CombFunc makes sequence-based functional predictions by querying multiple sequence databases for known motifs or clusters of charged residues.

For the non-multimeric and non-DNA bound model built from templates found by YASARA, the overall conformation of the C-terminal RHH domain of YbfE was modeled based on the RHH protein Yiif. Yiif is also uncharacterized, but the RHH domain is associated with DNA binding in the characterized proteins MetJ, NikR, Arc, and others [47-50]. The overall conformation of the RHH domain, shown in yellow in Figure 2.2a, shares strong structural similarity with the monomeric consensus RHH motif in green in Figure 2.2b. The RHH motif consists of an N-terminal β-strand followed by two α-helices. The N-terminal β-strand contains alternating hydrophilic and hydrophobic amino acids, the residues of which point in opposite directions. The hydrophilic residues interact with DNA bases and contribute to specificity. A loop of varying length connects the two α-helices, resulting in different dimerization and DNA binding conformations [51]. Together with the hydrophobic residues from the β-strand, several hydrophobic branched-chain amino acids form a hydrophobic core responsible for binding the DNA backbone when the protein is a homodimer. N- or C-terminal extensions that are present in an RHH sequence are often involved in interactions with other proteins or substrates that modulate the binding capabilities of the RHH domain through electrostatic interactions or conformational changes. For example, RHH protein MetJ [52] is a repressor of genes involved in methionine biosynthesis. A C-terminal extension binds S-adenosylmethionine and increases the
affinity of MetJ for DNA, lending to an overall mechanism of auto inhibition. YbfE has a 70+ amino acid highly-conserved N-terminal extension predicted to contain three alpha-helical regions and a long, potentially disordered domain. The RHH α-helices in the homology model are positioned at a wider angle relative to the consensus structure, while the RHH β-strand region extends upward and joins with the N-terminal domain, which extends behind and to the left of two RHH α-helices as pictured. Three residues including Arg71, Arg75, and Leu78 point outward from the β-strand region to form a hydrophobic core that binds the backbone of the DNA major groove when the protein is in a homodimeric form. The β-strand structure in the homology model is modeled as a coil instead of a ribbon; this might be because the β-ribbon structure represents an induced conformational change that occurs when the protein dimerizes.

The model built using NikR as a template only included the RHH domain and neglected the N-terminal domain. The amino acid sequence of NikR contains an RHH motif at the N-terminus of the protein, while a C-terminal extension contributes to nickel binding and specificity. Due to the opposite orientation of this extension and the lack of C-terminal residues after the RHH domain of YbfE, this model is well-suited for modeling the RHH domain but poorly suited to elaborating the structure of the N-terminal extension. The dimeric model shown in Figure 2.2c is highly similar to the consensus RHH structure in Figure 2.2b. The Arg99 backbone contacts residues in the coil connecting the two α-helices. The amine hydrogens interact with the electronegative phosphate groups of the DNA backbone at distances of 2-3 Å. The β-strand is better defined in the NikR model. The conserved hydrophobic residue Asp75 is again shown to point toward the core of the RHH2 domain, where it interacts non-specifically with the DNA backbone. The top two POOL-ranked residues Glu104 and Glu105 are predicted to bind Mg$^{2+}$ on the outward-facing side of the C-terminal helix (not pictured). Most of the top-
ranked POOL residues are present in the helices of the RHH domain for this model and may be important in dimerization interactions.

The N-terminal domain in the hybrid model, which shares sequence similarity with the transferase GPS, is modeled as a long α-helix beginning at the N-terminus of the protein. It is connected to a short α-helix by a 19-residue coil spanning positions 38-56 and extending behind the RHH domain. This coil alone contains four Arg residues that could interact with the DNA backbone during binding, and contains three of the top five POOL residues for the hybrid model (Table 2.4). Negatively-charged residues such as Glu40, Asp54, and Glu55 could interact with DNA bases to contribute to sequence specific binding.

The POOL results appear to reflect the bias contributed by the more prevalent domain of each model. The NikR template model consisted entirely of the C-terminal RHH domain, and POOL assigned high ranks to those residues involved in binding the DNA backbone, dimerization, and binding a Mg$^{2+}$ ion. The hybrid model included the enigmatic N-terminal domain, and the POOL results mainly comprised N-terminal residues. The coiled region is particularly attractive as a site that could contribute to the DNA sequence specific binding. It is also possible that some regulatory substrate binds in this region and affects the binding affinity of YbfE for DNA. The conserved catalytic and substrate binding motifs of GPS are not correlated with its sequence in the area used to model the coil, so a potential substrate is not apparent.

