

2014

## Pragmatic protein domain identification

Jeffrey R. Barrett  
*University of Wollongong*

Follow this and additional works at: <https://ro.uow.edu.au/theses>

**University of Wollongong**

**Copyright Warning**

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following: This work is copyright. Apart from any use permitted under the Copyright Act 1968, no part of this work may be reproduced by any process, nor may any other exclusive right be exercised, without the permission of the author. Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material.

Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

---

### Recommended Citation

Barrett, Jeffrey R., Pragmatic protein domain identification, Doctor of Philosophy thesis, School of Chemistry, University of Wollongong, 2014. <https://ro.uow.edu.au/theses/4141>

## **UNIVERSITY OF WOLLONGONG**

### **COPYRIGHT WARNING**

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.



# **Pragmatic Protein Domain Identification**

A thesis submitted in fulfilment of the requirements for the award  
of the degree

**Doctor of Philosophy**

from

**University of Wollongong**

by

**Jeffrey R. Barrett, B.Biotech(Hons)**

**Centre for Medical and Molecular Bioscience, School of  
Chemistry**

**2014**

## Certification

This Thesis is submitted in accordance with the regulations of the University of Wollongong in fulfilment of the Doctor of Philosophy (Chemistry). It does not include any material published by any other person, except where due reference is made in the text. The experimental work described in this thesis is original and has not been submitted for assessment in any other university.

---

Jeffrey R. Barrett  
February 4, 2014



## Acknowledgments

Thank you to my supervisor Professor Nicholas Dixon for your dedication, instruction and most of all your patience. I am also appreciative for the help and friendship of past and present Dixon lab members and associates, especially Dr Simon Brown, Dr Flynn Hill, Dr Hara Ioannou, Dr Slobodan Jergic, Dr Aaron Oakley and Dr Andrew Robinson. I would like to thank the following collaborators for their scientific input and access to facilities: Professor Jennifer Beck and Dr Claire Mason, Dr Celine Kelso, Dr Michelle Blayney, Dr Thitima Urathamakul of the Mass Spectrometry User Resource and Research Facility, University of Wollongong, Wollongong, NSW, Australia, and Dr Xun-Cheng Su, State Key Laboratory of Elemento-organic Chemistry, Nankai University, China.

I'm thankful for the endurance and understanding of my family and friends while I was completing this PhD. I'm above all grateful to Anthea for providing so much support while I completed my studies and to my parents Greg and Maree Barrett. Also to all of my friends who put up with not seeing or hearing from me for long stretches at a time: my friends from home, Corey Carnegie, Adam Greer, Adam Pincham, Melisa Tovey; my student comrades Grant and Debora Clark, Luke Collins, Patrick Constantinescu, Simon Cook, Ben Gooden, Michael Marthick, Jake Matic, Rocky, James Scifleet, James Tsatsaronis, and especially Peter Maamary, Karina Rovere and Steve Stone.

## Conference Abstracts

**Barrett, J.R.**, Robinson, A., Billingham, S., Dixon, N.E. (2009) Identification of soluble protein domain constructs. *Lorne Proteins 2009, 34th Lorne Conference on Protein Structure and Function, 8-12 February 2009, LORNE, VIC, Australia.*

**Barrett, J.R.**, Robinson, A., Dixon, N.E. (2009) Exploring protein domains: a system to identify soluble portions of multi-domain proteins. *Combio 2009, 6-10 December 2009, CHRISTCHURCH, New Zealand.*

**Barrett, J.R.**, Robinson, A., Dixon, N.E. (2011) Identification of soluble protein domain constructs. *Lorne Proteins 2011, 36th Lorne Conference on Protein Structure and Function, 6-10 February 2011, LORNE, VIC, Australia.*

# Abbreviations

<b><math>A_{260}</math></b>	absorbance at 260 nm
<b><math>A_{280}</math></b>	absorbance at 280 nm
<b><math>A_{600}</math></b>	absorbance at 600 nm
<b>bp</b>	base pair(s)
<b>BSA</b>	bovine serum albumin
<b><i>bla</i></b>	$\beta$ -lactamase gene
<b>C-terminus</b>	carboxyl-terminus
<b>CD</b>	circular dichroism
<b>CV</b>	column volumes
<b>Da</b>	Dalton(s)
<b>dATP</b>	deoxyadenosine 5'-triphosphate
<b>dCTP</b>	deoxycytidine 5'-triphosphate
<b>dCTP<math>\alpha</math>S</b>	deoxycytidine 5'- $\alpha$ -thiotriphosphate
<b>dGTP</b>	deoxyguanosine 5'-triphosphate
<b>dGTP<math>\alpha</math>S</b>	deoxyguanosine 5'- $\alpha$ -thiotriphosphate
<b>dTTP</b>	deoxythymidine 5'-triphosphate
<b>dNTP</b>	deoxynucleoside 5'-triphosphate
<b>DNA</b>	deoxyribonucleic acid
<b>DHFR</b>	the dihydrofolate reductase protein from <i>Escherichia coli</i>
<b>TMP</b>	trimethoprim

<b>DHF</b>	dihydrofolate
<b>DMSO</b>	dimethyl sulfoxide
<b>DnaG</b>	the bacterial DNA primase
<b><i>dnaG</i></b>	the gene encoding DnaG protein
<b>dsDNA</b>	double stranded DNA
<b>DTT</b>	dithiothreitol
<b>EDTA</b>	ethylenediaminetetraacetic acid
<b>EGFP</b>	enhanced green fluorescent protein
<b><i>egfp</i></b>	the gene encoding EGFP protein
<b>ESI</b>	electrospray ionization
<b>ESI-MS</b>	electrospray ionisation mass spectrometry
<b><i>ExoIII</i></b>	exonuclease III
<b>EzrA</b>	the <u>E</u> xtra <u>Z</u> -rings protein from <i>Staphylococcus aureus</i>
<b><i>ezrA</i></b>	the gene encoding EzrA protein
<b><i>folA</i></b>	the gene encoding DHFR protein
<b>FPLC</b>	fast protein liquid chromatography
<b>GFP</b>	green fluorescent protein
<b>GpsB</b>	guiding PBP1 shuttling protein
<b>GTP</b>	guanosine triphosphate
<b>HBD</b>	helicase binding domain (of DnaG)
<b>IMAC</b>	immobilised metal ion chromatography
<b>IPTG</b>	isopropyl- $\beta$ -D-thiogalactopyranoside
<b>MALDI</b>	matrix-assisted laser desorption/ionization
<b>MBP</b>	maltose-binding protein
<b>MCS</b>	multiple cloning site
<b>MS</b>	mass spectrometry
<b>MWCO</b>	molecular weight cut off

<b>N-terminus</b>	amino-terminus
<b>NMR</b>	nuclear magnetic resonance
<b>ORF</b>	open reading frame
<b>PAGE</b>	polyacrylamide gel electrophoresis
<b>PBP</b>	penicillin-binding protein
<b>PBP1</b>	penicillin-binding protein 2b
<b>PCR</b>	polymerase chain reaction
<b>PDB</b>	protein data bank
<b>RBS</b>	ribosome binding site
<b>RPD</b>	RNA polymerase domain (of DnaG)
<b>RPM</b>	revolutions per minute
<b>S1</b>	S1 nuclease
<b>SDS</b>	sodium dodecyl sulphate
<b>ssDNA</b>	single stranded DNA
<b>TE</b>	10 mM Tris-HCl, 1 mM EDTA, pH 8.0
<b>THF</b>	tetrahydrofolate
<b>TMP</b>	trimethoprim
<b>TOCSY</b>	TOtal Correlation Spectroscopy
<b>Tris</b>	tris(hydroxymethyl)-aminomethane
<b>ZBD</b>	zinc-binding domain (of DnaG)

# Abstract

Studying proteins is hard. Even in well studied model systems, some proteins are recalcitrant to production in useful amounts in soluble form. These proteins can be difficult to over-express and/or are not soluble and often the lack of protein solubility is blamed on poor/improper protein folding. However, in general, small proteins are easier to express in soluble form than large proteins.

