Expression, Cloning and Purification of Two Bacterial Transcription Factors of the IclR Family of Proteins

By

Brian Haichet Sithoo
B.Sc. University of Manitoba, 1998

October, 2003

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In

THE FACULTY OF GRADUATE STUDIES

Of

THE UNIVERSITY OF MANITOBA

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

Expression, Cloning and Purification of Two Bacterial Transcription Factors of the Ic1R Family of Proteins

BY

Brian Haichet Sithoo

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

MASTER OF SCIENCE

"To my wonderful family and friends, whose love and support sustain me..."

-Brian H. Sithoo

ACKNOWLEDGEMENTS

To my wonderful and loving parents Chin and Swee Chin Si-Thoo; I wish I could say everything for everything that you are and everything that you mean to me. All I can say is that without you, I would not be the man I am today nor would I be able to become the man you envision. I love you both with everything that I am.

I would like to express my thanks to Dr. Harry W. Duckworth and the other members of my M.Sc. committee, Drs. Frank Hruska and Pavel Dibrov. Thank you all for agreeing to be my committee for one, but for also providing valuable insight and advice for the preparation of my thesis. At this point, I would like to take this opportunity to extend my utmost and humblest gratitude to my M.Sc. supervisor, Dr. Harry W. Duckworth for his unwavering and limitless patience, guidance, wisdom and inspiration. Undertaking this thesis project was indeed a wonderful opportunity because I was not only able to learn a great deal about molecular biology, biochemistry and science in general, but I also gained many valuable life experiences. Dr. Duckworth, I cannot thank you enough for all that you have done for me and for all that you have been to me. You have been more than just a supervisor. You have been my teacher, my guide, my mentor and my friend. God bless you.

To Dr. Lynda J. Donald, I thank you kindly for teaching and training me so well. I am a better scientist and researcher as well as a better person because of you. Thank you for everything Lynda.

To Stasa Veroukis, Sudeep Shivakumar, Gillian Sadler, Charlton Cooper, Susan Cuvelier and David Stokell; I was fortunate to have such fine fellow lab members through out the years who were also my friends. Dave, we were friends before our time together in the Duckworth lab, but we will be friends forever. I cannot thank you enough for simply being my friend, but of course I could not have made it through without you.

Thanks to Dr. Joseph D. O'Neil for his generous contribution of the pET vectors used in this thesis work and to Drs. Gillian Henry and Darren Manley for many useful discussions.

To the Department of Chemistry at the University of Manitoba; this is the finest thing that has ever happened to me. Whereas my B.Sc. was first in my heart, now this occasion will take first place always. If I could take out of my life everything except my experiences at the University of Manitoba, I would still have a rich and full life.

To my dear friend Tamika Williams; we completed our undergraduate and graduate degrees together. Your friendship carried me though it all. May we indulge the next stage of our lives and careers successfully once again as the truest and best of friends.

To Pooran Jyoti Ghai; I had been a teaching assistant for the university for a number of years where I met you and so many others as such. I have never met a person as special and beautiful as you in my lifetime. Your friendship completes me as you have always been a most brilliant light in my life.

Last but not least, to my dear friends Jamie Galka and Khristofer Garcia; I met you both during my graduate studies. You are the finest friends a person could hope or ask for. May we remain the best of friends for an eternity and beyond.

ABSTRACT

Escherichica coli IclR (Isocitrate Lyase Repressor) protein is a 274 AA polypeptide that heads a family of homologous sequence-specific DNA-binding proteins, the so-called IclR family of homologues. This family of homologous proteins shares the following characteristics: 1) they function as transcriptional regulators in carbon metabolism, 2) they are 200 – 300 residues in length, 3) they are comprised of an N-terminal domain containing the helix-turn-helix DNA-binding motif and 4) a C-terminal domain containing a signal-dependent regulatory binding site (Nasser et al., 1994; Zhang et al., 2002). Most of these IclR family members remain as hypothetical gene products of open reading frames (ORFs) detected by complete genome sequencing projects. The annotation of well-characterized, partially characterized and uncharacterized IclR family members is an ongoing process and new members are constantly added from the efforts of sequence identity and homology searches.

E. coli IclR and another family member known as *E. coli* GclR (allantoin regulator) have been shown to bind interchangeably to each others target DNA sequences (Donald et al., 2001). Such discoveries strongly suggest that even in "simple" systems such as prokaryotic operons, these proteins do not function in isolation but rather interact with other DNA-binding proteins in oligomeric nucleoprotein complexes. This further suggests that both protein-DNA and protein-protein interactions are essential for their activities, and the allosteric regulations among binding partners give rise to observed cooperative interactions. However, though the target DNA sequences and the regulatory roles of both *E. coli* IclR and GclR have been determined (Donald et al., 2001), the structural information as well as the molecular details of biochemical function remains unclear. To understand the structural and functional characteristics of this family of transcriptional regulators, the individual members must be studied as a group. We are interested in the structure and function of DNA-binding proteins from the IclR family of homologues and their sequence-specific interactions with DNA.

This thesis investigation describes the isolation and purification and presents some early biochemical characterization of two uncharacterized *E. coli* polypeptides, YagI and YiaJ, from the IclR family of homologues using a combination of molecular biology, recombinant DNA technology and protein chemistry. The polymerase chain reaction (PCR) was used to generate two uncharacterized *E. coli* ORFs from the IclR family of homologous DNA-binding proteins, YagI and YiaJ. Synthetic DNA products were checked against the correct genomic sequences by automated DNA sequencing. The PCR amplified ORFs were then cloned and their respective protein products were expressed using the pET-Vector system™ of cloning and expression. Following purification by standard column chromatographic methods, the protein products were isolated and subjected to mass spectrometric analysis to verify the integrity of the proteins.

TABLE of CONTENTS

LIST of FIGURES

Figure **Page Number**

From Chapter II: MATERIALS and METHODS (Continued)

From Chapter III: RESULTS

Figure **Page Number**

From Chapter III: RESULTS (Continued)

Figure **Page Number**

From Chapter IV: DISCUSSION

LIST of TABLES

Table **Page Number**

From Chapter III: RESULTS (Continued)

From Chapter IV: DISCUSSION

LIST of ABBREVIATIONS

BAP	Bacterial Alkaline Phosphatase
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
BSA	Bovine Serum Albumin
CAD	Collision-Activated Dissociation
cAMP	Cyclic Adenosine Monophosphate
CAP	Catabolite Activator Protein
CIAP	Calf Intestinal Alkaline Phosphatase
CID	Collision-Induced Dissociation
Crp	Cyclic AMP Receptor Protein
(k)Da	(kilo)Dalton
DC	Direct Current Voltage
DDBJ	The Center for Information Biology and DNA Data Bank of Japan
dH_2O	Deionized Water
ddH_2O	Double Deionized Water
DHB	2,5-Dihydroxybenzoic Acid
DHDPS	Dihydrodipicolinate Synthetase
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleoside Triphosphate
DTNB	5,5'-Dithiobis(2-Nitrobenzoic Acid)
DTT	Dithiothreitol
EDTA	Ethylene Diamine Tetra Acetic Acid
EMBL	The European Molecular Biology Laboratory
ESI	Electrospray Ionization
EtBr	Ethidium Bromide
ExPASy	The Expert Protein Analysis System
GclR	Glyoxalate Carboligase Regulator
GdnHCl	Guanidine Hydrochloride
GPH	Glycoside-Pentoside-Hexuronide
H_2O	Water
HCl	Hydrochloric Acid
HMM	Hidden Markov Model
HPLC	High Performance Liquid Chromatography
HSF	Heat Shock Factor
HTH	Helix-Turn-Helix
IclR	Isocitrate Lyase Regulator
IPTG	Isopropyl-Beta-D-Thiogalactopyranoside
KCl	Potassium Chloride
LB	Luria Bertani
LiCl	Lithium Chloride
m/z	Mass-to-Charge Ratio
M_r	Molecular Weight

MALDI	Matrix Assisted Laser Desorption/Ionization
MS	Mass Spectrometry
NaCl	Sodium Chloride
NaOAc	Sodium Acetate
NaOH	Sodium Hydroxide
NCBI	The National Center for Biotechnology Information
NH_4HCO_3	Ammonium Bicarbonate
NMR	Nuclear Magnetic Resonance
Oct1-1	Octamer Transcription Factor 1
ORF	Open Reading Frame
PAGE	Polyacrylamide Gel Electrophoresis
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PIR	The Protein Information Resource
RF	Radio-Frequency Potential
RNA	Ribonucleic Acid
rpm	Revolutions Per Minute
SDS	Sodium Dodecyl Sulfate
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SIB	Swiss Institute of Bioinformatics
SIP	Shrimp Alkaline Phosphatase
T_m	Melting Temperature
TAE	Tris-Acetate-EDTA
TE	Tris-EDTA (buffer)
TIGR	The Institute for Genomic Research
TIM	Triosephosphate Isomerase
(QQ)TOF	(Quadrapole Quadrapole) Time-Of-Flight
UV	Ultraviolet
XRC	X-Ray Crystallography

Chapter I:

Introduction

1.0 Prokaryotic Genomics

Bacterial genomics refers to the DNA sequence analysis of an entire bacterial chromosome. The first complete prokaryotic genome to be sequenced was *Haemophilus influenzae* in 1995 (Fleischmann et al, 1995) and marked the first complete genetic map of a free-living organism (researchers had previously determined the complete genome sequence for several viruses, including cytomegalovirus, vaccinia and variola). However, a number of prokaryotic genomes have since been sequenced including most recently, *Candidatus Blochmannia floridanus* (Gil et al, 2003).

The preparation of these genomic sequences would not have been possible, especially in such relatively short periods of time, without impressive new computer software and hardware that, in addition to creating physical maps of the genomes, enable research teams to catalogue gene locations and many of their functions. In addition, other types of information typically made available include:

- Open reading frames, which are protein coding regions.
- Amino acid sequences, which can be determined from DNA sequence and genetic code
- Gene control sequences
- Evolutionary information from identification of sequences conserved between organisms, including prokaryotic-eukaryotic conservation

Access to a complete genome assists researchers in determining the genetic organization of living organisms as well as how they have evolved. In addition, such information should lead to the determination of the pathogenic determinants of the genome.

The *E. coli* K-12 MG1655 Genome

Nineteen hundred and ninety six marked a great year as the *Escherichia coli* genome had been fully sequenced and was made available to the scientific community and public (Blattner et al, 1997). This sequence was determined by the *E. coli* Genome Project at the University of Wisconsin-Madison where the entire sequence was independently determined for *E. coli* K-12 strain MG1655.

With the completion of the sequence, 4405 open reading frames were predicted. Although a number of these open reading frames have been identified with known genetic loci and subsequently annotated, the annotation of the genome is an ongoing process and many open reading frames are still only hypothetical. Ultimately, the desired result would be to make the genome sequence more useful by correlating it with experimental data.

Although gene annotation is important by itself, it only leads to speculation as to gene functions. *E. coli* proteins may be classified into over 20 functional groups and classes including those for: regulatory function, cell structure, biosynthesis, transport, energy metabolism, etc., yet many of these proteins are hypothetical or unclassified. Hence without question, there is no substitute for research leading to the determination of the structure and function of the actual protein products.

1.1 Protein Families

The numbers of proteins in existence are vast in number and diverse in structure and function. Newly discovered proteins resulting from ongoing genome sequencing projects are constantly added to information resources. However, many of the proteins are only hypothetical products of uncharacterized open reading frames and themselves

remain uncharacterized. Despite the vast complexity of proteins nearly all proteins share structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. Typically between 40% and 65% of the proteins found by genomic sequencing show significant sequence similarity to proteins with known function (Casari et al, 1996; Tatusov et al, 1996.) and usually a large fraction of them show similarity with each other (Tatusov et al, 1996; Brenner et al, 1995). Hence, for classification of newly discovered proteins as well as orderly management of previously determined sequences, it would be advantageous to organize known sequences into a collective database of information. In order to catalogue and annotate this information, proteins are classified into groups or families that reflect structural, functional and evolutionary relatedness. Therefore, many of the new entries are simply new members of existing protein families.

Several protein information databases exist for grouping related proteins into families where members belonging to a protein family are assigned via results of multiple sequence alignments and profiles (evolutionary relationship) as well as fold/domain and pattern searches (major structural similarities). Two popular databases of protein families are PROSITE (http://www.expasy.ch/prosite; Sigrist et al, 2002) and PFAM (http://pfam.wustl.edu; Bateman et al, 2002). In PROSITE, the constant and variable properties of groups of similar sequences are analyzed in order to derive a signature or consensus sequence for a protein family. This protein family signature distinguishes its members from all other unrelated proteins and is used as a model to search for new family members. PFAM protein families, on the other hand, are generated using multiple sequence alignments, profiles and Hidden Markov Models (HMMs) which are simply

statistical models of the sequence consensus of a homologous family. These HMMs, like the signature sequences of PROSITE, are used to relate new family members.

While both PROSITE and PFAM databases are very useful for identifying to which known protein family a new sequence belongs, it should be noted that a distinguishing feature between a pattern and a profile is that the former is usually confined to a small region with high sequence similarity whereas the latter attempts to characterize a protein family or domain over its entire length. In addition, although sequence patterns are very useful, there are a number of protein families as well as functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence. Examples of important functional domains which are weakly conserved are the globins and the SH2 and SH3 domains. In such domains there are only a few sequence positions which are well conserved. Therefore, members of a protein family must be considered in terms of sequence context (sequence identity) as well as structural context (structural similarity but not necessarily sequence similarity).

One of the pursuits in the laboratory of Dr. Harry W. Duckworth has been to gain a greater understanding of the molecular mechanisms behind the regulation of the glyoxylate cycle operon of *Escherichia coli*. The regulator protein for this operon, the isocitrate lyase regulator (IclR), represents a family of transcription factors namely the IclR family of transcriptional regulators, all of which have been identified by sequence similarity. Unfortunately, most of the family members exist solely as annotations on information databases with little to no biochemical characterization. Though IclR is a known regulator of the glyoxylate operon, its structure and the details of its function as a regulator remain a mystery.

The approach to understanding more about IclR, and hence the other IclR family members, has been to study multiple family members simultaneously. In this thesis, two members of the IclR family of transcriptional regulators discovered by sequence similarity searches will be presented and discussed.

1.2 The Interactions between Proteins and DNA

Proteins encoded by DNA are involved in almost all biological activities. Hence, in order to fully comprehend the genetic makeup of a living organism, the study of protein structure and function is vital.

Transcriptional Regulation of Genes in Prokaryotes

Through constant evolutionary changes, bacteria have developed a variety of ways for adapting to a myriad of environments. Naturally, a large number of genes are expressed at all times under all conditions of growth (constitutive genes), e.g., those encoding transfer RNA, ribosomal RNA, ribosomal proteins, RNA polymerase, enzymes catalyzing central metabolic pathways and so forth. However, for survival in different environments, bacteria have the ability to alter their gene activity such that gene products appropriate for the new environment are synthesized. These changes in gene activity are under the control of regulatory mechanisms and the genes controlled in response to the needs of a cell are called regulated genes. Inducible genes are those that are expressed at a high level only at certain times or only under certain defined conditions; at other times they are expressed at a very low level, e.g., enzymes for amino acid biosynthesis, enzymes for metabolism of specific energy sources. Conversely, gene products that decrease in concentration in response to a molecular signal are referred to as repressible.

Transcription and translation in prokaryotic cells are coupled both spatially and temporally, but in order to conserve cellular energy both processes must be carefully regulated. Since both processes occur simultaneously, it is more efficient to regulate the first step, i.e., the initiation of transcription. The key element of transcription is RNA polymerase, and therefore the regulation of transcription initiation is the regulation of the interaction of RNA polymerase with DNA. Hence, gene transcription is mediated and regulated by protein-DNA interactions.

Gene transcription can be controlled either by negative regulation or positive regulation where both types involve regulatory DNA-binding proteins. In negative regulation, a regulatory protein known as a repressor binds to a sequence of DNA called an operator and blocks transcription by interfering with RNA polymerase binding. In positive regulation, an activator is the regulatory protein which binds to a regulatory sequence of DNA referred to as an activator-binding site, facilitating the efficiency of both DNA-binding and transcription of RNA polymerase. The former method of regulation is the predominant form in prokaryotes while the latter is typical of eukaryotic organisms. However, examples of both have been well characterized in bacteria.

The Structure and Expression of the Prokaryotic Operon

In bacteria, the genes for construction of the components of a metabolic pathway are clustered in regulatory units on the bacterial chromosome and transcribed into a single mRNA. All of the regulatory elements involved in gene transcription of a particular metabolic pathway are organized into what is referred to as an operon. All of the genes of the operon are coordinately controlled.

The promoter is the site where RNA polymerase binds to the DNA prior to beginning transcription. There are two consensus sequences in prokaryotic promoters. The first is the -35 region (-35 bp upstream from the initiation site) which typically occurs as the sequence TTGACA; the second is -10 bp upstream from the initiation site of transcription and occurs as the consensus sequence TATAAT ("Pribnow box") which is responsible for identifying the precise nucleotide at which transcription begins. Structural genes that code for the enzymes of the metabolic pathway are usually adjacent to one another. They are transcribed into one long polycistronic mRNA molecule which is translated into distinct polypeptides that become the different enzymes of the metabolic pathway. The regulatory gene codes for a regulatory protein, either a repressor or activator. These regulatory proteins are capable of recognizing a specific sequence of base pairs within the DNA and binding to this region with high affinity.

Both activators and repressors utilize additional regulatory elements for proper functionality. These extra regulatory elements come in the form of molecular signals, i.e. a small ligand that binds to the regulatory protein acting as either a co-activator or co-repressor or as an inducer. Once bound, the ligands induce a conformational change either facilitating or debilitating the DNA-binding ability of the regulatory protein. Negative regulation can occur via an inducible or a repressible system. The difference between repressible and inducible systems is the result that occurs when the signal molecule binds to the repressor. With inducible systems, the binding of the signal molecule to the repressor greatly reduces the affinity of the repressor for the operator, the repressor is released and transcription proceeds. The *lac* operon is an example of an inducible system. In repressible systems, the binding of the signal molecule to the

repressor greatly increases the affinity of repressor for the operator, the repressor binds, and transcription stops. The *trp* operon is an example of such a system where tryptophan is the co-repressor that binds to the TrpR repressor protein. Similar effects are seen in positive regulation. The binding of a co-activator to its activator will enhance the ability of activator to bind its recognition site. The interaction between *E. coli* catabolite activator protein (CAP) and its co-activator, cyclic adenosine monophosphate (cAMP), is a well known example of positive regulation with induction for many *E. coli* promoters (Busby and Ebright, 1994 & 1999). Positive regulation can also occur with repression. For example, in the absence of exogenous fatty acids, FadR activates transcription of fatty acid biosynthetic genes (*fab*), by binding to an activator site near the promoter for these genes. In the presence of exogenous fatty acids, FadR binding is inactivated, repressing *fab* expression (DiRusso et al, 1993).

1.3 Sequence Specific DNA-Binding Proteins

As the primary regulators of gene expression, DNA-binding proteins and their interactions are crucial for higher order biological processes in development and disease. In addition, their regulatory functions provide a handle for genetic manipulation.

The interaction of a DNA-binding protein such as RNA polymerase or transcriptional regulators is dependent on the affinity of the protein for the binding site. This affinity will vary with different physiological conditions, as the concentration and nature of the protein changes but of course depends on the binding site itself, i.e. the exact sequence of nucleotide base pairs. Hence a protein with a specific function or purpose will bind to a specific sequence of DNA.

The specific DNA-binding sites for regulatory proteins are generally inverted repeats of a short DNA sequence (a palindrome) at which two (dimer) or four (tetramer) copies of a regulatory protein bind cooperatively. Most of the protein-DNA contacts that impart specificity are through hydrogen bonds. In particular, the surfaces on the DNA with which regulatory proteins must interact are mainly hydrogen-bond donor and acceptor groups exposed in the major groove of DNA. These hydrogen-bonding patterns allow discrimination between different base pairs. Therefore, DNA-binding proteins must have discrete DNA-binding domains capable of making the distinctions between different sequences of DNA. Only by a combination of structure-based mutagenesis, protein engineering and biophysical and structural analysis can one achieve a quantitative understanding of how the structure of these proteins dictates the recognition of specific DNA sites.

Of the 240-260 candidate *E. coli* DNA-binding proteins, only a small fraction have binding sites identified by DNA footprinting. However, known binding sites for a regulatory DNA-binding protein can be used to identify additional sites and thereby identify further genes regulated by that protein. Slight variations in a sequence can also be bound by different DNA-binding proteins, but the optimal binding site is of course the one that is closest to the consensus sequence. One can thus have a range of activity at different promoters/operators by having differences in DNA binding sites. The availability of complete bacterial genome sequences offers new opportunities to describe networks of regulatory interactions.

DNA-Bindng Protein Motifs

The interaction between regulatory proteins and their target DNA sequences involves several well-characterized protein-DNA interactions. These protein-DNA interactions fall in distinct structural motif classes. In prokaryotes, the most well-known example or most widely utilized motif for biochemical regulation rather, is known as the helix-turn-helix motif.

The Helix-Turn-Helix Motif

The helix-turn-helix (HTH) motif consists of two segments of alpha helix separated by a short extended amino acid chain referred to as a turn. One helix can fit into the major groove of DNA where contacts between the amino acid R-groups of this helix and the nucleotides of the groove determine the sequence specificity of the DNA binding. This motif is found in hundreds of DNA-binding proteins, including lambda repressor, tryptophan repressor, catabolite activator protein (CAP), octamer transcription factor 1 (Oct-1) and heat shock factor (HSF).

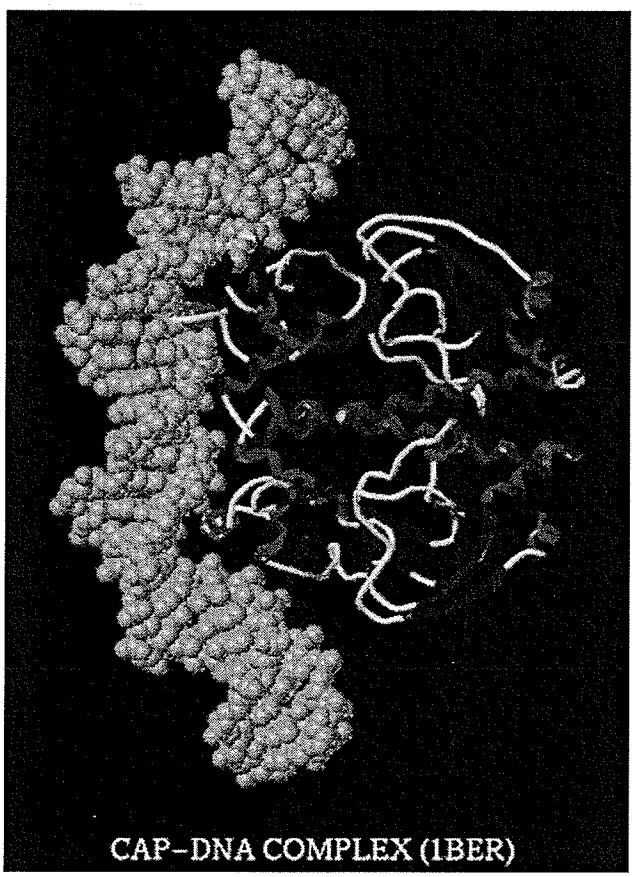


FIGURE 1.0: Catabolite Activator Protein. The CAP protein is shown here complexed with DNA via the helix-turn-helix motif (Parkinson et al, 1996).

Proteins that incorporate the helix-turn-helix motif recognize base sequences which are palindromic, i.e. possess an internal two-fold symmetry axis. As only one of the helices from the helix-turn-helix motif serves as the recognition helix, this implies that the full binding motif must require two helix-turn-helix motifs, i.e., one from each monomer must bind to each half of the palindromic DNA sequence. Therefore, the fully functional helix-turn-helix DNA-binding protein must at least be a dimer. For example, the two recognition helices of the cro protein are related by a two-fold axis passing through the central beta-sheet region of the dimer. Therefore, the recognition helices of the cro dimer fit into the major groove of the DNA and interact with each identical half of

the palindrome. Though only two polypeptides cooperate to produce a full binding motif, the lac repressor protein is fully functional as a tetramer.

Hence, the second helix of the helix-turn-helix motif has an important role in recognizing the DNA while the remainder of the structure serves to keep the two helices in the correct relative position for fitting in the major groove of DNA.

1.4 The IclR Family of Transcriptional Regulators

Many bacterial transcriptional regulators bind DNA using the helix-turn-helix motif, yet several different protein families of transcriptional regulators exist that are structurally and functionally distinct from one another. Therefore, a single DNA-binding motif is not satisfactory in classifying an entire protein family. In this thesis investigation, we were primarily concerned with the IclR family of transcriptional regulators.

General Characteristics

The head of all IclR family members, as the name suggests, is the isocitrate lyase regulator/repressor. IclR is a 274 amino acid residue DNA-binding protein found in *Escherichia coli* and *Salmonella typhimurium* that is responsible for the regulation of the glyoxylate bypass operon (*ace*BAK), which encodes isocitrate lyase, malate synthase as well as isocitrate dehydrogenase kinase/phosphorylase. IclR functions specifically as a repressor with a negative transcriptional effect on the *ace*BAK operon and is therefore referred to as the acetate operon repressor. It is believed that IclR binds to DNA using a helix-turn-helix motif found in its N-terminal domain in a signal-mediated fashion. Those residues associated with signal binding are believed to be in the C-terminal domain (Bateman et al, 2002). Therefore, members belonging to the IclR family not only possess

significant degrees of homology with respect to primary structure, but are also proposed to meet the following criteria:

1). Functions as a transcriptional regulator

2). 200 – 300+ residue length

3). N-terminal helix-turn-helix DNA-binding motif

4). C-terminal regulatory binding site

Members of the IclR family are electronically annotated on any one of a number of different protein databases as illustrated in the following table.

TABLE 1.0: Protein Databases with Annotated IclR Family Members

Database	URL	Accession No.	No. of IclR Family Members
PROSITE	http://www.expasy.ch/prosite	PS01051	48 proteins
PFAM	http://pfam.wustl.edu	PF01614	168 members
SMART	http://smart.embl-heidelberg.de/	SM00346	199 members

The discrepancy in the number of members belonging to the IclR protein family between the different databases is a consequence of the different methods of member prediction. As previously described, PROSITE attempts to identify a signature pattern for each family to which all members are assessed. The C-terminal signature consensus sequence used by PROSITE to annotate the IclR family can be seen in FIGURE 1.1.

A).
[GAD]-[YWA]-[GSTAC]-x-[DSV]-x(2)-E-x(6)-[CSA]-[LIV]-[GSA]-x(2)-[LIV]-[FYWLH]-[DNG]

B).

```
TM0065      GYAVDNEENEIGIMCVGVPIFD
Gylr_strco GYAADVEETWEGIASIAAPIHD
Gylr_strgr GWALDLEETWEGVASLAAPVHD
Iclr_ecoli GYSFDDEEHALGLRCLAACIFD
Iclr_salty GYSFDDEEHALGLRCVASCIYD
Kdgr_ecoli GYGEDNEEQEEGLRCIAVPVFD
Kdgr_erwch GYGEDNEEQEEGLRCIAVPVFD
Pcar_psepu GWCVVDQELEQGLRSIAVPIYD
Pcau_acica GWCYSSEEHELGVHALAVPIYG
Pobr_acica DYCLSTEEHELGLIAIAVPVLN
YagI_ecoli GYALDSEENEQGVRCVAVPVWN
Ybbu_ecoli GYTVDKEEHVVGLNCIASAIYD
Ycso_bacsu GYTVSYSELENYTAAIGAPIFN
Yfax_ecoli GWSYDNGEDYADVRCVAAPVFN
YiaJ-ecoli GAAMDREENELGVSCIAVPVFD
YiaJ_haein AYAMDREENELGVTCIACPIFD
YjhI_ecoli GWAIDNEESTYGAVCLSMPVFN
Mhpr_ecoli does not follow pattern
```

FIGURE 1.1: The C-terminal Consensus IclR Signature. A). The signature sequence for IclR family members consists of a 22 residue consensus located in the C-terminal region of each member. B). Examples of IclR family members subscribing to the consensus sequence. The MhpR protein of *E. coli* does not follow the signature pattern yet it is still included in the family due to its significant primary structure homology with IclR and other members (Sofia et al, 1994; PROSITE)

The PFAM and SMART databases use more reliable methods of multiple sequence alignments in order to accurately generate statistical models for assigning members. FIGURE 1.2 illustrates how members of the IclR family are annotated based on results of multiple sequence alignment.

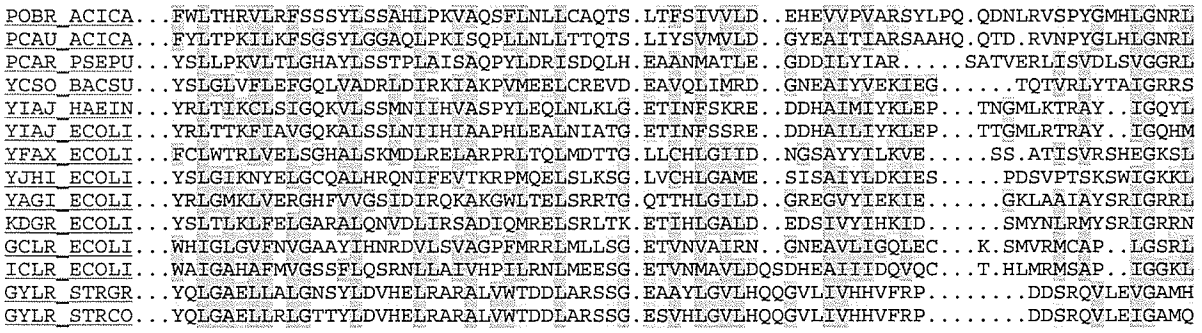


FIGURE 1.2: Multiple Sequence Alignment of IclR Family Members. Alignments are colored using the ClustalX scheme in Jalview (orange:glycine (G); yellow: Proline (P); blue: small and hydrophobic amino-acids (A, V, L, I, M, F, W); green: hydroxyl and amine amino-acids (S, T, N, Q); red: charged amino-acids (D, E, R, K); cyan: histidine (H) and tyrosine(Y)). PFAM seed (PF01614)

The protein database known as InterPro (http://www.ebi.ac.uk/interpro/) offers an amalgamation of the results of a number of protein databases including the three above-named databases. In addition to providing accession numbers for all integrated database cross-references, the IclR family members annotated on InterPro are conveniently listed with their corresponding Swiss-Prot and TrEMBL (http://ca.expasy.org/sprot/) accession numbers.

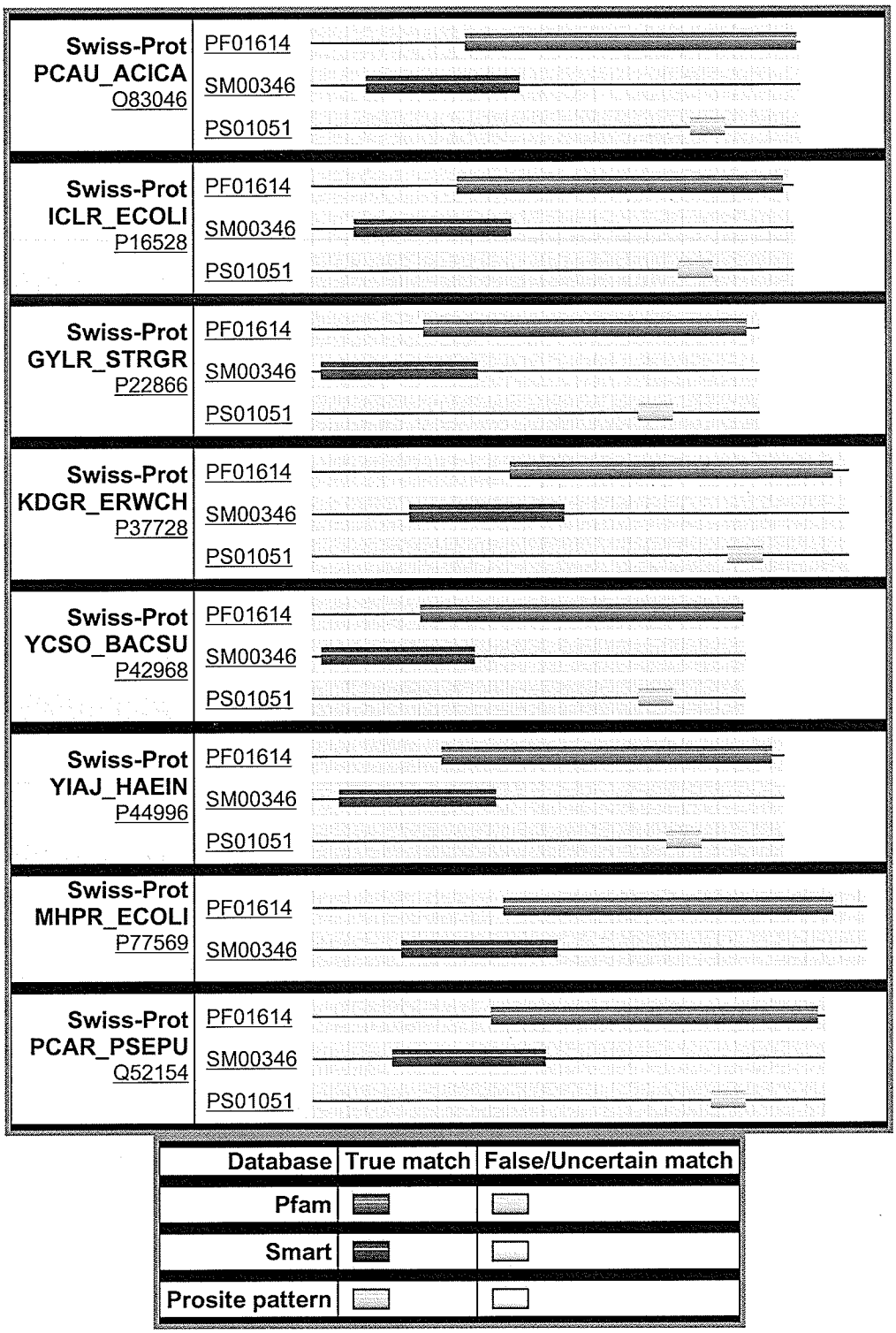


FIGURE 1.3: IclR Family Members from InterPro. While this figure only shows an excerpt from the total 195 members, it can be seen that members come from different organisms.

While the preceding figure only shows a small representation of the IclR family members annotated on the InterPro database, members of this protein family include various species of bacteria as well as archaea. The entire listing of IclR family members annotated on InterPro can be observed using the InterPro accession number IPR005471. As indicated in FIGURES 1.2 and 1.3 above, *E. coli* MhpR does not possess the C-terminal IclR family signature sequence (Sofia et al, 1994); however this protein does bear a significant sequence homology making it a member of the IclR family. There are actually a number of other IclR family members that do not possess this sequence and hence one can see how pattern prediction is not always a reliable method of member prediction and annotation. However, there are examples of proteins that do not share a significant degree of homology with other proteins resulting in a failed attempt at family member classification via multiple sequence or pairwise alignment. In these instances, the protein may only be categorized using a matching signature pattern e.g. YagF protein (see below).

Disregarding the variant annotation methodologies and to reiterate, common features found amongst the members of the IclR protein family include a 200 – 300+ amino acid residue length, an N-terminal helix-turn-helix motif for DNA-binding and a C-terminal regulatory substrate binding site (Bateman et al, 2002). However, it should be noted that the latter two features are only proposals based almost entirely on secondary structural predictions (detailed structural information regarding this class of transcriptional regulators is not available) and limited biochemical evidence (Kok et al, 1998). Unfortunately, though the IclR family of transcriptional regulators contains nearly 200 members, only one has been successfully analyzed by x-ray crystallography (Zhang

et al, 2002). This member is known as TM0065, from the organism *Thermotoga maritima* and serves as the only tangible link to a clearer understanding of the functional mechanisms behind this family (for description of structure, see Chapter 4: Results). By taking the approach of studying a multitude of IclR family members, hopefully the intimate details behind structure and regulatory function will be elucidated for this family of transcriptional regulators.

Functional Characteristics

Members of the IclR family of transcription regulators modulate signal-dependent expression of genes involved in carbon metabolism in bacteria and archaea (Zhang et al, 2002; Nasser et al, 1994). In addition to being involved in carbon metabolism, members of the IclR family have also been implicated in the control of expression of virulence and sporulation genes (Nomura et al, 1998; van Wezel et al, 2000). TABLE 1.1 contains a select number of IclR family members from various organisms with associated function.

TABLE 1.1: Various Examples of IclR Family Members

Protein Name	Organism(s)	Function	Available Reference(s)
A). IclR Family Members with Established or Partly Known Function			
IclR	*E. coli* *S. typhimurium*	repressor of the *ace*BAK operon (also known as glyoxylate bypass operon)	Reverchon et al, 1991 Galinier et al, 1990
KdgR	*E. chrysanthemi* *E. carotovara*	repressor of genes involved in pectinolysis and in pectinase secretion	
MhpR	*E. coli*	activator of the *mhp*RABCDEF operon coding for components of the 3-hydroxyphenylpropionate degradation pathway	
PcaR PcaU	*P. putida* and *A. calcoaceticus*	protocatechuate utilization genes activator	Eulberb et al, 1998 Kok et al, 1998
PobR	A. calcoaceticus	activator of the *pob*A gene	Kok et al, 1998
YiaJ	*E. coli*	repressor of *yia*KLMNOPQRS operon	Ibanez et al, 2000
GclR	*E. coli*	Negative regulator of allantoin and glyoxylate utilization operons	Donald et al, 2001
Pir	*E. chrysanthemi*	Induction of pectate lyase virulence genes	Nomura et al, 1998
B). IclR Family Members with Inferred Function			
ArcR^{+}	*H. salinarum*	possible regulator for *arc*RACB cluster; fermentative arginine degradation	Ruepp & Soppa, 1996
GylR^{+}	*S. griseus* *S. coelicolor*	possible activator protein for the *gyl*ABX glycerol operon	
PcaR^{+}	*R. opacus*	part of a protocatechuate catabolic gene cluster	
TM0065^{+}	*T. maritima*	variant Entner-Doudoroff pathway	Zhang et al, 2002
MhpR*	*C. testosteroni*	Homologue of *E. coli* MhpR	
YiaJ*	*H. influenzae*	Homologue of *E. coli* YiaJ	
KdgR	*E. coli*	Homologue of *Erwinia* KdgR	
C). Hypothetical IclR Family Members (amino acid sequence similarity to IclR)			
YagI	*E. coli*	unknown	
YfaX	*E. coli*	unknown	
YjhI	*E. coli*	unknown	
YcsO	*B. subtilis*	unknown	

*function is inferred through significant homology with existing proteins from group A
$^{+}$some biochemical evidence for function

Targeted IclR Family Members from *E.coli*

While nearly 200 members of the IclR family from various prokaryotic organisms have been catalogued, many remain annotated as uncharacterized, hypothetical proteins. Nine members of the IclR family come from *E. coli* alone, eight of which (see preceding Table 1.1) are found in *E. coli* K-12 MG1655 while the ninth, a 284 hypotheical protein with accession number AAN83212 (GenBANK), is only found in *E. coli* CFT073. As our laboratory was primarily concerned with the characterization of the *E. coli* IclR family members, where prevalent and exhaustive research on *E. coli* IclR and GclR continue, two additional members from *E. coli* were selected in an attempt to gain a further understanding of the structural and functional characteristics of the IclR family of transcriptional regulators as a whole. In particular, the *E. coli* YagI and YiaJ proteins were selected as targets for isolation and purification for the scope of this thesis investigation.