To begin experimental confirmation of the structure and function of YbfE, the protein needed to be expressed and purified. An N-terminal His-tag expression vector (pET15b) was selected for ease of purification by affinity chromatography. Both the 120aa-YbfE and 97aa-YbfE were expressed, but thus far only 97aa-YbfE has been successfully purified. The cell death
caused by YbfE overexpression, particularly of 120aa-YbfE, could result in the formation of inclusion bodies or cause protein degradation, which impedes purification. Because 97aa-YbfE still contains the functional RHH motif and part of the N-terminal domain, in vitro assays using the available protein were prioritized. Using the predicted structure and function for YbfE as a dimeric α-helical DNA binding protein, several techniques were selected which could offer experimental support of these predictions. Electrophoretic mobility shift assays are based on the observation that protein-DNA complexes normally possess less mobility than free DNA. [53] The technique offers a semi-quantitative approach to assay DNA binding. Relative band intensity can be quantified and used to calculate binding affinities [54]. The EMSAs done as part of this work showed YbfE-dsDNA binding at a stoichiometric ratio of 2:1 (protein:45 bp-oligonucleotide). This could correspond to roughly one dimer of YbfE per oligonucleotide, although the smearing observed in the lanes is indicative of additional binding states. Other RHH protein dimers have been shown to bind adjacent to one another at DNA sequence sub-site repeats, forming cooperative protein-protein contacts between dimers and covering varying lengths of DNA sequence such as occur near promoter regions [51]. The formation of YbfE dimers was observed by formaldehyde cross-linking, and α-helical secondary structure confirmed by circular dichroism. Thus, confirmation of three basic structural and functional predictions was achieved. However, with little information to guide in vitro assays, additional phenotypic characterization was sought.

Interestingly, genome microarray experiments done in UV-induced cells do not show ybfE transcript upregulation [23, 27] while Northern Blot analysis in cells damaged with mitomycin C or cells that are LexA deficient do show increased levels of transcript [26]. Our own observations do not show increased UV sensitivity in ΔybfE strains (data not shown),
though these strains are sensitive to alkylating agents (Figure 1.2). The \textit{ybfE} gene promoter has one of the highest heterology indices (HI) of 14.07 on a scale from 0 – 15, with 0 being the LexA binding consensus sequence [25, 26]. A higher heterology index is associated with lower binding affinity with LexA and a decreased threshold for SOS-induction. Genes expressed earlier in the SOS-response can be thought of as first line responses, while more tightly regulated genes such as \textit{sulA} (a cell division inhibitor) and \textit{umuDC} (a mutagenic TLS polymerase) have more dramatic effects on cells and are more useful in cases of acute or prolonged DNA damage. The \textit{uvrA} and \textit{uvrB} genes of the nucleotide-excision repair pathway are among the first to be upregulated after SOS induction [27, 55]. The UvrABC pathway has been shown to play an important role in SOS induction, increasing the rate of LexA degradation by helping to create ssDNA substrate for RecA polymerization [56]. Additionally, it results in a less prolonged SOS response by removing lesions [56]. The \textit{uvrC} promoter is not under LexA control [57-59], and it does not appear to be damage inducible. The early induction of YbfE suggested by its LexA box might coincide with induction of the NER response.

The overexpression lethality phenotype exhibited for YbfE, particularly 120aa-YbfE, has two important implications. Firstly, the increased killing by the long construct 120aa-YbfE compared to 97aa-YbfE supports a functional importance for the N-terminus. Secondly, the rescue of cells from overexpression lethality by \textit{uvrC} deletion implies an important functional role for UvrC in cell killing. UvrABC endonuclease has been shown to excise undamaged DNA [60], so YbfE could act on the pathway in such a way that non-specific NER activity is increased or suppressed. Given the presence of RHH motifs in transcription factors, YbfE binding to the promoter regions of the \textit{uvrABC} genes warrants investigation. YbfE could act as a regulator in a RecA-LexA-independent DNA damage pathway [26]. SOS-induced yet LexA-independent
regulation of the pathway would exert an additional level of molecular control over the NER pathway and promote crosstalk between damage response mechanisms.
CHAPTER 3 – Conclusions and Future Work
3.1 Conclusions