Protein evolution has produced many modular multi-domain proteins that are made from several smaller folded domains in series, as it is more efficient to fuse established functional domains together than to construct a large protein *de novo*. It has been known for a long time that distinct domains of large proteins can be more easily over-expressed than the full-length protein, and many full-length proteins have been studied by over-expressing their domains separately.

This Thesis presents a new pragmatic methodology for truncating and identifying soluble fragments of proteins. This new technique was used to identify previously unattainable soluble domain constructs of proteins of interest to our research group. The new technique for protein domain truncation uses exonuclease III to delete a protein gene in a specially constructed plasmid. Gene deletion can

be performed to result in truncation from either the amino- (N-) or carboxy- (C-) termini of a protein and makes a library of truncated protein genes. The truncated protein genes in these plasmids are fused to a downstream gene for either enhanced green fluorescent protein (EGFP) or dihydrofolate reductase (DHFR). The fused EGFP or DHFR gives cells expressing the fusion protein a distinct phenotype depending on whether the truncated protein-fusion is soluble or not. This technique allows pragmatic protein domain identification as soluble truncated proteins can be assumed to not include incomplete protein domains, and are thus truncated at a domain boundary.

Truncation of *Acinetobacter baylyi* DNA primase (DnaG) successfully produced genes for soluble expression of both the zinc binding domain (ZBD) and an RNA polymerase-helicase binding domain fusion (RPD-HBD) that were not able to be produced using existing methods. Several constructs for both the ZBD and RPD-HBD were confirmed to be soluble after purification and ZBD constructs were further investigated by circular dichroism and nuclear magnetic resonance (NMR) to show that they were folded. Examination of some ZBD proteins by mass spectrometry indicated that they bind zinc, which only occurs in the ZBDs of other species when they are correctly folded.

At the commencement of this Thesis no isolated domains of *Staphylococcus aureus* EzrA (extra Z-rings) or its homologues were known. A comprehensive library of genes encoding N- or C- terminally truncated EzrA were produced and used to identify soluble proteins at each of EzrA's 5 domain boundaries. One-dimensional and total correlation spectroscopy NMR experiments showed that most of the

truncated EzrA fragments were at least partially folded, and in the case of proteins for domains 1–2 (EzrA<sup>24–214</sup>), 4–6 (EzrA<sup>302–564</sup>) and 6 (EzrA<sup>484–564</sup>) were well folded without flexible ends. Although not all of the soluble fragments of EzrA were investigated, those that were allowed identification of protein domain boundaries for domains 2–6.

The technique we have developed and utilised for pragmatic protein domain identification allows a large number of shortened protein genes to be produced and easily screened to identify soluble protein constructs. Characterisation of purified truncated soluble proteins using techniques such as NMR allow identification of domains in previously uncharacterised proteins or that have been resistant to investigation by other methods.



# Contents

<b>Abbreviations</b>	<b>v</b>
<b>Abstract</b>	<b>viii</b>
<b>1 General introduction</b>	<b>1</b>
1.1 Problem: producing soluble proteins . . . . .	1
1.2 Protein structure and folding . . . . .	2
1.2.1 Protein folding . . . . .	3
1.2.2 Protein size and folding rates . . . . .	6
1.2.3 Protein synthesis and folding chaperones . . . . .	6
1.2.4 Passenger solubilisation . . . . .	13
1.2.5 Protein domain evolution . . . . .	14
1.3 Small proteins are often better suited for three-dimensional structure determination . . . . .	15
1.3.1 Nuclear magnetic resonance and protein size . . . . .	17
1.3.2 X-ray crystallography and protein size . . . . .	19
1.4 Protein domain identification . . . . .	19
1.4.1 Computational methods for protein domain identification . .	21
1.4.2 Experimental methods for protein domain identification . . .	24
1.5 Protein domain libraries . . . . .	25
1.6 Selecting for protein solubility . . . . .	26
1.6.1 Dihydrofolate reductase and solubility reporting . . . . .	27
1.6.2 Green fluorescent protein and solubility reporting . . . . .	30
1.7 Aims of this Thesis . . . . .	32
<b>2 General materials and methods</b>	<b>35</b>
2.1 Chemicals, reagents, enzymes and instruments . . . . .	35
2.2 Bacterial strains and transformation . . . . .	36
2.2.1 Routine growth of <i>Escherichia coli</i> . . . . .	36
2.2.2 Clonal isolation of transformants . . . . .	37
2.2.3 Transformation of <i>Escherichia coli</i> . . . . .	37

2.2.4	Long term storage of bacterial strains . . . . .	38
2.3	Molecular genetics . . . . .	39
2.3.1	Preparation of plasmid DNA . . . . .	39
2.3.2	Restriction digestion of DNA . . . . .	39
2.3.3	Preparation of oligonucleotides . . . . .	40
2.3.4	Amplification of DNA by polymerase chain reaction . . . . .	40
2.3.5	Colony PCR and storage of truncation libraries . . . . .	41
2.3.6	Electrophoresis of DNA . . . . .	42
2.3.7	Isolation of DNA reaction products by agarose gel electrophoresis . . . . .	42
2.3.8	Isolation of DNA reaction products by silica column . . . . .	43
2.3.9	Isolation of DNA reaction products by ethanol precipitation . . . . .	43
2.3.10	Ligation of DNA . . . . .	43
2.3.11	Dye terminator sequencing of DNA . . . . .	44
2.4	Estimation of DNA concentrations . . . . .	45
2.4.1	Estimation of DNA concentration by agarose gel electrophoresis . . . . .	45
2.4.2	Spectrophotometric determination of DNA concentration . . . . .	45
2.5	Protein over-expression and purification . . . . .	46
2.5.1	Protein expression by auto-induction . . . . .	46
2.5.2	Cell lysis by French press . . . . .	47
2.5.3	Clarification of bacterial lysates . . . . .	47
2.5.4	Protein purification using ÄKTA FPLC systems . . . . .	47
2.5.5	Determination of protein concentration . . . . .	48
2.5.6	Protein dialysis . . . . .	48
2.5.7	Storage of proteins . . . . .	48
2.5.8	Concentration of protein samples . . . . .	48
2.5.9	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis . . . . .	49
2.5.10	Mass spectrometry of proteins samples . . . . .	50
<b>3</b>	<b>Gene deletion and solubility selection</b>	<b>52</b>
3.1	Introduction . . . . .	52
3.1.1	Gene truncation using exonuclease III . . . . .	54
3.2	Aims . . . . .	56
3.3	A technique for gene truncation and solubility selection . . . . .	57
3.3.1	Gene fusion for solubility selection . . . . .	59
3.3.2	Layout of plasmids for gene truncation using exonuclease III . . . . .	60
3.4	Assembly of plasmids for gene truncation and solubility selection . . . . .	67
3.4.1	Strategy for construction of solubility selection plasmids . . . . .	67
3.4.2	Construction of pJB1703 . . . . .	69
3.4.3	Construction of pJB1704 . . . . .	70