YagI

The protein product of the *yag*I (b0272, f252) open reading frame found in section 25/400 (GenBank nucleotide sequence database accession no. AE000135) is a 252 AA residue protein named YagI. This protein shares a significant degree of homology with IclR.

Score = 123 bits (309), Expect = 4e-27
Identities = 75/248 (30%), Positives = 129/248 (51%), Gaps = 2/248 (0%)

```
4    IQSVERALQILDLFNEQATELKITDISKLMGLSKSTLHSLLKTLQLHGYIDQNPENGKYR  63
     +QS+ R L++L+    E     + +T++++  GL  ST H LL T+Q  G++ Q  E G +
37   VQSLTRGLKLLEWIAESNGSVALTELAQQAGLPNSTTHRLLTTMQQQGFVRQVGELGHWA  96

64   LGMKLVERGHFVVGSIDIRQKAKGWLTELSRRTGQTTHLGILD--GREGVYIEKIEGKLA  121
     +G      G   + S ++        L  L   +G+T ++ +LD    E + I++++
97   IGAHAFMVGSSFLQSRNLLAIVHPILRNLMEESGETVNMAVLDQSDHEAIIIDQVQCTHL  156

122  AIAYSRIGRRLPVHATAIGKVLIAWLGEAELNALLEGYQYTTFTPATLASREALMSALAQ  181
         + IG +LP+HA+  GK  +A L E ++  LL        +T ATL S   L   LAQ
157  MRMSAPIGGKLPMHASGAGKAFLAQLSEEQVTKLLHRKGLHAYTHATLVSPVHLKEDLAQ  216

182  TREQGYALDSEENEQGVRCVAVPVWNHESRVIAALSLSTLTSRVDDAELANFREQLQQAG  241
     TR++GY+ D EE+  G+RC+A  +++       AA+S+S   SR+ D  +  F   + +A
217  TRKRGYSFDDEEHALGLRCLAACIFDEHREPFAAISISGPISRITDDRVTEFGAMVIKAA  276

242  LALSRALG
       ++ A G
277  KEVTLAYG
```

FIGURE 1.4: Homology Between *E. coli* IclR and YagI Proteins. Using the BLAST (GenBANK) program specifically with the BLOSUM62 matrix, YagI protein (top sequence) was aligned with IclR protein (bottom sequence). Identities refer to those residues that are conserved between the two sequences and positives include identities as well as those residues that are similar with respect to physical properties (i.e side-chain structure, polarity, etc.).

In addition to a significant degree of homology to IclR, YagI also contains the proposed IclR family C-terminal signature (Sofia et al, 1994). For these reasons, it is believed that this 252 AA hypothetical protein is a transcriptional regulator that binds to DNA using a helix-turn-helix motif.

MPIIQSVERALQILDLFNEQATELKITDISKLMGLSKSTLHSLLKTLQLHGYIDQNP
ENGKYRLGMKLVERGHFVVGSIDIRQKAKGWLTELSRRTGQTTHLGILDGREGV
YIEKIEGKLAAIAYSRIGRRLPVHATAIGKVLIAWLGEAELNALLEGYQYTTFTPA
TLASREALMSALAQTREQGYALDSEENEQGVRCVAVPVWNHESRVIAALSLSTL
TSRVDDAELANFREQLQQAGLALSRALGYPA
HTH: predicted Helix-Turn-Helix motif
SGN: C-terminal IclR family signature

FIGURE 1.5: The Primary Structure of YagI Protein. YagI is the 252 residue hypothetical protein product of the YagI open reading frame found in section 25/400 of the *E. coli* K-12 MG1655 genome (GenBank accession number AE000135). Sequences highlighted in red and green represent a predicted helix-turn-helix motif (Dodd & Egan, 1990) and the C-terminal IclR family signature (Sofia et al, 1994) respectively.

The Hypothetical *yag*EFGH Operon

Analysis of the regions around the *yag*I gene reveals four neighboring genes immediately downstream of *yag*I progressing into section 24/400 of the *E. coli* genome (GenBank accession number AE000134). After conducting homology searches using various protein information resources described above, the protein products of these four genes were identified as members of protein families bearing significant homologies to functional enzymes.

From Section 24/400: o309 & o655

o309 (*yag*E): The protein product of this gene is 58% identical (no gaps) with the *E. coli* protein YjhH, both of which bear significant homology with the gene products of two *E. coli* genes namely, *dap*A (Dihydrodipicolinate synthetase, DHDPS) and *nan*A (N-acetylneuraminate pyruvate lyase (Mirwaldt et al, 1995; Izard et al, 1994). DHDPS is an enzyme discretely involved in lysine biosynthesis via a diaminopimelate pathway catalyzing the condensation of L-aspartate-β-semialdehyde and pyruvate to dihydropicolinic acid (Born & Blanchard, 1999). N-acetylneuraminate pyruvate lyase

catalyzes the condensation of N-acetyl-D-mannosamine and pyruvate to form N-acetylneuraminate.

o655 (*yag*F): The protein product of this gene is 67% identical (2 gaps) with the *E. coli* protein YjhG. Both of these proteins are significantly homologous to two evolutionary related dehydratases namely, the *ilv*D gene product dihydroxy-acid dehydratase and the *edd* gene product 6-phosphogluconate dehydratase (Egan et al, 1992). Dihydroxy-acid dehydratase catalyzes the dehydration of 2,3-dihydroxy-isovaleric acid into alpha-ketoisovaleric acid during the biosynthesis of isoleucine and valine. 6-phosphogluconate dehydratase catalyzes the dehydration of 6-phospho-D-gluconate into 6-phospho-2-dehydro-3-deoxy-D-gluconate in the Entner-Doudoroff pathway.

From Section 25/400: o460 & o536

o460 (*yag*G): The protein product of this gene is 43% identical (8 gaps) with the *E. coli* protein YicJ which belongs to the Glycoside-Pentoside-Hexuronide (GPH) cation symporter family (PFAM, http://pfam.wustl.edu/; Bateman et al, 2002). GPH cation symporters catalyze uptake of sugars (e.g. melibiose, lactose, raffinose, glucuronides, pentosides and isoprimeverose) in symport with a monovalent cation (e.g. Li^+, H^+ or Na^+). Examples of this family include the melibiose permeases of enteric bacteria, and the lactose permease of *Streptococcus thermophilus*.

o536 (*yag*H): The protein product of this gene is 52% identical (3 gaps) with the *xyn*B gene product of *Bacillus pumilus*. The *xyn*B gene encodes a protein known as xylan 1,4-β-D-xylosidase (Moriyama et al, 1987; Xu et al, 1991). This protein belongs to family 43/85 of a superfamily of O-glycosyl hydrolases; a broad spectrum of enzymes that

hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety (Henrissat & Bairoch, 1996).

Therefore taken together with the proposed transcriptional regulatory function of YagI, the proximity of these four genes to the *yag*I gene suggests that these elements constitute a hypothetical operon and the regulatory target of YagI.

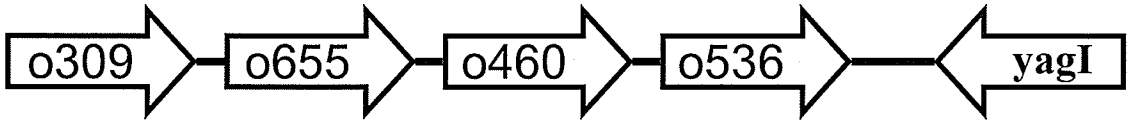


FIGURE 1.6: The Hypothetical *yag*EFGH Operon: Four genes, o309, o655, o460 and o536 are believed to be part of an operon regulated by the YagI gene product of *yag*I. The four target genes of YagI are oriented immediately downstream of *yag*I.

TABLE 1.2: Elements of the Hypothetical *yag*EFGH Operon

Gene Name(s)	Product	Swiss Prot/TrEMBL	Function
*yag*I (b0272)	YagI	P77300	putative regulator
*yag*E (o309)	YagE	P75682	putative lyase/synthase
*yag*F (o655)	YagF	P77596	putative dehydratase
*yag*G (o460)	YagG	P75683	putative permease
*yag*H (o536)	YagH	P77713	putative beta-xylosidase

Although this hypothetical operon is likely to be a functional operon that will be proved through studies and experimentation, it should be noted that *yag*I is found only in *E. coli* K-12. It is not found in other subspecies of *E. coli* nor is it found in other bacteria. Hence, the importance of this operon may not be significant for viability.

YiaJ

The YiaJ protein is the 282 residue gene product of the *yia*J gene found in section 325/400 of the *E. coli* genome (GenBank accession number AE000435). As this protein is of the typical amino acid length, significantly homologous with IclR and possesses the C-terminal IclR family pattern; it is classified as an IclR family member also believed to bind DNA using a helix-turn-helix motif.

Score = 109 bits (273), Expect = 6e-23
Identities = 73/225 (32%), Positives = 111/225 (48%), Gaps = 3/225 (1%)

```
24   QSLFRGLMLIEILSNYPNGCPLAHLSELAGLNKSTVHRLLQGLQSCGYVTTAPAAGSYRL  83
     QSL RGL L+E ++        L  L++ AGL  ST HRLL  +Q  G+V      G + +
38   QSLTRGLKLLEWIAESNGSVALTELAQQAGLPNSTTHRLLTTMQQQGFVRQVGELGHWAI  97

84   TTKFIAVGQKALSSLNIIHIAAPHLEALNIATGETINFS-SREDDHAILIYKLEPTTGML  142
           VG   L S N++ I  P L  L   +GET+N +   + DH  +I      T ++
98   GAHAFMVGSSFLQSRNLLAIVHPILRNLMEESGETVNMAVLDQSDHEAIIIDQVQCTHLM  157

143  RTRAYIGQHMPLYCSAMGKIYMAFGHPDYVKSYWESHQHEIQPLTRNTITELPAMFDELA  202
     R  A IG  +P++ S  GK ++A    + V      H+  +   T  T+     + ++LA
158  RMSAPIGGKLPMHASGAGKAFLAQLSEEQVTKL--LHRKGLHAYTHATLVSPVHLKEDLA  215

203  HIRESGAAMDREENELGVSCIAVPVFDIHGRVPYAVSISLSTSRL                247
       R+ G + D EE+ LG+ C+A  +FD H     A+SIS   SR+
217  QTRKRGYSFDDEEHALGLRCLAACIFDEHREPFAAISISGPISRI                260
```

FIGURE 1.7: Homology Between *E. coli* IclR and YiaJ Proteins. Using the BLAST (GenBANK ref) program specifically with the BLOSUM62 matrix, YiaJ protein (top sequence) was aligned with IclR protein (bottom sequence). Identities refer to those residues that are conserved between the two sequences and positives include identities as well as those residues that are similar with respect to physical properties (i.e side-chain structure, polarity, etc.).

MGKEVMGKKENEMAQEKERPAGSQSLFRGLMLIEILSNYPNGCPLAHLSELAGL
NKSTVHRLLQGLQSCGYVTTAPAAGSYRLTTKFIAVGQKALSSLNIIHIAAPHLEA
LNIATGETINFSSREDDHAILIYKLEPTTGMLRTRAYIGQHMPLYCSAMGKIYMAF
GHPDYVKSYWESHQHEIQPLTRNTITELPAMFDELAHIRESGAAMDREENELGVS
CIAVPVFDIHGRVPYAVSISLSTSRLKQVGEKNLLKPLRETAQAISNELGFTVRDD
LGAIT
HTH: predicted Helix-Turn-Helix motif
SGN: C-terminal IclR family signature

FIGURE 1.8: The Primary Structure of YiaJ Protein. YiaJ is the 282 residue protein product of the *yia*J open reading frame found in section 325/400 of the *E. coli* K-12 MG1655 genome (GenBank accession number AE000435). Sequences highlighted in red and green represent a predicted helix-turn-helix motif (Dodd & Egan, 1990) and the C-terminal IclR family signature (Sofia et al, 1994) respectively.

Unlike YagI, the YiaJ protein no longer remains classified as a putative regulator. Biochemical evidence has proven that it is in fact the transcriptional regulator for the *yia*KLMNOPQRS (*yia*K-S) operon (Ibanez et al, 2000a). In addition, YiaJ is found not only in *E. coli* species K12, CFT073, E2348, and 042, but also in several species of *Salmonella* (e.g. *enterica serovar typhimurium*, *typhimurium* LT2, *typhimurium* SL1344,

enteritidis), *Shigella flexneri*, *Haemophillus influenze*, *Pasteurella multocida*, and *Klebsiella oxytoca*.

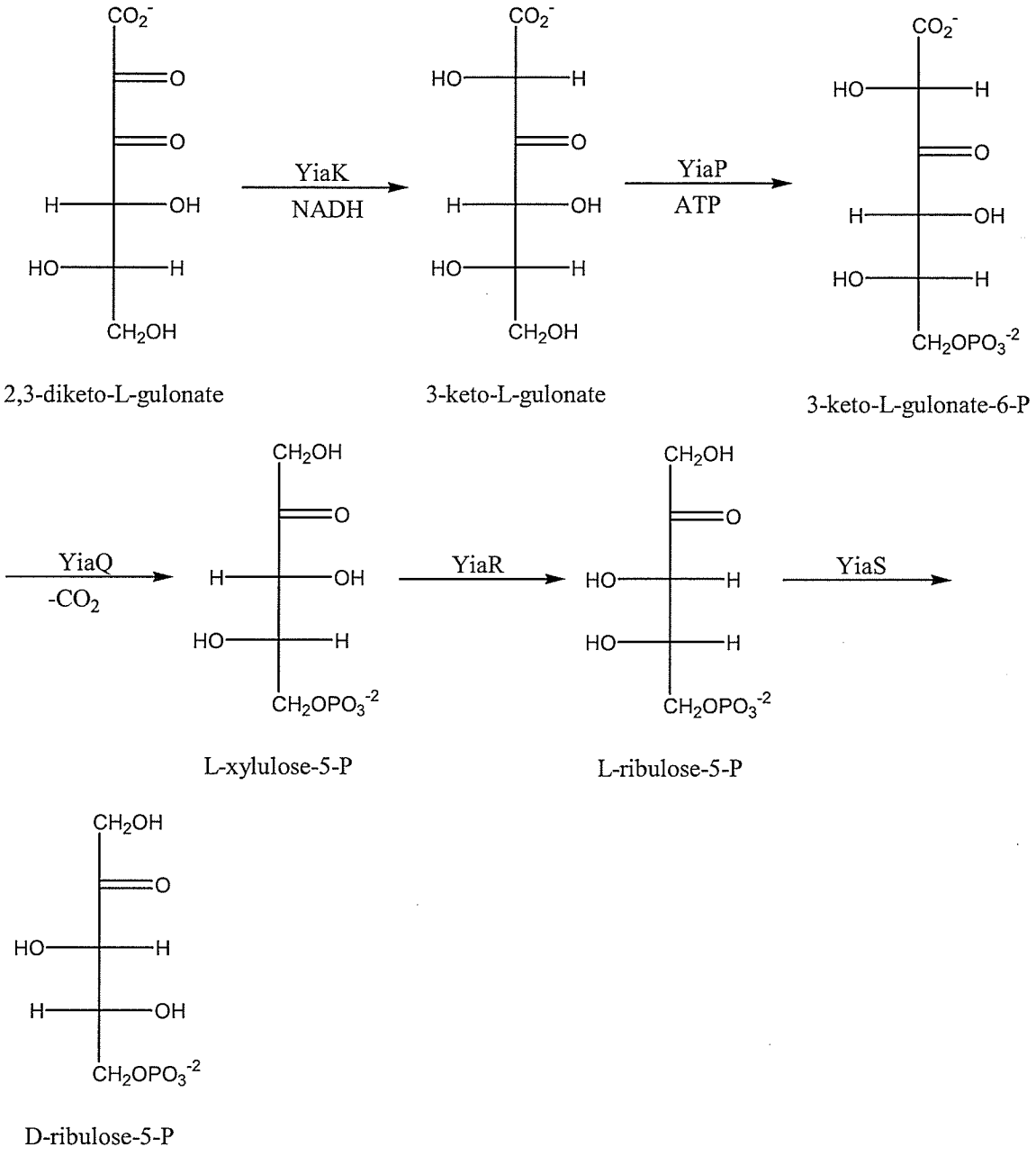
The *yia*K-S Operon

The *yia*K-S operon is a carbohydrate-metabolizing operon consisting of nine genes found along with *yia*J in section 325/400 of the *E. coli* genome. The *yiaK-S* genetic cluster is processed as a single transcriptional unit evidenced from observations of a mutant *E. coli* strain selected for its ability to grow on L-lyxose. This mutant strain, known as strain JA134, is deficient in the regulator gene due to inactivation by a genome rearrangement mediated by IS1 transposition (Badia et al, 1998). Inactivation of this regulatory gene, which was subsequently identified as *yia*J, leads to constitutive expression of the *yiaK-S* operon (Badia et al, 1998). The isolation of polycistronic mRNA of the structural genes corresponding to the full-length transcript of the *yiaK-S* system via Northern blot hybridization from JA134 supported evidence for the *yia*K-S operon (Ibanez et al, 2000a). Evidence supporting the transcription of the *yiaK-S* genes as a single unit was also demonstrated by the properties of multiple *lacZ* fusions with the structural genes of the *yia*K-S operon. The results of the *lac*Z fusions also indicated a unique functional promoter for the structural genes located upstream of *yiaK* (Ibanez et al, 2000a). In addition, analysis of mRNA led to the identification of the *yia*J promoter, the putative *yia*J regulatory binding site and two putative cyclic AMP receptor protein (Crp) consensus sites, Crp site 1 and Crp site 2 (from gel mobility shift assays) as well as a potential integration host factor (IHF)-binding site. When *lacZ* transcriptional fusions were introduced into the genetic background of the mutant JA134 strain, it was revealed that *yiaK-S* was modulated by the integration host factor and by the cyclic AMP receptor

protein (Crp) activator complex. The *lacZ* transcriptional fusions of different fragments of the *yia*K-S promoter demonstrated that binding of cAMP-Crp to the Crp site 1 was essential for *yia*K-S expression (Ibanez et al, 2000a).

Although the *yia*K-S operon has been partially characterized at both the genetic and biochemical levels (Sanchez et al, 1994; Badia et al, 1998; Ibanez et al, 2000a & 2000b; Yew & Gerlt, 2002), a specific role for this system has yet to be determined; the reason being that the substrate and hence the physiological function of the operon remains unknown. Four of the protein products of this system have been functionally characterized (Sanchez et al, 1994; Ibanez et al; 2000a &2000b; Yew & Gerlt, 2002). They are 2,3-diketo-L-gulonate reductase (YiaK), 3-keto-L-gulonate kinase (YiaP, LyxK), 3-keto-L-gulonate 6-phosphate decarboxylase (YiaQ, SgbH), and L-ribulose 5-phosphate 4-epimerase (YiaS, SgbE). A functional isomerase activity has been predicted for YiaR, but with less confidence as its primary structure is significantly homologous to several unknown proteins, all of which have been classified as putative isomerases (Ibanez et al, 2000b). However, it is believed that YiaR functions specifically as a 3-epimerase to complement the 4-epimerase activity of YiaS and the kinase activity of YiaP in the metabolism of the ketose L-xylulose (Sanchez et al, 1994; Ibanez et al, 2000b; Yew & Gerlt, 2002). In particular, YiaP, YiaR and YiaS are responsible for transforming L-xylulose into D-xylulose-5-phosphate, an intermediate metabolite of the pentose phosphate pathway (Ibanez et al, 2000b; Campos et al, 2002). L-xylulose itself, however, cannot be taken up by cells for utilization; instead it must be generated endogenously, e.g., via L-rhamnose isomerase activity on L-lyxose (Ibanez et al, 2000b). L-xylulose is phosphorylated via the kinase activity of YiaP to form L-xylulose-5-phosphate. It is now

believed that a 3-epimerase activity could convert the L-xylulose-5-phosphate into the L-ribulose-5-phosphate substrate of YiaS. Finally, the action of YiaS would convert the L-ribulose-5-phosphate into the D-xylulose-5-phosphate metabolite of the pentose phosphate pathway. The 3-epimerase activity of YiaR is an assumption largely based on a 56% sequence identity to a protein with a demonstrated L-xylulose 5-phosphate 3-epimerase activity known as UlaE (Yew & Gerlt, 2002). However, attempts to illustrate this function have so far been unsuccessful (Ibanez et al, 2000b & Yea & Gerlt, 2002). FIGURE 1.9 describes a possible pathway for the enzymes of the *yia*K-S operon.



2,3-diketo-L-gulonate

3-keto-L-gulonate

3-keto-L-gulonate-6-P

L-xylulose-5-P

L-ribulose-5-P

D-ribulose-5-P

FIGURE 1.9: Hypothetical Pathway for the Formation of D-ribulose. A possible pathway for the formation of D-ribulose from the precursor carbohydrate 2,3-diketo-L-gulonate via the actions of YiaK, YiaP, YiaQ, YiaR and YiaS from ((Ibanez et al, 2000b & Yew & Gerlt, 2002).

The functions of YiaL, YiaM, YiaN and YiaO are also still unclear, as only putative functions have been assigned to each. Hence, only a portion of the genes comprising the *yia*K-S operon attribute to the metabolism of L-xylulose which is itself an

intermediate resulting from L-lyxose. However, it is believed that all nine gene products of the *yia*K-S operon are responsible for the metabolism of a carbohydrate other than L-lyxose yet, this substrate still remains unknown. Attempts at identifying the elusive substrate are continually being made. The following figure and table illustrates the genetic organization and summarizes the components making up the *yia*K-S operon respectively.

FIGURE 1.10: The *yia*K-S Operon. The *yia*K-S (short for *yia*KLMNOPQRS) is a multigene operon consisting of nine structural genes. It has its own operator that is divergently transcribed from that of its repressor YiaJ

TABLE 1.3: Elements of the *yia*K-S Operon

Gene Name(s)	Product	Swiss Prot/TrEMBL	Function
*yia*J (b3574)	YiaJ	P37671	Repressor of the *yiaK-S* operon.
*yia*K(b3575)	YiaK	P37672	-putative dehydrogenase -2,3-diketo-L-gulonate reductase
*yia*L(b3576)	YiaL	P37674	putative lipase
*yia*M(b3577)	YiaM	P37674	unknown
*yia*N(b3578)	YiaN	P37675	putative membrane protein
*yia*O(b3579)	YiaO	P37676	putative solute-binding transport protein
*yia*P (b3580) *lyxK*	YiaP LyxK	P37677	-L-xylulose kinase -3-keto-L-gulonate kinase
*yia*Q(b3581) *sgbH*	YiaQ SgbH	P37678	-probable 3-hexulose 6-phosphate synthase -3-keto-L-gulonate 6-phosphate decarboxylase
*yia*R(b3582) *sgbU*	YiaR SgbU	P37679	-3-Epimerase -probable 3-hexulose-6-phosphate isomerase
*yia*S(b3583) *sgbE*	YiaS SgbE	P37680	-L-ribulose-5-phosphate 4-epimerase -putative epimerase/aldolase

As previously described, orthologues of *yia*J are found in *Haemophilus influenzae* (HI1032), *Pasteruella multocida* (PM1257) and *Klebsiella oxytoca* (YiaJ). In addition,

an operon with a similar organization to that of *E. coli yia*K-S is found adjacent to the *yia*J orthologues of these other bacteria.

TABLE 1.4: Orthologues of *yia*J and their Corresponding Operons

E. coli		*H. influenzae*		*P. multocida*		*K. oxytoca*	
genome segment: 325/400		genome segment: 98/163		genome segment: 131/204		GenBank No.: AF282849	
Gene	bp	Gene	bp	Gene	bp	Gene	bp
*yia*J	1569-721	HI1032	118-924*	PM1257	139-948**	*yia*J	1068-250
*yia*K	1770-2768	HI1031	10991-9996	PM1256	13719-12721	*yia*K	1286-2284
*yia*L	2780-3247	N/P		PM1255	12711-12235	*yia*L	2342-2806
*yia*M	3365-3838	HI1030	9921-9436	PM1254	12238-11756	N/P	
*yia*N	3841-5115	HI1029	9439-8162	PM1253	11759-10479	N/P	
*yia*O	5132-6118	H11028	8131-7145	PM1252	10459-9473	N/P	
N/P		N/P		PM1251	9395-7878	N/P	
N/P		N/P		PM1250	7834-6788	N/P	
N/P		N/P		PM1249	6748-5639	N/P	
N/P		N/P		PM1248	5576-4692	N/P	
N/P		N/P		N/P		*yia*Xl	2820-3767
N/P		N/P		N/P		*yia*X2	3819-5138
*lyx*K	6122-7618	HI1027	7086-5629	PM1247	4679-3228	*lyx*K	5216-6721
*sgb*H	7615-8277	HI1024	4016-3339+	PM1246	3209-2565	*yia*Q	6718-7380
*sgb*U	8237-9130	HI1026	5615-4755	PM1245	2540-1680	*yia*R	7373-8233
*sgb*E	9124-9819	HI1025	4761-4066	PM1244	1722-991	*yia*S	8227-8943

N/P - not present in genome
* - Genome segment 99/163
** - Genome segment 132/204
+ - Out of order compared to other organisms

Regulation of *yia*K-S by YiaJ

Upstream of the *yia*K-S operon, the *yiaJ* gene was identified as the gene which encodes the transcriptional regulator for this operon (Badia et al, 1998). Although over 80 carbohydrate and derivative test compounds have been examined, no inducer molecule of the *yia*K-S operon has been identified. Expression of the *yia*K-S operon is only seen in the *yia*J-deficient mutant strain JA134.

The promoter region of the *yia*J gene has been found to overlap with the divergent *yia*K-S promoter (Ibanez et al, 2000a). Whereas the *yia*K-S promoter strongly responds

to glucose catabolite repression through the Crp site 1, the *yia*J promoter responds to the same conditions with increased expression (Ibanez et al, 2000a). Thus, the transcription of the *yia*K-S operon is coordinately repressed in the presence of glucose by both the effect of the absence of cAMP-Crp binding to the promoter and the increase in YiaJ repressor synthesis. The mechanism for the *yia*J activation is believed to be based upon the absence of the Crp-cAMP complex from Crp site 1 (i.e. unhindered *yia*J transcription). Hence, in the presence of glucose, catabolite repression of *yia*K-S and increase in YiaJ repressor synthesis coordinate the full mechanism of *yia*K-S repression. Conversely, in the absence of glucose, cAMP-Crp complex binds to Crp site 1 activating *yia*K-S expression and obstructing the *yia*J-encoded repressor expression and synthesis (Ibanez et al, 2000a). Although a mechanism has been proposed involving the Crp site 1, the function of Crp site 2 remains unknown in spite of its observed binding abilities. Nevertheless, the transcriptional regulation of expression of the *yia*K-S operon is believed to be repressed by YiaJ and modulated by IHF and Crp.

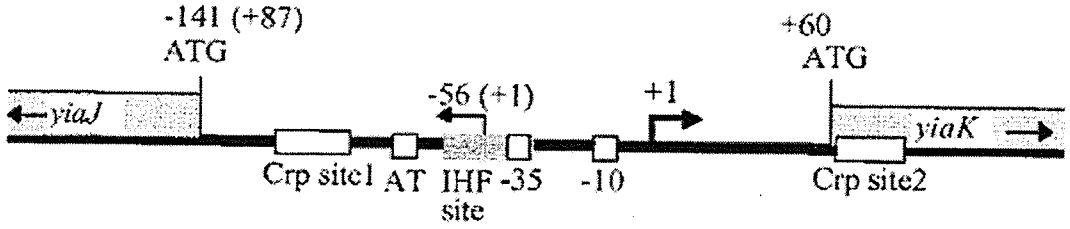


FIGURE 1.11: The *yia*K-*yia*J Promoter Region. Upstream of the start site of the *yia*K-S operon the -10 and -35 consensus, high AT content region, IHF site, and Crp-binding sites are located by the corresponding boxes (Ibanez et al, 2000a).

1.5 Thesis Objectives

The two aforementioned *E. coli* IclR homologues detected by sequence similarity searches, YagI and YiaJ, were selected for study. The project began with generating artificial clones of the open reading frames *yag*I and *yia*J using classical as well as modern molecular biology techniques. Once successful clones were produced, the gene products of the subjects under study were isolated and purified for initial biochemical characterization of structure and function. The aim of this research was to obtain further understanding of the IclR family of transcriptional regulators by studying individual members collectively.

Chapter II:

Materials and Methods

2.0 Molecular Biology Databases: The Importance of Databases

Molecular biology databases play an essential role in the research and development of computational biology (otherwise known as bioinformatics), that is the application of mathematical and computer science methods to solving problems in molecular biology that require large scale data, computation, and analysis (Baker et al., 1999). Throughout the course of this thesis investigation, the following databases were routinely employed:

1). The Center for Information Biology and DNA Data Bank of Japan (DDBJ)
http://www.cib.nig.ac.jp

2). The European Molecular Biology Laboratory (EMBL)
http://www.embl-heidelberg.de

3). The Expert Protein Analysis System (ExPASy)
http://www.expasy.ch

4). The National Center for Biotechnology Information (NCBI)
http://www.ncbi.nlm.nih.gov

5). The Protein Information Resource (PIR)
http://pir.georgetown.edu

6). The Institute for Genomic Research (TIGR)
http://www.tigr.org/

The availability of such databases is a powerful resource for the researcher as it provides access to the most current and comprehensive sequence information.

NCBI and ExPASy: GENOMICS & PROTEOMICS

NCBI

The NCBI database is the home of an integrated collection of databases and services (*PubMed, GenBANK, Structure, Genome PopSet, ProbeSet, OMIM, Books, UniSTS*, etc.). The databases include information regarding nucleotide and protein sequences, macromolecular structures, and whole genomes. Navigation at NCBI is accomplished through *Entrez,* a search and retrieval engine that cross-references information from databases at NCBI.

The most well known of the NCBI databases is the National Institute of Health's (NIH) GenBANK database, an annotated collection of all publIcly available DNA sequences (Benson et al., 2000). Each GenBank entry contains a multitude of information on the sequence and its biochemical features (e.g. protein translations for coding regions), the source organism, as well as bibliographic references and a link to MEDLINE.

A particularly convenient and useful tool available at NCBI is BLAST® (Basic Local Alignment Search Tool), a set of similarity-search programs designed to explore all of the available sequence databases irrespective of the nature of the query (i.e. protein or DNA). In particular, a nucleotide or polypeptide sequence is entered as a query and the search tool identifies, aligns and retrieves those sequences bearing a significant degree of homology.

ExPASy

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server from the Swiss Institute of Bioinformatics (SIB) is a dedicated molecular biology service with an

emphasis on data relevant to the analysis of protein sequences and structures. Like NCBI, ExPASy serves as a metaserver for a subset of databases such as the *SWISS-PROT, PROSITE, SWISS-2DPAGE, SWISS-3DIMAGE, ENZYME, CD40Lbase* and *SeqAnalRef* databases. In addition, like NCBI and most other databases for that matter, ExPASy also allows queries through other cross-referenced databases (such as EMBL/GenBank/DDBJ, Medline, etc). However, arguably the greatest feature of this particular server is that it allows access to many analytical tools for the identification and characterization of proteins. In particular, from a simple input of the primary amino acid sequence of a given protein, similarity searches, pattern and profile searches, post-translational modification prediction, primary structure analysis, secondary as well as tertiary structure prediction, transmembrane regions detection, and sequence alignment are all possible.

2.1 Growth Media

The primary growth medium used throughout the course of this study was Luria Bertani (LB) bacteriological medium, a rich nutrient medium for the maintenance and propagation of *E. coli.* LB broth base (10 grams select peptone, 5 grams yeast extract, and 10 grams sodium chloride pH-adjusted to 7.5 and autoclaved per one litre of deionized water) was used for liquid *E. coli* cultures, whereas LB-Agar (10 grams select peptone, 5 grams yeast extract, 10 grams sodium chloride, and 15 grams bacteriological agar pH-adjusted to 7.5 and autoclaved per one litre of deionized water) was used to prepare agar plates for isolation of *E. coli* bacterial colonies.

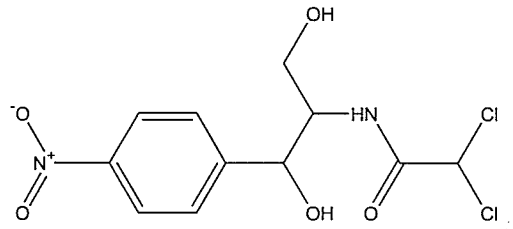
2.2 Antibiotics

Both LB broth base and LB-Agar were supplemented with different antibiotics or combinations of antibiotics throughout the course of experimentation. In particular, the three antibiotics tetracycline, kanamycin and chloramphenicol were combined with LB media throughout the course of this study. Chloramphenicol was used to select for the BL21 (DE3) pLysS cell line, tetracycline for the Novablue cell line, and kanamycin for cells transformed with pET-24b(+) and pET-28b(+) vector DNA. Chloramphenicol was used at a final concentration of 34 μg/mL tetracycline at a final concentration of 12.5 μg/mL, and kanamycin at a final concentration of 30 μg/mL.

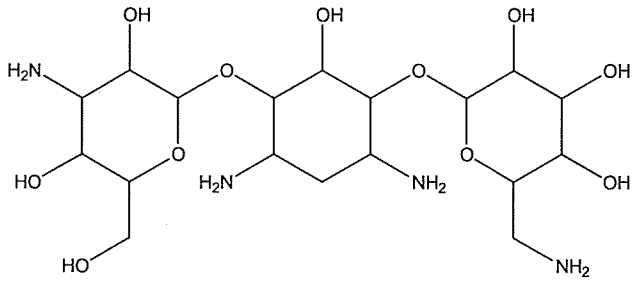
On the other hand, the antibiotic rifampicin was implemented to optimize protein expression. This particular antibiotic was only used in LB liquid media at a final concentration of 200μg/mL.

TABLE 2.0: Antibiotics Incorporated in LB Media.

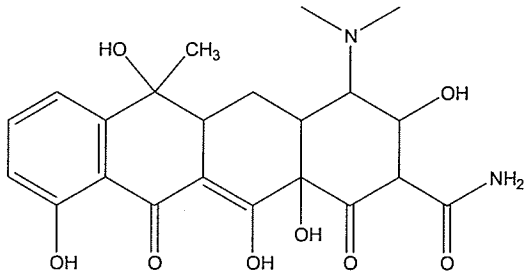
Antibiotic	Mode of Action	Supplemented LB
Chloramphenicol (34 μg/mL)	- protein synthesis inhibitor Binds to 23S rRNA of 50s subunit of the ribosome preventing attachment of amino-acyl tRNA to binding region (inhibits transpeptidation during protein synthesis)	LB broth and agar
Kanamycin (30 μg/mL)	- protein synthesis inhibitor - Binds irreversibly to 30S ribosomal subunit of the bacterial 70S ribosome and blocks the initiation complex - Causes misreading of mRNA so incorrect amino acids are inserted into the polypeptide leading to nonfunctional or toxic peptides	LB broth and agar
Tetracyline (12.5 μg/mL)	- protein synthesis inhibitor - Inhibits protein synthesis (elongation) by preventing binding of aminoacyl-tRNA to the 30S ribosomal subunit	LB broth and agar
Rifampicin (200 μg/mL)	- nucleic acid synthesis inhibitor - Inhibits initiation of mRNA synthesis by binding to β-subunit of bacterial DNA-dependent RNA polymerase	LB broth only



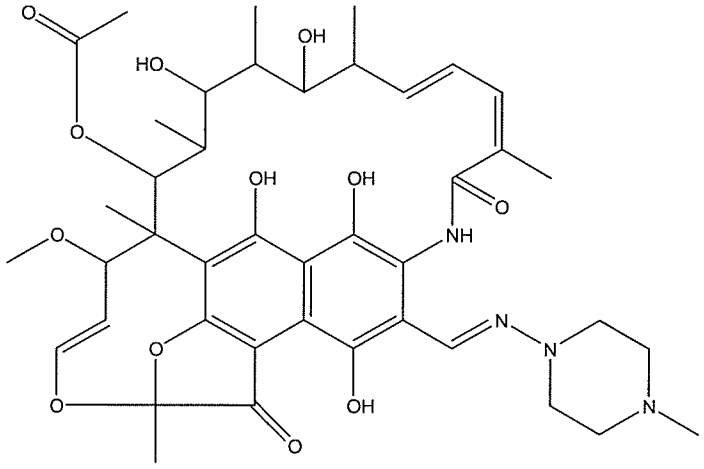
Chloramphenicol



Kanamycin



Tetracycline



Rifampicin

FIGURE 2.0: Antibiotics Employed in the Study. Chloramphenicol and tetracycline were used to select for BL21 (DE3) pLysS and Novablue cell lines respectively. Kanamycin was used as a selective agent for transformed cells and rifampicin was used to enhance protein expression.

2.3 Nucleic Acid Determination via Spectrophotometry

DNA, RNA, oligonucleotides and even mononucleotides can be measured directly in aqueous solutions in a diluted or undiluted form using spectrophotometry. Aqueous buffers with low ion concentrations (e.g. TE buffer) are ideal for this method. In particular, assuming that an absorption of $OD_{260} = 1.0$ is equivalent to approximately 50 µg/ml dsDNA, approximately 37 µg/ml ssDNA, 40 µg/ml RNA or approximately 30 µg/ml for oligonucleotides (primers), it is possible to determine the concentration of a sample of nucleic acid. In addition, the purity of a nucleic acid solution may be estimated as interference by contaminants can be recognized by the calculation of a ratio.

The ratio A_{260}/A_{280} is used to estimate the purity of nucleic acid, since proteins absorb at 280 nm. Pure DNA should have a ratio of approximately 1.8, whereas pure RNA should give a value of approximately 2.0. Absorption at 230 nm reflects contamination of the sample by substances such as carbohydrates, peptides, phenols or aromatic compounds. In the case of pure samples, the ratio A_{260}/A_{230} should be approximately 2.2.

All spectrophotometric measurements made in this study were performed on a Milton Roy Spectronic 3000 Array spectrophotometer.

2.4 Oligonucleotides

Initially, the method of selection from EMBL3 Lambda *E. coli* genomic libraries was chosen as the method for selecting wild type genomic DNA containing the *E. coli* YagI open reading frame (ORF). However, this method was abandoned for the more direct method of generating synthetic DNA using the polymerase chain reaction (PCR).

Following successful experimentation with the *yag*I ORF, the PCR method was then adopted for the *E. coli yiaJ* ORF. In particular, the correct 5'→3' DNA sequences for both *yag*I and *yia*J were obtained from GenBANK and analyzed for the proper design of PCR primers. PCR primer oligonucleotides as well as sequencing primers were obtained commercially through GibcoBRL Life Technologies for *yag*I whereas PCR primers were obtained through Sigma-Genosys for *yia*J amplification. All oligonucleotide primers used in this study were obtained as desalted-lyophilized material on a 50 nM scale.

Stock concentrations of oligonucleotides were prepared by reconstituting the lyophilized oligonucleotides in TE buffer (10 mM Tris-HCl, pH 8.0, 1mM EDTA) at concentrations greater than 10μM and stored at –20°C. To prevent interference from EDTA, aliquots from stocks were diluted to desired working concentrations with Nanopure™ water (ddH_2O). Concentrations of oligonucleotides were determined via spectrophotometric methods as previously described.