The \textit{ybfE} gene is LexA-regulated and therefore undergoes transcriptional upregulation following SOS-induction in response to DNA damage. Our current inquiry into the function of the uncharacterized \textit{ybfE} gene was sparked by the observation that its deletion increases cell sensitivity to alkylating agents. To better understand the role \textit{ybfE} plays as part of the SOS response, we first used computational tools to generate sequence-based predictions about the structure and function of the gene product. We found that the amino acid sequence of YbfE is conserved in Gammaproteobacteria, and its different occurrences across species comprise the PRK11675 superfamily. Inconsistent annotation of the 5′ start of the \textit{ybfE} gene in biological databases prompted examination of two putative open reading frames: one encoding a 120 residue protein (120aa-YbfE), and one encoding a 97 residue protein (97aa-YbfE) lacking 23 N-terminal residues compared with 120aa-YbfE. The C-terminal region of PRK11675 superfamily proteins contains a ribbon-helix-helix motif implicated in dimerization and DNA-binding.

Secondary structure prediction and homology modeling suggest a predominantly \(\alpha\)-helical secondary structure for YbfE. To validate these predictions experimentally, the \textit{ybfE} gene was cloned into vectors for overexpression in the 97aa-YbfE and 120aa-YbfE forms. Overexpression of 97aa-YbfE yielded a stable protein that bound dsDNA in EMSAs, and this protein was shown to have \(\alpha\)-helical secondary structure using circular dichroism. Formaldehyde cross-linking showed the formation of YbfE dimers in solution. These experimental findings agree with the computational predictions. Overexpression of 97aa-YbfE or 120aa-YbfE is lethal to cells, with 120aa-YbfE overexpression causing \(10^3\) decreased survival compared to 97aa-YbfE induced at the same concentration of IPTG. This constitutes additional evidence for a 120 amino acid \textit{ybfE} gene product. Deletion of the \textit{uvrC} gene rescues cells from the effects of YbfE overexpression,
implicating the NER endonuclease UvrC as a factor in YbfE-induced lethality. It is hypothesized that YbfE affects the NER pathway, possibly as a transcription factor. Additional studies to examine this hypothesis are warranted.

3.2 Ongoing and Future Work

To better understand the effects of YbfE overexpression lethality, quantification of the expressed protein is necessary. This can be done by αHis-tag Western blot on cell extract from IPTG-induced AB1157 pLYbfE24N cultures. Experiments to examine the hypothesis of uvrC transcriptional regulation are underway using fluorescent transcriptional reporters [61]. In these experiments AB1157 pCA24N, AB1157 pShYbfE24N, and AB1157 pLYbfE24N are transformed with a plasmid containing the gene encoding GFP downstream of the uvrC promoter. An increase in the detected fluorescence correlating with induction of YbfE overexpression would indicate transcriptional upregulation and thus YbfE binding specificity for the uvrC promoter sequence can be tested by EMSA. To further examine this hypothesis, αUvrC Western blotting could detect an increase in expression of UvrC in response to YbfE overexpression. DNA extracted from cells in which YbfE is overexpressed can be assayed for strand breaks caused by non-specific endonuclease activity. Finally, co-overexpression of the downstream NER proteins UvrD or DNA polymerase I could be tested for rescue of lethality by offering additional means of repairing non-specific incisions.

Some other immediate goals include cloning the ybfE gene coding 120aa-YbfE into a non-tagged expression vector such as pET11T and developing a purification protocol. Difficulty purifying 120aa-YbfE could be caused by the observed overexpression lethality, and alleviated by expression in a ΔuvrC strain. This would allow in vitro examination of the structural and functional characteristics of the untagged full-length protein. DNA binding experiments such as
the EMSAs should be repeated using a variety of DNA substrates including ssDNA of varying length, hemi-methylated and fully methylated DNA, primer-template pairs, and specifically chosen sequences. A crystal structure of YbfE binding to DNA would be highly desirable as well.

Preliminary data from fluorescence quenching experiments suggests an interaction between YbfE and RecA. This experimental data needs replication. YbfE should also be assayed for protein-protein interactions with NER proteins.

YbfE variants based on computational predictions are also under production in the pCA24N constructs. Several variant constructs have already been made and need to be assayed for changes in the overexpression-lethal phenotype. Key residues for binding or activity could thus be implicated, and the variant of interest cloned into an expression vector for in vitro characterization following purification.

In conclusion, we have shown that the SOS-regulated ybfE gene plays a role in DNA damage tolerance. The ybfE gene product binds to DNA and forms dimers in solution, as predicted by its RHH motif. The N-terminal domain also plays in important role in YbfE’s function, which awaits further experimentation as outlined here.