3.4.4	Construction of pJB1705 . . . . .	70
3.4.5	Construction of pJB1706 . . . . .	72
3.4.6	Construction of pJB1707 . . . . .	72
3.4.7	Construction of enhanced green fluorescent protein expression plasmids with biotinylation or Ktag purification tags . . . . .	75
3.4.8	Construction of dihydrofolate reductase expression plasmids with biotinylation or Ktag purification tag . . . . .	77
3.4.9	Construction of gene-truncation plasmids . . . . .	77
3.4.10	Truncation apparatus . . . . .	81
3.5	Conclusion . . . . .	81
<b>4</b>	<b>Domain identification in <i>Acinetobacter baylyi</i> DNA primase</b>	<b>83</b>
4.1	DNA primase and DNA replication . . . . .	83
4.1.1	Preliminary work on <i>Acinetobacter baylyi</i> DNA primase . . .	86
4.1.2	The DNA primase zinc-binding domain . . . . .	87
4.1.3	Few soluble constructs for bacterial zinc-binding domains are known . . . . .	88
4.1.4	The zinc-binding domain of <i>Acinetobacter baylyi</i> DNA primase is followed by a novel sequence insertion . . . . .	90
4.2	Aims . . . . .	90
4.3	Materials and methods . . . . .	92
4.3.1	Plasmids for gene truncation of <i>Acinetobacter baylyi</i> DNA primase . . . . .	92
4.3.2	Library preparation . . . . .	93
4.3.3	Truncation mutation identification . . . . .	97
4.3.4	Removal of EGFP sequence from fusion genes . . . . .	97
4.3.5	Removal of His <sub>6</sub> -tag and EGFP sequence from fusion proteins	98
4.3.6	Expression, solubility examination and purification of truncation mutants . . . . .	99
4.3.7	Circular dichroism of truncated protein . . . . .	101
4.3.8	Nuclear magnetic resonance analysis of <i>Acinetobacter baylyi</i> DNA primase zinc-binding domain . . . . .	103
4.3.9	Protein crystallography of DnaG <sup>1-165</sup> and DnaG <sup>1-165</sup> . . . .	104
4.3.10	Homology modelling of the <i>Acinetobacter baylyi</i> DNA primase zinc-binding and RNA polymerase domains . . . . .	105
4.4	Results . . . . .	105
4.4.1	C-terminal truncation of <i>Acinetobacter baylyi</i> DNA primase	105
4.4.2	Examination of soluble N-terminal DNA primase mutants .	115
4.4.3	Purification of soluble N-terminal fragments of DNA primase	117
4.4.4	Circular dichroism of the DNA primase zinc-binding domain	119

4.4.5	Examination of the foldedness of C-terminally truncated DNA primase by nuclear magnetic resonance spectroscopy . . . . .	119
4.4.6	N-terminal fragments of DNA primase bind zinc . . . . .	122
4.4.7	Crystallisation of DnaG <sup>1-165</sup> and DnaG <sup>1-170</sup> . . . . .	125
4.4.8	N-terminal truncation of <i>Acinetobacter baylyi</i> DNA primase . . . . .	125
4.4.9	Solubility of N-terminally deleted DNA primase mutants . . . . .	138
4.4.10	Modelled protein structures of the <i>Acinetobacter baylyi</i> zinc-binding and RNA polymerase domains . . . . .	140
4.4.11	N-terminal DNA primase deletion mutants and predicted protein structure . . . . .	145
4.5	Discussion . . . . .	150
4.5.1	C-terminal deletions of DNA primase . . . . .	151
4.5.2	N-terminal deletions of DNA primase . . . . .	156
4.5.3	DNA primase truncation mutations and modelled protein structures . . . . .	157
4.5.4	General discussion . . . . .	158
<b>5</b>	<b>Soluble domain constructs of <i>Staphylococcus aureus</i> septation ring formation regulator (EzrA)</b>	<b>160</b>
5.1	Introduction . . . . .	160
5.1.1	Cell division and Z-rings . . . . .	160
5.1.2	EzrA has two Z-ring regulating roles . . . . .	162
5.1.3	Cell wall synthesis . . . . .	164
5.1.4	EzrA helps coordinate cell wall synthesis at the site of cell division . . . . .	165
5.1.5	The importance of EzrA in cell division in <i>Staphylococcus aureus</i> . . . . .	166
5.1.6	EzrA as a potential protein-protein interaction hub . . . . .	169
5.1.7	Little is currently known about the structure of EzrA . . . . .	171
5.2	Aims . . . . .	172
5.3	Materials and methods . . . . .	173
5.3.1	Preparation of <i>Staphylococcus aureus</i> <i>ezrA</i> truncation libraries	173
5.3.2	Preparation of EzrA truncation-His <sub>6</sub> tagged expression plasmids . . . . .	175
5.3.3	Expression, examination of protein solubility and purification of truncation mutants . . . . .	176
5.3.4	Nuclear magnetic resonance analysis of truncated EzrA . . . . .	178
5.3.5	Plasmids for over-expression of wild type and biotinylated <i>Staphylococcus aureus</i> cell division protein FtsZ . . . . .	178
5.3.6	Purification of biotinylated <i>Staphylococcus aureus</i> FtsZ . . . . .	180
5.4	Results . . . . .	184

5.4.1	Genetic truncation of <i>Staphylococcus aureus</i> EzrA . . . . .	184
5.4.2	Examination of solubility of truncated EzrA mutants . . . . .	194
5.4.3	Purification of soluble N- and C-terminal fragments of EzrA . . . . .	198
5.4.4	Examination of the foldedness of N- and C-terminally truncated <i>Staphylococcus aureus</i> EzrA . . . . .	204
5.4.5	Probing the EzrA–FtsZ interaction using affinity pull-down . . . . .	211
5.5	Discussion . . . . .	215
5.5.1	Truncated EzrA and solubility . . . . .	215
5.5.2	Structural insights into EzrA truncations . . . . .	217
5.5.3	The interaction between EzrA and FtsZ . . . . .	221
5.5.4	General discussion . . . . .	223
<b>6</b>	<b>Concluding remarks</b>	<b>225</b>
6.0.5	Caveats . . . . .	226
6.0.6	Usefulness of this method . . . . .	227
	<b>Bibliography</b>	<b>229</b>
	<b>Appendices</b>	<b>253</b>
<b>A</b>	<b>Apparatus for generating exonuclease III libraries</b>	<b>254</b>
<b>B</b>	<b>Library population prediction</b>	<b>256</b>
<b>C</b>	<b>Oligonucleotides</b>	<b>258</b>
<b>D</b>	<b>Enzyme buffers</b>	<b>260</b>
<b>E</b>	<b>Mass spectra and NMR analysis of truncated EzrA proteins</b>	<b>261</b>