TABLE 2.1: Oligonucleotides Used in this Study

Oligonucleotide	Purpose/ Application	Properties M_r (g/mol) Tm (°C)	Sequence (5'-3')
Doc Xho1 P1	*yag*I PCR Primer	9931.5 74.0	CTGCAGGACTCGAGCCGGTCTAACGGCTCAGG
Doc Nde1 P2	*yag*I PCR Primer	12051.9 76.0	CTCGATAGATCATATGCCGATTATTCAGTCTGTTGAACG
pETDOCSP1	Sequencing Primer	5229.4 56.0	GCAGCAGCCAACTCAGC
pETDOCSP2	Sequencing Primer	5275.5 54.0	GCTCGCGGAAATTAGCC
pETDOCSP3	Sequencing Primer	5298.4 58.0	GGGCGTAGCCTTGCTCG
pETDOCSP4	Sequencing Primer	5267.4 58.0	CACCTTGCCGATGGCGG
pETDOCSP5	Sequencing Primer	5237.4 62.0	CCGCCGGGACAGCTCCG
pETDOCSP6	Sequencing Primer	5312.5 52.0	GCTGCAGGGTTTTTAGC
Dopey Hind3 P1	*yiaJ* PCR Primer	8634.7 82.0	CACTTGTCCAAAGCTTTATGTGATCGCG
Dopey Nco1 P2	*yia*J PCR Primer	8481.6 86.0	CAGGACGCGACACCATGGGGAAAGAAG
T7 Promoter Primer	Sequencing Primer	6205.1 56.0	TAATACGACTCACTATAGGG
T7 Terminator Primer	Sequencing Primer	5914.9 58.0	GCTAGTTATTGCTCAGCGG

2.5 The pET System

The plasmid for expression by T7 RNA polymerase or pET *E. coli* expression system is a widespread practice of in vivo bacterial cloning and expression. This system, created by Moffat and Studier, combines a bacteriophage T7 RNA polymerase-based expression vector with a genetically engineered host bacterium (Studier and Moffatt, 1986, Rosenberg et al., 1987, Studier et al., 1990, Dubendorff and Studier,1991). These pET vectors are available from a number of commercial suppliers including Stratagene, Promega and Invitrogen to name a few. However, the company known as Novagen in particular, has developed a variety of derivatives of the original pET vectors and currently offers the largest selection of pET vectors to date. The pET System™ is the

most powerful system yet developed for the cloning and expression of recombinant proteins in *E. coli* (Novagen, 1999).

The pET system was used to express two different protein sequences of interest in this study and involved the two pET vectors pET-24b+ and pET-28b+ and the two cell lines Novablue and BL21(DE3)pLysS.

pET-24b(+) and pET-28b(+) Vectors

Two different pET vectors were used to express protein from two different inserts. In particular, the *yag*I ORF was cloned into the pET-24b(+) vector (5310 bp) whereas the *yia*J ORF was cloned into the pET-28b(+) vector (5369 bp). Although the pET-28b(+) and pET-24b(+) vectors are essentially identical, the main difference between them is that the former has an *Nco*1 restriction endonuclease cleavage site that is absent in the latter. The main features of both vectors are as follows:

- Kanamycin resistance marker
- pBR322 origin of replication (all pET vectors are derivatives of the pBR322 vector)
- f1 origin of replication (allows single stranded vector to be produced when co-infected with M13 helper phage)
- *lacI* gene (*lac* repressor protein)
- T7 transcription promoter (specfic for phage T7 RNA polymerase)
- *lac* operator region 3' to the T7 promoter
- highly efficient ribosome binding site (rbs) from the phage T7 major capsid protein
- multiple cloning site (polylinker region) downstream of the T7 promoter (a gene of interest can thus be cloned to be in-frame with the transcribed region downstream from the T7 promoter)

The pET vectors also have features that not only facilitate easy subcloning, but permit easy detection and purification. In particular, both of the pET-24b(+) and pET-28b(+) vectors contain regions within the multiple cloning site encoding different stretches of peptide. When an insert is cloned into these sites, the result is expression of

the cloned insert as a fusion protein. The following figure depicts the various fusion tags found in the multiple cloning site of the pET-24b(+) and pET-28b(+) vectors.

A). pET-24b(+) Fusion Tag Regions

```
 Nde1   Nhe1                                                BamH1   EcoR1   Sac1
CAT ATG GCT AGC ATG ACT GGT GGA CAG CAA ATG GGT CGG GAT CCG AAT TCG AGC TCC
    Met Ala Ser Met Thr Gly Gly Gln Gln Met Gly Arg Asp Pro Asn Ser Ser Ser
                        T7-Tag
  Sal1   Hind3   Not1    Xho1
GTC GAC AAG CTT GCG GCC GCA CTC GAG CAC CAC CAC CAC CAC CAC TGA
Val Asp Lys Leu Ala Ala Ala Leu Glu His His His His His His End
                                    C-terminal His-Tag
```

B). pET-28b(+) Fusion Tag Regions

```
 Nco1
ACC ATG GGC AGC AGC CAT CAT CAT CAT CAT CAC AGC AGC GGC CTG GTG CCG CGC GGC AGC
    Met Gly Ser Ser His His His His His His Ser Ser Gly Leu Val Pro Arg Gly Ser
                          N-terminal His-Tag                Thrombin Cleavage Site
  Nde1   Nhe1                                                BamH1   EcoR1   Sac1
CAT ATG GCT AGC ATG ACT GGT GGA CAG CAA ATG GGT CGG GAT CCG AAT TCG AGC TCC
His Met Ala Ser Met Thr Gly Gly Gln Gln Met Gly Arg Asp Pro Asn Ser Ser Ser
                        T7-Tag
  Sal1   Hind3   Not1    Xho1
GTC GAC AAG CTT GCG GCC GCA CTC GAG CAC CAC CAC CAC CAC CAC TGA
Val Asp Lys Leu Ala Ala Ala Leu Glu His His His His His His End
                                    C-terminal His-Tag
```

FIGURE 2.1: Fusion Tag Regions of the pET-24b(+) and pET-28b(+) Vectors. A). pET-24b(+) has a T7-Tag® (the 11 amino acid leader sequence of gene 10) as well as a C-terminal His-Tag (histidine hexapeptide). B). pET-28b(+) has both an N-terminal as well as C-terminal His-Tag, a thrombin cleavage site and the T7-Tag.

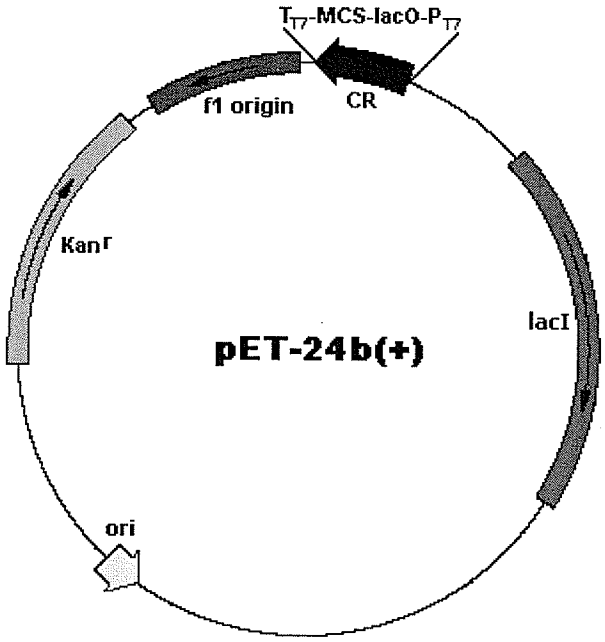
Unfortunately, the different tags can sometimes have unfavorable effects on the behavior of fusion proteins. For example, the tags may favor solubility in some cases and impede it in others because the extraneous peptide sequences may affect the native conformation of target proteins. Hence, although the fusion tags are most often beneficial, there are instances where they are not. It is therefore desirable to avoid expressing target inserts as fusion proteins all together.

Referring to FIGURE 2.1 above, it can be seen that both pET-24b(+) and pET-28b(+) vectors have a unique restriction site in the initiation codon of the cloning region (*Nde*I for pET-24b(+) and *Nco*1 for pET-28b(+)). Therefore, cloning into the *Nde*1 (pET-24b(+)) and *Nco*1 (pET-28b(+)) sites in these vectors allow this codon to be

replaced by the initiation codon of the insert, resulting in expression of the insert with no fusion. As our aim was to synthesize protein specimens as accurate replicas of the naturally occurring templates, we cloned our inserts such that the expressed protein targets were not fusion proteins. Furthermore, the inserts were constructed such that both the true initiation and termination codons of the *yag*I and *yia*J ORFs would be fully functional in the recombinants.

The use of the two different pET vectors was necessary in order to satisfy the requirements of the cloning process. In particular, the successful cloning and expression of a target insert requires that the restriction sites selected from the multiple cloning site must not be present in said insert. The *yag*I ORF was the first to be successfully cloned into and expressed from the pET-24b(+) vector using the pET Expression System®. However, as there is an *Nde*1 restriction endonuclease recognition site internal to the *yia*J ORF, the pET-24b(+) could not be used to prepare the *yia*J ORF using our methods. Instead, the pET-28b(+) vector was employed to clone and express *yia*J. FIGURES 2.2 and 2.3 detail the features of the pET-24b(+) and pET-28b(+) vectors.

A).



Legend:

f1 origin - f1 origin of replication
Kan^r - Kanamycin resistance marker
lacI - lac repressor gene
ori - pBR322 origin
CR - Cloning Region
lacO - lac operator
MCS - Multiple Cloning Site
P_{T7} - T7 promoter
T_{T7} - T7 terminator

B).

T7 Promoter Primer

T7 Promoter — lac operator

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATT

rbs — NdeI — NheI

TTGTTTAACTTTAAGAAGGAGATATACATATGGCTAGCATGACTGGTGGACAGCAAATG

BamH1 — EcoR1 — SacI — SalI — HindIII — NotI — XhoI

GGTCGGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCA

CCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTG

T7 Terminator

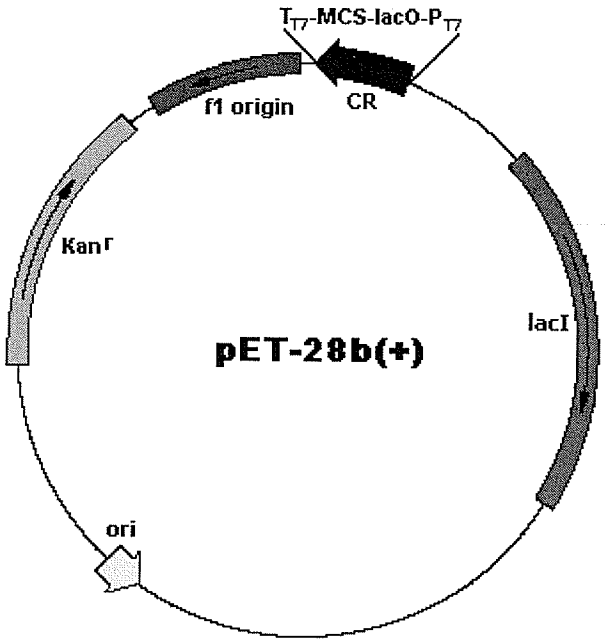
CCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGT

T7 Terminator Primer

TTTTTG

FIGURE 2.2: The pET-24b(+) Vector. A). Features of the pET-24b(+) vector include: the f1 origin of replication (allows production of single stranded DNA upon helper phage infection), kanamycin resistance, the gene for the lac repressor, the pBR322 origin, and cloning region. The cloning region includes the lac operator, a multiple cloning site as well as both the T7 promoter and terminator. B). An expanded view showing the features of the cloning region in detail.

A).



Legend:
f1 origin - f1 origin of replication
Kan^r - Kanamycin resistance marker
lacI - lac repressor gene
ori - pBR322 origin
CR - Cloning Region
lacO - lac operator
MCS - Multiple Cloning Site
P_{T7} - T7 promoter
T_{T7} - T7 terminator

B).

T7 Promoter Primer

T7 Promoter — lac operator

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATT

rbs — NcoI

TTGTTTAACTTTAAGAAGGAGATATACCATGGGCAGCAGCCATCATCATCATCATCACA

NdeI — NheI

GCAGCGGCCTGGTGCCGCGCGGCAGCCATATGGCTAGCATGACTGGTGGACAGCAAATG

BamH1 — EcoR1 — SacI — SalI — Hind3 — NotI — XhoI

GGTCGGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCA

CCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTG

T7 Terminator

CCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGT

T7 Terminator Primer

TTTTTG

FIGURE 2.3: The pET-28b(+) Vector. A). Features of the pET-28b(+) vector include: the f1 origin of replication, kanamycin resistance, the gene for the lac repressor, the pBR322 origin, and cloning region. The cloning region includes the lac operator, a multiple cloning site as well as both the T7 promoter and terminator. B). An expanded view showing the features of the cloning region in detail.

2.6 Cell Lines

Novablue: A Host for Cloning

Before gene transcription occurs, the pET recombinant must first be established in a non-expression *E. coli* host, i.e., one that does not contain the T7 RNA polymerase gene. This cloning stage is necessary as basal levels of gene transcription can result in plasmid instability due to the potential toxicity of target proteins to the cells. Using suitable cloning hosts, the desired result would be high yields of high quality plasmid DNA as well as hosts for maintaining the recombinants (frozen transformant stocks). The JM109, DH5α, XL1-Blue and Novablue *E. coli* strains are particularly useful as they are all *recA*$^-$ *endA*$^-$ for their genotypes, have high transformation efficiencies and yield large quantities of excellent quality plasmid. However, Novablue was selected for our purposes because it possesses additional advantageous characteristics.

Aside from being *recA*$^-$ *endA*$^-$ the *E. coli* K-12 NovaBlue strain also carries tetracycline resistance which facilitates a higher degree of selection stringency. If so desired, mutagenesis is also made possible using this strain due to an F factor that allows helper phage infection and hence the production of single stranded plasmid DNA (for plasmids carrying the f1 origin of replication).

Once the proper pET-recombinant is established, it is then transformed into an expression host containing a chromosomal copy of the T7 RNA polymerase gene, which are as previously stated, most commonly DE3 lysogens.

BL21(DE3) pLysS: A Host for Expression

In DE3 lysogens, the T7 RNA polymerase gene is under the control of the *lacUV5* promoter. In theory, the gene can only be transcribed if IPTG is added to activate the

*lac*UV5 promoter and turn on synthesis of T7 RNA polymerase. This in turn transcribes the gene in the pET recombinant. In practice, however, there is always a basal level of transcription in the uninduced state and hence the *E. coli* RNA polymerase does make a small amount of T7 RNA polymerase without induction. This causes a small amount of the cloned gene to be expressed from the pET recombinant, which may have detrimental effects on host cell growth. As well, gene products may be susceptible to enzymatic proteolysis from proteases present in certain host strains. Hence, to achieve a higher level of stringency and to avoid the proteolysis problem, the *E. coli* host expression strain, BL21(DE3)pLysS was incorporated.

As with all DE3 lysogens, BL21(DE3) and the BL21(DE3)pLysS hosts permit induced expression of T7 RNA polymerase. However, as these *E. coli* BL21 strains are derived from *E. coli* B, they naturally lack the Lon protease and are engineered to be deficient for the OmpT protease. The BL21(DE3)pLysS strain has the added advantages of chloramphenicol resistance, which allows for a higher stringency of selection following transformation, and it contains the pLysS plasmid, which carries the gene encoding T7 lysozyme. T7 lysozyme is important for two reasons. First, T7 lysozyme is a natural inhibitor of T7 RNA polymerase, but it does not affect the level of expression achieved following induction with IPTG. Hence, basal expression of a cloned gene in uninduced cells is further minimized. Secondly, T7 lysozyme normally attacks the cell wall, but since it is enclosed within the cell membrane, it cannot harm the cell wall. However, if the cell membrane were to lose its integrity, then the cell wall would certainly be susceptible to lysis. As will be seen later, T7 lyzozyme carries a further advantage of facilitating the preparation of cell extracts.

2.7 Competent Cells and $CaCl_2$ Transformation

Competent Cells

Three milliliters of an overnight culture supplemented with the appropriate antibiotic(s) is added to 100 mL of fresh LB-antibiotic media and grown to mid-log phase (A_{600} = 0.6-0.8). The cells are chilled on ice for 5 minutes then centrifuged for 5 min. at 5K after which the supernatant is removed. The cell pellets are then resuspended in ice-cold sterile 100 mM $CaCl_2$ (0.5 volumes/mL of original culture volume) and incubated on ice for 30 minutes. The cells are centrifuged again for 5 minutes at 5K and resuspended in fresh ice-cold sterile 100 mM $CaCl_2$ (50 µL/mL of original culture volume). Several 100µL aliquots were prepared from this point as transformation samples (stored at 4°C for 1 week if not in immediate use) as well as glycerol stocks.

Storage of Strains: Glycerol Stocks

Stocks of host cell lines and pET recombinants were routinely prepared and stored as 10% glycerol stocks. Following the second resuspension of cells with 100 mM $CaCl_2$ as described above, 100 µL 10% glycerol stocks were prepared, flash frozen in liquid nitrogen, and stored at -70°C.

Bacterial Transformation

To 100 µL aliquots of competent cells are added, 0.5-5 µL of DNA, followed by gentle mixing and incubation on ice for 5 min. The transformation mixture is then heat shocked at 42°C for 2 min. and then immediately placed on ice for 5 min. Five hundred microlitres of antibiotic free LB broth is added to the heat-shocked cells, which are then allowed to recover in a shaking incubator at 37°C for 1 hour. The volume of cell solution is reduced by centrifuging the transformed cells and removing supernatant such that the

combined volume of cells and supernatant upon resuspension is only 100μL. This 100μL volume of transformed cells is then spread plated onto fresh LB-antibiotic agar plates and incubated at 37°C overnight.

2.8 Alkaline-Lysis Plasmid Preparation

Both pET-24b+ and pET-28b+ vectors were obtained from Dr. J.D. O'Neil and transformed into Novablue cells. High-yield stocks of each vector were obtained through a standard alkaline-lysis plasmid preparation as follows (Sambrook et al., 1989). In order to perform the plasmid preparation using this method, the three following solutions were required:

BUFFFER 1
50mM Glucose
10mM EDTA
25mM Tris-HCl
pH 8.0
Store at 4.0°C

BUFFER 2 (Make up fresh)
0.2M NaOH
1.0% SDS

BUFFER 3
3M sodium acetate, pH 4.8

One hundred millilitres of LB-Tetracycline-Kanamycin are inoculated with a 3 mL overnight culture of Novablue transformed cells and are incubated at 37°C in a shaking incubator overnight (12hrs.). These 100 mL cultures were then centrifuged at 3000 rpm for 10 minutes. The supernatant is drained off and pellets are completely resuspended in 10 mL of ice cold Buffer 1. This solution is allowed to stand for 5 minutes on ice. Twenty millilitres of FRESH Buffer 2 is added and gently mixed by inversion. This solution is allowed to stand for 10 minutes on ice. Fifteen millilitres of

ice cold Buffer 3 is added and gently mixed by inversion. This solution is allowed to stand for 10 minutes on ice after which it is centrifuged at 10,000 rpm for 10 minutes. The supernatant is collected and centrifuged until pelleting from cellular debris is no longer visible. RNase A (10mg/mL) is added to a final concentration of 20 ug/mL and the supernatant is incubated at 37°C for a minimum of 1 hr. Isopropanol is added at a volume of 0.6, mixed by inversion and the solution is refrigerated at 4°C overnight. The solution is then centrifuged at 10,000 rpm for 10 minutes and the supernatant is discarded. The pellet is taken up in 5 mL of 0.1% SDS/40mM Tris-HCl/0.1M NaOAc/1 mM EDTA/pH 8.0. Five millilitres of phenol/chloroform/isoamyl alcohol (25:24:1) is added and the mixture is vortexed. The mixture is allowed to stand for 5 minutes with frequent agitation. The solution is then centrifuged at 5000 rpm for 2 minutes. The aqueous layer is transferred to a clean vessel and the phenol/chloroform/isoamyl alcohol extraction is repeated. The aqueous layer is then extracted twice with chloroform. Isopropanol is added at a volume of 0.6, mixed gently and the solution is refrigerated at 4°C overnight. The solution is centrifuged at 10,000 rpm for 10 minutes after which the supernatant is discarded and the pellet is resuspended in 500 uL of Nanopure water then transferred to a microfuge tube. An equal volume of 10M LiCl is added to the solution, mixed gently and allowed to stand at room temperature for 30 minutes. The solution is placed in a microfuge and spun for 10 minutes at maximum velocity. The supernatant is then transferred to a fresh tube containing 600 uL of isopropanol and mixed. The solution is allowed to stand for 30 minutes to overnight after which 1 mL of 70% ethanol is added and mixed gently. The solution is centrifuged for 10 minutes after which the supernatant is discarded and the pellet is allowed to air dry. The pellet is then dissolved

in 100 uL of TE or Nanopure™ water (ddH_2O). The purity and concentration of the plasmid preparations were assessed by measuring the absorbances at 260 nm and 280 nm. When required (sequencing), QIAgen DNA purification filters were implemented for an extra pure finish. From this preparation, as high as 5 mg/mL DNA concentrations were achieved.

2.9 Agarose and Polyacrylamide Gel Electrophoresis

Agarose Gels

DNA was routinely visualized in 1% agarose slab gels prepared from electrophoresis-grade agarose (Fischer Scientific). Either of two DNA molecular size standards was commonly electrophoresed alongside DNA samples. In particular, GIBCO's *1Kb PLUS DNA Ladder*™ (100, 200...500, 650, 850, 1000, 1650, 2000, 3000...12000 bp fragments) or *λ DNA/Hind III Fragments* (564, 2027, 2322, 4361, 6557, 9416, 23130 bp fragments) standards were used. DNA samples were loaded with loading dye (10 mM EDTA, 40% sucrose, 0.25% bromphenol blue) into 8" x 10" (20 sample wells; up to 40 μL volume) or 2.7" x 4" (10 sample wells; up to 30 μL volume) slab gels. Electrophoresis was then performed in Tris-Acetate-EDTA (TAE) buffer (40 mM Tris, 1 mM EDTA pH adjusted to 7.7 with glacial acetic acid) containing 20 μg/mL of ethidium bromide (EtBr) at 50 volts and 25 volts. Following electrophoresis, excess EtBr was rinsed off with water and results were visualized with UV light.

Polyacrylamide Gels

Protein samples were routinely analyzed for size and purity using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and staining with Coomassie Brilliant Blue (Sigma Aldrich). In particular, all samples were prepared and

electrophoresed under denaturing conditions (heat, SDS and β-mercaptoethanol) in standard Laemmli (Laemmli, 1970) 12% polyacrylamide gels (5% stacking gel; 12% resolving gel). A three-component protein mixture comprised of bovine serum albumin (65kDa), carbonic anhydrase (29kDa) and lysozyme (14kDa) served as the molecular weight standard. Proteinaceous samples (10 – 20 μL) were mixed with equal volumes of 2x SDS sample buffer (0.5 M Tris-HCl (pH 6.8), 20% glycerol, 4% SDS, 10 mg/ml Bromphenol Blue, and 2% β-mercaptoethanol), vortexed vigorously, and heated to upwards of 70°C. Denatured samples were then loaded into thin-casted 3.32" x 4" polyacrylamide gels and eletrophoresed at 150-200 volts in Tris-Glycine-SDS PAGE running buffer (0.025 M Tris base, 0.192 M glycine, and 1% SDS pH adjusted to 8.3 w. HCl). Following electrophoresis, gels were removed from moulds and fixed in 0.25% isopropanol:acetic acid for 10 minutes. The fixing/destaining solution was then replaced with Coomassie Brilliant Blue staining solution (fixing/destaining solution with 0.25% Coomassie Brilliant Blue) and staining was allowed to occur anywhere from 2 hours to overnight. Staining solution was then removed; the gels were rinsed with deionized water (dH_2O) and placed into fixing/destaining solution for several hours. Protein bands were then visible as blue bands against a transparent background. Gels were then dehydrated and preserved by wrapping in cellophane.

2.10 Synthesis of *yag*I and *yia*J Genes

Oligonucleotide Design

In order to successfully amplify the *yag*I and *yia*J inserts via the polymerase chain reaction (PCR), it was necessary to not only take into consideration the basic requirements of PCR primer design (i.e. >G-C vs. <A-T content, melting temperature of

each primer, secondary structure potential, primer dimerization etc.), but the primers had to be engineered with unique restriction sites.

To begin, the correct DNA sequences of *yag*I and *yia*J were obtained from GenBANK. The sequences were then analyzed for compatible restriction endonuclease recognition sites for use with the pET-24b(+) and pET-28b(+) vectors. In order to clone and express the *yag*I insert without any fusions, it was necessary to clone the insert into the pET-24b(+) vector. As previously mentioned, the *yia*J insert would have to be cloned into and expressed from the pET-28b(+) vector due to a naturally occurring internal *Nde*1 site present in the ORF.

In order to avoid expressing target proteins as fusion proteins, *Nco*1 must be one of the restriction sites used for pET-28b(+) and *Nde*1 must be one of the restriction sites used for pET-24b(+). So long as none of the other restriction sites in the multiple cloning site are internal to the inserts, any one of them may be selected as the other engineered restriction site. However, when introducing a mutation into a target DNA sequence, it is always desirable to introduce as few base-pair changes as possible. The restriction endonucelase recognition sites that would require the least number of base pair changes to the natural reading frames of the *yag*I and *yia*J ORFs were then selected. Once the restricion endonuclease recognition site selections were made, the proper primers could then be designed.

As PCR is based on template-directed 5' to 3' extension of oligonucleotide primers, both primers will read 5' to 3', but each primer must be complementary to its target template strand and flanking the region of DNA to be amplified. In particular, the primer containing the natural methionine initiation start codon of the target gene will be

complementary to the sense strand (sense primer) and the other primer containing the natural stop codon will be complementary to the antisense strand (antisense primer).

The success of PCR is due largely in part, to high temperatures for stringency hence the need for thermostable DNA polymerases. The melting temperature (T_m), i.e. the temperature at which 50% of a given oligonucleotide and its perfect complement are in duplex, is especially crucial to the stringency and success of PCR.

In the absence of destabilizing agents, such as formamide, DMSO or urea, T_m depends on the oligonucleotide concentration (high oligonucleotide concentrations favor hybrid formation, which results in a higher melting temperature) and salt contentration (high ionic strength results in a higher T_m as cations stabilize the DNA duplexes). However, the most important factor is the sequence of the oligonucleotide itself as a GC-rich sequence versus an AT-rich sequence has a higher melting temperature. This is due to a G-C base pair having three hydrogen bonds as opposed to only two in an A-T base pair.

The simplest equation for calculating T_m for oligonucleotides is the Wallace rule (Wallace et al., 1979):

Tm (°C) = [(# G + C) x 4] + [(# A + T) x 2]

High melting temperatures (> 70°C) for each primer were selected and the primers were then either lengthened or shortened to obtain a desirable melting temperature.

yagI

Upon obtaining the *yagI* ORF located in section 25/400 of the *E. coli* genome from GenBANK, analysis of the ORF revealed that it would be possible to introduce two

unique restriction sites, namely *Nde*1 and *Xho*1, to facilitate cloning into the pET-24b(+)

E. coli vector.

A).
*Xho*1 restriction endonuclease recognition site – CTCGAG

*Nde*1 restriction endonuclease recognition site – CATATG

B).
CGACTACTTCACCTACGAGCCGGTCTAACGGCTCAGGCCGGGTAGCCCAGCGCGCGCGAG
AGCGCGAGCCCGGCCTGCTGAAGCTGCTCGCGGAAATTAGCCAGCTCCGCGTCGTCCACG
CGGGAGGTCAGCGTCGACAGGCTCAGGGCGGCGATGACGCGGGACTCGTGGTTCCACACC
GGCACCGCCACGCAGCGCACGCCCTGCTCGTTCTCTTCGCTGTCCAGGGCGTAGCCTTGC
TCGCGGGTCTGCGCCAGGGCGCTCATTAAGGCTTCGCGAGACGCGAGGGTGGCGGGCGTA
AAGGTAGTGTACTGATAGCCCTCCAGCAGGGCGTTCAGCTCGGCCTCGCCCAGCCAGGCA
ATCAACACCTTGCCGATGGCGGTGGCGTGCACCGGCAGGCGGCGGCCGATGCGTGAATAG
GCGATGGCGGCCAGCTTGCCTTCAATCTTCTCGATATAGACCCCTTCACGCCCGTCCAGG
ATCCCCAGATGGGTGGTCTGCCCGGTCCGCCGGGACAGCTCCGTCAGCCAGCCTTTTGCC
TTCTGCCGAATATCGATGGAGCCCACGACAAAATGGCCGCGCTCGACCAGCTTCATGCCG
AGGCGATACTTGCCGTTCTCCGGGTTCTGATCGATATAGCCGTGAAGCTGCAGGGTTTTT
AGCAGCGAGTGGAGGGTACTCTTGCTCAGCCCCATCAGTTTGCTGATGTCGGTGATCTTA
AGCTCGGTGGCCTGCTCGTTGAACAGGTCGAGGATCTGCAACGCACGTTCAACAGACTGA
ATAATCGGCATAATGCTGGCATGTCCACGC

FIGURE 2.4: Designing the *yag*I Insert. A). Xho1 and Nde1 restriction endonuclease recognition sequences in blue and red respectively represent sites for insertion of unique restriction endonuclease recognition B). The 5'-3' *yag*I antisense open reading frame: small caps represents required base-pair mutations for insertion; underlined triplet represents start (CAT) and stop (TCA) codons.

It was then decided that the following two PCR primers would be used to generate the synthetic *yag*I ORF:

Doc Xho1 P1:

CTGCAGGACTCGAGCCGGTCTAACGGCTCAGG (antisense primer);

Tm = 74°C

Doc Nde1 P2:

CTCGATAGATCATATGCCGATTATTCAGTCTGTTGAACG (sense primer);

Tm = 76°C

yiaJ

The *yia*J ORF located in section 325/400 of the *E. coli* genome was also obtained from GenBANK and analyzed for introduction of unique restriction endonuclease recognition sites. It was decided that Nco1 and Hind3 recognition sites would be introduced such that cloning into the pET-28b(+) vector could be achieved.

A).
Hind3 restriction endonuclease recognition site – AAGCTT

Nco1 restriction endonuclease recognition site – CCATGG

B).
CCCTACTGCTTTGATCCAACGAAGGAAATGTTAAAGTTTTGGCACTTGTCCAAAGCgTTA
TGTGATCGCGCCCAGGTCATCCCTGACAGTAAATCCCAGTTCATTAGAAATAGCCTGCGC
GGTTTCACGCAGTGGTTTCAGGAGATTTTTCTCTCCCACCTGTTTCAGACGTGATGTCGA
AAGCGAAATCGACACGGCGTACGGCACCCGCCCATGAATATCAAACACCGGAACAGCAAT
ACAGGAGACGCCGAGTTCGTTTTCTTCTCTGTCCATCGCCGCTCCGCTTTCACGAATGTG
CGCCAGTTCGTCGAACATCGCGGGCAGCTCGGTAATGGTATTGCGGGTTAACGGCTGGAT
CTCATGCTGATGGCTTTCCCAGTATGACTTCACGTAGTCCGGGTGACCAAACGCCATGTA
GATCTTGCCCATTGCGGAACAGTAGAGCGGCATATGCTGGCCAATATAGGCACGGGTTCG
CAGCATCCCGGTTGTGGGTTCCAGCTTATAAATCAAAATAGCGTGATCGTCTTCGCGGCT
GGAGAAGTTAATGGTTTCACCAGTGGCGATGTTCAGTGCCTCAAGATGCGGAGCGGCGAT
ATGAATGATATTCAGCGAAGACAGCGCCTTCTGCCCGACGGCAATAAATTTGGTGGTCAG
GCGATAACTCCCTGCGGCAGGCGCGGTGGTCACATAGCCGCAGGACTGTAATCCCTGCAA
TAAGCGATGGACGGTACTCTTATTTAAACCAGCCAGCTCCGAAAGATGCGCCAACGGACA
ACCGTTTGGATAGTTGCTCAAAATCTCAATCAGCATCAACCCGCGAAACAGACTCTGGCT
TCCGGCTGGACGCTCTTTTTCCTGCGCCATCTCGTTCTCTTTTTTTCCCATCACTTCTTT
CCCCATTTTGTCGCGTC

FIGURE 2.5: Designing the *yia*J Insert. A). sequences in blue and red represent sites for insertion of unique restriction endonuclease recognition B). The *yia*J 5'-3' antisense open reading frame small caps represent required base-pair mutations for insertion; underlined triplets represents start (CAT) and stop (TTA) codons.

The two following primers were subsequently designed to synthetically amplify the *yia*I gene via PCR.

Dopey Hind3 P1:

CACTTGTCCAAAGCTTTATGTGATCGCG (antisense primer);

Tm = 82°C

Dopey Nco1 P2:

CAGGACGCGACACCATGGGGAAAGAAG (sense primer);

Tm = 86°C

The Polymerase Chain Reaction

The polymerase chain reaction (PCR; Mullis et al., 1986) exploits certain features of DNA replication. In particular, the technique involves a thermostable DNA polymerase catalysis (maximal catalytic activity at 75 °C to 80°C) of single-stranded template-directed DNA synthesis from nucleotide triphosphates. Genomic DNA serves as the template, two primers (one complementary to each of the strands) each having a free 3' hydroxyl are required to initiate synthesis, and magnesium ion is necessary for the enzyme to successfully catalyze the reaction. A number of thermostable DNA polymerases are available for the experimenter interested in PCR, the most commonly used one being *Taq* DNA polymerase.

Taq DNA polymerase is a thermostable enzyme from the thermophilic eubacterium *Thermus aquaticus*. It is relatively cheap and is therefore the enzyme of choice for routine and control experiments. However, one of the most discussed characteristics of thermostable polymerases is their error rate (mutation frequency per base pair per duplication). As would be expected, polymerases lacking 3'-5' exonuclease activity, such as *Taq*, generally have higher error rates than the polymerases with exonuclease activity otherwise known as proofreading DNA polymerases. TABLE 2.2 lists some different thermostable DNA polymerases and their associated error rates.

TABLE 2.2: Thermostable DNA Polymerases and Associated Error Rates

Thermostable DNA Polymerase	Source Organism	Error Rate(s) (Base substitutions/bp)
Taq	*Thermus aquaticus*	1.1×10^{-4} (Tindall and Kunkel, 1988) 2.1×10^{-4}(Keohavang and Thilly, 1989) 7.2×10^{-5}(Ling et al., 1991) 8.9×10^{-5}(Cariello et al., 1991) 2.0×10^{-5}(Lundberg et al., 1991) 1.1×10^{-4} (Barnes, 1992)
Vent	*Thermococcus litoralis*	2.4×10^{-5}(Cariello et al., 1991) 4.5×10^{-5}(Ling et al., 1991) 5.7×10^{-5} (Matilla et al., 1991)
Replinase	*Thermus flavis*	1.03×10^{-4} (Matilla et al., 1991)
Pfu	*Pyrococcus furiosus*	1.6×10^{-6} (Lundberg et al., 1991)

Proofreading or high fidelity DNA polymerases are enzymes with a very low rate of nucleotide misincorporation. The use of high fidelity DNA polymerases in the polymerase chain reaction (PCR) is essential for reducing the introduction of amplification errors in PCR products that will be cloned, sequenced and expressed.

As *Pfu* polymerase purportedly has the lowest profiled average error rate of any other DNA polymerase (Cline et al., 1996), it so became the thermostable DNA polymerase of choice for the amplification of both the YagI and YiaJ open reading frames. In addition to its high fidelity, *Pfu* polymerase is also one of the most thermostable DNA polymerases known as it can tolerate temperatures exceeding 95°C, enabling it to PCR amplify GC-rich targets (Cline et al., 1996).

One hundred microlitre reactions included: 32 pmoles of each primer, 1 μg of genomic template DNA, 250 nmoles of each dNTP, 10 uL of 10x *Pfu* Optimized buffer (Stratagene) 2.5 units of *Pfu* Polymerase and double deionized water (ddH_20) to 100 uL. Samples for PCR amplification were then placed into an ML-100 TM Programmable Thermal Cycler (MJ Research Inc.) using the cycling parameters outlined in TABLE 2.3.

TABLE 2.3: Cycling Parameters for the Polymerase Chain Reaction

Step	Number of Cycles	Temperature (°C)	Duration
1). Initial Denaturation	1	94	3 min.
2). A. Denaturation B. Annealing C. Extension	25	94 55 72	2 min. 1 min. 1 min.
3). Final Extension	1	72	10 min.
4). Finish	1	4	INDEFINITE

Following completion of the PCR amplification, products were analyzed for size and purity using 1% agarose gels.

2.11 Subcloning of DNA Fragments: Vector-Insert Preparation

Subcloning of DNA restriction fragments utilizes most of the basic techniques required for the cloning of DNA: restriction of DNA, phenol extraction to remove enzymes and protein contaminants, ethanol precipitation to concentrate DNA, ligation of DNA fragments, and the introduction of DNA into *E. coli*.

Restriction Endonucleases

A variety of restriction endonucleases was used for restriction mapping as well as ligation. All restriction endonucleases, obtained through GIBCO BRL Lifetechnologies, were supplied with a buffer (REact buffers) designed to optimize the cleavage efficiency of that enzyme. For a single digest, the typical reaction included DNA, 2-10 units of enzyme, enzyme buffer, 0.1mg/mL acetylated BSA and ddH_2O to a final volume of 20 µL (Maniatis et al). Where double digests (cleavage by two different restriction endonucleases simultaneously) or cleavage with the same enzyme was required twice, the components were pro-rated to a reaction volume of 40 µL. Since glycerol concentrations over 5% in the final reaction can sometimes inhibit enzyme activity, enzyme volumes larger than 1/10 of the reaction volume were avoided. All digests were performed at

37°C for 1 hour. Where recovery of DNA was required for further enzymatic manipulation(s), DNA was purified and recovered by phenol extraction and ethanol precipitation; otherwise, the digests would be electrophoresed in 1% agarose gels, stained with EtBr and viewed with UV light.

Phenol Extraction

Phenol is typically used in molecular biology applications to denature and remove proteins from nucleic acid solutions. An equal volume of TE-buffered Phenol/Chloroform/Isoamyl Alcohol (24:24:1) is added to the sample to be deproteinated then shaken or vortexed to mix. The resulting mixture is centrifuged for 2 minutes at high velocity to separate the organic and the aqueous phases. The denatured proteins may be visible as a milky white interface layer between the two phases. The aqueous phase (which contains the DNA) is carefully removed and transferred into a fresh tube. The extraction of the aqueous phase is repeated until there is no longer any visible material at the interface. The aqueous phase is then twice extracted with an equal volume of chloroform containing 4% isoamyl alcohol (this step helps remove residual protein, as well as phenol, which might remain from the phenol extraction). The aqueous phase is then recovered after centrifugation and ethanol is used to precipitate the DNA.

Ethanol Precipitation of Nucleic Acids

Ethanol precipitation is an easy way to concentrate nucleic acids, or to transfer them into a new set of buffer and salts. To an aqueous solution containing DNA, 0.1 volumes of 3M NaOAc, pH 4.8 is added and mixed well. Two volumes of absolute ethanol, which has been chilled to -20°C, are then added and mixed. The ethanol precipitation is then incubated at -20°C for 6 hours or at -70°C for 30 min, or in a dry-ice

ethanol bath or liquid nitrogen for 5 to 10 minutes. The tube containing precipitated nucleic acid is then centrifuged for 10 minutes in a microfuge. Following centrifugation, all of the supernatant is carefully removed and ice cold 70% ethanol is added to the tube, which is then centrifuged again. The pellet is allowed to air dry after using a capillary pipette to remove all traces of moisture. The pellet now contains relatively salt-free DNA, which may then be solubilized in a buffer of choice (usually TE).

Vector Preparation

Two micrograms of each vector were doubly digested with the proper restriction endonucleases (*Nde*1 and *Xho*1 for pET-24b+; *Nco*1 and *Hind*3 for pET-28b+). Following restriction endonuclease digestion, a phenol/chloroform extraction was performed to inactivate and remove the enzymes. The digested vectors were then recovered by ethanol precipitation. In order to ensure successful ligation and cloning, the digested vectors were then subjected to alkaline phosphatase treatment.

Alkaline Phosphatase Treatment

A common difficulty in achieving successful cloning, is the recircularization of the vector during ligation without insertion of any cloned DNA. However, an excellent preventative measure is to treat the restriction endonuclease digested vector with a phosphatase, which removes the 5' phosphate group from the vector. Since ligation reactions have an absolute requirement for the presence of a 5' phosphate, this treatment effectively blocks the religation of the vector, thereby reducing the background of non-recombinants, as the required phosphate group can only come from the DNA insert.

The terminal 5' phosphates can be removed from DNA by treatment with bacterial alkaline phosphatase (BAP), shrimp alkaline phosphatase (SIP), or calf

intestinal alkaline phosphatase (CIAP). Both SIP and CIAP have the considerable advantage in that following the phosphatase reaction, they can be heat inactivated. BAP, on the other hand, is heat-resistant. For most purposes, therefore, SIP and CIAP are the preferred enzymes. CIAP was the enzyme of choice for this study.

Although many workers claim to obtain excellent results simply by adding alkaline phosphatase to restriction enzyme digestions during the digestion, the buffer can hardly be at an optimum in this case. However, apparently enough dephosphorylation occurs to greatly reduce the recircularization problem. Nevertheless, following restriction endonuclease digestion, a phenol/chloroform extraction was performed (rather than heat inactivation) and the recovered DNA following ethanol precipitation was redissolved in the optimal buffer supplied with CIAP (GIBCO BRL).

Insert Preparation: Synthetic *yag*I & *yia*J PCR Products

Following successful PCR amplification of *yag*I and *yia*J, both products were purified using a QIAgen PCR purification kit and the concentrations of each were determined spectrophotometrically. Two micrograms of each *yag*I and *yia*J were digested with the proper restriction endonucleases. *yag*I was designed to be cloned into pET-24b(+), hence it was so digested with the same restriction endonucleases that were used to cut this vector i.e. *Nde*1 and *Xho*1. As pET-28b(+) was cut with *Nco*1 and *Hind*3 restriction endonucleases, so to was the *yia*J PCR product. Following digestion, both of the digested inserts were subjected to phenol/chloroform extractions and ethanol precipitation.

Ligation

The ligation of the insert into the linearized vector involves the formation of phosphodiester bonds between adjacent 5'-phosphate and 3'-hydroxyl residues. This reaction can be catalyzed by two different ligases: *E. coli* DNA ligase and bacteriophage T4 DNA ligase (ability to join blunt-ended DNA fragments as well as sticky ends). T4 DNA ligase from Gibco BRL was the enzyme of choice used in this study.

If the concentration of insert DNA is substantially lower than that of the vector, the ligation efficiency becomes very low. Hence, in order to obtain a high yield of the correct recombinants, ligation reactions with a molar ratio of insert to vector DNA of 3:1 were used. However, higher yields of the correct recombinant were expected as both the vector and insert had been prepared using two different restriction enzymes. The ligation mixture consisted of ligase reaction buffer, 30 fmol of vector DNA and 90 fmol of insert DNA, T4 DNA ligase and autoclaved ddH_2O to a final volume of 20 µL. The ligation reactions were performed at room temperature for 5 minutes in 0.5 mL microfuge tubes, where the contents were mixed gently and then centrifuged briefly to bring the contents to the bottom of the tube. Following the 5-minute incubation, 1, 2, 5 and 10 µL aliquots of the ligation mixture were used to transform 100 µL of competent Novablue cells.

2.12 Initial Cloning and Expression

Recombinant Isolation & Purification

Transformed Novablue cells were spread plated onto LB-Tetracycline-Kanamycin agar plates. A single recombinant colony was then streaked onto a fresh LB-Tetracycline-Kanamycin agar plate to produce multiple isolates of the same clone. A 100

mL plasmid preparation was then used to generate stock DNA for analysis via restriction mapping and automated DNA sequencing.

Verification of DNA Sequence

In order to ascertain the integrity of the recombinants pET-*yag*I (pET-24b+/*yag*I) and pET-*yia*J (pET-28b+/*yia*J), but more importantly the inserts, restiction mapping followed by sequencing at an automated sequencing facility was used to validate our efforts.

Restriction Mapping

Using the software programs MacVector 4.0 and Gene Runner v.3.0, theoretical restriction maps were generated for both pET-*yag*I and pET-*yia*J. Proper restriction endonuclease digests, both single and double digests, were then selected and prepared. Following enzymatic digestion, the samples were loaded into 1% agarose gels alongside molecular weight markers, cut vector, and insert standards. Following electrophoresis, results were observed under UV light, documented and compared against the theoretical restriction maps. Upon successful restriction mapping, samples of ultra clean DNA were sent away to automated-sequencing facilities.

Automated DNA Sequencing

In order for successful DNA sequencing, the DNA samples must be extremely pure. Although pET-*yag*I and pET-*yia*J DNA obtained from the alkaline lysis plasmid preparations was relatively pure as evidenced by approximate ratios of 1.8 (A_{260}/A_{280}) and 2.2 (A_{260}/A_{230}), QIAgen filter columns were used as an extra measure to ensure the DNA was satisfactory for the process. As well, DNA from plasmid preparations was suspended in TE where the automated sequencing services required DNA to be dissolved

in ddH_2O (DNA bound to the filter columns was eluted using ddH_2O). All sequencing was performed using an ABI PRISM™ 310 Gene Analyzer.

pET-*yagI*

Six samples of pET-*yagI* with accompanying sequencing primers (see Table 3.2 above) were used to sequence the *yagI* insert. As each run only sequences approximately 250 bases (with only 100-200 bases of reliable sequence), the sequencing primers were designed to overlap such that the entire range of the 750 plus-sized *yagI* insert could be sequenced. The samples were prepared in 200 μL thin-walled PCR tubes and included 300 ng of double-stranded pET-*yagI* DNA, 5 pmol of primer and ddH_2O to a total volume of 3 μL.

pET-*yiaJ*

A change in automated sequencing facilities allowed for a more convenient and economical sequencing service. In particular, we were only required to supply the DNA to be sequenced as the facility had their own supply of sequencing primers including both the T7 Promoter and Terminator primers (see Table 3.2 above). In addition, the sequencing at this facility also allowed a longer stretch of DNA to be sequenced (>500 bases).

The pET vectors contain regions of DNA that are referred to as the T7 Promoter primer, 5'-TAATACGACTCACTATAGGG-3' (includes the entire T7 promoter) and Terminator primer, 5'-CCGCTGAGCAATAACTAGC-3' (immediately following the T7 terminator) regions (see figures 3.5 and 3.6 above). Hence with the two above-mentioned sequencing primers, the entire length of an insert may be sequenced bi-directionally.

Expression of Recombinant DNA

Once the integrity of the pET-*yag*I and pET-*yia*J recombinants had been verified, competent BL21(DE3)pLysS cells were then transformed with the recombinant DNA and transformants were isolated on LB-chloramphenicol-kanamycin agar plates. Upon establishing the recombinant plasmids in BL21(DE3)pLysS, expression of target DNA followed thereafter.

One hundred millitres of LB-chloramphenicol-kanamycin was inoculated with a 3 mL overnight culture of transformed BL21(DE3)pLysS and incubated in a shaking incubator at 37°C to an OD_{600} of 0.6 – 0.8 (mid log phase). Two 1 L volumes of fresh LB-chloramphenicol-kanamycin broth were then inoculated using 40 mL aliquots of the 100 mL culture and incubated in the same manner until an OD_{600} of 0.6 – 0.8 was reached. Induction was then achieved by the addition of isopropyl-beta-D-thiogalactopyranoside (IPTG) to the growing cultures to a final concentration of 1 mM and allowing the cells to grow for a further 30 min. Following the 30 min. induction period, rifampicin (inhibits transcription by the host RNA polymerase) was added to the growing cultures to a final concentration of 200 μg/mL and the cultures were allowed to continue growth at 37°C in a shaking incubator.

Following three hours of growth after the addition of rifampicin, the cells were chilled on ice for 5 minutes and then harvested by centrifugation at 5000 rpm for 5 min. at 4°C using large collection bottles (250 mL volume) and a GSA rotor in a Sorval centrifuge. The cell pellets resulting from 1 L of culture were then washed by resuspension in 250 mL (0.25 culture volume) of ice-cold 20 mM Tris-HCl (pH 8.0), and centrifuged as before. The supernatant was then removed and the cell pellets were then

flash frozen using a liquid nitrogen bath. The frozen pellets were then thawed immediately or stored at -20°C until needed.

2.13 Isolation and Purification of Recombinant Protein

Using BL21(DE3)pLysS cells as a host expression strain provides a convenient method for extracting the target proteins following expression of the recombinant DNA. In particular, once the protein-rich cells are harvested and flash frozen in liquid nitrogen following expression, the cell membrane of the host cells is sheared from the freezing. Thawing the frozen cell mass then permits the T7 lyzozyme (from the pLysS plasmid) to act upon the cell walls releasing the contents of the cells without resorting to sonication or a French Press.

Frozen cell pellets were thawed using chilled 20 mM Tris-HCl, 50 mM KCl, 1 mM EDTA, 1mM dithiothreitol (DTT) pH 7.8. To reduce viscosity, genomic DNA was sheared by sequentially passing the ruptured cell solution through 16, 20 and 26-guage needles. The lysed cells were then centrifuged at 12K rpm for 60 min. to remove cell debris. The supernatant was collected and kept chilled on ice prior to purification via column chromatography.

BioREX-70™: Ion Exchange Chromatography

As the YagI and YiaJ proteins are hypotheical DNA-binding proteins from the IclR family of homologues, these proteins have a greater proportion of basic amino acid residues than acidic residues. Hence, purification using a cation exchange resin served as a first measure to purify the target proteins. In particular, the cation exchanger BioREX-70 (BioRAD), a weakly acidic cation exchange resin, was used to purify YagI and YiaJ proteins.

Analytical grade BioREX-70 resin contains carboxylic acid exchange groups on a macroreticular acrylic polymer lattice. This resin is a popular choice for the purification and fractionation of proteins and peptides due to its large capacity for high molecular weight molecules and chemical inertness towards neutral molecules and anions. The high porosity of BioREX-70 allows large protein molecules to penetrate the pores and interact with the numerous exchange sites located throughout the matrix.

Proteinaceous supernatant following centrifugation of lysed cells containing target protein was loaded directly onto a 2.5 x 15 cm BioREX-70 column (10 – 12g of resin) that had been equilibrated with 20 mM Tris-HCl, 50 mM KCl, 1 mM EDTA, 1mM DTT pH 7.8. Five hundred millilitres of the same buffer was then used to elute unbound proteins.

Proteins bound to the matrix were released using a linear salt gradient. An elution reservoir consisting of 500 mL of low salt buffer (20 mM Tris-HCl, 50 mM KCl, 1 mM EDTA, 1mM DTT pH 7.8) connected to 500 mL of high salt buffer (20 mM Tris-HCl, 0.5 M KCl, 1 mM EDTA, 1mM DTT pH 7.8) by a salt bridge was attached to the top of the BioREX-70 column. A fraction collector was set up to collect 150-drop fractions (approx. 7.5 mL) where the column, elution reservoir and fraction collector were all kept at 4°C in a cold room. Once the reservoir had been depleted, fractions were qualitatively tested for protein content using Bio-Rad Protein Assay Solution, a modified Bradford (Bradford, 1976) assay solution. In particular, a colorimetric test was performed by mixing 50 μL of fraction with 150 μL of a 1 in 5 dilution of Bio-Rad Protein Assay Solution. An immediate appearance of the color blue indicates the presence of protein

where the intensity of the blue is directly proportional to the concentration of protein present (i.e. deeper blue for higher protein concentration).

Fractions suspected of containing protein were double-checked by monitoring the OD_{280} to create an elution profile. Meanwhile, to determine the salt concentration required to elute the YagI and YiaJ proteins from the BioREX- 70 column, a calibration curve was generated from a set of standards (20 mM Tris-HCl, 1 mM EDTA, pH 7.8; 0 M KCl, 0.1 M KCl…0.5 M KCl) equilibrated at the same temperature as the fractions containing proteins using a conductivity meter.

SDS-PAGE was then used to identify protein content as well as purity in the fractions. Protein fractions containing target protein were then pooled together and concentrated via vacuum dialysis before loading on to a size-exclusion column.

Concentration of Protein by Vacuum Dialysis

Both YagI and YiaJ target proteins could only be concentrated via vacuum dialysis (concentration using an Amicon filtration chamber had been attempted, but proteins precipitated rapidly). The concentration apparatus involved a 250 mL side-arm flask connected to a vacuum line, 8 - 10 cm of 10 mm 12 –14K molecular weight limit dialysis tubing (Fischer Scientific), and rubber tubing. A glass connector penetrating a bore through rubber stopper connects the dialysis tubing to the rubber tubing line.

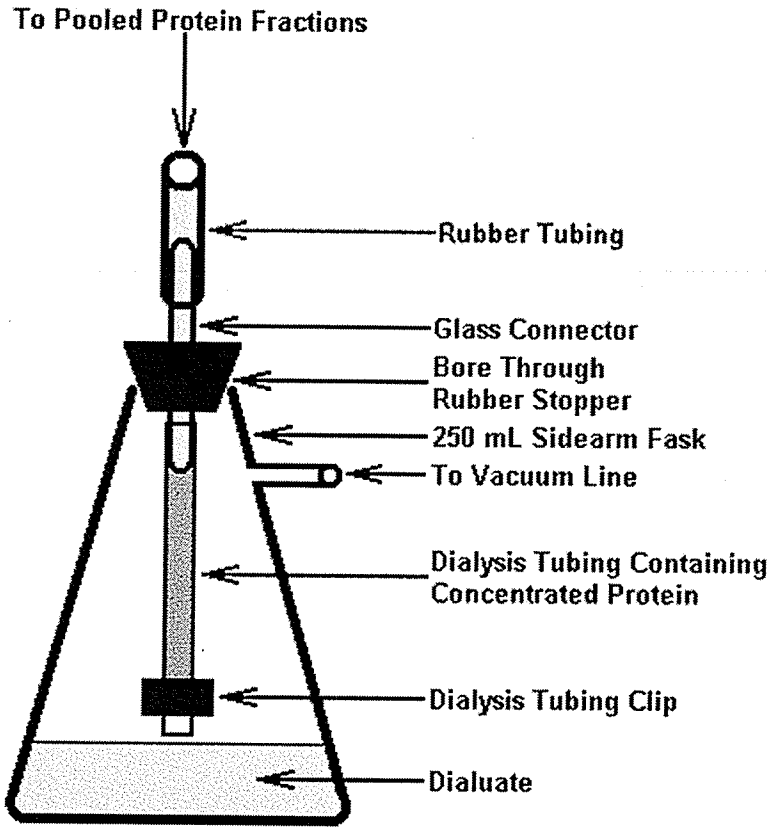


FIGURE 2.6: Concentration of Protein via Vacuum Dialysis. Target proteins were concentrated in a gentle fashion using a simple apparatus set up involving a 250 mL sidearm flask connected to a vacuum line, dialysis tubing and a line running to pooled protein fractions.

Sephadex G-100™: Size Exclusion Chromatography

As a further measure to purify target protein isolated from the BioREX-70 cation exchange column, size exclusion, or more specifically, gel filtration chromatography was employed.

Sephadex (Pharmacia Biotech) is an epichlorohydrin cross-linked dextran gel commonly used for the fractionation of macromolecules. The separation matrix consists of porous gel beads of different sizes with various pore sizes permitting the separation of ranges of low to high molecular weight compounds. We used Sephadex G-100, which has a fractionation range of M_r 1,000 to 100,000 Da.

Concentrated protein eluate from the BioREX-70 column was loaded onto a 4 x 110 cm Sephadex G-100 column with 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA, 1 mM DTT; pH 7.8 as the column buffer. A fraction collector was programmed to collect 200-drop fractions after which protein content was assessed via a Bradford assay and elution profile in the same manner previously described. Fractions containing protein were analyzed with SDS-PAGE and then pooled and concentrated via vacuum dialysis.

2.14 Verification of Protein Integrity

From SDS-PAGE, it certainly seemed that both the YagI and YiaJ proteins had been successfully isolated and purified using the combination of ion exchange and size exclusion chromatographic methods we selected. However, it remained necessary to determine whether the pET Expression System had been successful in generating accurate protein specimens. Fortunately, access to mass spectrometric methods allowed us to verify the integrity of the isolated YagI and YiaJ proteins.

ESI-TOF Mass Spectrometry

The technique of electrospray ionization (ESI) mass spectrometry introduces desolvated ions into a high vacuum environment from an atmospheric-pressure stream of droplets of polar molecules in a mixed aqueous/organic solvent. The stream of droplets is generated by passing the output of a syringe pump or the eluant of high performance liquid chromatography (HPLC) through a fine stainless steel tip held at a high voltage i.e. a fine spray of highly charged droplets is created in the presence of a strong electric field. Vaporization of these charged droplets results in the production of singly or multiply-charged ions in the gas phase. The number of charges retained by an analyte can depend on such factors as the composition and pH of the electrosprayed solvent as well as the

chemical nature of the sample. For small molecules (< 2000 Da), ESI typically generates singly or doubly charged ions; while for large molecules (> 2000 Da) the ESI process typically gives rise to a series of multiply charged species. Once ionization has taken place, a mass analyzer (e.g. ion trap, quadrupole, etc.) is used to generate a mass spectrum, but because mass spectrometers measure the mass-to-charge (m/z) ratio, the resultant ESI mass spectrum contains multiple peaks corresponding to the different charged states.

In the experiments described in this thesis, the mass analyzer used in this case was a time-of-flight (TOF) mass analyzer. Time-of-flight mass analysis, with respect to mass spectrometry, is based on accelerating a set of ions to a detector with the same amount of energy. In a TOF-based mass spectrometer, ions formed in the presence of an electrical field are accelerated from this source region into a longer field-free region (referred to as the drift region), into the mass spectrometer, and finally onto a suitable detector. Because the ions have the same energy yet different masses, they reach the detector at different times. Hence, the time the ions require to travel through the drift region to the detector (time of flight) is dependent solely on mass and the time spectrum is easily converted into a mass spectrum.

ESI allows for very sensitive analysis of small, large and labile molecules such as peptides, proteins, organometallics, oligosaccharides, and polymers directly from solution. Unfortunately, ESI-MS is limited by the need for relatively high purity samples and by a poor tolerance for electrolytes and detergents. Hence, it was necessary to remove the buffer components of the aqueous YagI and YiaJ protein samples in order to analyze the samples using the ESI-MS method.

In order to remove the buffer components of the YagI and YiaJ aqueous protein solutions, 2 – 3 nmoles (100 – 150 μL volume) of purified protein samples were placed in waterbugs (Orr et al.,1995) and dialyzed against several changes of Nanopure™ water. Precipitated protein was then resolubilized upon acidifying with acetic acid to a final concentration of 5%. Denatured protein samples were diluted to a 20 μM monomer concentration and methanol was added to a concentration of 40%. Mass spectral analysis was then performed using ESI-TOF MS.

2.15 Determination of Molar Absorptivity

In order to utilize the Beer-Lambert law i.e. $A = \varepsilon Cl$ for determining the concentration of a protein solution, the molar absorption coefficient, ε, of the protein must be known. The molar extinction coefficients (molar absorptivity) of the YagI and YiaJ proteins were determined using the Edelhoch method (Edelhoch, 1967; Gill and von Hippel, 1989).

The absorbance of most protein solutions above 275 nm depends on just three chromophores: the R-groups of tryptophan and tyrosine residues and the disulfide bonds of cystine (Wetlaufer, 1962). The Edelhoch method is a spectrophotometric technique that measures the change in absorbance (above 275 nm) of a protein solution resulting from the ionization of the R-groups of tyrosine residues at different pH. It is convenient and accurate, and the most reliable experimental method for determining ε (Pace et al, 1995). The basic assumption of the Edelhoch method is that ε values determined for model compounds for tryptophan, tyrosine, and cystine in 6 M guanidine hydrochloride (GdnHCl) can be used to approximate the ε values for the same chromophores in a protein unfolded in 6 M GdnHCl. However, as the YagI and YiaJ proteins were prepared

with DTT present throughout the course of preparation, the absorbance due to cystine was ignored. In contrast to cystine disulfide bonds, individual cysteine residues do not contribute significantly to the absorbance above 275 nm (Bailey, 1968). In addition, cystine absorbance is only a small contribution in comparison to tyrosine and tryptophan.

TABLE 2.4: Molar Extinction Coefficients for Protein Chromophores (Edelhoch, 1967)

Chromophore	Absorption Wavelength			
	280	288	295	300
Tyrosine	1280 $M^{-1}cm^{-1}$	385 $M^{-1}cm^{-1}$	2480 $M^{-1}cm^{-1}$	2270 $M^{-1}cm^{-1}$
Tryptophan	5690 $M^{-1}cm^{-1}$	4815 $M^{-1}cm^{-1}$		
Cystine	120 $M^{-1}cm^{-1}$	70 $M^{-1}cm^{-1}$		

We used 7.5 M GdnHCl (prepared in 20 mM Tris-HCl, pH 7.8) and 10 M NaOH to perform the Edelhoch experiment. Using dH_2O as a blank, the OD_{280} of 800 µL of 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA, 1 mM DTT; pH 7.8 (reference) was measured after which 200 µL of purified protein (YagI or YiaJ) was added, mixed and measured again. This diluted sample of protein was referred to as sample A. Next, the OD_{280} and OD_{288} of 800 µL of 7.5 M GdnHCl was measured after which 200 µL of purified protein (YagI or YiaJ) was added and mixed. The resulting solution was referred to as sample B and the absorbance of this solution was measured at 280, 288, 295 and 300 nm (the readings at 280 and 288 nm are used to calculate tryptophan precisely and tyrosine approximately). Finally, 10 µL of 10 M NaOH was added to solution B to increase the pH to greater than 12 pH units. This solution was referred to as sample C and the absorbance of this solution was measured at 295 and 300 nm (the readings at 295 and 300 nm after ionization of tyrosine residues will yield more exact values for tyrosine). The following equations were then used to calculate epsilon values for YagI and YiaJ proteins:

A). $\Delta OD_{280} = OD_{280}$(sample A) - OD_{280}(reference) = A ($A = \varepsilon cl$)

B). $\Delta OD_{295} = OD_{295}$(sample C) x 1.01 – OD_{295}(sample B)

C). $\Delta OD_{300} = OD_{300}$(sample C) x 1.01 – OD_{300}(sample B)

D). $[Tyr] = (\Delta OD_{295} + \Delta OD_{300})/(2480 + 2270)$

E). $[Trp]^{280} = [OD_{280} - [Tyr](1280)]/5690$

F). $[Trp]^{288} = [OD_{288} - [Tyr](385)]/4815$

G). $[Trp] = ([Trp]^{280} + [Trp]^{288})/2$

H). [protein] = [([Trp]/no. of Trp residues) + ([Tyr]/no. of Tyr residues)]/2 = c ($A = \varepsilon cl$)

I). $\varepsilon = A/cl$

2.16 Modification of Reactive Cysteine Residues

In proteins, thiol groups (also called mercaptans or sulfhydryls) are present in cysteine residues. Thiols can also be generated by selectively reducing cystine disulfides with reagents such as DTT or 2-mercaptoethanol. For peptides containing cysteine residues, two simple experiments allow one to determine whether or not one or more of these residues are available for reaction (i.e. not buried or involved in a disulfide bond). In particular, peptides containing free sulfhydryl residues can be quantitated with Ellman's reagent (Ellman, 1959), 5,5'-dithiobis(2-nitrobenzoic acid) (DTNB) and identified with 2-mercaptoethanol.

Reaction with DTNB

Ellman's reagent is an important reagent for quantitating thiols in proteins, cells and plasma by absorption measurements. The reaction of DTNB with sulfhydryl groups produces a protein with a modified cysteine and 5-mercapto-2-nitrobenzoic acid which

dissociates at above neutral pH liberating the yellow *p*-nitrothiophenol anion chromophore (absorption maximum 412 nm = 13,600 $cm^{-1}M^{-1}$).

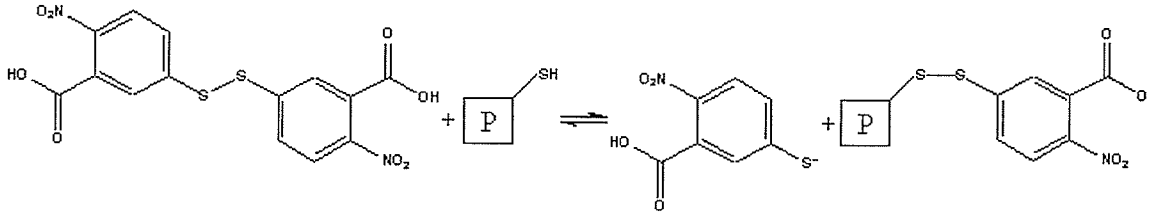


FIGURE 2.7: Reaction of a Protein with Ellman's Reagent. A protein containing one or more reactive cysteines undergoes a chemical modification when reacted with 5,5'-dinitrobis-(2-nitrobenzoic acid) otherwise known as DTNB or Ellman's reagent. The reaction yields a nitrophenol, which dissociates to a p-nitrothiophenolate species (a very strong chromophore).

The progress of this reaction may be monitored spectrophotometrically and providing that the molar concentration of protein is known, the number of moles of reactive sulfhydryl group per protein molecule may be determined (the reaction is stoichiometric).

To perform the experiment, a mixture of 100 µl of 1 mM DTT, 100 µl of KCl and 700 µL of 0.02 M Tris-HCl, 1 mM disodium EDTA, pH 7.8 was prepared in a 1 mL cuvette. To this mixture, 100 µL of stock protein solution (YagI or YiaJ) was added and using 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA, 1 mM DTT; pH 7.8 as a blank, the change in absorbance at 412 nm was monitored over a timed period of 2, 4…10, 15…60 min. A control reaction without the presence of DTNB was performed by mixing 20 µL 1 M KCl, 160 µL 0.02 M Tris-HCl, 1 mM disodium EDTA, pH 7.8 and 20 µL stock protein, diluting a 20 µL aliquot of this mixture into 980 µL of 0.02 M Tris-HCl, 1 mM disodium EDTA, pH 7.8 and repeating the measurements.

Reaction with 2-mercaptoethanol

The amino acid cysteine is modified in the presence of the reducing agent 2-mercaptoethanol as shown in FIGURE 2.8.

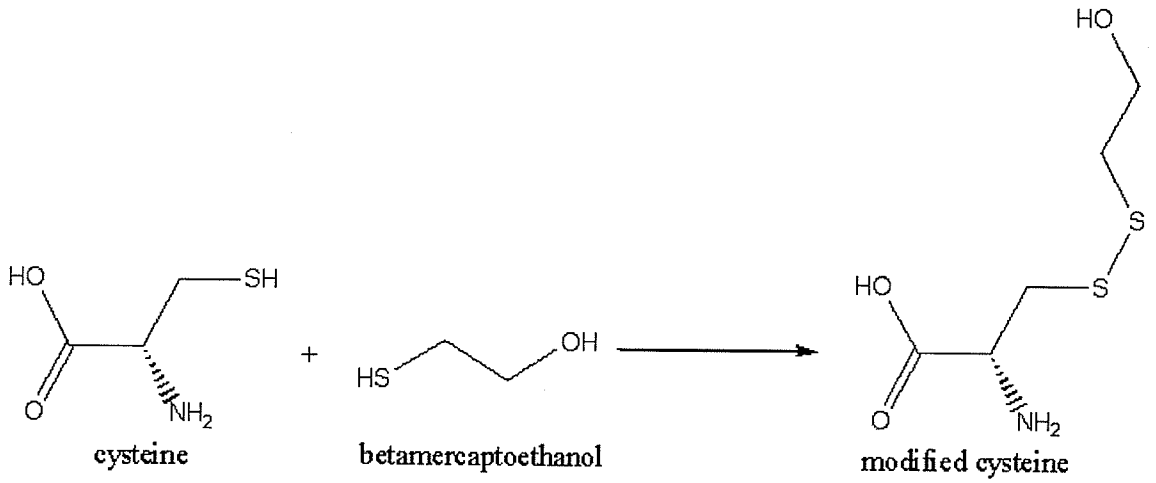


FIGURE 2.8: Modicifation of Cysteine with 2-mercaptoethanol. Cysteine undergoes chemical modification in the presence of the reducing agent, 2-mercaptoethanol.

Proteins containing one or more reactive cysteine residues will undergo this chemical modification allowing the positions of the modification to be determined. To allow the exchange to occur, 100 μL of pure protein at a concentration of 30 μM is incubated at room temperature for 1 hr. in the presence of 10 μL of 2-mercaptoethanol.

As a control, a second 100 μL aliquot of protein sample at the same concentration is reserved without having been incubated in the presence of 2-mercaptoethanol.

2.17 Identification of Reactive Cysteine Residues

Proteins treated with 2-mercaptoethanol will then contain one or more 2-Amino-3-(2-hydroxy-ethyldisulfanyl)-propionic acid moieties (modified cysteine). The identity of the reactive cysteine(s) can then be fingerprinted via mass spectrometry analysis. In particular, by comparing the tryptic digest of the protein sample containing the modified cysteine residue(s) against the tryptic digest of the control protein sample, a fragment

with 76 Da of extra mass would then identify the cysteine within said fragment and hence the identity of the reactive cysteine in the primary structure of the protein.

Before tryptic digestion can occur, protein samples (100 μL) are dialysed using waterbugs against a 1 L volume of 0.1 M NH_4HCO_3 over several hours (4hrs. at room temp. OR overnight at 4°C). Protein samples (100 μL) are then dialysed against 1 L of 10 mM NH_4HCO_3 in the same container over several hours and for a final time in a fresh 1 L volume of 10 mM NH_4HCO_3. Half of each sample (50 μL) is then reserved for tryptic digestion while the other half is subjected to preliminary ESI-MS analysis by adding acetic acid to 2% and methanol to 50% final concentration.

Enzymatic Trypsin Digestion

Trypsin is a proteolytic enzyme that cleaves after lysine and arginine residues. The resulting fragments, so-called tryptic fragments, comprise the protein, and represent the primary structure of a protein. In particular, each peptide fragment resulting from a proteolytic digestion with trypsin will have a characteristic mass.

Protein samples were digested using trypsin at a ratio of 1 μg trypsin/1 mg protein and incubating at 37 °C for several hours.

Identification of Reactive Cysteines via MALDI-MS

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) of peptide/protein involves mixing samples for analysis with a large molar excess (1:100 to 1:1000 molar ratio) of a matrix solution, which consists of weak organic acids capable of absorbing laser radiation. This mixture is then applied to a sample disk, which is most commonly made of stainless steel, and the matrix and analyte are allowed to co-crystallize. Once co-crystallization has occurred, the MALDI disk is introduced into the

mass spectrometer and irradiated under high vacuum using either a UV or an IR laser source. With each laser shot, the matrix material absorbs the incident radiation and undergoes a photo-ionization event, which results in a thermal plume of excited matrix and analyte molecules/ions. In this plume, ion-molecule reactions in the gas phase involving proton transfer occur between matrix and analyte. Hence, the matrix serves as a proton donor, acting to ionize the analyte in positive ionization mode. In MALDI-TOF MS, the ions are then accelerated via a strong electrical field and travel through a drift region toward a detector plate. The interval from firing the laser until the resulting ions strike the time-of-flight detector is a function of mass to charge ratio for each given peptide/protein-derived species.

MALDI-TOF MS has high sensitivity, covers a broad mass range, and unlike its ESI counterpart, shows tolerance to millimolar salt concentrations and suitability for the analysis of relatively complex mixtures. Hence, MALDI-TOF MS offers a sensitive, accurate and rapid procedure for determination of peptide/protein mass. However, one limitation of MALDI-TOF MS is that the measured signal intensity is not linear with the quantity of introduced sample, and not all components are ionized with the same efficiency, and therefore the method is not applicable for determining relative concentrations of peptides/proteins in a mixture.

When the sample is an enzymatic digest, it is frequently possible to identify an unknown protein constituent by computerized comparison of the peptide masses against a database of predicted fragments from all known protein sequences. In the case of a modification of a reactive cysteine with 2-mercaptoethanol, the result would be a mass increase of 76 mass units and hence any tryptic fragment containing such a modified

cysteine would be recognized in this way. In order to identify the exact location of the chemical modification in the primary structure of the protein, the peptide fragment containing the chemical modification must be sequenced using tandem mass spectrometry otherwise known as MS/MS (Busch et al, 1988; Biemann, 1988 & 1992; Papayannopoulos, 1995).

Tandem mass spectrometry is accomplished using two stages of mass analysis. In the first, an ion of interest is preselected with a mass analyzer. This is subsequently fragmented, and the second mass analyzer is used to obtain a fragment ion spectrum. Fragmentation is induced, for instance, by collision with an inert gas such as argon or helium via a process known as collision-induced dissociation (CID; also known as also known as collision-activated dissociation (CAD)) in a region of the mass spectrometer referred to as the collision cell. A popular choice for the mass analyzers used in tandem mass spectrometry is the quadrupole which consists of four precisely parallel rods with a direct current (DC) voltage and a superimposed radio-frequency (RF) potential, such that by simply scanning a pre-selected radio-frequency field, one effectively scans a mass range.

Drs. Werner Ens and Kenneth Standing have developed a revolutionary mass spectrometer at the University of Manitoba referred to as the matrix assisted laser desorption/ionization quadrupole quadrupole time-of-flight mass spectrometer (MALDI-QQTOFMS; Loboda et al., 2000). Fortunately for our research group, we were able utilize this novel approach to analyze the tryptic digests of protein samples and subsequently identify modified cysteine residues (Sadler, 2000).

The MALDI-QQTOFMS was used to identify reactive cysteine residues by co-crystallizing 0.6 μL of trypsin-digested protein samples with an equal volume of 2,5-dihydroxybenzoic acid (DHB) matrix on a stainless steel probe.

2.18 Secondary Structural Predictions

Using the software program PepTool version 2.0 (BioTools Incorporated), secondary structural predictions were made for *E. coli* IclR, YagI and YiaJ proteins as well as for *T. maritima* IclR. PepTool uses an algorithm of structural prediction based on existing information of known folding motifs in combination with the cumulative results of the following prediction algorithms.

1. Motif-Based Secondary Structure Prediction

This procedure predicts secondary structure based on primary sequence patterns contained in the information databases within the software program. It is an extension of the methods first proposed by Rooman and Wodak (Rooman & Wodak,1988; Rooman et al, 1990) for identifying and incorporating well established sequence/structure patterns in secondary structure prediction schemes (BioTools Inc.).

2. Homology-Based Secondary Structure Prediction

This procedure searches for short stretches of homologous sequences within the protein of interest and compares them to known protein structures. This algorithm is based on a number of related proposals simultaneously offered by several authors (Nishikawa & Ooi, 1986; Sweet, 1986; Levin & Garnier, 1986).

3. Hydrophobic-Moment-Based Secondary Structure Prediction

This procedure determines the secondary structure for any given protein sequence on the basis of hydrophobic periodicities. It has its origins with the Fourier analysis of

hydrophobicity profiles as first proposed by Eisenberg et al in 1984 (Eisenberg et al, 1984). According to Eisenberg, stretches of residues with hydrophobic periodicities in the range of 90 to 120 degrees (corresponding to a hydrophobic residue every three to four residues) are typically found in alpha-helices, while stretches of amino acids with hydrophobic periodicities of 160 to 180 degrees (corresponding to alternating hydrophobic and hydrophilic residues) are typically found in beta strands. In contrast to the statistical techniques of Chou and Fasman (Chou & Fasman, 1974, 1978), it is an approach that is based on well-established physico-chemical principles (BioTools Inc.).

4. GOR Secondary Structure Prediction

Commonly called the GOR method (after the initials of authors Garnier, Osguthorpe, and Robson), this procedure predicts secondary structure on the basis of parameters obtained from information theory. It is based on a series of proposals originally put forward by these investigators in the 1970s (Garnier et al, 1978). It is very much a statistical technique, not unlike the Chou-Fasman approach, except that it takes into account the positional preferences of amino acids within helices, beta strands and coils (BioTools Inc.).

5. Chou and Fasman Secondary Structure Prediction

This procedure uses a modification of the original Chou and Fasman algorithm to predict secondary structure for protein sequences. The Chou-Fasman algorithm is based on statistically observed propensities of all 20 amino acids to occur in various protein secondary structures. For instance, valine and isoleucine have high beta-strand probabilities, alanine and glutamic acid have high helical probabilities, and glycine and proline tend to appear in random coil structures (BioTools Inc.).

PepTool then determines a consensus secondary structure for the protein sequence of interest based on the culmination of scores from the homology, hydrophobic moment, GOR, Chou-Fasman and Motif-based structural predictions. The residue specific scores for each method are weighted according to its expected prediction accuracy where the homology-based technique has the strongest weighting and the Chou-Fasman technique has the weakest.

The Protein Explorer Interface

Protein Explorer (http://www.umass.edu/microbio/chime/explorer/) is a widely used freeware program for visualizing the three-dimensional structures of protein, DNA and RNA macromolecules, and their interactions and binding of ligands, inhibitors, and drugs. Protein Explorer offers the same features and benefits of RasMol(Mac), but it is a more user-friendly software interface application than its predecessor. Using Protein Explorer and any Brookhaven National Laboratories (Bernstein et al., 1977) Protein Data Bank (PDB; http://www.rcsb.org/pdb/; Berman et al., 2000) file, one can perform many tasks and operations for studying the molecular structure of biological macromolecular crystal structures in greater detail.

As previously mentioned, the only IclR family member that has been successfully analyzed by x-ray crystallography to date is the *Thermotoga maritima* IclR protein (TM0065; Zhang et al., 2002). Hence, the pdb file containing the crystallographic data of this protein was obtained and analyzed for structural comparisons with *E. coli* YagI, YiaJ and IclR proteins.

Chapter III:

Results

3.0 PCR Amplification of *yag*I and *yia*J

The polymerase chain reaction was used to generate large amounts of synthetic *yag*I and *yia*J DNA for cloning purposes. The thermostable DNA polymerase *Pfu,* which has a very low eror rate, was used to minimize the frequency of possible mutation, and the PCR primer pairs used for the amplification of each ORF were designed such that the PCR products could be easily cloned into a pET expression vector. As a result, the amplification reaction the PCR products would not only contain the entire length of the ORF, but two unique restriction sites with exogenous DNA flanking the start and stop codons (to accommodate restriction endonuclease digestion).

***yag*I**

Natural *E. coli yag*I is a 759 bp ORF found in section 25/400 of the *E. coli* K-12 MG1655 genome. Using the two PCR primers (5' – 3'): CTCGATAGATCATATGCCGATTATTCAGTCTGTTGAACG and CTGCAGGACTCGAGCCGGTCTAACGGCTCAGG, a 799 bp synthetic PCR product was generated (5'CTCGATAGATCATATG...TGAGCCGTTAGACCGGCTCGAGTCC TGCAG 3'; sequence between ATG and TGA represents complete *yag*I amplification).

This 799 bp PCR product contains an *Nde*1 restriction endonuclease recognition site (CATATG) at the natural initiation codon and an *Xho*1 restriction endonuclease recognition site (CTCGAG) downstream of the natural termination codon of the *yag*I ORF.

Following amplification, 5% of the amplification reaction (5 μL) was examined on a 1% agarose gel. A single, intense band in the correct size range (~ 800 bp) indicated that the amplification was indeed successful.

***yia*J**

Natural *E. coli yia*J is an 849 bp ORF found in section 325/400 of the *E. coli* K-12 MG1655 genome. With the following pair of PCR primers (5'-3'), CAGGACGCGACACCATGGGGAAAGAAG and CACTTGTCCAAAGCTTTATGTGATCGCG, an 878 bp synthetic PCR product was generated (5' CAGGACGCGACACC**ATG**...**TAA**AGCTTTGGACAAGTG 3'; sequence between **ATG** and **TAA** represents complete *yia*J amplification).