# List of Tables

<b>2</b>	<b>General materials and methods</b>	<b>35</b>
2.1	Bacterial strains . . . . .	36
2.2	Antibiotics . . . . .	37
2.3	Extension parameters for thermostable DNA polymerases . . . . .	41
2.4	Sequencing primers . . . . .	45
<b>3</b>	<b>Gene deletion and solubility selection</b>	<b>52</b>
3.1	Gene deletion and solubility selection plasmids . . . . .	80
<b>4</b>	<b>Domain identification in <i>Acinetobacter baylyi</i> DNA primase</b>	<b>83</b>
4.1	Randomly sequenced <i>dnaG</i> C-terminally truncated plasmids . . . . .	108
4.2	Sequenced C-terminally deleted <i>Acinetobacter baylyi</i> DnaG constructs	111
4.3	Plasmids for expression of putatively soluble DNA primase N-terminal fragments . . . . .	116
4.4	Randomly sequenced N-terminally truncated DNA primase plasmids	128
4.5	Sequenced N-terminally deleted <i>A. baylyi</i> DnaG . . . . .	129
4.6	Plasmids for expression of putative soluble DNA primase C-terminal fragments . . . . .	138
4.7	Putatively soluble N-terminally deleted mutants of DNA primase .	147
4.9	Putatively soluble C-terminally deleted mutants of DNA primase .	149
<b>5</b>	<b>Soluble domain constructs of <i>Staphylococcus aureus</i> septation ring formation regulator (EzrA)</b>	<b>160</b>
5.1	Essentiality of proteins suspected of interacting with EzrA . . . . .	170
5.2	Concentration of EzrA mutants for analysis by nuclear magnetic resonance . . . . .	179
5.3	N-terminally deleted <i>Staphylococcus aureus</i> EzrA mutants . . . . .	187
5.4	C-terminally deleted <i>S. aureus</i> EzrA . . . . .	191

5.5 Identity of putatively soluble truncated <i>Staphylococcus aureus</i> <i>ezrA</i> plasmids . . . . .	195
--	-----

<b>Appendices</b>	<b>253</b>
-------------------	------------

C.1 Oligonucleotides used in this work . . . . .	258
D.1 Enzyme buffers . . . . .	260
E.1 Mass spectrometry of His <sub>6</sub> -tagged EzrA truncated proteins . . . . .	262

# List of Figures

<b>1</b>	<b>General introduction</b>	<b>1</b>
1.1	Protein folding energy landscape . . . . .	4
1.2	Protein folding rate is related to protein size . . . . .	7
1.3	Crystal structure of the GroE complex . . . . .	10
1.4	Protein domain fusion is a process whereby genes encoding two independent protein domains recombine . . . . .	16
1.5	Bottlenecks in structural biology pipelines . . . . .	17
1.6	Size distribution of published protein structures . . . . .	20
1.7	Domain fragmentation . . . . .	22
1.8	Trimethoprim blocks the dihydrofolate binding site in dihydrofolate reductase . . . . .	28
1.9	Chromophore maturation in green fluorescent protein . . . . .	31
<b>3</b>	<b>Gene deletion and solubility selection</b>	<b>52</b>
3.1	Distribution of exonuclease III truncation reaction products . . . . .	56
3.2	Methodology for uni-directional gene truncation . . . . .	58
3.3	General layout of gene truncation and solubility reporting plasmids . . . . .	60
3.4	Cloning site for N-terminal gene truncation and solubility reporting plasmids . . . . .	61
3.5	Cloning site for C-terminal gene truncation and solubility reporting plasmids . . . . .	63
3.6	Alternate cloning site for C-terminal gene truncation and solubility reporting plasmids . . . . .	64
3.7	Removal of purification and solubility reporter tags from truncated genes of interest . . . . .	66
3.8	Removal of solubility reporting gene fusion from truncated genes of interest . . . . .	68
3.9	Strategy for construction of protein purification and solubility selection plasmids . . . . .	69



3.10	Construction of pJB1705 . . . . .	71
3.11	Construction of pJB1706 . . . . .	73
3.12	Construction of pJB1707 . . . . .	74
3.13	Construction of pJB1708 . . . . .	76
3.14	Construction of pJB1709 . . . . .	78
3.15	Gene truncation plasmids . . . . .	79
<b>4</b>	<b>Domain identification in <i>Acinetobacter baylyi</i> DNA primase</b>	<b>83</b>
4.1	DNA replication fork . . . . .	84
4.2	The domains of DNA primase . . . . .	85
4.3	Bacterial primase zinc-binding domains . . . . .	88
4.4	Domain location and known soluble constructs of DNA primase . .	89
4.5	Zinc-binding domain extension in <i>Moraxellaceae</i> . . . . .	91
4.6	Plasmids for gene truncation of <i>dnaG</i> . . . . .	93
4.7	Methodology for C-terminal truncation of <i>dnaG</i> . . . . .	106
4.8	DNA primase C-terminally deleted green fluorescent mutants . . .	110
4.9	Green fluorescence phenotype of <i>Acinetobacter baylyi</i> DnaG ZBD truncations . . . . .	110
4.10	Soluble over-expression of DNA primase C-terminal deletion mutants	116
4.11	Purification of DNA primase C-terminally truncated proteins . . .	118
4.12	Circular dichroism of DNA primase zinc-binding domain-extension proteins . . . . .	120
4.13	One dimensional nuclear magnetic resonance spectrum of DnaG <sup>1-165</sup>	121
4.14	Two dimensional TOCSY NMR spectra of DnaG <sup>1-165</sup> -His <sub>6</sub> and DnaG <sup>1-170</sup> -His <sub>6</sub> . . . . .	123
4.15	Purified untagged DnaG <sup>1-165</sup> and DnaG <sup>1-170</sup> . . . . .	124
4.16	Positive ion electrospray mass spectra of denatured and native DNA primase zinc-binding domain-extension mutants . . . . .	124
4.17	Methodology for N-terminal truncation of DNA primase . . . . .	126
4.18	Over-expression and protein solubility of DNA primase N-terminally deleted mutants . . . . .	139
4.19	Structural model of DNA primase zinc-binding domain . . . . .	144
4.20	Structural model of DNA primase RNA polymerase domain . . . .	146
4.20	N-terminal breakpoints of truncated DNA primase . . . . .	147
4.20	C-terminal breakpoints of truncated DNA primase . . . . .	149
4.21	Soluble truncated DNA primase . . . . .	152