This 878 bp PCR product contains an *Nco*1 restriction endonuclease recognition site (CCATGG) at the natural initiation codon and a *Hind*3 restriction endonuclease recognition site (AAGCTT) at the natural termination codon of the *yia*J ORF.

The amplification reaction was also deemed successful as a single and intense band of the correct size range (~900 bp) was observed following 1% agarose gel electrophoresis.

A).
5'ggacatgccagcatt**ATG**CCGATTATTCAGTCTGTTGAACGTGCGTTGCAGATCCTC
GACCTGTTCAACGAGCAGGCCACCGAGCTTAAGATCACCGACATCAGCAAAC
TGATGGGGCTGAGCAAGAGTACCCTCCACTCGCTGCTAAAAACCCTGCAGCT
TCACGGCTATATCGATCAGAACCCGGAGAACGGCAAGTATCGCCTCGGCATG
AAGCTGGTCGAGCGCGGCCATTTTGTCGTGGGCTCCATCGATATTCGGCAGA
AGGCAAAAGGCTGGCTGACGGAGCTGTCCCGGCGGACCGGGCAGACCACCC
ATCTGGGGATCCTGGACGGGCGTGAAGGGGTCTATATCGAGAAGATTGAAGG
CAAGCTGGCCGCCATCGCCTATTCACGCATCGGCCGCCGCCTGCCGGTGCAC
GCCACCGCCATCGGCAAGGTGTTGATTGCCTGGCTGGGCGAGGCCGAGCTGA
ACGCCCTGCTGGAGGGCTATCAGTACACTACCTTTACGCCCGCCACCCTCGCG
TCTCGCGAAGCCTTAATGAGCGCCCTGGCGCAGACCCGCGAGCAAGGCTACG
CCCTGGACAGCGAAGAGAACGAGCAGGGCGTGCGCTGCGTGGCGGTGCCGG
TGTGGAACCACGAGTCCCGCGTCATCGCCGCCCTGAGCCTGTCGACGCTGAC
CTCCCGCGTGGACGACGCGGAGCTGGCTAATTTCCGCGAGCAGCTTCAGCAG
GCCGGGCTCGCGCTCTCGCGCGCGCTGGGCTACCCGGCC**TGA**gccgttagaccggctc
gtaggtgaagtagtcg3'
B).
5'ctaccatcaggacgcgacaaa**ATG**GGGAAAGAAGTGATGGGAAAAAAAGAGAACGAG
ATGGCGCAGGAAAAAGAGCGTCCAGCCGGAAGCCAGAGTCTGTTTCGCGGGT
TGATGCTGATTGAGATTTTGAGCAACTATCCAAACGGTTGTCCGTTGGCGCAT
CTTTCGGAGCTGGCTGGTTTAAATAAGAGTACCGTCCATCGCTTATTGCAGGG
ATTACAGTCCTGCGGCTATGTGACCACCGCGCCTGCCGCAGGGAGTTATCGC
CTGACCACCAAATTTATTGCCGTCGGGCAGAAGGCGCTGTCTTCGCTGAATAT
CATTCATATCGCCGCTCCGCATCTTGAGGCACTGAACATCGCCACTGGTGAAA
CCATTAACTTCTCCAGCCGCGAAGACGATCACGCTATTTTGATTTATAAGCTG
GAACCCACAACCGGGATGCTGCGAACCCGTGCCTATATTGGCCAGCATATGC
CGCTCTACTGTTCCGCAATGGGCAAGATCTACATGGCGTTTGGTCACCCGGAC
TACGTGAAGTCATACTGGGAAAGCCATCAGCATGAGATCCAGCCGTTAACCC
GCAATACCATTACCGAGCTGCCCGCGATGTTCGACGAACTGGCGCACATTCG
TGAAAGCGGAGCGGCGATGGACAGAGAAGAAAACGAACTCGGCGTCTCCTG
TATTGCTGTTCCGGTGTTTGATATTCATGGGCGGGTGCCGTACGCCGTGTCGA
TTTCGCTTTCGACATCACGTCTGAAACAGGTGGGAGAGAAAAATCTCCTGAA
ACCACTGCGTGAAACCGCGCAGGCTATTTCTAATGAACTGGGATTTACTGTCA
GGGATGACCTGGGCGCGATCACA**TAA**cgctttggacaagtgccaaa3'

FIGURE 3.0: DNA Sequences of *yag*I and *yia*J Open Reading Frames. A). The *yagI* open reading frame from section 25/400 of the *E. coli* K-12 MG1655 genome B). The *yiaJ* open reading frame from section 325/400 of the *E. coli* K-12 MG1655 genome. Codon **ATG** is the methionine initiation codon, **TGA** & **TAA** indicate termination codons. Bases in lowercase represent flanking DNA sequences illustrating where the primers selected for the amplification of each ORF interact with genomic DNA.

3.1 Production of Recombinant Plasmids

The synthetic *yag*I and *yia*J PCR products were ligated into pET-24b(+) and pET-28b(+) vectors respectively. Competent Novablue cells were transformed with the recombinants, pET-*yag*I and pET-*yia*J and plated onto LB-tetracycline-kanamycin agar plates. The selection of recombinants was fairly stringent as the ligation was performed after using two different restriction endonuclease digestions on the inserts and vectors, alkaline phosphatase treatment on the vectors and two different antibiotics. Tetracycline resistance is attributed solely to the host cell (Novablue), but kanamycin resistance can only be achieved if an intact recombinant is present. Hence, the recombination appeared to be successful as several transformants were obtained. However, before protein expression could occur, the integrity of the recombinants needed to be verified. Recombinant DNA was isolated and purified using a standard alkaline lysis preparation of plasmid DNA. The DNA was then examined first by restriction mapping followed by DNA sequencing at an automated DNA sequencing facility.

Restriction Endonuclease Digestion Patterns

Theoretical restriction maps of each resulting recombinant were prepared using MacVector 4.0 and Gene Runner v. 3.0.

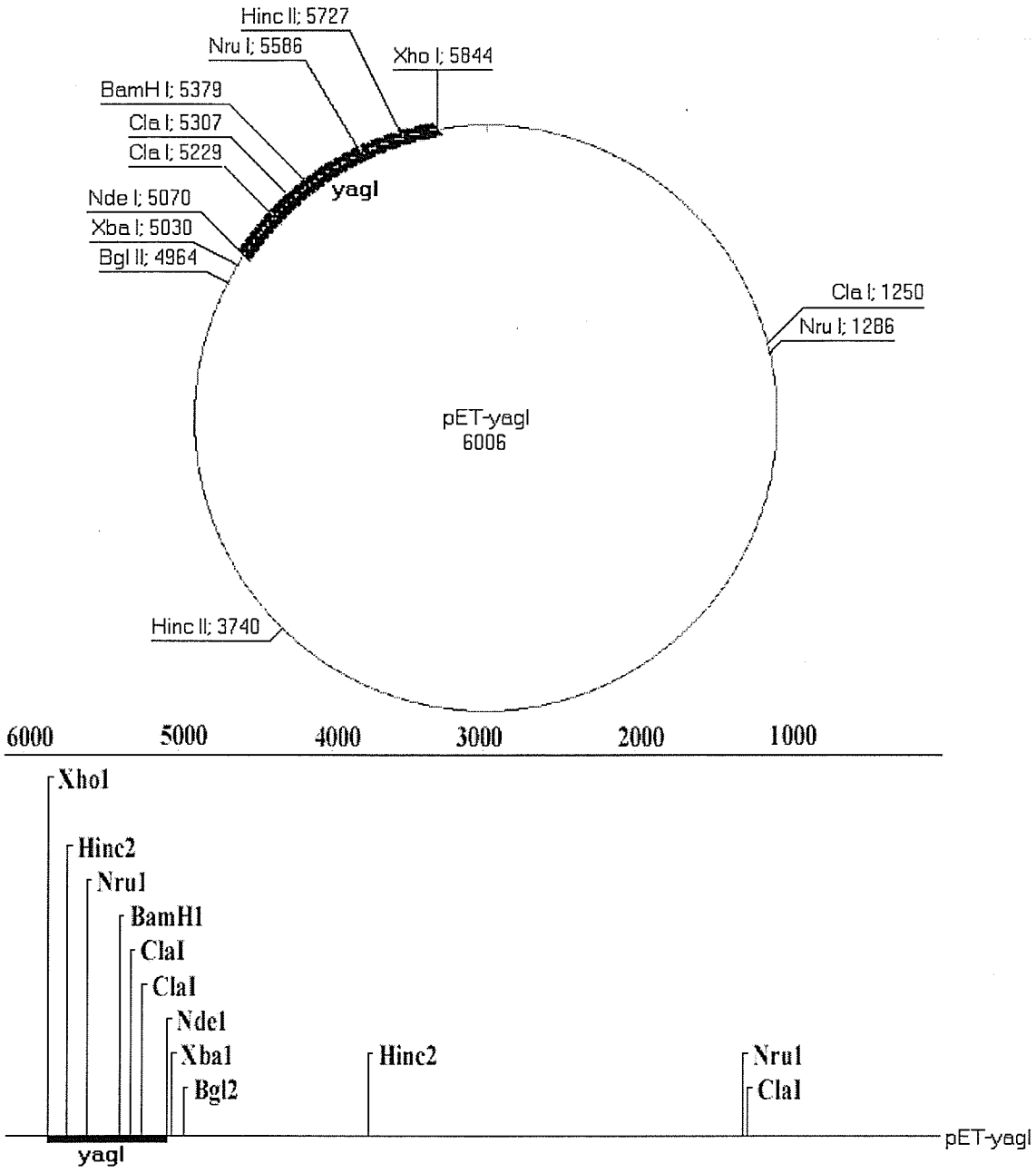


FIGURE 3.1: Theoretical Restriction Map of the pET-*yag*I Recombinant. By analyzing the 6006 bp pET-*yag*I sequence for *Bam*H1, *Bgl*2, *Cla*1, *Hinc*2, *Nde*1, *Nru*1, *Xba*1 and *Xho*1 restriction endonuclease recognition sites, a theoretical restriction map is generated so that a comparison can be made with the actual experimental results.

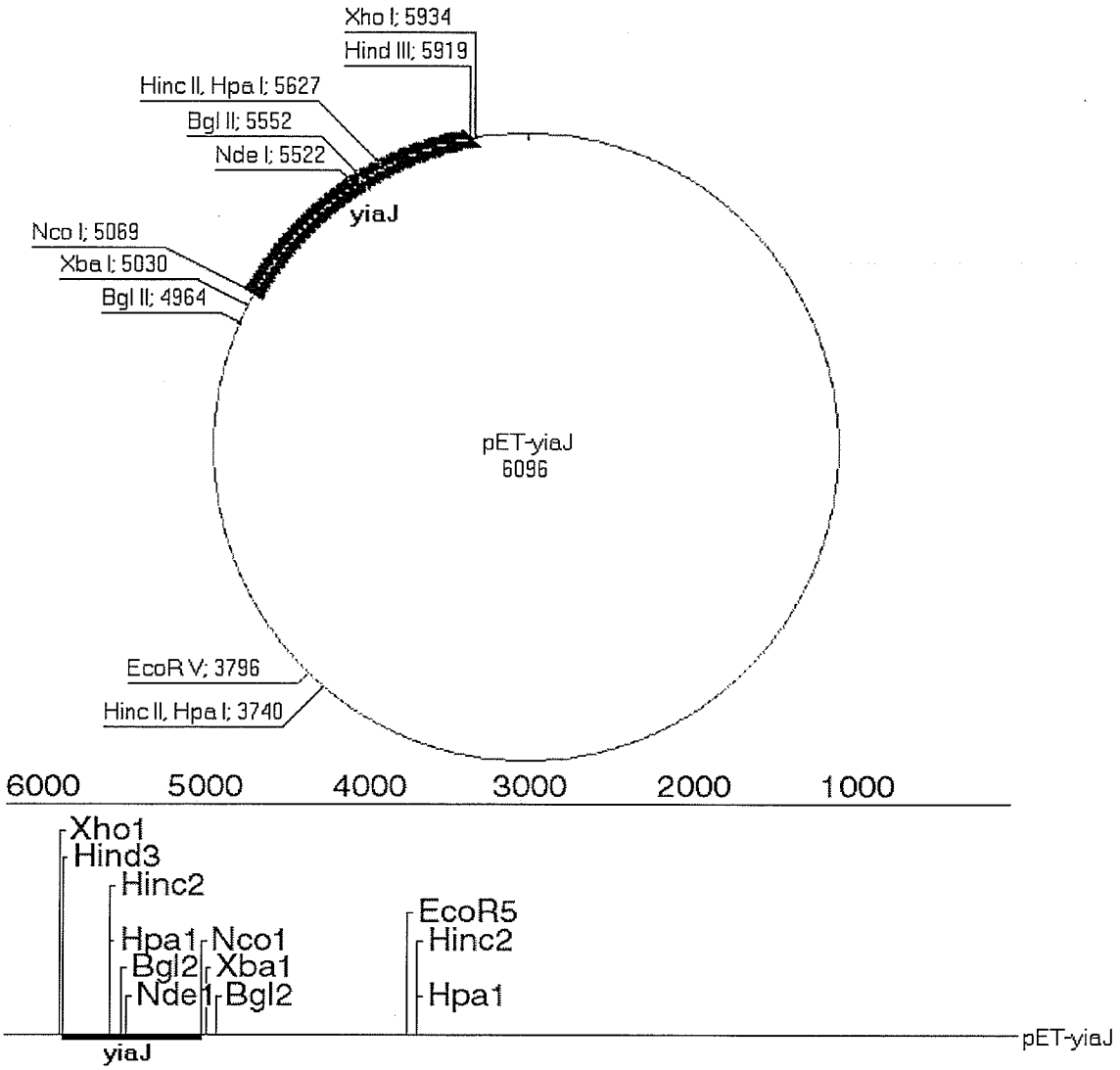


FIGURE 3.2: Theoretical Restriction Map of the pET-*yia*J Recombinant. By analyzing the 6096 bp pET-*yia*J sequence for *Bgl*2, *EcoR*5, *Hinc*2, *Hind*3, *Hpa*1, *Nco*1, *Nde*1, *Xba*1 and *Xho*1 restriction endonuclease recognition sites, a theoretical restriction map is generated so that a comparison can be made with the actual experimental results.

Double and single digests using the restriction endonucleases selected to generate the theoretical restriction maps were then performed on pET-*yag*I and pET-*yia*J DNA. The results of the digestions were examined on 1% agarose gels where the fragments resulting from each digestion were expected to match with those predicted from the theoretical maps providing that both the PCR amplifications and ligations were successful.

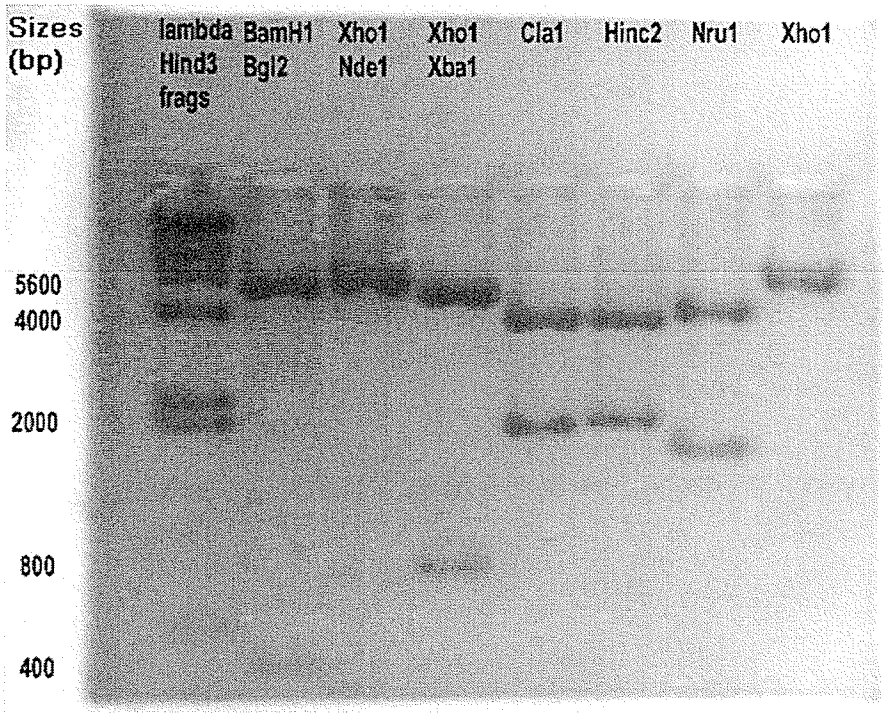


FIGURE 3.3: Agarose Gel Showing the Results of Restriction Endonuclease Digestion on pET-*yag*I DNA. Single and double digests were conducted using enzymes selected from the theoretical restriction map. The *lambda Hind3 fragments* molecular weight ladder was used to approximate fragment sizes. The *Xho*1/*Nde*1 double digestion shows a faint band at ~800 bp corresponding to the *yag*I insert.

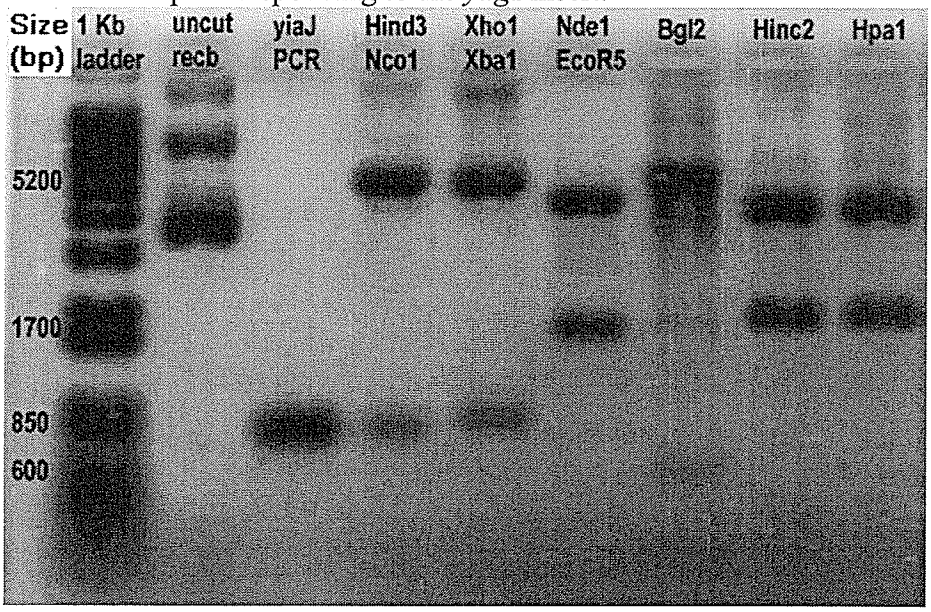


FIGURE 3.4: Agarose Gel Showing the Results of Restriction Endonuclease Digestion on pET-*yia*J DNA. Single and double digests were conducted using enzymes selected from the theoretical restriction map. The *1 Kb DNA Ladder* molecular weight ladder was used to approximate fragment sizes and uncut recombinant and the PCR *yia*J amplification product were used as references. The *Hind*3/*Nco*1 double digestion shows a band at ~850 bp corresponding to the *yia*J insert.

All of the restriction endonuclease digestion patterns resulting from the single and double matched perfectly to those predicted from the theoretical restriction maps. Of particular interest were the *Nde*1/*Xho*1 and *Nco*1/*Hind*3 double digestions on pET-*yag*I and pET-*yia*J DNA respectively. The former pair of digestions results in the liberation of the *yag*I insert from pET-24b(+) vector as seen in FIGURE 3.3 where the latter pair regenerates the *yia*J insert and pET-28b(+) vector as seen in FIGURE 3.4.

Following the success of the restriction mapping, DNA samples were prepared for sequencing at an automated DNA-sequencing facility.

Automated DNA Sequencing Data

pET-*yag*I and pET-*yia*J DNA samples were prepared for DNA sequencing using QIAgen filtration columns to ensure that the DNA was extremely pure and in the proper matrix (ddH_2O) as required by the sequencing process.

pET-*yag*I DNA was sequenced unidirectionally using six sequencing primers to cover the entire scope of the YagI insert. Using the 5'→3' YagI antisense sequence as a template, the results of the six sequencing primers were compared and analyzed for mutations. All results read 5'→3' in the antisense frame.

The 5'→3' *yag*I antisense sequence:

TCAGGCCGGGTAGCCCAGCGCGCGCGAGAGCGCGAGCCCGGCCTGCTGAAGCTGCTCGCGGAAATTAGCCAG
CTCCGCGTCGTCCACGCGGGAGGTCAGCGTCGACAGGCTCAGGGCGGCGATGACGCGGGACTCGTGGTTCCA
CACCGGCACCGCCACGCAGCGCACGCCCTGCTCGTTCTCTTCGCTGTCCAGGGCGTAGCCTTGCTCGCGGGT
CTGCGCCAGGGCGCTCATTAAGGCTTCGCGAGACGCGAGGGTGGCGGGCGTAAAGGTAGTGTACTGATAGCC
CTCCAGCAGGGCGTTCAGCTCGGCCTCGCCCAGCCAGGCAATCAACACCTTGCCGATGGCGGTGGCGTGCAC
CGGCAGGCGGCGGCCGATGCGTGAATAGGCGATGGCGGCCAGCTTGCCTTCAATCTTCTCGATATAGACCCC
TTCACGCCCGTCCAGGATCCCCAGATGGGTGGTCTGCCCGGTCCGCCGGGACAGCTCCGTCAGCCAGCCTTT
TGCCTTCTGCCGAATATCGATGGAGCCCACGACAAAATGGCCGCGCTCGACCAGCTTCATGCCGAGGCGATA
CTTGCCGTTCTCCGGGTTCTGATCGATATAGCCGTGAAGCTGCAGGGTTTTTAGCAGCGAGTGGAGGGTACT
CTTGCTCAGCCCCATCAGTTTGCTGATGTCGGTGATCTTAAGCTCGGTGGCCTGCTCGTTGAACAGGTCGAG
GATCTGCAACGCACGTTCAACAGACTGAATAATCGGCAT

Sequencing Primer 1 results:

ttgttagcagccggatctcagtggtggtggtggtggtgCTCGAGCCGGTCTAACGGC
TCAGGCCGGGTAGCCCAGCGCGCGCGAGAGCGCGAGCCCGGCCTGCTGAAGCTGCTCGCGGAAATTAGCCAG
CTCCGCGTCGTCCACGCGGGAGGTCAGCGTCGACAGGCTCAGGGCGG

Sequencing Primer 2 results:

GTCAGCGTCGACAGGCTCAGGGCGGCGATGACGCGGGACTCGTGGTTC
CACACCGGCACCGCCACGCAGCGCACGCCCTGCTCGTTCTCTTCGCTGTCCAGGGCGTAGCCTTGCTCGCG
GGTCTGCGCCAGGGCGCTCATTAAGGCTTCGCGAGACGCGAGGGTGGCGGGCGTAAAGGTAGTGTACTGATA
GCCCTCCAGCAGGGCGTTCAGCTCGGCCTCGCCCAGCCAGG

Sequencing Primer 3 results:

CATTAAGGCTTCGCGAGACGCGAGGGTGGCGGGCGTAAAGGTAGTGTACTGAT
AGCCCTCCAGCAGGGCGTTCAGCTCGGCCTCGCCCAGCCAGGCAATCAACACCTTGCCGATGGCGGTGGCG
TGCACCGGCAGGCGGCGGCCGATGCGTGAATAGGCGATGGCGGCCAGCTTGCCTTCAATCTTCTCGATATAG
ACCCCTTCACGCCCGTCCAGGATCCCCAGATGGGTGG

Sequencing Primer 4 results:

GGCGGCGGCCGATGCGTGAATAGGCGATGGCGGCCAGCTTGCCTTCAATCTTCTCGATATA
GACCCCTTCACGCCCGTCCAGGATCCCCAGATGGGTGGTCTGCCCGGTCCGCCGGGACAGCTCCGTCAGCCA
GCCTTTTGCCTTCTGCCGAATATCGATGGAGCCCACGACAAAATGGCCG

Sequencing Primer 5 results:

TGCCTTCTGCCGAATATCGATGGAGCCCACGACAAAATGGCCGCGCTCGACCAGCTTCATGCCG
AGGCGATACTTGCCGTTCTCCGGGTTCTGATCGATATAGCCGTGAAGCTGCAGGGTTTTTAGCAGCGAGTGG
AGGGTACTCTTGCTCAGCCCCATCAGTTTGCTGATGTCGGTGATCTTAAGCTCGGTGGCCTGCTCGTTGAAC
AGGTCGAGGATCTGCAACGCACGTTCAACAGA

Sequencing Primer 6 results:

CTCTTGCTCAGCCCCATCAGTTTGCTGATGTCGGTGATCTTAAGCTCGGTGGCCTGCTCGTTGA
ACAGGTCGAGGATCTGCAACGCACGTTCAACAGACTGAATAATCGGCATATGtatatctccttcttaaagtt
aaacaaaattatttctagaggggaattgttatccgctcacaattcccctatagtgagtcgtattaatttcgc
gggatcgagatctcgatcctctacgccg

FIGURE 3.5: Results of the Automated Sequencing of pET-*yag*I Recombinant DNA. Sequences in blue represent regions of overlap between successive sequencing primers. Sequences in lowercase represents pET-24b(+) vector DNA

A change in sequencing services allowed for the pET-*yia*J DNA to be sequenced bidirectionally using the T7 promoter and terminator primers. The change also permitted longer spans of DNA to be sequenced.

Using both the 5'→3' antisense and sense sequences of *yia*J as templates, the results from the sequencing of pET-*yia*J were compared.

The 5'→3' *yia*J Antisense Reading Frame:

TTATGTGATCGCGCCCAGGTCATCCCTGACAGTAAATCCCAGTTCATTAGAAATAGCCTGCGCGGTTTCACG
CAGTGGTTTCAGGAGATTTTTCTCTCCCACCTGTTTCAGACGTGATGTCGAAAGCGAAATCGACACGGCGTA
CGGCACCCGCCCATGAATATCAAACACCGGAACAGCAATACAGGAGACGCCGAGTTCGTTTTCTTCTCTGTC
CATCGCCGCTCCGCTTTCACGAATGTGCGCCAGTTCGTCGAACATCGCGGGCAGCTCGGTAATGGTATTGCG
GGTTAACGGCTGGATCTCATGCTGATGGCTTTCCCAGTATGACTTCACGTAGTCCGGGTGACCAAACGCCAT
GTAGATCTTGCCCATTGCGGAACAGTAGAGCGGCATATGCTGGCCAATATAGGCACGGGTTCGCAGCATCCC
GGTTGTGGGTTCCAGCTTATAAATCAAAATAGCGTGATCGTCTTCGCGGCTGGAGAAGTTAATGGTTTCACC
AGTGGCGATGTTCAGTGCCTCAAGATGCGGAGCGGCGATATGAATGATATTCAGCGAAGACAGCGCCTTCTG
CCCGACGGCAATAAATTTGGTGGTCAGGCGATAACTCCCTGCGGCAGGCGCGGTGGTCACATAGCCGCAGGA
CTGTAATCCCTGCAATAAGCGATGGACGGTACTCTTATTTAAACCAGCCAGCTCCGAAAGATGCGCCAACGG
ACAACCGTTTGGATAGTTGCTCAAAATCTCAATCAGCATCAACCCGCGAAACAGACTCTGGCTTCCGGCTGG
ACGCTCTTTTTCCTGCGCCATCTCGTTCTCTTTTTTTCCCATCACTTCTTTCCCCAT

T7 Terminator Sequencing Results:

cttccttcgggctttgttagcagccggatctcagtggtggtggtggtggtgctcgagtgcggccgcAAGCTT
TATGTGATCGCGCCCAGGTCATCCCTGACAGTAAATCCCAGTTCATTAGAAATAGCCTGCGCGGTTTCACGC
AGTGGTTTCAGGAGATTTTTCTCTCCCACCTGTTTCAGACGTGATGTCGAAAGCGAAATCGACACGGCGTAC
GGCACCCGCCCATGAATATCAAACACCGGAACAGCAATACAGGAGACGCCGAGTTCGTTTTCTTCTCTGTCC
ATCGCCGCTCCGCTTTCACGAATGTGCGCCAGTTCGTCGAACATCGCGGGCAGCTCGGTAATGGTATTGCGG
GTTAACGGCTGGATCTCATGCTGATGGCTTTCCCAGTATGACTTCACGTAGTCCGGGTGACCAAACGCCATG
TAGATCTTGCCCATTGCGGAACAGTAGAGCGGCATATGCTGGCCAATATAGGCACGGGTTCGCAGCATCCCG
GTTGTGGGTTCCAGCTTATAAATCAAAATAGCGTGATCGTCTTCGCGGCTGGAGAAGTTAATGGTTTCACCA
GTGGCGATGTTCAGTGCCTCAAGATGCGGAGCGGCGATATGAATGATATTCAGCGAAGACAGCGCCTTCTGC
CCGACGGCAATAAATTTGGTGGTCAGGCGATAACTCCCTGCGGCAGGCGCGGTGGTCACATAGCCGCAGGAC
TGTAATCCCTGCAATAAGCGATGGACGGTACTCTTATTTAAACCAGCCAGCTCCGAAAGATGCGCCAACGGA
CAACCGTTTGGATAGTTGCTCAAAATCTCAATCAGCATCAACCCGCGAAACAGACTCTGGCTTCCG
*sequence in lowercase represents pET-28b(+) vector DNA

FIGURE 3.6: Results from the Automated Sequencing of pET-*yia*J Recombinant DNA. pET-*yia*J recombinant DNA was sequenced bidirectionally i.e. using two primers complementary to each end of the recombinant (T7 Promoter & T7 Terminator primers), it was possible to sequence the entire length of the *yia*J PCR product insert from both ends. The sequence shown here is that which was obtained using the T7 Terminator primer as the sequencing primer.

The 5'→3' *yia*J Sense Reading Frame:

ATGGGGAAAGAAGTGATGGGAAAAAAAGAGAACGAGATGGCGCAGGAAAAAGAGCGTCCAGCCGGAAGCCAG
AGTCTGTTTCGCGGGTTGATGCTGATTGAGATTTTGAGCAACTATCCAAACGGTTGTCCGTTGGCGCATCTT
TCGGAGCTGGCTGGTTTAAATAAGAGTACCGTCCATCGCTTATTGCAGGGATTACAGTCCTGCGGCTATGTG
ACCACCGCGCCTGCCGCAGGGAGTTATCGCCTGACCACCAAATTTATTGCCGTCGGGCAGAAGGCGCTGTCT
TCGCTGAATATCATTCATATCGCCGCTCCGCATCTTGAGGCACTGAACATCGCCACTGGTGAAACCATTAAC
TTCTCCAGCCGCGAAGACGATCACGCTATTTTGATTTATAAGCTGGAACCCACAACCGGGATGCTGCGAACC
CGTGCCTATATTGGCCAGCATATGCCGCTCTACTGTTCCGCAATGGGCAAGATCTACATGGCGTTTGGTCAC
CCGGACTACGTGAAGTCATACTGGGAAAGCCATCAGCATGAGATCCAGCCGTTAACCCGCAATACCATTACC
GAGCTGCCCGCGATGTTCGACGAACTGGCGCACATTCGTGAAAGCGGAGCGGCGATGGACAGAGAAGAAAAC
GAACTCGGCGTCTCCTGTATTGCTGTTCCGGTGTTTGATATTCATGGGCGGGTGCCGTACGCCGTGTCGATT
TCGCTTTCGACATCACGTCTGAAACAGGTGGGAGAGAAAAATCTCCTGAAACCACTGCGTGAAACCGCGCAG
GCTATTTCTAATGAACTGGGATTTACTGTCAGGGATGACCTGGGCGCGATCACATAA

T7 Promoter Sequencing Results:

ggataacaattcccctctagaaataattttgtttaactttaagaaggagatatacc
ATGGGGAAAGAAGTGATGGGAAAAAAAGAGAACGAGATGGCGCAGGAAAAAGAGCGTCCAGCCGGAAGCCAG
AGTCTGTTTCGCGGGTTGATGCTGATTGAGATTTTGAGCAACTATCCAAACGGTTGTCCGTTGGCGCATCTT
TCGGAGCTGGCTGGTTTAAATAAGAGTACCGTCCATCGCTTATTGCAGGGATTACAGTCCTGCGGCTATGTG
ACCACCGCGCCTGCCGCAGGGAGTTATCGCCTGACCACCAAATTTATTGCCGTCGGGCAGAAGGCGCTGTCT
TCGCTGAATATCATTCATATCGCCGCTCCGCATCTTGAGGCACTGAACATCGCCACTGGTGAAACCATTAAC
TTCTCCAGCCGCGAAGACGATCACGCTATTTTGATTTATAAGCTGGAACCCACAACCGGGATGCTGCGAACC
CGTGCCTATATTGGCCAGCATATGCCGCTCTACTGTTCCGCAATGGGCAAGATCTACATGGCGTTTGGTCAC
CCGGACTACGTGAAGTCATACTGGGAAAGCCATCAGCATGAGATCCAGCCGTTAACCCGCAATACCATTACC
GAGCTGCCCGCGATGTTCGACGAACTGGCGCACATTCGTGAAAGCGGAGCGGCGATGGACAGAGAAGAAAAC
GAACTCGGCGTCTCCTGTATTGCTGTTCCGGTGTTTGATATTCATGGGCGGGTGCCGTACGCCGTGTCGATT
TCGCTTTCGACATCACGTCTGAAACAGGTGGGAGAGAAAAATCTCCTGAAACCACTGCGTGAAACCGCGCAG
GCTATTTCTAATGAACTGGGATTTACTGTCAGGGATGACCTGGGCGCG
*sequence in lowercase represents pET-28b(+) vector DNA

FIGURE 3.7: Results from the Automated Sequencing of pET-*yia*J Recombinant DNA. Shown here is the sequence obtained using the T7 Promoter primer as the sequencing primer for pET-*yia*J recombinant DNA.

From the results of the sequencing, the accuracy of the PCR amplification and ligation of both the YagI and YiaJ inserts was verified, as no mutations were found in either case.

3.2 Expression and Purification of Recombinant YagI and YiaJ

Following the verification of the integrity of the pET-*yag*I and pET-*yia*J recombinants, competent BL21 (DE3) pLysS cells were transformed (separately) with the recombinants and plated onto LB-chloramphenicol-kanamycin plates. Colony isolates were then grown to 1 L batch volumes in LB-chloramphenicol-kanamycin broth; IPTG to

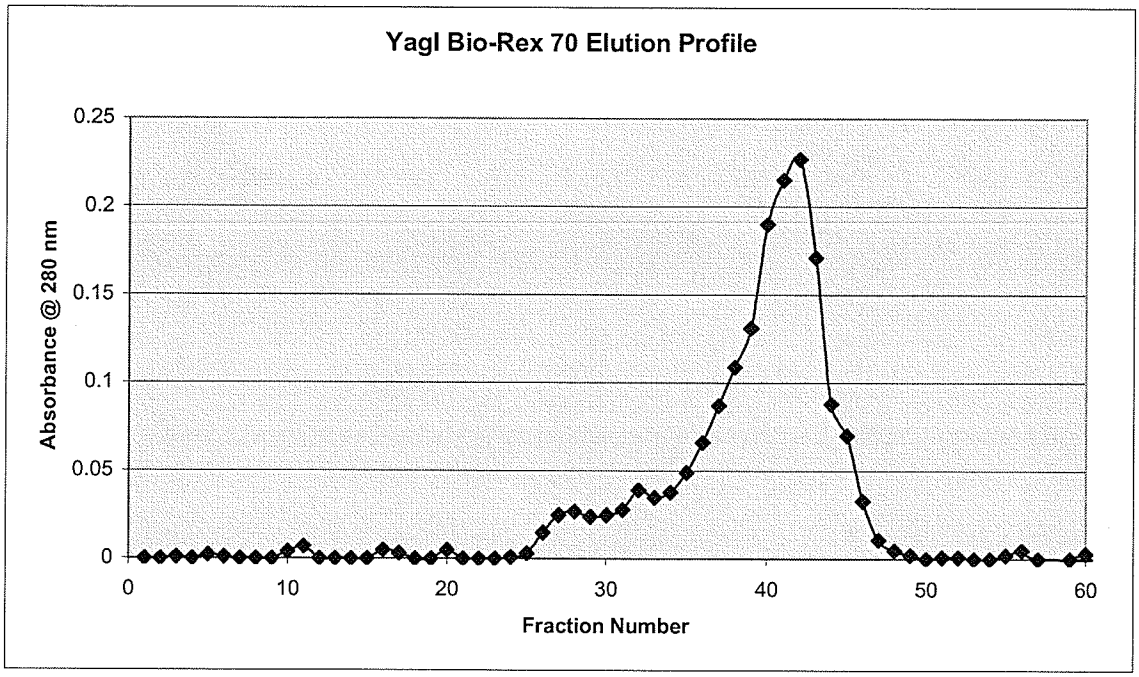
a final concentration of 1 mM induced protein expression. Protein-rich cells were harvested, flash frozen in liquid nitrogen and then thawed using 20 mM Tris-HCl, 50 mM KCl, 1 mM EDTA, 1mM DTT pH 7.8 (lysis/equilibration buffer). The lysed cell solution was then centrifuged to remove cellular debris.

BioREX-70 Cation Exchange Chromatography

Proteinaceous supernatant obtained following lysis and centrifugation was applied directly to a 2.5 x 15 cm BioREX-70 column equilibrated in lysis/equilibration buffer. Unbound proteins were removed from the BioREX-70 resin by washing with 500 mL of lysis/equilibration buffer. Proteins bound to the BioREX-70 matrix were eluted from the column using a linear salt gradient consisting of a low salt buffer reservoir (20 mM Tris-HCl, 50 mM KCl, 1 mM EDTA, 1mM DTT pH 7.8) and a high salt reservoir (20 mM Tris-HCl, 0.5 M KCl, 1 mM EDTA, 1mM DTT pH 7.8) connected by a salt bridge. Several 150-drop fractions (approx. 7.5 mL) were collected using a fraction collector after, and qualitatively tested for protein content using the Bio-Rad Protein Assay solution in a spot test. Typically, following elution from BioREX-70 columns, a stretch of 10-20 fractions was found to contain protein using this colorimetric method. However, the absorbance values of all fractions were also measured at 280 nm to verify the presence of protein, the results of which are depicted in FIGURE 3.8 below.

The colorimetric and spectrophotometric methods of evaluation both indicated the presence of one major protein constituent in BioREX-70 eluate resulting from both YagI and YiaJ-derived protein preparations.

A).



B).