<b>5 Soluble domain constructs of <i>Staphylococcus aureus</i> septation ring formation regulator (EzrA)</b>	<b>160</b>
5.1 Cell division and the divisome . . . . .	161
5.2 EzrA spacing and FtsZ polymer size . . . . .	163
5.3 Coordination of cell wall synthesis at the divisome . . . . .	166
5.4 Known phenotypes of <i>ezrA</i> mutants . . . . .	172
5.5 Plasmids for gene truncation of <i>Staphylococcus aureus ezrA</i> . . . . .	174
5.6 Plasmid for over-expression of N-terminally biotinylated <i>Staphylococcus aureus</i> FtsZ . . . . .	180
5.7 Methodology for 5'- and 3'-truncation of <i>Staphylococcus aureus ezrA</i> . . . . .	185
5.8 EzrA truncations . . . . .	194
5.9 Over-expression and protein solubility of EzrA N- and C-terminally deleted mutants . . . . .	196
5.10 Purification of N- and C-terminally deleted EzrA mutants . . . . .	199
5.11 N-terminal proteolysis of EzrA <sup>425–564</sup> . . . . .	202
5.12 N-terminal proteolysis of EzrA <sup>24–126</sup> , EzrA <sup>24–128</sup> and EzrA <sup>24–129</sup> . . . . .	203
5.13 Two-dimensional TOCSY NMR fingerprint region in spectra of C-terminally deleted EzrA mutants . . . . .	206
5.14 Two-dimensional TOCSY NMR fingerprints of N-terminally deleted EzrA mutants . . . . .	209
5.15 Screening for FtsZ interaction using EzrA fragment libraries . . . . .	212
5.16 Identification of FtsZ–EzrA interaction by pull-down assay . . . . .	214
5.17 Overview of soluble truncated <i>Staphylococcus aureus</i> EzrA mutant proteins . . . . .	216
5.18 Preliminary crystallography of EzrA <sup>24–214</sup> . . . . .	217
5.19 Cytoplasmic EzrA domain architecture . . . . .	219
5.20 FtsZ pull-down of purified truncated EzrA proteins . . . . .	222
 <b>Appendices</b>	 <b>253</b>
A.1 Apparatus for generating exonuclease III libraries . . . . .	255
B.1 Distribution of truncation lengths for an ideal uni-directional exonuclease III truncation library . . . . .	256
B.2 Distribution of truncation lengths for an ideal uni-directional exonuclease III truncation library . . . . .	257
E.1 Mass and NMR spectral analysis of EzrA <sup>277–564</sup> . . . . .	263
E.2 Mass and NMR spectral analysis of EzrA <sup>280–564</sup> . . . . .	264
E.3 Mass and NMR spectral analysis of EzrA <sup>302–564</sup> . . . . .	265
E.4 Mass and NMR spectral analysis of EzrA <sup>381–564</sup> . . . . .	266

E.5	Mass and NMR spectral analysis of EzrA <sup>24-139</sup>	267
E.6	Mass and NMR spectral analysis of EzrA <sup>443-564</sup>	268
E.7	Mass analysis of EzrA <sup>453-564</sup>	269
E.8	Mass analysis of EzrA <sup>476-564</sup>	269
E.9	Mass and NMR spectral analysis of EzrA <sup>484-564</sup>	270
E.10	Mass and NMR spectral analysis of EzrA <sup>24-97</sup>	271
E.11	Mass and NMR spectral analysis of EzrA <sup>24-126</sup>	272
E.12	Mass and NMR spectral analysis of EzrA <sup>24-128</sup>	273
E.13	Mass and NMR spectral analysis of EzrA <sup>24-129</sup>	274
E.14	Mass and NMR spectral analysis of EzrA <sup>24-139</sup>	275
E.15	Mass and NMR spectral analysis of EzrA <sup>24-214</sup>	276
E.16	Mass and NMR spectral analysis of EzrA <sup>24-238</sup>	277
E.17	Mass and NMR spectral analysis of EzrA <sup>24-476</sup>	278
E.18	Mass and NMR spectral analysis of EzrA <sup>24-564</sup>	279

# Chapter 1

## General introduction

### 1.1 Problem: producing soluble proteins

Production of soluble, correctly folded protein is usually the first step for *in vitro* investigation of biochemical and three-dimensional properties of a protein. Many straightforward methodologies for cheaply producing large quantities of a protein of interest have been developed using *Escherichia coli* and other bacteria as a host. Protein over-expression in bacterial hosts provides an advantage as protein translation in bacteria can be an order of magnitude faster than in eukaryotes and several protein over-expression techniques have been developed for directing massive production of a target protein. However, many proteins are not initially soluble or correctly folded when over-expressed.

## 1.2 Protein structure and folding

Proteins are regularly composed of distinctly folded domains, commonly ranging in size from 100 to 150 amino acids in length (11 to 17 kDa) (Berman *et al.*, 1994; Xu and Nussinov, 1998) which can clearly be seen by the clustering of proteins in multiples of this range in gel electrophoresis in species such as *Escherichia coli* ( $\sim 14$  kDa; Savageau, 1986). Protein folding is a complex process with conflicting energetic factors that scale differently; entropic loss from restricting conformational space of the folded polymer must be compensated by enthalpic intra-molecular interactions and exclusion of water from hydrophobic regions (Xu and Nussinov, 1998).

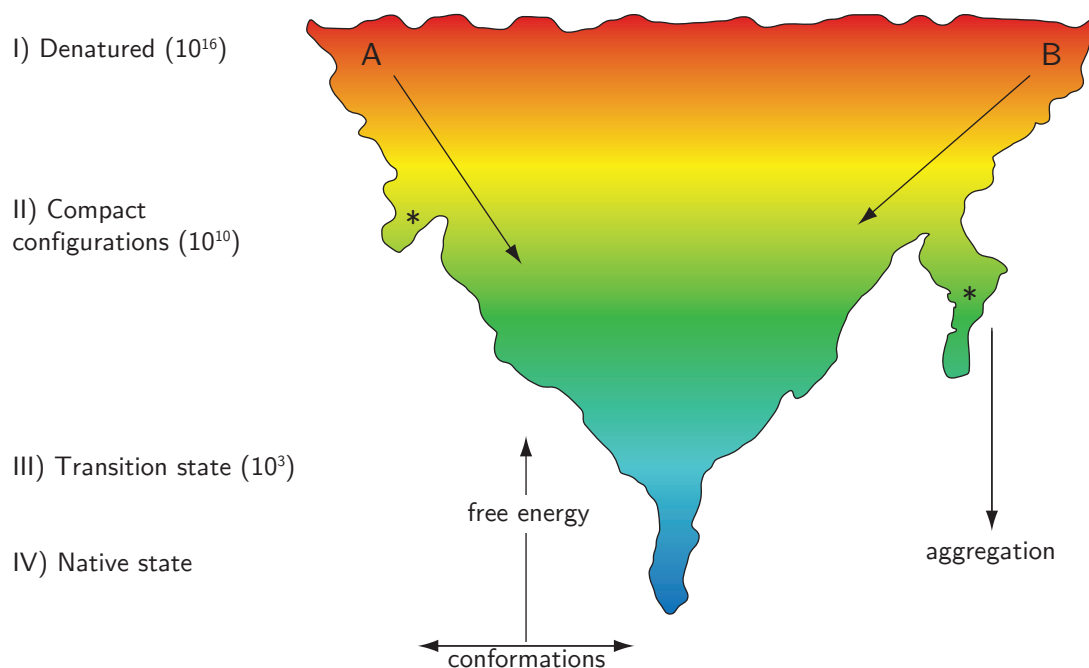
Entropic energy loss from protein folding scales with protein length in a quadratic function, whereas enthalpic intra-molecular interactions scale linearly (Xu and Nussinov, 1998). These two conflicting sequence length dependencies result in a “sweet spot” of protein domain length. Proteins that are very short do not exclude water or have sufficient enthalpy from inter-residue contacts to balance the entropy loss from folding, while for very large domains, the cost of configurational entropy increases beyond the capacity of intra-molecular interactions (Xu and Nussinov, 1998).