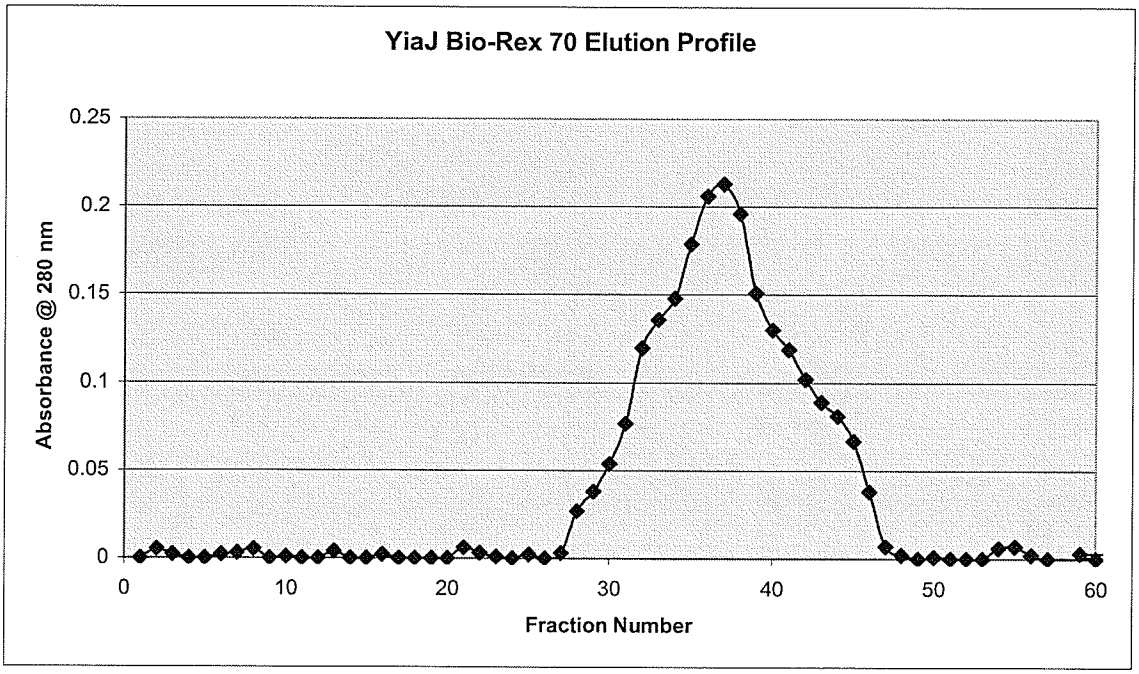


FIGURE 3.8: Elution Profiles of: A) YagI and B) YiaJ Proteins from BioREX-70 Columns. Following cell lysis of induced cells containing pET-*yagI*/pET-*yiaJ* recombinant, a proteinaceous solution is obtained. This solution is applied directly to a BioREX-70 cation exchange column and bound protein is eluted using a KCl salt gradient. Spectrophotometric evaluation of eluate from the ion exchange resin indicates the presence of one major protein peak in each case.

The pET Expression System, upon first glance, seemed successful in generating the YagI and YiaJ proteins. However, though there was an abundance of protein present, whether or not it was target protein remained to be verified. If, however, this major protein product was indeed target protein, the BioREX-70 cation exchange resin would have proved to be an excellent first measure in isolation and purification of YagI and YiaJ proteins.

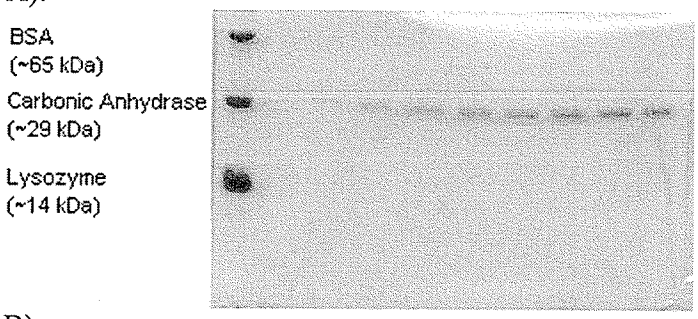
To test our suspicions of having successfully obtained YagI and YiaJ protein by our methods, the contents of the protein containing fractions were evaluated using SDS-PAGE.

SDS-PAGE Analysis of BioREX-70 Fractions

Fractions of BioREX-70 eluate suspected of containing target protein were analyzed using SDS-PAGE. Samples for SDS-PAGE were prepared by mixing 10 - 20 μL of fraction with an equal volume of 2x SDS-PAGE sample buffer, vortexing and heating to >70°C. Samples were then loaded into thin-cast 12% SDS-polyacrylamide gels alongside a molecular weight ladder consisting of BSA (65 kDa), carbonic anhydrase (29 kDa) and lysozyme (14 kDa), and electrophoresed at 150 – 200 volts. After fixing the gels, staining in Coomassie Brilliant Blue stain and then destaining, the results were highly indicative of having indeed isolated and purified YagI and YiaJ proteins. The presumed target proteins were visualized as distinct bands with a similar mobility to the 29 kDa molecular weight standard band. In the case of YagI (27, 707 Da), all fractions suspected of containing said protein showed a major protein band just below the 29 kDa carbonic anhydrase molecular weight standard. In the case of YiaJ (30, 935 Da),

prominent bands above the carbonic anhydrase band were present in those fractions suspected of containing YiaJ protein.

A).



B).

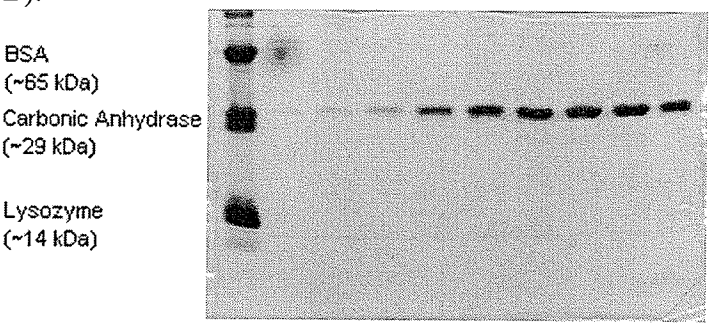


FIGURE 3.9: 12% SDS-PAGE Gels of BioREX-70 Fractions. A). The *yag*I ORF predicts a protein of 27, 707 Da. The prominent band just below the 29 kDa carbonic anhydrase band appearing in the tested fractions strongly indicates that YagI protein has been isolated. B). YiaJ protein is predicted to be 30, 935 Da. A prominent band just above the 29 kDa carbonic anhydrase marker appearing in tested fractions, in turn strongly indicates successful isolation of YiaJ protein.

As evidenced from the gels, the target proteins seemed to be very pure at this stage of the purification. However, as a final measure of purification, the fractions were pooled, concentrated, and then fractionated using size-exclusion chromatography. Before concentrating the proteins, the salt concentration required to elute the YagI and YiaJ proteins needed to be determined

Buffer Conditions for Isolation and Purification

Using a conductivity meter and six measuring standards (20 mM Tris-HCl, 1 mM EDTA, pH 7.8; 0 M KCl, 0.1 M KCl…0.5 M KCl) a calibration curve was prepared from

which it was determined that 0.35 M KCl was the salt concentration necessary to elute both YagI and YiaJ proteins bound to BioREX-70 cation exchange resin. Consequently, YagI and YiaJ proteins were stored in a buffer consisting of 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA and 1 mM DTT at pH 7.8.

Concentration via Vacuum Dialysis

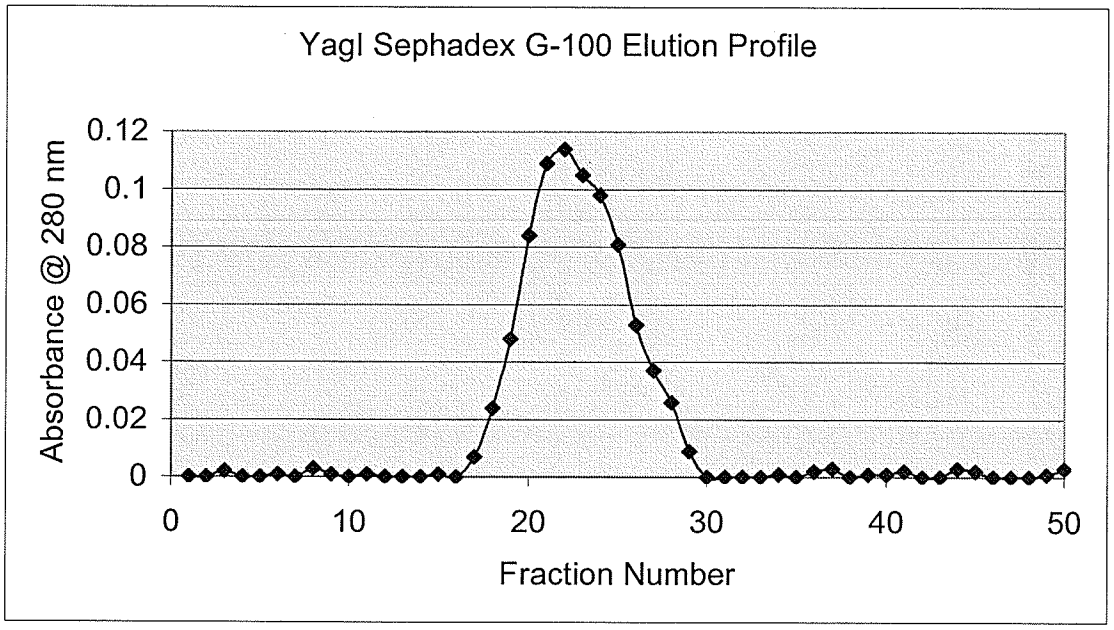
Fractions identified by SDS-PAGE as containing target protein were pooled and concentrated via vacuum dialysis. Although vacuum dialysis is notoriously slow, it was the only manner in which the YagI and YiaJ protein solutions could be concentrated as attempts at using an Amicon cell resulted in rapid precipitation.

Efforts were made to increase the solubility of both proteins under study such as substituting NaCl for KCl in the buffer and by varying both the Tris-HCl and KCl concentrations (i.e. varying the ionic strength). However, both YagI and YiaJ proteins could rarely be concentrated above 2 mg/mL without the occurrence of precipitation regardless of any attempts to alter the buffer conditions.

Sephadex G-100 Gel Filtration Chromatography

Following the concentration of pooled fractions containing target protein, 5 – 10 mL of concentrate (5 – 20 mg of protein) was loaded onto a Sephadex G-100 gel filtration column equilibrated with 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA, 1 mM DTT; pH 7.8 buffer. Several 200-drop fractions were collected and evaluated using the colorimetric and spectrophotometric methods previously described. Elution profiles and 12% SDS-PAGE gels were also prepared from fractions resulting from the size-exclusion chromatography.

A).



B).

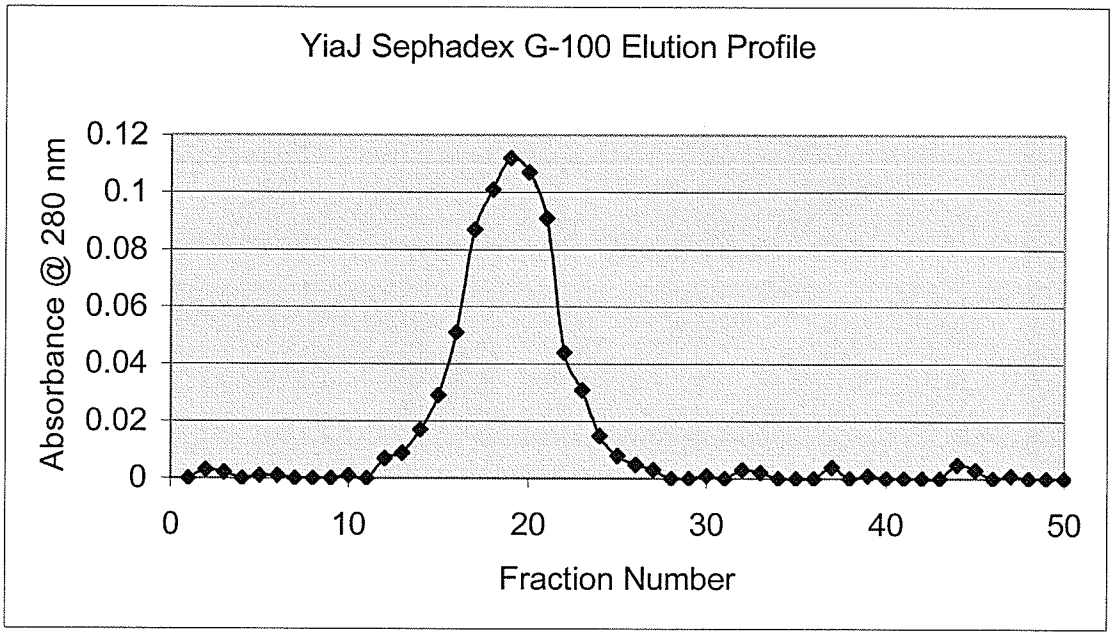
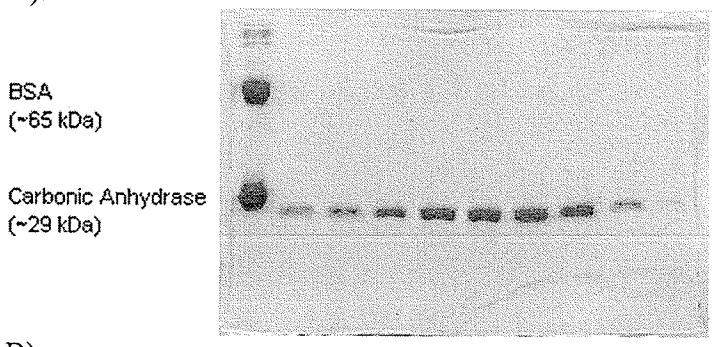


FIGURE 3.10: Elution profiles for A) YagI and B) YiaJ Proteins Obtained from Sephadex G-100 Chromatography. Following concentration of pooled BioREX-70 fractions containing target protein, concentrated samples are loaded onto Sephadex G-100 columns and resolved. As in BioREX-70 fractions, only one major protein peak is present illustrating successful purification and isolation.

A).



B).

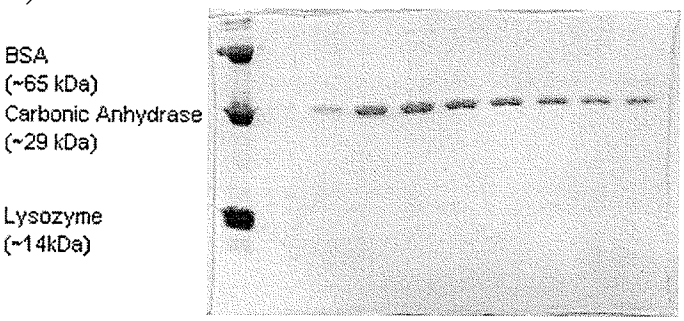


FIGURE 3.11: SDS-PAGE Analysis of Sephadex G-100 Fractions Containing: A) YagI and B) YiaJ Proteins.

Using this combination of ion-exchange (Bio-REX 70) and size-exclusion (Sephadex G-100) chromatography, we were able to successfully purify the YagI and YiaJ proteins for characterization studies.

3.3 Mass Spectrometry

If the methionine codons selected as the start of the protein sequences of interest were correctly assigned, YagI should be 252 residues in length with a corresponding molecular mass of 27837 Da and YiaJ should be 282 residues in length with a corresponding mass of 31066 Da. However, as these are the hypothetical protein products of open reading frames resulting from the *E. coli* genome sequencing project, they have yet to be extensively studied and characterized. The occurrences of post-translational modifications (which would be detected as variations from the expected molecular mass) are certainly possible for both of the proteins under study. Fortunately,

access to mass spectrometry would not only reveal whether or not we had successfully synthesized the correct protein, but would also give the first glimpses into the structural details of YagI and YiaJ proteins. ESI-TOF mass spectrometry was used to determine the mass of protein products obtained from the pET-Expression System.

ESI-MS determines the charge state and molecular mass of the ions from electrospray ionization mass spectra. However, due to the multiple charge states that can result for a single ion species, the raw data from ESI-MS typically appear as a charge envelope. The interpretation of ESI mass spectra becomes especially difficult when dealing with mixtures of components as each component can produce several peaks where even a simple two-component mixture may produce more than 20 peaks. For this reason, a special algorithm referred to as the deconvolution algorithm is routinely used in the analysis and interpretation of multiply charged mass spectral data of biopolymers (primarily spectra of mixtures or spectra with low signal-to-noise ratios). In particular, after the charge states are identified the deconvolution algorithm transforms an ESI-MS spectrum of relative abundance versus mass-to-charge ratio into a subsequent spectrum of relative abundance versus mass. Hence, each sequence of multiply charged ion peaks (charge envelope corresponding to one sample component) in the acquired mass spectrum is converted into a single peak corresponding to the molecular mass.

Samples for an ESI-TOF mass spectrometer were prepared from BioREX-70 fractions of both the YagI and YiaJ proteins. Removing the buffer components of the aqueous protein solutions typically resulted in precipitation of protein, but samples were resolubilized using acetic acid at a 5% concentration. Hence, all samples analyzed using ESI-TOFMS were denatured protein samples. Both protein species yielded a charge

envelope, which upon deconvolution produced single peaks with a molecular mass corresponding to the purified proteins.

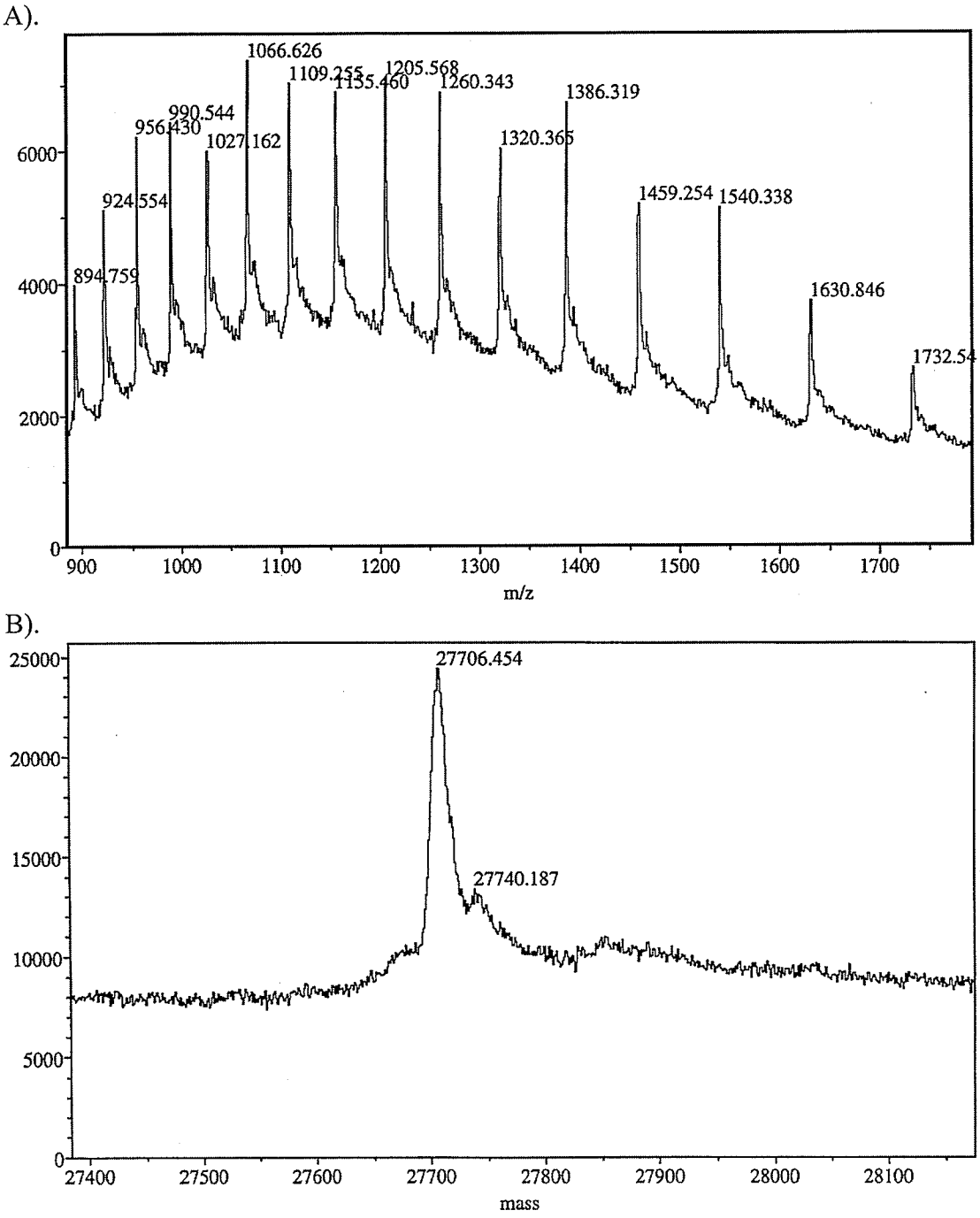


FIGURE 3.12: Raw ESI-TOF MS Analysis of YagI Protein. A) Charge envelope of YagI monomer B) Protein peaks detected from concentrated BioREX-70 fractions after applying the deconvolution algorithm.

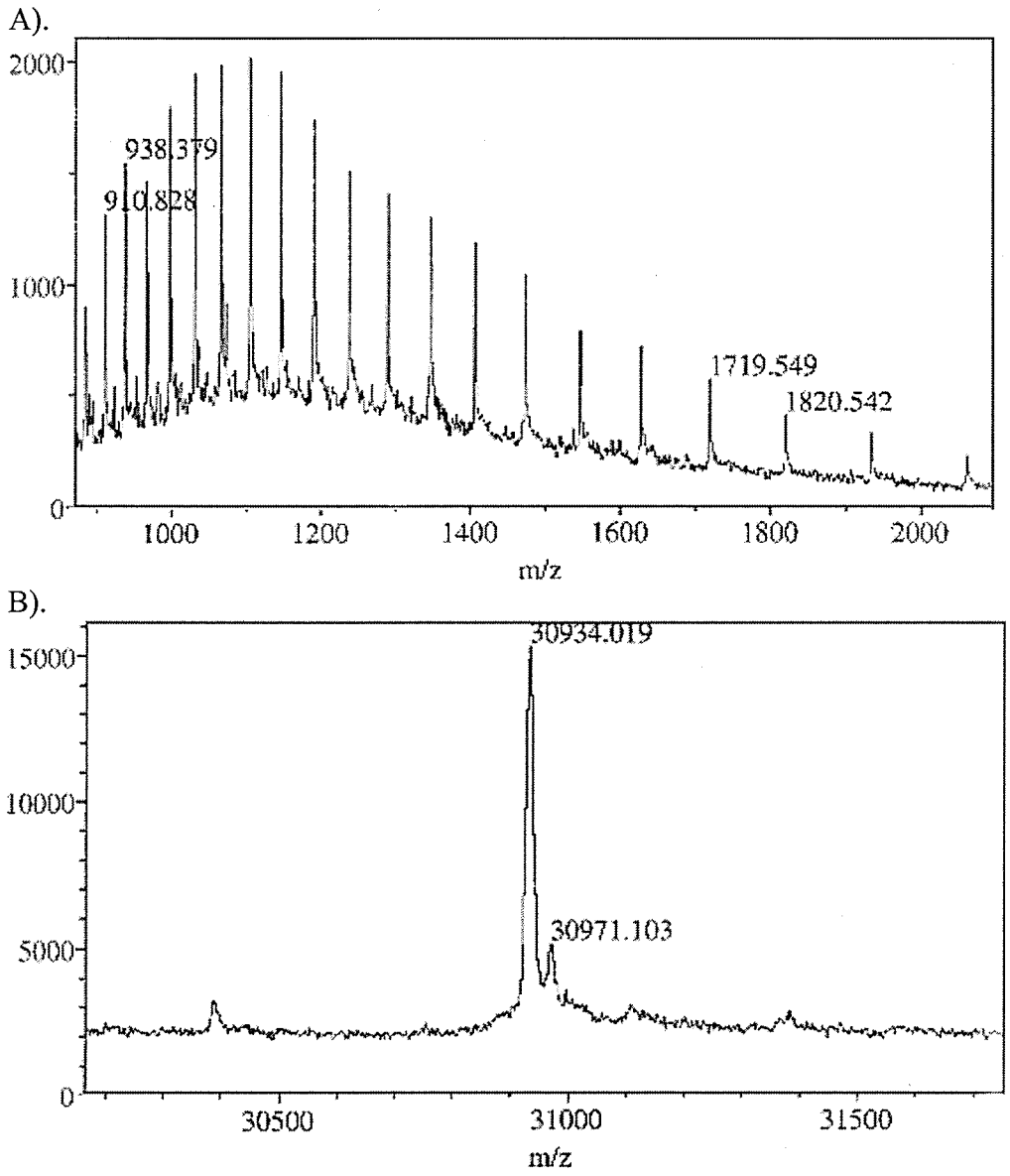


FIGURE 3.13: Raw ESI-TOFMS Analysis of YiaJ Protein. A). Charge envelope of YiaJ monomer B). Protein peaks detected from concentrated BioREX-70 fractions after applying the deconvolution algorithm.

Our ESI-TOFMS spectra verified that YagI and YiaJ proteins were successfully generated, isolated and purified in cases where the only post-translational modification appears to be the N-terminal removal of the methionine initiation codon.

TABLE 3.0: ESI-TOF MS Experimental Results for YagI and YiaJ Proteins

Protein	Theoretical Mass (Da)	Experimental Mass (Da)
YagI	27837 (w. N-terminal Met) **27706** (w/o N-terminal Met)	**27706**
YiaJ	31066 (w. N-terminal Met) **30935** (w/o N-terminal Met)	**30934**

Upon successful production of the YagI and YiaJ proteins, characterization of the proteins began with determination of the molar extinction coefficients.

3.4 Molar Extinction Coefficients for YagI and YiaJ

The molar extinction coefficient of a protein may be estimated from knowledge of its amino acid composition (Gill & von Hippel, 1989). From the molar extinction coefficients of tyrosine, tryptophan and cystine (as previously mentioned, absorbance due to cystine disulfide bonds do not contribute appreciably at wavelengths >260 nm) at a given wavelength the extinction coefficient of a denaturated protein can be computed using the equation:

$\varepsilon(Prot) = No.Tyr \times \varepsilon(Tyr) + No.Trp \times \varepsilon(Trp) + No.Cystine \times \varepsilon(Cystine)$

The absorbance (optical density) may then be calculated using the following formula:

$A(Prot) = \varepsilon (Prot) / Molecular\ weight\ of\ protein$

The conditions for which these equations are valid are: pH 6.5, 6.0 M guanidinium hydrochloride, 0.02 M phosphate buffer. Under these conditions, the protein primary structure analysis tool known as ProtParam (http://ca.expasy.org/tools/protparam.html) was used to predict ε values for YagI and YiaJ proteins. The results of the computation are summarized in TABLES 3.1 and 3.2.

TABLE 3.1: YagI Extinction Coefficients Predicted from ProtParam

	276 nm	278 nm	279 nm	280 nm	282 nm
ε (M^{-1} cm^{-1})	27800	28000	27740	27310	26400
Abs.1 mg/mL*	0.999	1.006	0.996	0.981	0.948

* - The absorbance resulting from a concentration of 1 mg/mL

TABLE 3.2: YiaJ Extinction Coefficients Predicted from ProtParam

	276 nm	278 nm	279 nm	280 nm	282 nm
ε (M^{-1} cm^{-1})	20190	19854	19350	18730	17840
[2]	19900	19600	19110	18490	17600
Abs.1 mg/mL	0.650	0.639	0.623	0.603	0.574
[2]	0.641	0.631	0.615	0.595	0.567

[2] - The first set of values were computed assuming ALL Cys residues appear as half cystines, whereas the second set of values assumes that NONE do. This assumption is ignored in YagI as there is only one Cys residue in the primary structure.

The experimental molar extinction coefficients (ε_M) for YagI and YiaJ proteins were also determined by the method of Edelhoch. The absorbances of native purified proteins were first measured at 280 nm. Proteins were then denatured using 7.5 M guanidine hydrochloride and absorbance measurements were made at 280, 288, 295 and 300 nm. The pH of denatured protein solutions was then raised by adding 10 M NaOH (to deprotonate the tyrosine phenol) and the resulting absorbances were collected at 295 and 300 nm. The results of the Edelhoch experiment are reported in the following table.

TABLE 3.3: Experimentally Determined Molar Extinction Coefficients at 280 nm

Protein	Theoretical ε (M^{-1} cm^{-1})	Experimental ε (M^{-1} cm^{-1})
YagI	27310	31076
YiaJ	18610*	20935

* - the theoretical ε in this table is reported as the average of the two values appearing in TABLE 3.2 above for comparison to the experimental value.

Using the Edelhoch method, we were able to determine molar absorptivities for both the YagI and YiaJ proteins comparable to those that were predicted for each protein.

3.5 YiaJ Contains a Reactive Cysteine

Of the two proteins under investigation in this study, only YiaJ was found to contain at least one reactive cysteine residue. This was first evidenced from the reaction of protein with 5,5'-dithiobis(2-nitrobenzoic acid).

Reaction with DTNB

Purified YiaJ protein was reacted with DTNB and the production of the yellow *p*-nitrothiophenol anion chromophore resulting from the reaction was monitored spectrophotometrically at intervals for 1 hour. Using the molar extinction coefficient for the nitrothiophenolate anion (13,600 M^{-1} cm^{-1}), the molar concentration of nitrothiophenolate accumulated per unit time was plotted in order to determine the ratio of reactive cysteine residues per mole of protein.

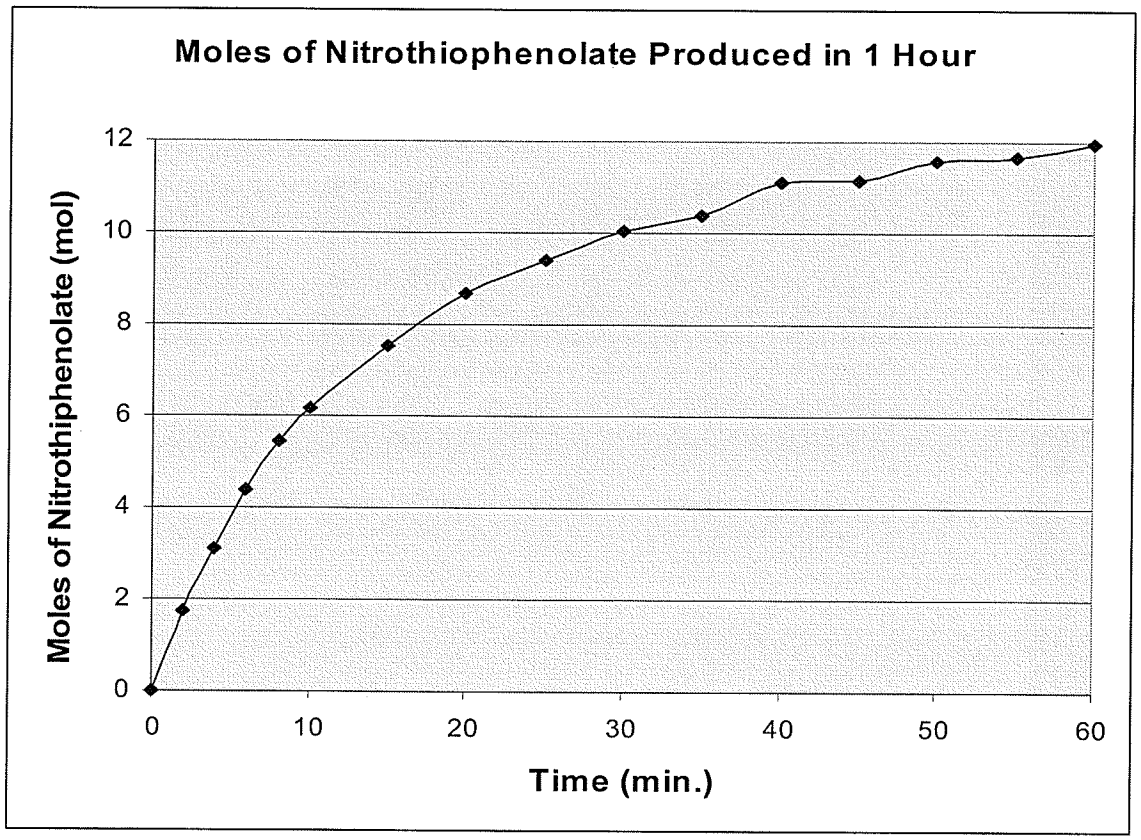


FIGURE 3.14: Molar Concentration of Nitrothiophenolate Accumulated in 1 hour upon Reaction of DTNB with YiaJ Protein.

It was determined that only one of the four cysteines in the YiaJ protein is available for reaction. The identity of this reactive cysteine was determined by performing a different covalent modification involving the reducing agent, 2-mercaptoethanol. As previously mentioned, the YagI protein failed to react with DTNB and hence its one cysteine residue was concluded to be buried in the three dimensional structure and unavailable for reaction.

Reaction with 2-mercaptoethanol

In order to determine the location of the reactive cysteine residue within the primary structure of the YiaJ protein, a covalent modification involving 2-mercaptoethanol was performed. In this reaction, YiaJ protein was incubated with 2-mercaptoethanol and as a control, an equal amount of protein was reserved without reacting with this reducing agent. In the allotted time for reaction/exchange (1 hour), any cysteines available for reaction would undergo the chemical modification. Following the reaction, both the control and modified protein samples were dialyzed against ammonium bicarbonate buffer to prepare for tryptic digestion and mass spectral analysis.

3.6 Identification of the Reactive Cysteine

ESI-TOFMS Verification of the Covalent Modification

In order to verify whether or not the covalent modification of a reactive cysteine residue had occurred, ESI-TOFMS was used as an initial check before proceeding with the actual identification of reactive cysteines in the primary structure of the protein. Both a control sample and one that had reacted with 2-mercaptoethanol were analyzed. The resulting spectra are provided in FIGURES 3.15 and 3.16.

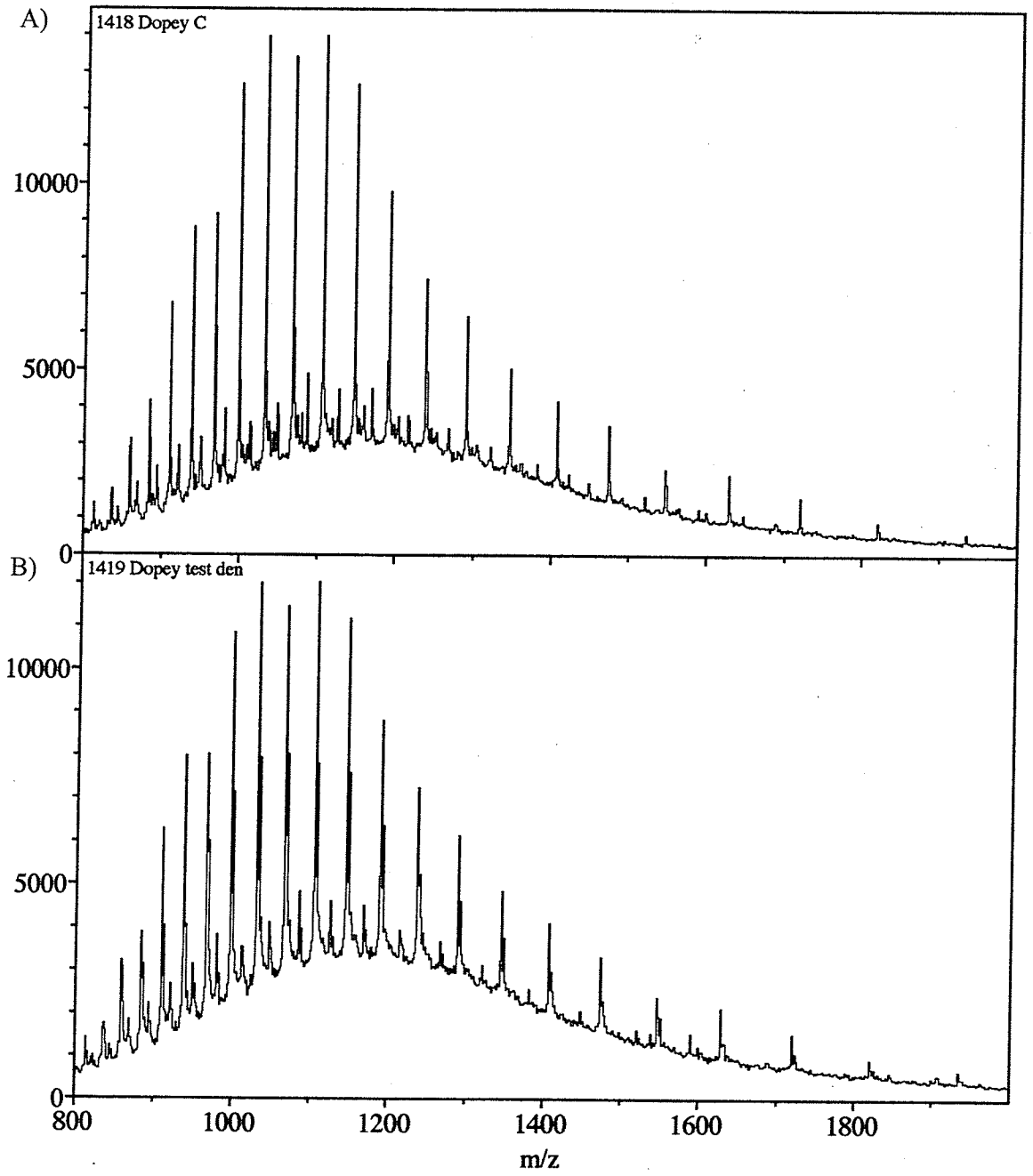


FIGURE 3.15: ESI-MS Raw Data for Modified Cysteine. A) Charge envelope resulting from the control sample B) Charge envelope resulting from YiaJ reacted with 2-mercaptoethanol.

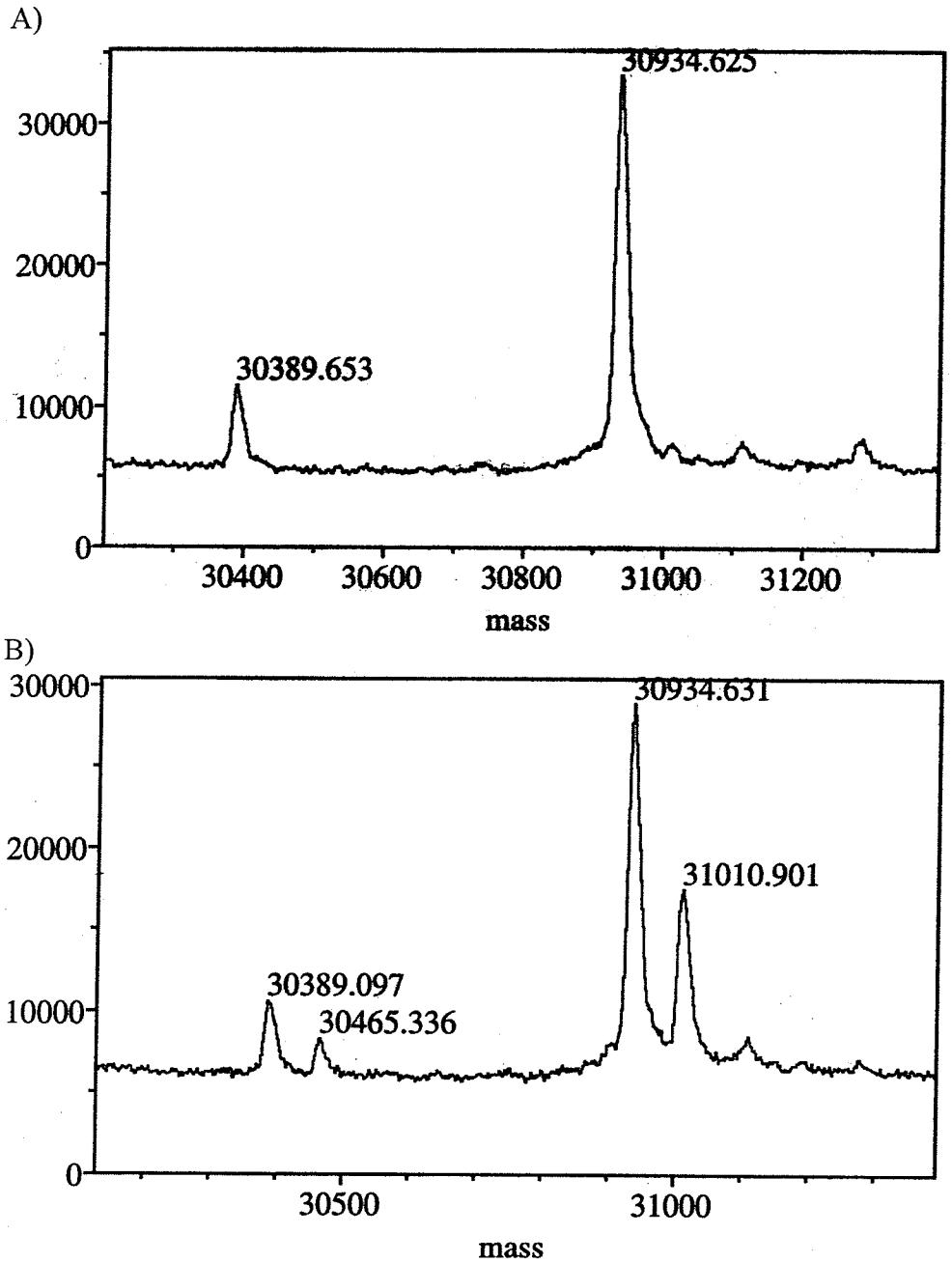


FIGURE 3.16: Transformed Spectra for YiaJ Protein Containing Covalent Modification. Deconvolution algorithm transforms envelope of multiply charged peaks into one peak corresponding to molecular mass of protein. A) Control sample B) Covalent modification

As expected, there were no covalent modifications in the control sample as illustrated in FIGURES 3.15 A) and 3.16 A) However, in the case of YiaJ incubated with 2-mercaptoethanol, a mass increase of 76 Da was detected indicating that only one of the four cysteines in the primary structure of YiaJ is available for reaction.