### 1.2.1 Protein folding

Protein folding is a intricate process, where an initially unfolded, high entropy polypeptide can attain a productively folded, low entropy state on a time-scale of milliseconds to seconds (Dinner *et al.*, 2000). Protein folding is believed to progress in a somewhat directed manner, whereby partially folded intermediates form transiently, and these transient intermediates then combine to form a thermodynamically favourable native state (Dobson *et al.*, 1998). In the polar intra-cellular environment, association between hydrophobic residues is strong, leading to folded protein conformations with dense hydrophobic cores and hydrophilic surfaces.

A semi-directed pathway for protein folding is required to compensate for the Levinthal paradox, where, if protein folding occurred solely by randomly sampling all the possible conformations, each protein molecule would require astronomical time scales to fold to its native shape ( $10^{11}$  years for a 100 amino acid protein; Levinthal, 1969; Dinner *et al.*, 2000). Semi-directed protein folding can be explored through the concept of an energy landscape which shows the relationship between steps of folding and overall free energy (Figure 1.1; Dill and Chan, 1997; Dinner *et al.*, 2000; Dobson, 2003). The many energetic factors contributing to protein folding occur both within local protein regions and across the larger distances through a protein, with each interaction contributing either positively or negatively to the free energy of the protein.

Initially the high entropy unfolded protein is in high conformational flux. As new,



**Figure 1.1: Protein folding energy landscape.** Denatured, high free energy protein with many conformations progresses to a native state protein with low free energy and few conformations. Points I–IV show the progression of a 27-mer protein from unfolded through to a folded protein and the number of available conformations (Dinner *et al.*, 2000). Local free energy minima are indicated with \*. With increasing depth of free energy minima, re-entry to the protein folding pathway becomes less favourable and may prevent the protein from folding to more compact and low energy conformations. Figure adapted from Dinner *et al.* (2000).

small, transiently folded regions form, favourable intra-molecular interactions reduce the available conformations of the protein chain. These sterically restrained conformations have a lower overall free energy and lead to an incremental process, whereby as native-fold contacts are formed, they progressively lower the overall numbers of possible conformations and protein free energy until a transition state occurs, from which the protein collapses quickly into the native fold (Dinner *et al.*, 2000; Dobson, 2003). However, protein folding does not always lead to stable, natively folded conformations. Stalled, partially folded intermediates may become trapped in regions of local free energy minima, restricting further productive folding.

For example, two different theoretical protein folding trajectories are shown in Figure 1.1. Trajectories A and B progress from an unfolded, high free energy state and progress through many intermediates, to a lower free energy conformation. During the folding pathways, a localised region of low free energy is experienced (at \*), and the accessible conformations are restricted; the intermediate cannot continue to productively fold. In order for folding along trajectory A or B to progress further, an increase in free energy is required to access more conformations. Trajectory A has a small local free energy minimum and the protein here can more easily re-enter the folding pathway and progress to the native fold than the one following trajectory B, which has a very low local energy minimum. If the energy required to re-enter the folding pathway is too great the intermediate state may become trapped and folding will stall (Bollen *et al.*, 2004; Wu *et al.*, 2007).

When a protein is in a partially folded state, hydrophobic residues destined for the hydrophobic core are solvent exposed. If a partially folded protein does not progress to its low free energy native conformation, aggregation can decrease the local free energy through hydrophobic, inter-molecular and solvent excluding interactions with other protein molecules. The resulting dense aggregates of non-natively folded polypeptides are stabilised by hydrophobic interactions and a severe reduction in surface area. Consequently, once a protein aggregate has formed, unfavourable, large energy inputs are required to release the proteins to allow productive protein folding to recommence. As protein size increases, proportionately more partially folded states are encountered, and the risk of



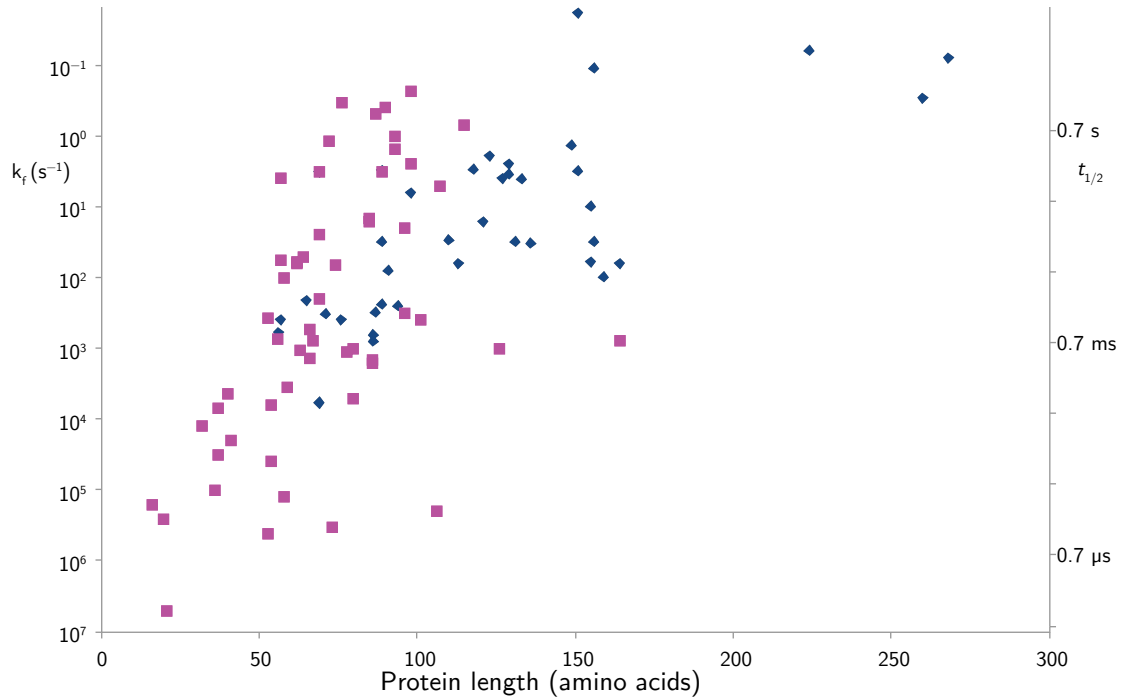
protein aggregation increases (Jahn and Radford, 2005; Brockwell and Radford, 2007).

### 1.2.2 Protein size and folding rates

Protein folding generally proceeds more slowly as protein size increases (Figure 1.2; Fersht and Daggett, 2002; Ivankov and Finkelstein, 2004; Ouyang and Liang, 2009) and this is due to several factors. Increased protein size leads to a larger entropy loss upon folding, resulting in less favourable native folds and more complex protein folding trajectories. These proteins can spend longer times in intermediate folded states which increases the likelihood of aggregation. Larger proteins also pose a larger aggregation potential as increased domain size necessitates a larger number of hydrophobic core residues (Kiraga *et al.*, 2007). Protein domains in larger proteins fold with a large degree of autonomy from each other, which is useful for attaining faster protein folding rates (Figure 1.2), but the close proximity of domains does not allow folding to occur as quickly as for protein domains in isolation.

### 1.2.3 Protein synthesis and folding chaperones

The densely packed environment within cells promotes macromolecular association and encourages protein aggregation (van den Berg *et al.*, 1999; Jahn and Radford, 2008). Cellular processes have developed to mitigate aggregation of unfolded proteins by chaperoning aggregation prone nascent proteins and aiding in



**Figure 1.2: Protein folding rate is related to protein size.** Data for proteins for which both in-water folding rates and protein structures are known, compiled from Ivankov and Finkelstein (2004) and Ouyang and Liang (2009). Single domain,  $\blacksquare$ ; multi-domain,  $\blacklozenge$ . Primary Y-axis shows in-water folding rates,  $k_f$  ( $\text{s}^{-1}$ ); secondary Y-axis half-folding time;  $t_{1/2}$  (s).

productive protein folding (Hartl, 1996; Hartl *et al.*, 2011). Protein folding chaperones in general act to sequester hydrophobic regions present in nascent protein chains. By inhibiting inter-molecular interactions between immaturely folded proteins, aggregation through these regions is reduced. Folding chaperones function by one of two modes: I) chambered chaperones provide an environment promoting protein folding; and II) passive chaperones hold hydrophobic regions of nascent chains while folding occurs.

Protein synthesis in *E. coli* occurs in polysomal ribosome clusters, where many unfolded polypeptides exit ribosomes in close proximity and where the hydrophobic sections of the new polypeptides pose an aggregation risk to

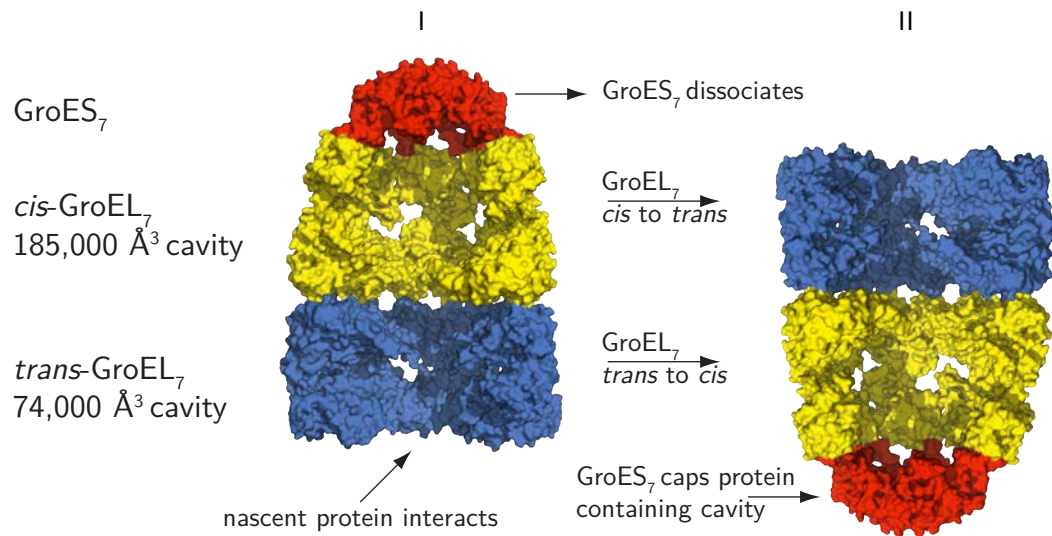
each other. For this reason two proteins with chaperone capacity are found to localise near the ribosome exit to help mitigate aggregation of newly synthesised polypeptides. Both DnaK, which associates with nascent chains (Gaitanaris *et al.*, 1994), and the peptidyl-prolyl *cis-trans* isomerase trigger factor, which localises to the mouth of the ribosome exit tunnel (Stoller *et al.*, 1995; Hesterkamp *et al.*, 1996), act to chaperone new proteins in the cell.

After sequestering hydrophobic protein folding intermediates, DnaK passively supports protein folding through cyclically releasing and binding partially folded proteins in an ATP hydrolysis dependent manner (Deuerling *et al.*, 1999; Teter *et al.*, 1999). The half-time for DnaK cycling is on the order 2 minutes (Liberek *et al.*, 1995), which provides sufficient time for polar regions to productively interact while aggregation prone hydrophobic regions are protected. The cyclic nature of DnaK action allows intra-molecular hydrophobic interactions to occur when appropriate. *E. coli* DnaK can identify partially folded proteins by the presence of its binding motif that occurs on average once every 36 amino acids in the *E. coli* proteome (Rüdiger *et al.*, 1997). These recognition sequences distinguish partially folded states from natively folded proteins as mature protein states do not contain the DnaK binding sequence on their surfaces and therefore can not interact with DnaK.

In addition to passive protein chaperones, *E. coli* produces the GroE chaperone protein complex, which takes an active role in protein folding by forming a large, empty, hydrophilic chamber where a protein may fold without inter-molecular interactions (Xu *et al.*, 1997; Hartl *et al.*, 2011). The GroE complex is comprised of

two GroEL heptamers in alternate, ATP-dependent interconverting conformations which vary significantly in cavity size and hydrophobicity (Figure 1.3). The *trans*-cavity, which initially interacts with nascent protein chains, has a small (85,000 Å<sup>3</sup>) hydrophobic cavity while the *cis*-conformation has a large (175,000 Å<sup>3</sup>) hydrophilic cavity (Fenton *et al.*, 1994; Hlodan *et al.*, 1995; Itzhaki *et al.*, 1995; Xu *et al.*, 1997). The hydrophobic nature of the GroEL *trans*-cavity promotes its association with hydrophobic, partially folded proteins, while the smaller cavity acts as a gatekeeper to restrict the size of the encapsulated proteins to below 70 kDa; the *trans*-GroEL cavity accommodates maximally a globular folded 60 kDa protein (Xu *et al.*, 1997; Fenton and Horwich, 2003). After the *trans*-GroEL cavity binds an unfolded polypeptide, a heptamer of GroES seals the GroEL cavity, triggering a shift to the large, hydrophilic *cis*-cavity conformation which provides a suitable protected, polar and dilute environment for protein folding (Saibil *et al.*, 1993; Xu *et al.*, 1997; Horwich *et al.*, 2009). Dissociation of the GroES cap allows the folded protein to exit GroEL.