With regards to FIGURES 3.15 and 3.16, the importance of the deconvolution algorithm should be reiterated as more than one charge envelope was present i.e. a mixture of analyte. A contaminating protein was present resulting in two charge envelopes in FIGURE 3.16 A) and four charge envelopes in FIGURE 3.16 B). When the deconvolution algorithm was applied to the control and to the 2-mercaptoethanol-reacted YiaJ samples, the transformed spectra resulted in two peaks, one of which was YiaJ (30934 Da) in FIGURE 3.16 A), and four peaks in FIGURE 3.16 B). Two of the four peaks in FIGURE 3.16 B) had masses of 30934 Da and 31010 Da, which correspond to YiaJ protein, and YiaJ protein covalently modified with 2-mercaptoethanol respectively (a difference of 76 mass units between the two). The two other peaks in FIGURE 3.16 B) yielded masses of 30389 Da and 30465 Da. It is believed that this artifact is a truncated form of the YiaJ protein possibly resulting from partial proteolysis. The true identity of this protein remains unknown, but it was a very minor component of the sample. The mass difference between the two peaks (76 Da) is indicative that this unknown protein also contains at least one reactive cysteine. Although both proteins resulted in multiply charged states and both reacted with 2-mercaptoethanol yielding subsequent multiply charged species, the deconvolution algorithm allowed the mixture to be easily resolved into four distinct peaks.

Having shown that there is one reactive cysteine in the YiaJ protein, it was then a matter of identifying the precise location of this reactive cysteine residue in the primary structure of YiaJ. This was accomplished using a combination of enzymatic proteolysis and MALDI-MS.

Tryptic Digests

Following verification that the YiaJ protein had indeed been covalently modified by 2-mercaptoethanol, the now covalently modified protein was partially digested with trypsin, a proteolytic enzyme that cuts the primary amino acid sequence after lysine and arginine residues.

1 MGKEVMGKKE	11 NEMAQEKERP	21 AGSQSLFRGL	31 MLIEILSNYP	41 NGCPLAHLSE	51 LAGLNKSTVH
61 RLLQGLQSCG	71 YVTTAPAAGS	81 YRLTTKFIAV	91 GQKALSSLNI	101 IHIAAPHLEA	111 LNIATGETIN
121 FSSREDDHAI	131 LIYKLEPTTG	141 MLRTRAYIGQ	151 HMPLYCSAMG	161 KIYMAFGHPD	171 YVKSYWESHQ
181 HEIQPLTRNT	191 ITELPAMFDE	201 LAHIRESGAA	211 MDREENELGV	221 SCIAVPVFDI	231 HGRVPYAVSI
241 SLSTSRLKQV	251 GEKNLLKPLR	261 ETAQAISNEL	271 GFTVRDDLGA	281 IT	

*K = lysine residues; R = arginine residues

FIGURE 3.17: Trypsin Cleavage Sites in the YiaJ Protein Sequence. Proteolytic trypsin cleaves peptide/protein sequences after lysine(K) and arginine(R) residues yielding tryptic fragments of various lengths.

MALDI-QQTOFMS Analysis of Tryptic Fragments

The tryptic digest of 2-mercaptoethanol-reacted YiaJ protein was prepared for MALDI-QQTOFMS analysis using 2,5-Dihydroxybenzoic acid (DHB) as the matrix where an equal volume (0.6 μL) of sample and matrix were mixed. The following mass spectrum of tryptic fragments was obtained.

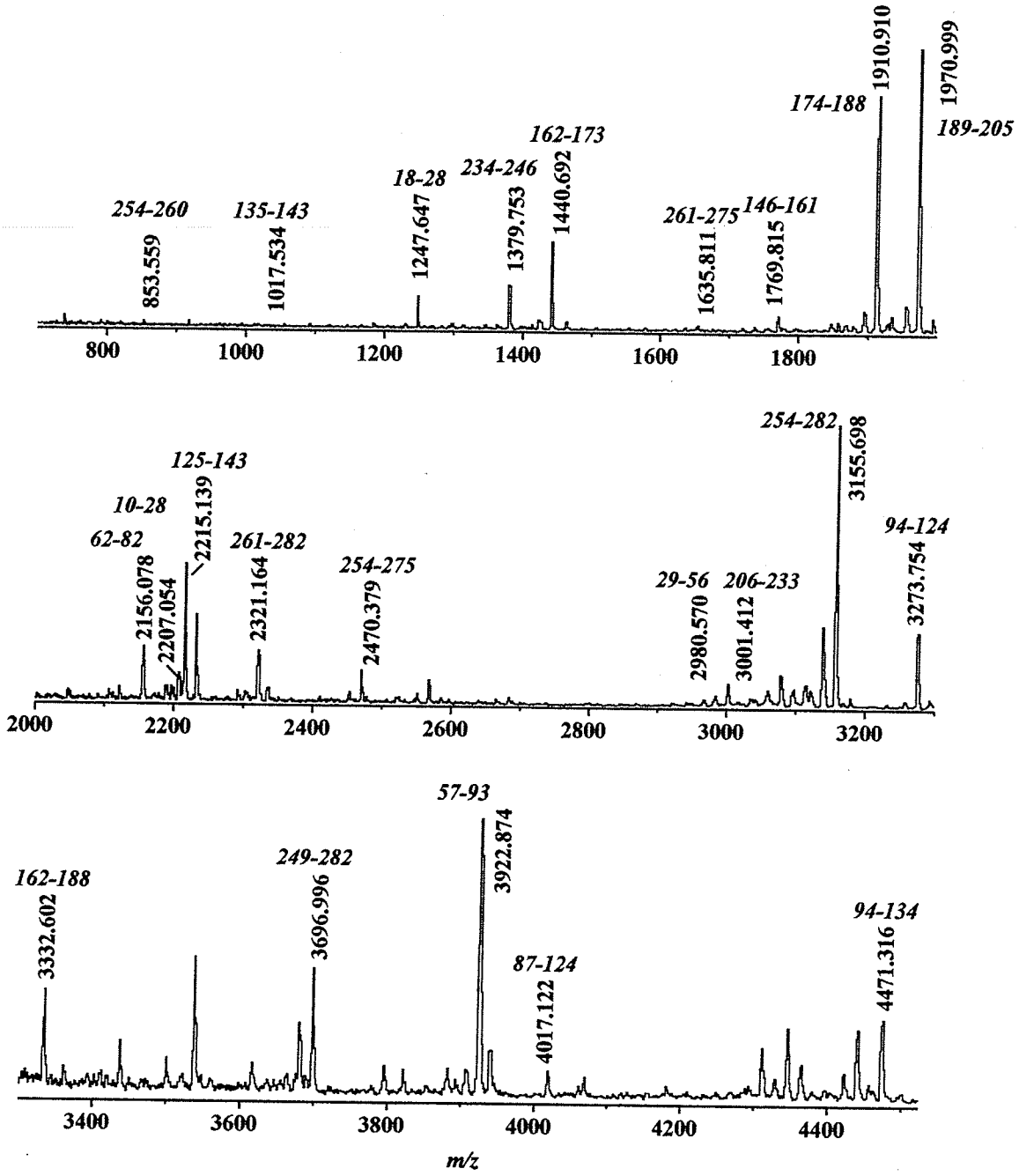


FIGURE 3.18: MALDI-QQTOFMS Mass Spectrum of Trypsin-Digested Covalently Modified YiaJ Protein. YiaJ contains one reactive cysteine which was covalently modified upon incubating the protein in 2-mercaptoethanol. To identify this particular cysteine residue, the covalently modified protein was digested with trypsin and then subjected to MALDI-QQTOFMS analysis.

The tryptic fragments detected by the mass spectrometer are summarized in TABLE 3.4.

TABLE 3.4: Tryptic Fragments Detected with MALDI-QQTOFMS

Tryptic Digest	Peptide Sequence	Mass (Da)
254-260	NLLKPLR	853.559
135-143	LEPTTGMLR	1017.534
18-28	ERPAGSQSLFR	1247.647
234-246	VPYAVSISLSTSR	1379.753
162-173	IYMAFGHPDYVK	1440.692
261-275	ETAQAISNELGFTVR	1635.811
146-161	AYIGQHMPLYCSAMGK	1769.815
174-188	SYWESHQHEIQPLTR	1910.910
189-205	NTITELPAMFDELAHIR	1970.999
62-82	LLQGLQSCGYVTTAPAAGSYR	2156.078
10-28	ENEMAQEKERPAGSQSLFR	2207.054
125-143	EDDHAILIYKLEPTTGMLR	2215.139
261-282	ETAQAISNELGFTVRDDLGAIT	2321.164
254-275	NLLKPLRETAQAISNELGFTVR	2470.379
29-56	GLMLIEILSNYPNGCPLAHLSELAGLNK	2980.570
206-233	ESGAAMDREENELGVSCIAVPVFDIHGR	3001.412
254-282	NLLKPLRETAQAISNELGFTVRDDLGAIT	3155.698
94-124	ALSSLNIIHIAAPHLEALNIATGETINFSSR	3273.754
162-188	IYMAFGHPDYVKSYWESHQHEIQPLTR	3332.602
249-282	QVGEKNLLKPLRETAQAISNELGFTVRDDLGAIT	3696.996
57-93	STVHRLLQGLQSCGYVTTAPAAGSYRLTTKFIAVGQK	3922.874
87-124	FIAVGQKALSSLNIIHIAAPHLEALNIATGETINFSSR	4017.122
94-134	ALSSLNIIHIAAPHLEALNIATGETINFSSREDDHAILIYK	4471.316

To identify the reactive cysteine in YiaJ, a fragment containing a modified cysteine would be detected as a peptide with an increase of 76 mass units. Upon focusing in on the m/z region between 2150 and 2240 from FIGURE 3.18 above, two peaks can be seen with masses of 2156 Da and 2232 Da (a mass difference of 76 mass units!).

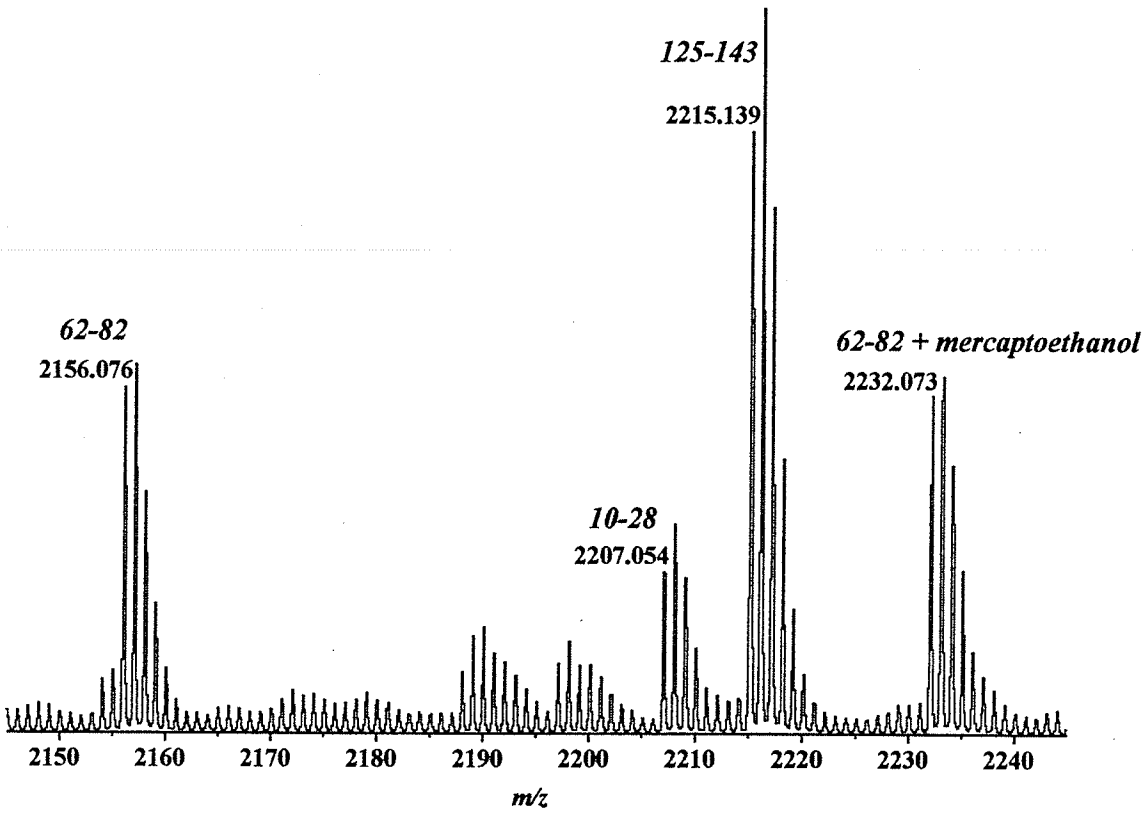


FIGURE 3.19: Locating the Tryptic Fragment Containing the Reactive Cysteine in YiaJ Protein. The m/z region between 2150 and 2240 from FIGURE 3.18 is examined with closer inspection. Two peaks are of particular interest where one has a mass of 2156.076 Da (corresponds to the tryptic fragment containing residues 62-82 of the primary amino acid sequence of YiaJ) and the other has a mass of 2232 Da (a difference of exactly 76 mass units).

The two ions corresponding to masses of 2156 Da and 2232 Da were then selected and individually analyzed using MS/MS to determine the exact position of the modified cysteine residue in the 62-82 tryptic fragment.

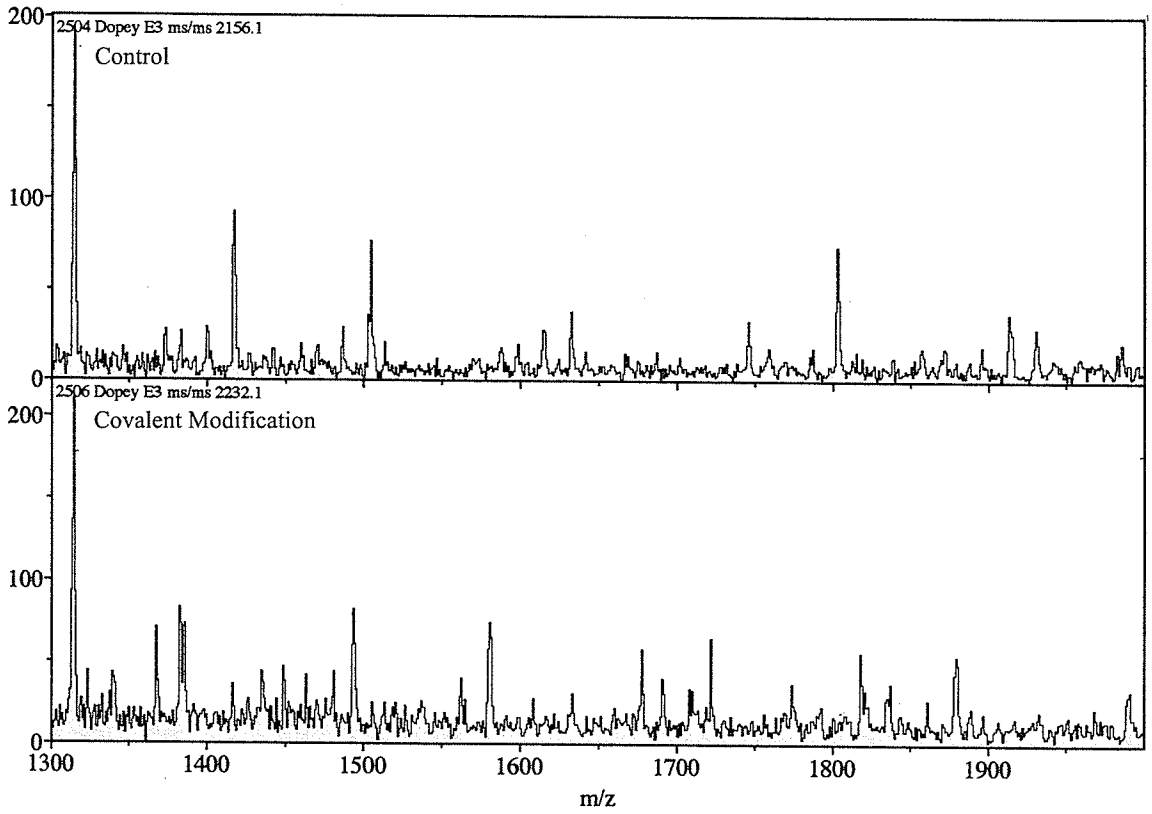


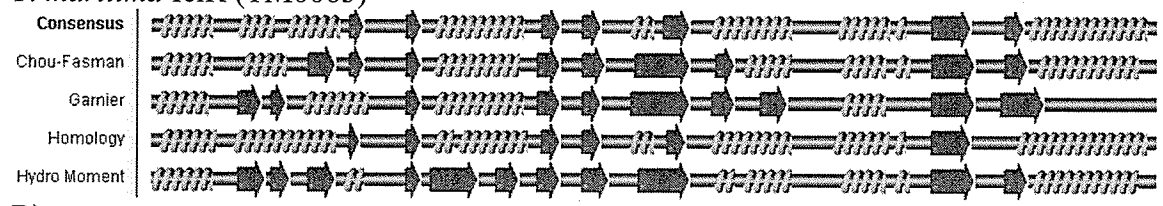
FIGURE 3.20: MS/MS of Selected Ions Identifying the Location of the Reactive Cysteine Residue in YiaJ Protein. After selecting ions from corresponding to masses of 2156.1 and 2232.1 Da (refer to FIGURE 3.19), MS/MS was performed on each ion to identify the sequence of peptide containing the modified cysteine residue.

Hence, the partial sequencing by MS/MS verifies that of the four cysteines in the primary structure of YiaJ protein (cysteine-43, cysteine-69, cysteine-156, and cysteine-222), only cysteine-69 is located at a reactive surface position.

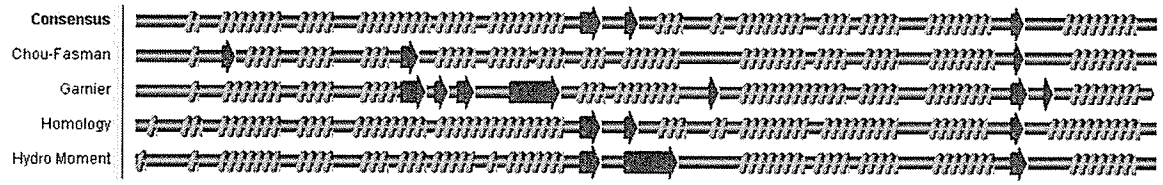
3.7 Secondary Structural Predictions

Using the software program PepTool version 2.0, the following secondary structural predictions were made for *T. maritima* IclR, *E.coli*: IclR, YiaJ, and YagI proteins.

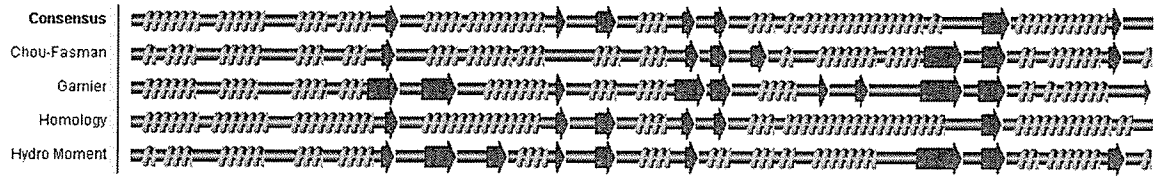
A).
T. maritima IclR (TM0065)



B).
E. coli IclR



C).
E. coli YiaJ



D).
E. coli YagI

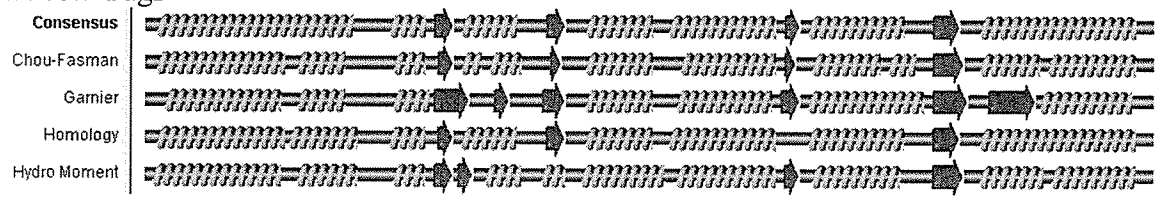


FIGURE 3.21: Secondary Structural Predictions for A) *T. maritima* IclR, B) *E. coli* IclR, C) *E. coli* YiaJ and D) *E. coli* YagI Proteins. The software PepTool v.2.0 (BioTools Incorporated) was used to make secondary structural predictions of the above-named proteins. PepTool uses a combination of popular secondary structural prediction algorithms to generate a consensus sequence. Straight bars in grey represent regions of random coil, orange coils represent regions of alpha helices and arrows in blue represent regions of beta strand.

Of the four IclR-type proteins listed above, only *E. coli* YiaJ and *T. maritima* IclR bare a resemblance with respect to secondary structural organization. However, according to sequence alignments, the identity between these two proteins is low.

TM0065 IclR vs. IclR (27% Identity)

```
Identities = 69/248 (27%), Positives = 146/248 (58%), Gaps = 3/248 (1%)

TM0065:1   MNTLKKAFEILDFIVKNPGDVSVSEIAEKFNMSVSNAYKYMVVLEEKGFVLRKKD-KRYV 59
           + +L +  ++L++I ++ G V+++E+A++  +  S  ++ +  ++++GFV +  +   +
IclR:  37   VQSLTRGLKLLEWIAESNGSVALTELAQQAGLPNSTTHRLLTTMQQQGFVRQVGELGHWA 96

TM0065:60  PGYKLIEYGSFVLRRFNIRDIAHDHLVDIMKRTGETVHLILKD--GFEGVYIDKVEGEQS 117
            G      GS  L+  N+  I H  L ++M+ +GETV++ + D    E + ID+V+
IclR:  97   IGAHAFMVGSSFLQSRNLLAIVHPILRNLMEESGETVNMAVLDQSDHEAIIIDQVQCTHL 156

TM0065:118 IPMVSRLGMKVDLYSTASGKSILAFVPEKELKEYLKIVELKPKTPNTITNPRVLKRELEK 177
           + M + +G K+ ++++ +GK+ LA + E+++ + L    L   T  T+ +P  LK +L +
IclR:  157 MRMSAPIGGKLPMHASGAGKAFLAQLSEEQVTKLLHRKGLHAYTHATLVSPVHLKEDLAQ 216

TM0065:178 IRKRGYAVDNEENEIGIMCVGVPIFDHNGYPVAGVSISGVARKFTEEKIEEYSDVLKEKA 237
            RKRGY+ D+EE+ +G+ C+   IFD +  P A +SISG   + T++++ E+  ++ + A
IclR:  217 TRKRGYSFDDEEHALGLRCLAACIFDEHREPFAAISISGPISRITDDRVTEFGAMVIKAA 276

TM0065:238 EEISRKLG 245
           +E++   G
IclR:  277 KEVTLAYG 284
```

FIGURE 3.22: BLAST Alignment of *T. maritima* IclR and *E. coli* IclR.

TM0065 vs. YagI (31% Identity)

```
Identities = 78/247 (31%), Positives = 150/247 (60%), Gaps = 1/247 (0%)

TM0065:1   MNTLKKAFEILDFIVKNPGDVSVSEIAEKFNMSVSNAYKYMVVLEEKGFVLRK-KDKRYV 59
           + ++++A +ILD   +   ++ +++I++   +S S  +  +  L+  G++ +  ++ +Y
YagI:  4   IQSVERALQILDLFNEQATELKITDISKLMGLSKSTLHSLLKTLQLHGYIDQNPENGKYR 63

TM0065:60  PGYKLIEYGSFVLRRFNIRDIAHDHLVDIMKRTGETVHLILKDGFEGVYIDKVEGEQSIP 119
            G KL+E G FV+   +IR  A   L ++ +RTG+T HL + DG EGVYI+K+EG+ +
YagI:  64  LGMKLVERGHFVVGSIDIRQKAKGWLTELSRRTGQTTHLGILDGREGVYIEKIEGKLAAI 123

TM0065:120 MVSRLGMKVDLYSTASGKSILAFVPEKELKEYLKIVELKPKTPNTITNPRVLKRELEKIR 179
             SR+G ++ +++TA GK ++A++ E EL   L+  +    TP T+ +   L   L + R
YagI:  124 AYSRIGRRLPVHATAIGKVLIAWLGEAELNALLEGYQYTTFTPATLASREALMSALAQTR 183

TM0065:180 KRGYAVDNEENEIGIMCVGVPIFDHNGYPVAGVSISGVARKFTEEKIEEYSDVLKEKAEE 239
           ++GYA+D+EENE G+ CV VP+++H    +A +S+S +  +  + ++  + + L++
YagI:  184 EQGYALDSEENEQGVRCVAVPVWNHESRVIAALSLSTLTSRVDDAELANFREQLQQAGLA 243

TM0065:240 ISRKLGY 246
           +SR LGY
YagI:  244 LSRALGY 250
```

FIGURE 3.23: BLAST Alignment of *T. maritima* IclR and *E. coli* YagI

TM0065 vs. YiaJ (28% Identity)

```
Identities = 71/248 (28%), Positives = 120/248 (47%), Gaps = 4/248 (1%)

TM0065:3   TLKKAFEILDFIVKNPGDVSVSEIAEKFNMSVSNAYKYMVVLEEKGFVLRKKDK-RYVPG 61
           +L +   +++ +   P    ++ ++E   ++ S  ++ +  L+  G+V        Y
YiaJ:  25  SLFRGLMLIEILSNYPNGCPLAHLSELAGLNKSTVHRLLQGLQSCGYVTTAPAAGSYRLT 84

TM0065:62  YKLIEYGSFVLRRFNIRDIAHDHLVDIMKRTGETVHLILKDGFEGVYIDKVEGEQSIPMV 121
            K I  G   L   NI  IA  HL  +   TGET++   ++    + I K+E    +
YiaJ:  85  TKFIAVGQKALSSLNIIHIAAPHLEALNIATGETINFSSREDDHAILIYKLEPTTGMLRT 144

TM0065:122 -SRLGMKVDLYSTASGKSILAFVPEKELKEYLKI--VELKPKTPNTITNPRVLKRELEKI 178
            + +G  + LY +A GK  +AF     +K Y +    E++P T NTIT    +  EL  I
YiaJ:  145 RAYIGQHMPLYCSAMGKIYMAFGHPDYVKSYWESHQHEIQPLTRNTITELPAMFDELAHI 204

TM0065:179 RKRGYAVDNEENEIGIMCVGVPIFDHNGYPVAGVSISGVARKFTEEKIEEYSDVLKEKAE 238
           R+ G A+D EENE+G+ C+ VP+FD +G     VSIS    +  +   +     L+E A+
YiaJ:  205 RESGAAMDREENELGVSCIAVPVFDIHGRVPYAVSISLSTSRLKQVGEKNLLKPLRETAQ 264

TM0065:239 EISRKLGY 246
            IS +LG+
YiaJ:  265 AISNELGF 272
```

FIGURE 3.24: BLAST Alignment of *T. maritima* IclR and *E. coli* YiaJ

The Crystal Structure of *Thermotoga maritima* IclR (TM0065)

Although the sequence identity between *T. maritima* IclR and the selected *E. coli* proteins is a mere 27% for *E. coli* IclR and YiaJ and 31% for *E. coli* YagI, *T. maritima* IclR is the only IclR family member that has been successfully studied by the method of x-ray crystallography and hence our only basis for structural modeling.

Using the Protein Explorer software interface to study the crystallographic data for the *T. maritima* IclR protein, hydrogen bond distances were measured between the atoms of the participating residues. The primary sequence of the *T. maritima* IclR protein was then organized into elements of secondary structure, i.e., alpha helices and beta-strands. Ten alpha helices (A-J) and 8 beta strands in two antiparallel beta sheets (A&B) exist in one chain of the *T. maritima* IclR protein.

TABLE 3.5: Elements of Secondary Structure in *T. maritima* IclR

Start-End Residues	Sequence	Secondary Structure	ID	
Thr3-Lys16	**TLKKAFEILDFIVK**	α-helix	Helix A	
Val23-Lys29	**VSEIAEK**	α-helix	Helix B	
Val34-Glu46	**VSNAYKYMVVLEE**	α-helix	Helix C	
Val50-Arg52	**VLR**	β-strand	Strand 1	Sheet A
Tyr58-Pro60	**YVP**	β-strand	Strand 2	
Lys63-Arg74	**KLIEYGSFVLRR**	α-helix	Helix D	
Ile77-Lys90	**IRDIAHDHLVDIMK**	α-helix	Helix E	
Val96-Lys101	**VHLILK**	β-strand	Strand 3	Sheet B
Glu105-Val112	**EGVYIDKV**	β-strand	Strand 4	
Lys127-Asp129	**KVD**	β-strand	Strand 5	
Ala134-Phe142	**ASGKSILAF**	α-helix	Helix F	
Glu145-Ile154	**EKELKEYLKI**	α-helix	Helix G	
Pro166-Arg181	**PRVLKRELEKIRKR**	α-helix	Helix H	
Ala184-Asn187	**AVDN**	β-strand	Strand 6	Sheet B
Ile194-Phe202	**IMCVGVPIF**	β-strand	Strand 7	
Pro208-Val217	**PVAGVSISGV**	β-strand	Strand 8	
Ala218-Lys220	**ARK**	α-helix	Helix I	
Glu223-Lys243	**EEKIEEYSDVLKEKAEEISRK**	α-helix	Helix J	

FIGURE 3.25 outlines the secondary structural organization of the *T. maritima* IclR protein.

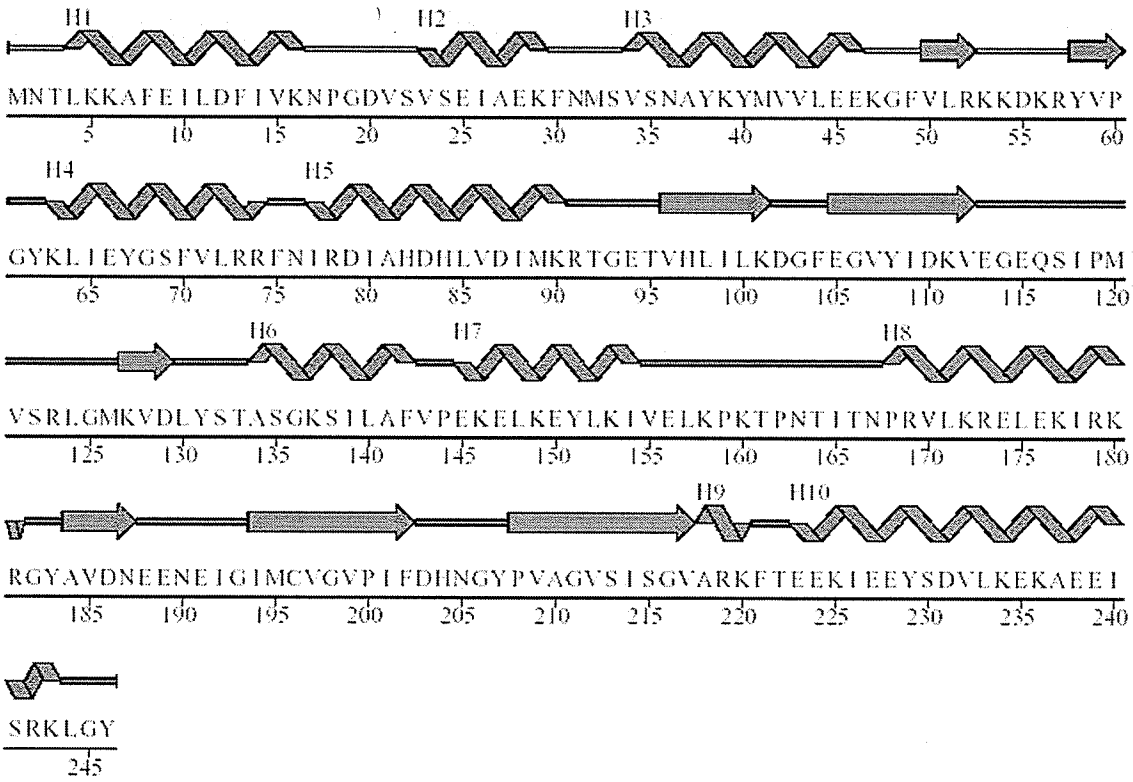


FIGURE 3.25: Secondary Structural Organization of the *T. maritima* IclR Protein. In this figure, the primary structure is translated into elements of secondary structure i.e. alpha helices and beta-strands (PDB ID: 1mkm; http://www.pdb.org)

Upon comparison of FIGURE 3.25 with FIGURE 3.21 A) above, it can be seen that PepTool accurately predicts the secondary structure of *T. maritima* IclR protein. The actual results of the PepTool prediction are appreciably accurate as shown in TABLE 3.6.

TABLE 3.6: Comparison of Predicted and Actual Secondary Structure in *T.maritima* IclR

PepTool Sequence	Predicted Secondary Structure	Published Sequence	Published Secondary Structure
Thr3-Val15	α-Helix A	Thr3-Lys16	Helix A
Ser22-Phe30	α-Helix B	Val23-Lys29	Helix B
Val34-Glu46	α-Helix C	Val34-Glu46	Helix C
Gly48-Arg52	β-Strand 1	Val50-Arg52	Strand 1
Lys63-Glu66	β-Strand 2	Tyr58-Pro60	Strand 2
Phe70-Arg91	α-Helix D	Lys63-Arg74 Ile77-Lys90	Helix D Helix E
Val96-Leu100	β-Strand 3	Val96-Lys101	Strand 3
Glu105-Asp110	β-Strand 4	Glu105-Val112	Strand 4
Ile118-Arg123	α-Helix E	No Match	No Match
Met126-Ser132	β-Strand 5	Lys127-Asp129	Strand 5
Ser138-Leu157	α-Helix F	Ala134-Phe142 Glu145-Ile154	Helix F Helix G
Pro168-Arg181	α-Helix G	Pro166-Arg181	Helix H
Tyr183-Val185	α-Helix H	Ala184-Asn187	Strand 6
Ile192-Ile201	β-Strand 6	Ile194-Phe202	Strand 7
Ala210-Ile214	β-Strand 7	Pro208-Val217	Strand 8
Gly216-Gly245	α-Helix I	Ala218-Lys220 Glu223-Lys243	Helix I Helix J

The results in TABLE 3.6 illustrate how accurately PepTool has predicted the secondary structural organization of the *T. maritima* IclR protein. The only discrepancies are a) a lack of secondary structure in residues 118-123 and b) the presence of residues 184-187 in a beta strand rather than in an alpha helix. The methods of secondary structural prediction used in PepTool clearly indicate that it is a reliable resource at least for the *T. maritima* IclR protein. However, the results of the secondary structural predictions for *E. coli* IclR, YagI and YiaJ in FIGURE 3.21 should not be discounted. As previously mentioned, it can be seen that *E. coli* YiaJ and the *T. maritima* IclR protein bare a similar secondary structural organization. Of course the only way to know for certain would be to obtain a crystal structure for *E. coli* YiaJ but for now, we have a model for which structural comparisons may be made with some assurance.

Chapter IV:

Discussion

4.0 YagI and YiaJ: Targeted *E. coli* IclR Family Members

In this thesis investigation, two members from the isocitrate lyase regulator (IclR) family of proteins were selected for isolation and purification in an attempt to gain a further understanding of the structural and functional characteristics of the IclR family of transcriptional regulators as a whole. Nine members of the IclR family come from *Escherichia coli* (1 from *E. coli* CFT073 and 8 from *E. coli* K-12 MG1655) alone. The two proteins selected for study, YagI and YiaJ, come from *E. coli* K-12 MG1655. Their amino acid compositions are outlined in the following table.

TABLE 4.0: Amino Acid Composition of YagI and YiaJ Proteins

Amino Acid	No. in YagI	% in YagI	No. in YiaJ	%in YiaJ
Ala A	26	10.32%	26	9.22%
Arg R	17	6.75%	14	4.96%
Asn N	7	2.78%	11	3.90%
Asp D	8	3.17%	8	2.84%
Cys C	1	0.40%	4	1.42%
Glu E	20	7.94%	22	7.80%
Gln Q	14	5.56%	10	3.55%
Gly G	20	7.94%	21	7.45%
His H	6	2.38%	11	3.90%
Ile I	16	6.35%	20	7.09%
Leu L	36	14.29%	31	10.99%
Lys K	11	4.37%	13	4.61%
Met M	4	1.59%	10	3.55%
Phe F	4	1.59%	7	2.48%
Pro P	6	2.38%	13	4.61%
Ser S	16	6.35%	21	7.45%
Thr T	16	6.35%	17	6.03%
Trp W	3	1.19%	1	0.35%
Tyr Y	8	3.17%	10	3.55%
Val V	13	5.16%	12	4.26%
Acidic	34		41	
Basic	37		41	
Polar	124		141	
Non-Polar	128		141	
TOTAL	**252**		**282**	

4.1 Recombinant DNA Production

With the entire *E. coli* K-12 MG1655 sequence available, it was a straightforward task to design artificial PCR primers for the direct amplification of the open reading frames of *yag*I (section 25/400; GenBank nucleotide sequence database accession no. AE000135) and *yia*J (section 325/400; GenBank accession number AE000435) from the genomic DNA of *E. coli* K-12 MG1655. Two PCR primers (forward & reverse) were designed for each target amplicon, but in addition to typical PCR primer design parameters (Tm, [$MgCl_2$], primer length, etc.) each primer was designed with a unique restriction endonuclease recognition site, not already present in the target DNA sequence to be amplified. This modification facilitated cloning of the artificial genes into expression vectors as well as a higher degree of stringency for recombinant selection.

Using traditional molecular biology techniques of cloning (restriction endonuclease digestion, alkaline phosphatase treatment, ligation, etc.) we were able to insert our synthetic PCR products into selected pET-vectors™ (pET-24b+ for *yag*I, pET-28b+ for *yia*J). The selected pET-vectors contained a kanamycin resistance cassette and recombinant DNA was first transformed into competent Novablue cells (tetracycline resistance). Recombinants were then selected on kanamycin/tetracycline LB-agar plates, and plasmid DNA was isolated from selected recombinants using a combination of alkaline lysis and QIAgen filter columns.

Automated DNA sequencing of the plasmids revealed that our PCR amplification under selected conditions was successful for both target genes with no insertions, deletions or mismatched base pairs. Once the correct DNA sequence was verified, we

then transformed the recombinant DNA into an expression host cell line known as BL21 (DE3) pLysS.

4.2 Protein Isolation and Purification

The selected preparation/storage buffer and purification method for *yag*I and *yia*J proteins were based on previous successful isolation of IclR and GclR proteins in our laboratory (Donald et al., 1996; Donald et al., 2001).

BioREX-70 Cation Exchange Chromatography

Using the pET expression system™, the cloned *yagI* and *yiaJ* genes were induced using 1 mM IPTG once host BL21(DE3) pLysS cells were grown to mid-log phase. The cells were then harvested, flash frozen using liquid nitrogen and then thawed with buffer. This freeze-thawing breaks the cell membrane and causes the release of cell contents including the T7 lysozyme product of the pLysS. The crude extract was then first applied to a BioREX-70 cation exchange resin. As the YagI and YiaJ proteins were expected to be DNA-binding proteins with a greater proportion of basic amino acid residues, the BioREX-70 column indeed proved to be a very effective measure of protein purification evidenced from samples of eluate examined under SDS-PAGE.

Buffer Conditions

Unfortunately, a recurring problem of the IclR-type proteins examined our laboratory is one of low solubility. Even with high salt concentrations of 0.5 M KCl, YagI and YiaJ proteins could not be concentrated beyond 2 mg/mL. After experimenting with different salt and Tris-HCl concentrations as well as trying NaCl in place of KCl, we found that a buffer of 20 mM Tris-HCl, 0.35 M KCl, 1 mM EDTA, pH 7.8 was best for a working protein concentration of 1 mg/mL.

Sephadex G-100 Size-Exclusion Chromatography

With an already high degree of protein purification, the concentrated eluate from the BioREX-70 chromatography was then applied to a Sephadex G-100 size-exclusion column. Following this step, we estimated that protein purity was >90% as judged by SDS-PAGE.

The purified YagI and YiaJ proteins were then concentrated using vacuum dialysis (Amicon pressure concentration was attempted, but protein rapidly precipitated out of solution) to working concentrations of 1 mg/mL.

4.3 Mass Spectrometry

ESI-TOF MS

Automated DNA sequencing provided a direct way of verifying the integrity of our PCR and recombinant DNA constructs. The integrity of our YagI and YiaJ proteins was verified using mass spectrometry. In particular, Dr. Lynda J. Donald, who graciously performed all mass spectrometry experimentation presented or mentioned in this thesis, prepared 2-3 nmole 100 μL pure protein samples dialyzed against at least three changes of 1 L of Nanopure™ water. All protein in the sample was recovered by direct addition of 2-5 μL of acetic acid and then diluting to 20 μM monomer and 50% methanol (v/v). These denatured protein samples were then characterized with electrospray ionization time-of-flight mass spectrometry (ESI-TOF MS). This method of mass spectral analysis is extremely beneficial as protein samples could be analyzed directly in buffer regardless of the high salt concentrations in our buffers. Analysis of both proteins in this way revealed that we had indeed isolated and purified *E. coli* YagI

and YiaJ proteins with the only post-translational modification being removal of the N-terminal methionine in both cases (a very common occurrence in *E. coli* proteins).

MALDI-TOF MS

In a separate experiment, it was found that YiaJ protein was able to react with DTNB in a time-dependent manner (YagI failed to undergo reaction with DTNB under any condition). This implies that one of the four cysteines in YiaJ (cysteine-43, cysteine-69, cysteine-156, or cysteine-222) is reactive or surface-exposed.

To identify which cysteine residue had undergone chemical modification with DTNB, tryptic digests of the chemically-modified YiaJ protein were subjected to matrix assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS). Using MS/MS, it was found that only cysteine-69 is located at a reactive surface position.

4.4 Molar Absorptivity

Following successful isolation and purification of YagI and YiaJ proteins, our next approach to characterization of said proteins was to determine the molar extinction coefficients, ε for both proteins via the method of Edelhoch (Edelhoch, 1967; Gill and von Hippel, 1989). Using the respective molar extinction coefficients for the following chromophores, we were able to determine molar extinction coefficients for YagI and YiaJ proteins comparable to predicted values.

TABLE 4.1 Molar Extinction Coefficients for Protein Chromophores

Chromophore	Absorption Wavelength			
	280	288	295	300
Tyrosine	1280 $M^{-1}cm^{-1}$	385 $M^{-1}cm^{-1}$	2480 $M^{-1}cm^{-1}$	2270 $M^{-1}cm^{-1}$
Tryptophan	5690 $M^{-1}cm^{-1}$	4815 $M^{-1}cm^{-1}$		
Cystine	120 $M^{-1}cm^{-1}$	70 $M^{-1}cm^{-1}$		

The molar extinction coefficients determined experimentally for YagI and YiaJ protein are presented in the following table along with other experimental vs. theoretical results.

TABLE 4.2 Theoretical vs. Experimental YagI and YiaJ

Protein	No. of Amino Acids	Theoretical pI	Theoretical Mass (Da)	Experimental Mass (Da)	Theoretical ε (M^{-1} cm^{-1})	Experimental ε (M^{-1} cm^{-1})
YagI	252/251*	7.03/7.48*	27837.9/ 27706.7*	27706	27310	31076
YiaJ	282/281*	6.42/6.44*	31066.6/ 30935.4*	30934	18610	20935

* following removal of N-terminal methionine

From the results of the preceding table, paying particular attention to the theoretical mass vs. the experimentally determined mass (via mass spectrometry), both the YagI and YiaJ proteins were successfully produced by our methods. As previously mentioned, the only post-translational modification resulting from the isolation is the removal of the N-terminal methionine in both cases.

4.5 Are YagI and YiaJ HTH DNA-Binding Proteins?

Unfortunately, the IclR family members typically have poor solubility and to date only the TM0065 protein of *Thermotoga maritima* has been successfully crystallized for X-ray crystallography. Although TM0065 is an IclR family member and its predicted helix-turn-helix motif (Dodd & Egan, 1990) was fairly conserved in what was found experimentally (refer to section IV), the alignment of this HTH motif is poor with other predicted HTH motifs. Disregarding this one anomaly, several IclR family members have been found to show a strong degree of conservation particularly between residues predicted to be involved in the HTH motif.

The following representative figure illustrates a strong degree of conservation between the predicted HTH motifs of different IclR family members with YagI and YiaJ.

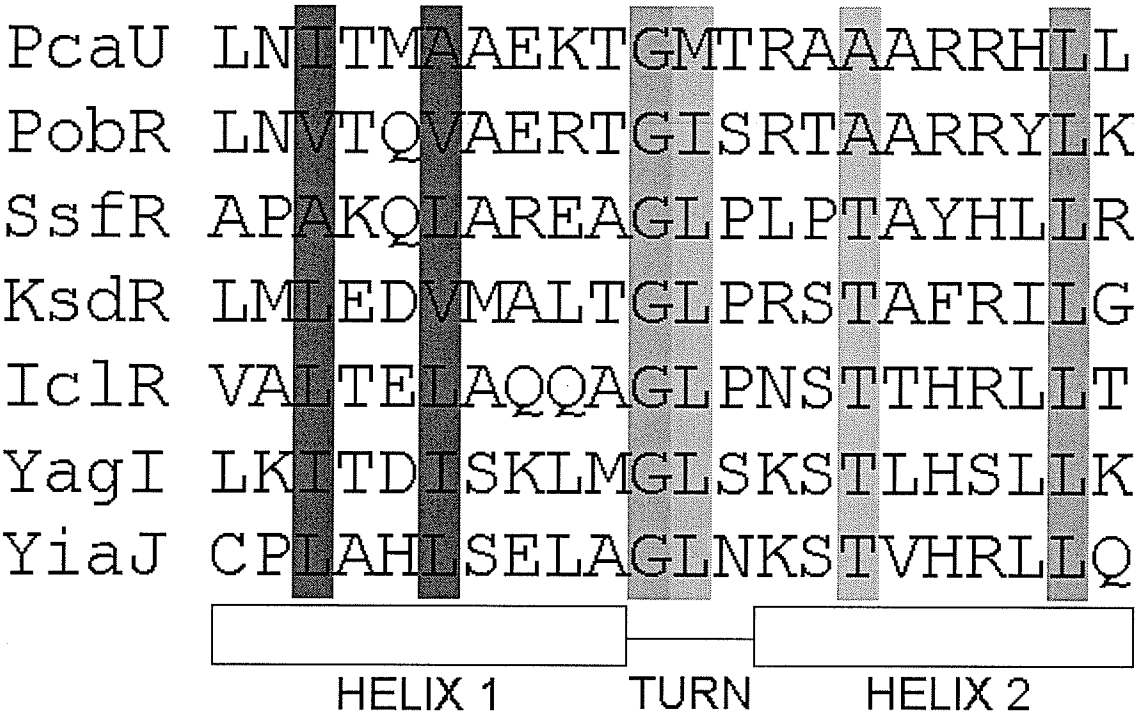


FIGURE 4.0: Putative HTH Motif and Comparison with other DNA-binding Proteins. For each protein, the 22-amino-acid alignment that constitutes the HTH motif is presented. The DNA-binding proteins which are selected for this alignment with *E. coli* YagI and YiaJ are PobR, the transcriptional activator of the *Acinetobacter calcoaceticus pobA* gene, IclR, the repressor of the *E. coli aceBAK* operon, PcaU, a regulatory protein of *Acinetobacter* sp. strain ADP1, KsdR, a transcriptional regulator of *Arthrobacter simplex* and SsfR, a sporulation protein from *Streptomyces griseus*. Amino acids that are absolutely conserved are highlighted in green, those that are strongly conserved are highlighted in cyan and those that are similar with respect to chemical identity are highlighted in red.

Although obtaining three-dimensional structures may be difficult for the IclR family members, determining the target DNA sequences of these regulatory proteins is not so difficult and has been done for IclR, GclR, YiaJ and a number of others (Donald et al, 1996; Pan et al 1996). In general, the IclR family members bind to their target DNA sequences as dimers, hence the recurring findings of palindromic sequences.

Obtaining important structural information for the IclR family members has proven to be very difficult. However, it is certainly as important if not more so, to determine the biochemical functionality of each member. This is the comprimise, for now we must accept uncertainty in structural information to obtain certainty in functional elucidation.

4.6 *Thermotoga maritima* IclR as a Model

When the results of the predicted secondary structure for *T. maritima* IclR (via PepTool version 2.0) were compared with the actual results published in the Brookhaven National Laboratories Protein Data Bank, the prediction proved to be extremely accurate. Hence, the validity of the predicted secondary structures of *E. coli* IclR, YagI and YiaJ proteins from PepTool were accepted without scrutiny.

The sequence similarities between *T. maritima* IclR and the selected *E. coli* proteins in this thesis investigation are low (27% for *E. coli* IclR and YiaJ and 31% for *E. coli* YagI). However, the secondary structural organization of *E. coli* YiaJ and *T. maritima* IclR appears to be quite similar. Despite this similarity, the tertiary and quaternary structures may in fact be quite different. The only way to know with certainty would be to obtain more structural information. Nevertheless, *T. maritima* IclR is the only IclR family member that has been studied by the method of x-ray crystallography and for this reason; it is the basis of structural comparisons with among all other IclR family members.

Dimeric *T. maritima* IclR

The *T. maritima* IclR protein was crystallized as a dimer of which each monomer consists of an N-terminal DNA-binding domain with a HTH motif and a C-

terminal putative signal-binding domain. These two domains are linked by an alpha helix and a short loop and the dimerization interface is formed exclusively by the DNA-binding domain (Zhang et al., 2002). In the dimer, the C-terminal domains do not contact each other.

FIGURE 4.1: The *T. maritima* IclR Dimer. The *T. maritima* IclR protein was crystallized as a dimer. The dimer contacts are formed entirely between the DNA-binding domains of each monomer. The C-terminal domains do not participate in dimer formation making for an asymmetric dimer (Zhang et al., 2002).

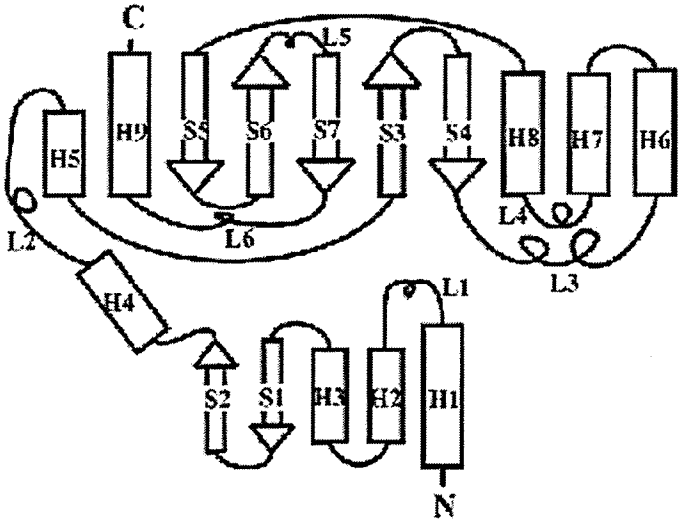


FIGURE 4.2: Topology of the *T. maritima* IclR Protein. Each monomer of *T. maritima* IclR consists of an N-terminal DNA-binding domain with a HTH motif and a C-terminal putative signal-binding domain. The HTH DNA-binding motif of the N-terminal domain is composed of two small, consecutive alpha-helices (H2 and H3 and the 4-residue turn connecting them). The HTH motif is followed by a short alpha-helix and a beta-hairpin (formed by S1&S2), which connects to the C-terminal alpha/beta domain by an alpha helix, H4 and a short loop, L2 (Zhang et al., 2002).

The dimerization interface comprises the alpha-helices H1 and H4. The two H1 alpha-helices from the N-terminal domains form a hydrophobic interface which includes four aromatic side chains (two from each monomer) and a number of additional hydrophobic interactions (e.g. Ala7-Leu4) and H-bonds (e.g.Asp12-Tyr67; Zhang et al., 2002).

The Putative Signal-Binding Pocket of *T. maritima* IclR

The C-terminal putative regulatory and signal-binding domain is composed of six beta-strands forming one anti-parallel beta-sheet, five alpha-helices, and one short 3_{10} helix. This beta-sheet is highly curved forming a half barrel that divides the five alpha-helices into two subgroups. Alpha-helices H4 and H5 lie on the outside of the curved sheet while alpha-helices H6, H7, and H8 lie on the inside, forming an $\alpha/\beta/\alpha$ structure resembling the triosephosphate isomerase (TIM) barrel (Pajudas & Palau, 1999; Wierenga, 2001; Zhang et al., 2002). The H6 alpha-helix inserts into the half TIM barrel and is parallel to the sheet surface. Helices H7 and H8 close one end of the half barrel. Two different perspectives of the *T. maritima* IclR monomer are provided in FIGURE 4.3.

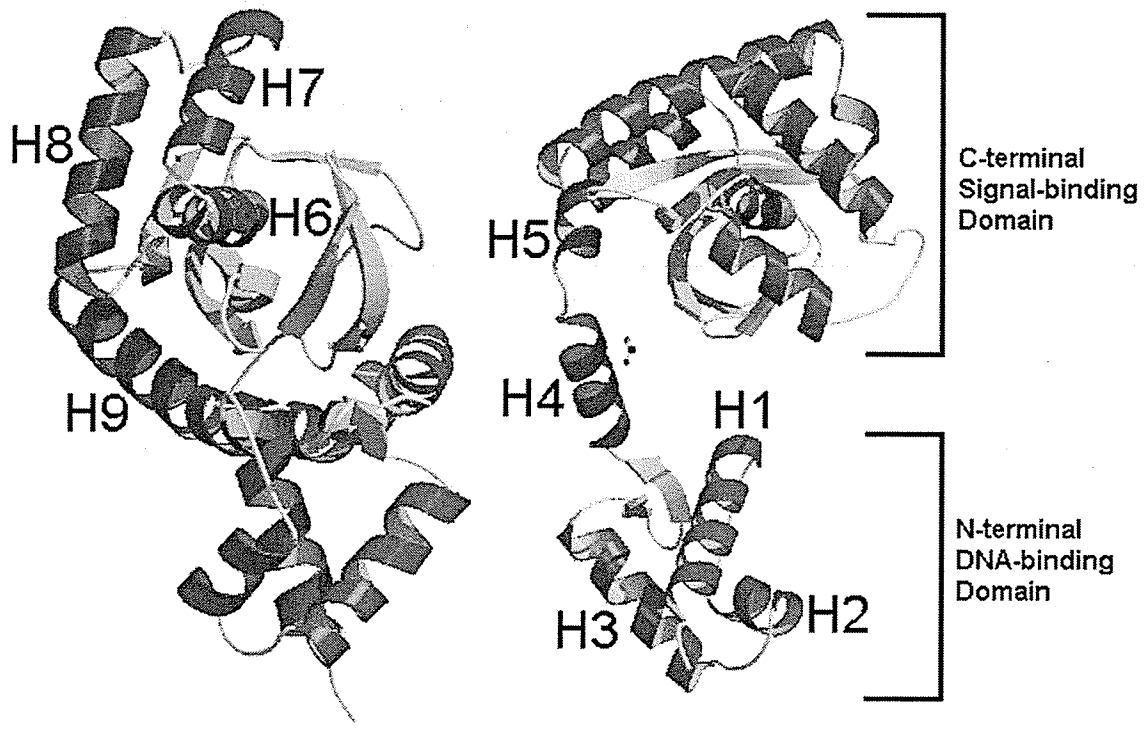


FIGURE 4.3: *T. maritima* IclR Monomers. The two different views of monomer illustrate the α/β/α half TIM barrel, C-terminal signal-binding and N-terminal DNA-binding domains reported by Zhang et al. (PDB ID: 1mkm; Zhang et al., 2002).

This half TIM barrel arrangement closely resembles the structure of an actin-binding protein known as profilin (Zhang et al., 2002). In each protein, there exists a pocket within the beta-sheet formed by the inner surface of the concave beta-sheet, a loop (L3) and a small alpha helix (H6) in *T. maritima* IclR (Zhang et al., 2002). The residues in PobR that mediate inducer (Kok et al., 1998) as well as residues that are reasonably conserved among IclR family members are found within this pocket (Zhang et al., 2002).

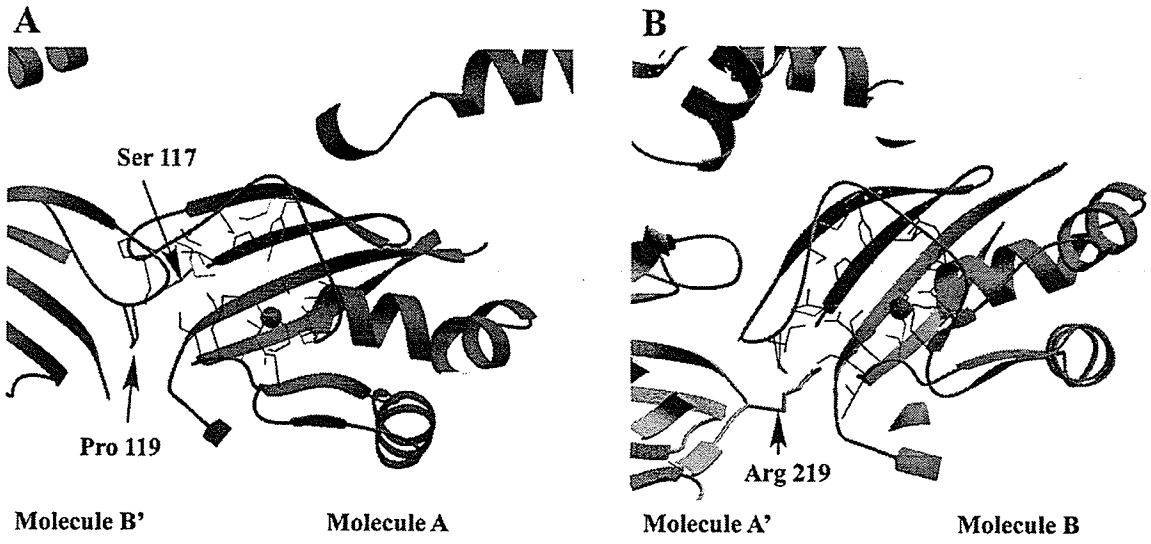


FIGURE 4.4: The Putative Signal-binding Pocket of *T. maritima* IclR. The putative signal-binding pockets are formed between the two signal-binding domains from two different molecules of *T. maritima* IclR, A and B' (*A*) (or molecules B and A' (*B*)).

T. maritima IclR Binds DNA as a Tetramer

Although the C-terminal ligand binding domains do not participate in dimer formation, these domains are available for interaction with corresponding domains in neighboring asymmetric dimers resulting in the formation of tetramers (Zhang et al., 2002). Interestingly enough, some of the residues found within the signal-binding region are those same residues, i.e. residues 114-118 and 214-220 that mediate tetramer formation in *T. maritima* IclR. The proximity of the putative signal-binding pocket to the region involved in tetramerization suggests that signal molecule binding and tetramerization may be linked.

The N-terminal DNA-binding domains do not participate in tetramer formation; hence tetramerization is exclusively due to C-terminal domain interaction.

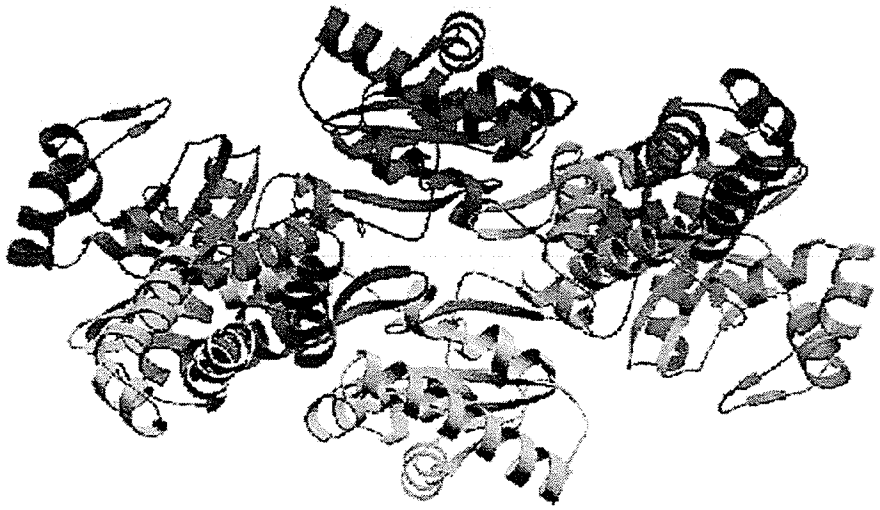


FIGURE 4.5: The *T. maritima* IclR Tetramer. The C-terminal ligand binding domains of *T. maritima* IclR monomers do not participate in dimer formation, but are responsible for the tetramerization of two identical asymmetric dimers. The tetramer forms by interaction between the signal domains of two asymmetric dimers. The N-terminal DNA-binding domains are not involved in tetramerization.

However, using size-exclusion chromatography, Zhang et al. report that *T. maritima* IclR exists as a dimer in solution and not a tetramer. It has been previously suggested (by way of mass spectrometry) that *E. coli* IclR and GclR repressor proteins exist as dimers and tetramers in solution (Donald et al., 2001). To explain the case of *T. maritima* IclR dimeric and tetrameric species, it has been hypothesized that *T. maritima* IclR exists as a dimer in the absence of signal molecule and target DNA, but adopts a tetrameric structure when bound to DNA in the absence of signal molecule (Zhang et al., 2002). In particular, it is believed that *T. maritima* IclR binds DNA weakly as a dimer in the absence of signal molecule. However, in the presence of two adjacent binding sites, *T. maritima* IclR cooperatively forms a high-affinity, stable tetrameric complex that represses transcription (Zhang et al., 2002). Binding of the as yet unknown signal molecule in the pocket then causes a disruption of the tetramer, decreased affinity for DNA, and increased transcription. Supporting evidence for this hypothesis has been

previously shown for *E. coli* IclR where the binding of ligands (e.g. phosphoenol pyruvate) to IclR inhibits DNA binding (Cortay et al., 1991).

Although the structure of the *T. maritima* IclR protein has been determined, it is only a homologue with an as yet undetermined function. Determining the biochemical role of this protein would still not explain the molecular mechanisms behind the signal-dependent regulation employed by the IclR family of homologues. It is rather daunting that a concise and well-received model explaining this signal-dependent transcriptional (de)repression still remains to be discovered. At best, it is understood that DNA-binding (and hence regulation) proceeds in a signal-dependent manner, where the binding of signal de-represses transcription by modulating either DNA binding, receptor multimerization, or the interaction of the repressor with the transcriptional machinery (Kok et al., 1998; Donald et al., 1996; Zhang et al., 2002). Even more discouraging is the fact that in such a large family of proteins, only one member has been studied with structural certainty. However, this lack of functional and structural detail presents many opportunities for research because the knowledge base for this family of transcriptional regulators remains unclear.

4.7 Future Endeavors

At the outset of this thesis investigation, *yagI* and *yiaJ* were annotated as hypothetical genes with corresponding hypothetical protein products. From our results both YagI and YiaJ proteins were isolated and purified allowing some early structural characterization of said proteins. The groups of Juan Aguilar and Josepha Badia have presented early results supporting the functional characterization of YiaJ as the transcriptional regulator for the *yia*KLMNOPQRS operon. Although a possible operon

has been identified for YagI, biochemical evidence is required to support the theory of the regulatory effect of YagI on this operon if any. A good start would be to determine the target DNA sequence of YagI using DNase protection experiments as well as the SELEX method of Turk and Gold (Turk & Gold, 1990).

Mutations in the primary structure and their effect on the function of YagI and YiaJ would prove to be very interesting as residue specificities could be determined for both DNA-binding ability as well as catabolite substrate binding. We believe that these proteins bind to DNA as a dimer using a HTH motif, hence it would be useful to test mutations (via site-directed mutagenesis) that would affect DNA-binding either by selecting residues within the HTH motif or residues responsible for dimerization. As previously mentioned, it has also been proposed that there is a C-terminal regulatory domain responsible for binding a catabolic substrate (Kok et al., 1998; Donald et al., 1996; Zhang et al., 2002). To determine whether this is true or not is an experiment unto itself, and would require one to perform mutations in this region.

Knockout experiments of YagI and YiaJ could also prove to be interesting with respect to the biochemical repercussions (or lack thereof) of mutants deficient/degenerate in said proteins.

Further insight into the three dimensional structures of YagI and YiaJ would definitely be important, but as these proteins have low solubility and require high salt concentrations for storage, the optimal conditions for preparing satisfactory crystals for x-ray crystallography may prove elusive for some time. Determining structural information for these proteins may not have to be left up to x-ray crystallography alone as studies

with nuclear magnetic resonance (NMR) spectroscopy could provide a solution. However, again such a technique would require higher protein solubility.

In short, there are a myriad of experiments and studies that can be concocted to solve the structure AND function of the YagI and YiaJ proteins. In this thesis investigation, our main goal was to purify and isolate *E. coli* YagI and YiaJ using a combination of molecular biology and protein chemistry techniques. We successfully purified both proteins by our methods and managed to present some early biochemical characterization, yet our knowledge remains incomplete. IclR family members as a whole remain weakly understood at best. Only by making a concerted effort into studying various members simultaneously can we shed more light on our understanding of this growing protein family.

References

Badia J., Ibanez E., Sabate M., Baldoma L., and Aguilar J. *J Biol Chem* 273:8376-8381. (1998)

Bailey, J.E.. UV absorption characteristics of N-acetyl methyl esters of the aromatic amino acids, cystine, and N-acetylcysteine. In: Sober HA, ed. *Handbook of Biochemistry.* Cleveland, Ohio: Chemical Rubber Co. p.B-18 (1968)

Bairoch, A. and. Apweiler, R. *Nucleic Acids Research*, Vol. 28, No. 1, pp. 45-48. (2000)

Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R. and Brass, A. *Bioinformatics*, 15: 510-520 (1999)

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M.A. *Nucleic Acids Research.* 28: 19-23 (2000)

Barker, W.C., Garavelli, J.S., Hou, Z., Huang, H., Ledley, R.S., McGarvey, P.B., Mewes, H-W., Orcutt, B.C., Pfeiffer, F., Tsugita, A., Vinayaka, C.R., Xiao, C., Yeh, L-S. L. and Wu, C. *Nucleic Acids Research.* 29: 29-32. (2001)

Barnes, W.M. *Gene* 112: 29-35 (1992)

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L.L. *Nucleic Acids Research*, 30(1):276-280. (2002)

Benson, D., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. *Nucleic Acids Research*, Vol. 28, No. 1, pp.15-18. (2000)

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. *Nucleic Acids Research*, 28 pp. 235-242 (2000)

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. *J. Mol. Biol.*, 112, 535–542. (1977)

Biemann, K., Biomed. *Env. Mass Spectrom.* 12, 99-111 (1988)

Biemann, K., *Annu. Rev. Biochem.* 61, 977-1010 (1992)

Blattner F.R., Plunkett G., 3rd, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y. *Science* 277:1453-1474. (1997)

Born, T.L. and Blanchard, J.S. *Curr Opin Chem Biol.* 5: 607-13 (1999)

Brenner S.E., Hubbard T., Murzin A., Chothia C. *Nature* 378:140. (1995)

Busby S and Ebright R.H. *Cell* 79:743–746 (1994)

Busby S and Ebright R.H. *J. Mol. Biol.* 293, 199-213 (1999)

Busch, K.L., Glish, G.L.and McLuckey, S.A *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry*, VCH Publishers, Inc., New York. (1988)

Campos, E., Aguilar, J., Baldoma, L., and Badia, J. *J Bacteriol* 184: 6065-6068. (2002)

Cariello, N.F., Swenberg, J.A., and Skopek, T.R. *Nucleic Acids Research.* 19:4193-4198. (1991)

Casari G., De Daruvar A., Sander C., Schneider R. *Trends Genet.* 12:244-245. (1996)

Chou, P.Y. and Fasman, G.D. *Biochemistry.* 13(2): 222-45. (1974)

Chou, P.Y. and Fasman, G.D. *Annu. Rev. of Biochem.* 47: 251-76. (1978)

Cline, J., Braman, J.C. and Hogrefe, H.H. *Nucleic Acids Research* 24, 3546–3551. (1996)

Cortay, J. C., Negre, D., Galinier, A., Duclos, B., Perriere, G., and Cozzone, A. J. *EMBO J.* 10, 675-679. (1991)

DiRusso, C.C., Metzger, A.K., Heimert, T.L. *Mol Microbiol.* Jan;7(2):311-22. (1993)

Dodd, I.B. and Egan, J.B. *Nucleic Acids Research* 18: 5019-5026. (1990)

Donald, L. J., Chernushevich, I. V., Zhou, J., Verentchikov, A., Poppe-Schriemer, N., Hosfield, D. J., Westmore, J. B., Ens, W., Duckworth, H. W., and Standing, K. G. *Protein Sci.* 5, 1613-1624. (1996)

Donald, L. J., Hosfield, D. J., Cuvelier, S. L., Ens, W., Standing, K. G., and Duckworth, H. W. *Protein Sci.* 10, 1370-1380. (2001)

Dubendorff, J.W. and Studier, F.W. *J. Mol. Biol.* 219, 45–59. (1991)

Eckert, K.A., and Kunkel, T.A. *Nucleic Acids Res*earch 18: 3739-3744. (1990)

Eckert, K.A., and Kunkel, T.A. *PCR Methods Appl.* 1: 17-24. (1991)

Edelhoch, H. *Biochemistry.* 6:1948-1954. (1967)

Egan ,S.E., Fliege, R., Tong, S., Shibata, A., Wolf, R.E. Jr., and Conway, T. *J Bacteriol* Jul;174(14):4638-46. (1992)

Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. *Journal of Molecular Biology.* 179(1): 125-42. (1984)

Ellman, G.L. Tissue sulfhydryl groups. *Arch Biochem Biophys* 82: 70-77. (1959)

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A., R Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., Venter, J. C. *Science*. 269(5223): 496-512. (1995)

Galinier A., Negre D., Cortay J.C., Marcandier S., Maloy S.R., Cozzone A.J. *Nucleic Acids Research*. 18:3656-3656. (1990)

Garnier, J., Osguthorpe, D.J., and Robson, B. *Journal of Molecular Biology*. 120(1): 97-120. (1978)

Gil, R., Silva, F.J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Holldobler, B., Van Ham, R.C., Gross, R., Moya, A. *Proc. Natl. Acad. Sci. U.S.A.* 100 (16), 9388-9393. (2003)

Gill, S.C. and von Hippel, P.H. *Analytical Biochem.* 182, 319-326. (1989)

Hanlon, N. and Barford, D. *Protein Sci.*, 7, 508-511. (1998)

Harrison, S.C. and Aggarwal, A.K. *Annu Rev Biochem.* 59: 933-969. (1990)

Henrissat,B. and Bairoch,A. *Biochem. J.* 316: 695-696. (1996)

Heyduk, T., Lee, J., Ebright, Y., Blatter, E., Zhou, Y. & Ebright, R. *Nature*, 364, 548-549. (1993)

Ibanez, E., Campos, E., Baldoma, L., Aguilar, J., Badia, J. *J Bacteriol* 182:4617-4624. (2000a)

Ibanez, E., Gimenez, R., Pedraza, T. Baldoma, L., Aguilar J., and Badia, J. *J Bacteriol* 182:4625-4627. (2000b)

Izard, T., Lawrence, M.C., Malby, R.L., Lilley, G.G., and Colman, P.M.; *Structure*; 2:361-369. (1994)

Joneson, T., White, M. A., Wigler, M. H., Bar-Sagi, D. *Science* 271, 810-812. (1996)

Keohavong, P. and Thilly, W.G. *Proc Natl Acad Sci USA* 86: 9253-9257. (1989)

Kolb, A., Igarashi, K., Ishihama, A., Lavigne, M., Buckle,M. & Buc, H. *Nucleic. Acids Research*. 21, 319-326. (1993)

Kong, H., Kucera, R.B., and Jack, W.E. *J Biol Chem* 268: 1965-1975. (1993)

Krek, W., Ewen, M.E., Shirodkar, S., Arany, Z., Kaelin, W.G. Jr & Livingstone, D.M. *Cell*, 78, 161-172. (1994)

Levin, J.M., Robson, B. and Garnier, J. *FEBS Letters*. 205(2): 303-8. (1986)

Ling, L.L., Keohavong, P., Dias, C., and Thilly, W.G. *PCR Methods Appl.* 1: 63-69. (1991)

Loboda, A.V., Krutchinsky, A.N., Bromirski, M., Ens, W. and Standing, K.G. *Rapid Commun. Mass Spectrom.* 14, 1047-1057. (2000)

Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A. & Mathur, E.J. *Gene* 108:1-6. (1991)

Malan, T., Kolb, A., Buc, H. and McClure, W. *J. Mol. Biol.* 180: 881-909. (1984)

Matilla, P., Korpela, J., Tenkanen, T., and Pitkanen, K. *Nucleic Acids Research*. 19:4967-4973. (1991)

Mirwaldt, C., Korndorfer, I., and Huber, R. *J Mol Biol*;246:227-239. (1995)

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich, H. *Cold Spring Harbor Symposia on Quantitative Biology* 51: 263-273. (1986)

Nasser, W., Reverchon, S., Condemine, G., and Robert-Baudouy, J. *J.Mol.Biol.* 236, 427-440. (1994)

Nishikawa, K. and Ooi, T. *Biochmica et Biophysica Acta*. 871(1): 45-54. (1986)

Noumura, K., Nasser, W., Kawagishi, H. & Tsuyumu, S., *Proc. Natl. Acad. Sci. USA* Vol. 95, pp. 14034–14039. (1998)

Novagen, pET System Manual 8th Edition, Novagen Literature (1999)

Oehler, S., Amouyal, M., Kolkhof, P., von Wilcken-Bergmann, B. and Muller-Hill, B. *EMBO J* 13: 3348-3355. (1994)

Pace, C.N. , Vajdos, F. , Fee, L. , Grimsley, G. and Gray, T. *Protein Sci.*, 4, 2411-2423. (1995)
Pan, B., Unnikrishnan, I., and LaPorte, D. C. *J Bacteriol*. 178(13), 3982-4. (1996)

Papayannopoulos, J. *Mass Spectrom. Rev.* 14, 49-73. (1995)

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. E. and Berman, H. M. *Journal of Molecular Biology*. 260(3): 395-408. (1996)

Pujadas, G. and Palau, J. *Biologia Bratislava*. 54(3): 231-254. (1999)

Ren, Y. L., Garges, S., Adhya, S. & Krakow, J. *Proc. Natl Acad. Sci. USA*, 85, 4138-4142. (1988)

Reverchon S., Nasser W., Robert-Baudouy, J. *Mol. Microbiol.* 5:2203-2216. (1991)

Roberts, S. G. E. & Green, M. R. *Nature* 371, 717-720. (1994)

Rooman, M.J. and Wodak, S.J. *Nature*. 335(6185): 45-9. (1988)

Rooman, M.J., Rodriguez, J. and Wodak, S.J. *Journal of Molecular Biology*. 213(2): 327-36. (1990)

Rosenberg, A. H., Lade, B. N., Chui, D., Lin, S., Dunn, J. J. and Studier, F. W. *Gene* 56, 125. (1987)

Ruepp,A. and Soppa,J. *J. Bacteriol.* 178 (16), 4942-4947. (1996)

Sadler, G. *Studying Chemical Modification and Limited Proteolysis of E. coli Citrate Synthase by Mass Spectrometry*. University of Manitoba M.Sc. Thesis. (2000)

Sambrook, J., Fritsch, E.F. and Maniatis, T. *Molecular Cloning: A Laboratory Manual*, 2nd Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. (1989)

Sanchez J.C., Gimenez R., Schneider A., Fessner W.D., Baldoma L., Aguilar J., Badia, J. *J Biol Chem* 269:29665-29669. (1994)

Sigrist C.J., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A., Bucher P. *Brief Bioinform*. 3:265-274. (2002)

Sofia H.J., Burland V., Daniels D.L., Plunkett G., 3rd, Blattner F.R. *Nucleic Acids Research* 22:2576-2586. (1994)

Straney, D., Straney, S. & Crothers, D. *J. Mol. Biol.* 206, 41-57. (1989)

Studier, F.W. *J. Mol. Biol.* 219: 37 – 44. (1991)

Studier, F.W. and Moffatt, B.A. *J. Mol. Biol.* 189: 113. (1986)

Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. *Meth. Enzymol.* 185: 60–89. (1990)

Sweet, R.M. *Biopolymers.* 25(8): 1565-77. (1986)

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. and Gojobori, T. *Nucleic Acids Research.* No. 28, Vol.1, pp. 24-26. (2000)

Tatusov R.L., Mushegian A. R., Bork P., Brown N., Hayes W. S., Borodovsky M., Rudd K. E., Koonin E. V. *Curr. Biol.* 6:279-291. (1996)

Tindall, K.R., and Kunkel, T.A. *Biochemistry* 27: 6008-6013. (1988)

van Wezel G. P., van der Meulen, J., Wamoto, S., Luiten, R.G.M., Koerten, H.K., and Kraal, B. *Journal of Bacteriology.* p. 5653–5662. (2000)

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T., and Itakura, K. *Nucleic Acids Res.* 6, 3543 (1979)

Wetlaufer, D. B. *Adv. Protein Chem.* 17:303-390. (1962)

Wierenga, R.K. *FEBS Lett.* 492(3):193-8. (2001)

Yamamoto, K. and Ishihama, A. *Molecular Microbiology.* 47: 183. (2003)

Yew, W.S. and Gerlt, J.A. *J Bacteriol.* 184:302-306. (2002)

Zhang, R., Kim, Y., Skarina, T., Beasley, S., Laskowski, R., Arrowsmith, C., Edwards, A., Joachimiak, A., and Savchenko, A. *J.Biol. Chem.* 277, 21. (2002)