The half-time between the GroEL *cis*-conformational change and protein release is 10 s (Fenton *et al.*, 1996; Rye *et al.*, 1997; Taguchi *et al.*, 2001; Hartl *et al.*, 2011), allowing even very long folding trajectories to proceed. In fact the extended folding time allows thermodynamic sampling of multiple protein conformations (Clarke, 1996; Ellis and Hartl, 1996). GroEL also appears to act as a passive protein folding chaperone, as experiments supplementing *E. coli* with a GroEL mutant incapable of forming the *cis*-form, can still promote correct folding of proteins, even in cases where the proteins are too large to be encapsulated by *trans*-GroEL (82 kDa; Chaudhuri *et al.*, 2001).



**Figure 1.3: Crystal structure of the GroE complex.** The GroE complex is composed of three heptameric structures, GroES<sub>7</sub> (red); *cis*-GroEL<sub>7</sub> (yellow) and *trans*-GroEL<sub>7</sub> (blue). I) Nascent folded client proteins first interact with hydrophobic *trans*-GroEL<sub>7</sub> and *cis*-GroES<sub>7</sub> dissociates from the GroE complex. II) A large conformational change interconverts the *trans*- and *cis*- GroEL<sub>7</sub> cavities. GroES<sub>7</sub> then re-binds over the new hydrophilic *cis*-GroEL<sub>7</sub> cavity, trapping the client protein in the hydrophilic environment (PDB: 1AON; Xu *et al.*, 1997).

#### 1.2.3.1 Chaperone capacity in growing *Escherichia coli*

Under normal growth conditions, *E. coli* clearly has sufficient chaperone capacity to fulfil cellular needs. However, protein over-expression in the laboratory can very quickly overwhelm the cellular chaperone apparatus and the regulatory networks that maintain balanced protein folding. Laboratory strains of *E. coli*, when grown at 37°C in minimal medium, contain on average  $3.52 \times 10^6$  proteins ( $2.34 \times 10^{-13}$  g protein/cell; average protein mass of 40 kDa; Bremer and Dennis, 1996). With a doubling time of 40 min, average protein synthesis is around 1,500 proteins per second. Some, but my no means all of these proteins require chaperone assistance to fold correctly.

Under standard conditions, there are around 12,000 DnaK molecules within an *E. coli* cell (Oh and Liao, 2000; Tomoyasu *et al.*, 2002; Lu *et al.*, 2006). With an ATP hydrolysis cycle half-life of 2 min (Liberek *et al.*, 1995), there is insufficient DnaK capacity inside of the cell to accommodate the  $\sim 180,000$  proteins produced during this time. DnaK is implicated in productive folding of 20% of nascent proteins by mass. This is biased towards larger proteins (Teter *et al.*, 1999), likely due to an increased number of DnaK binding sites. In fact, large proteins appear to rely more on DnaK to fold productively; the majority of proteins greater than 60 kDa are detectable in aggregates following DnaK depletion (Deuerling *et al.*, 1999, 2003).

Under similar conditions, *E. coli* has a GroEL 14-mer concentration of about 1,580 per cell (Lorimer, 1996; Vanbogelen *et al.*, 2005). Accounting for the 10 s half-time in which GroEL encapsulates a nascent protein within the *cis*-cavity, and assuming only a single binding cycle with GroEL, it is apparent that *E. coli* has an order of magnitude fewer available GroEL complexes than newly produced proteins. However, not all nascent proteins synthesised in *E. coli* interact with the GroE complex to fold. Few proteins less than 20 kDa interact with GroEL during folding as these proteins most often fold quickly and efficiently. It is the larger, multi-domain proteins that require the help of GroEL (Houry *et al.*, 1999). It has been shown that about 300 (12%) of the *E. coli* cytosolic proteins interact with GroEL during normal culture conditions, accounting for up to 30% of the chaperone capacity. These clients are mostly between 20 and 60 kDa in size (Houry *et al.*, 1999; Kerner *et al.*, 2005; Fujiwara *et al.*, 2010).

Since the DnaK and GroE complex chaperone capacities of *E. coli* are capable of folding around 100, and 150 new protein molecules per second respectively, these cells therefore have the capacity to fold, at most, just 10% of total cell protein synthesis (assuming one molecule of chaperone is required per protein). Over-expression of a large multi-domain protein can therefore quickly become problematic and overload cellular chaperone capacity. In addition to over-expression of chaperone-requiring proteins overwhelming normal chaperone capacity, over-expression of poorly soluble protein often sequesters cellular chaperones to inclusion bodies as aggregation occurs while the chaperone is bound to its client (Allen *et al.*, 1992; Carrio *et al.*, 1998).

*E. coli* has developed feedback mechanisms to cope with environmental conditions which burden protein folding. However, these regulatory mechanisms are not optimised for artificial over-expression of a poorly soluble protein. The chaperones DnaJ and DnaK participate in a sequestration feedback loop where free chaperone sequesters the sigma factor responsible for chaperone gene transcription (El-Samad *et al.*, 2005). Unfortunately, chaperone up-regulation during high level protein over-expression loads produces or leads to maximally about a two fold increase in levels of both GroEL (Cheng and Lee, 2010) and DnaK (Oh and Liao, 2000; Tomoyasu *et al.*, 2002; Lu *et al.*, 2006). This is insufficient to cope with problematic proteins during their over-expression. Co-over-expression of cellular chaperones can improve yields of soluble proteins during over-expression but can be of limited use (Goloubinoff *et al.*, 1989; Cole, 1996; Thomas *et al.*, 1997; Machida *et al.*, 1998).

### 1.2.4 Passenger solubilisation

In some cases amino-terminal (N-) fusions of proteins such as thioredoxin (LaVallie *et al.*, 1993), glutathione S-transferase (Smith and Johnson, 1988; Nygren *et al.*, 1994), maltose-binding protein (MBP; di Guan *et al.*, 1988; Bach *et al.*, 2001; Salema and Fernández, 2013), *E. coli* N-utilizing substance A (NusA; Davis *et al.*, 1999), periplasmic protein disulfide isomerase (DsbA; Zhang *et al.*, 1998), protein A (Nilsson *et al.*, 1987; Samuelsson *et al.*, 1994), and ubiquitin (Power *et al.*, 1990) can be used to improve the solubility of a passenger fusion protein. However, these solubility-promoting fusion partner proteins are not equal in their ability to improve soluble passenger protein yield (Kapust and Waugh, 1999), and success can vary depending on the identity of the passenger protein (Hammarström *et al.*, 2002; Raran-Kurussi and Waugh, 2012). In some cases, they do not promote folding to native, active conformations (Louis *et al.*, 1991; Saavedra-Alanis *et al.*, 1994; Sachdev and Chirgwin, 1998; Raran-Kurussi and Waugh, 2012).

The role of passenger solubilising fusion proteins is not always clear, and may not be universal, but in the exceptionally useful case of MBP, it appears to have a chaperone-like effect, where MBP interacts with unfolded intermediates and prevents inter-molecular aggregation during high level over-expression (Richarme and Caldas, 1997; Kapust and Waugh, 1999; Bach *et al.*, 2001; Fox *et al.*, 2001; Raran-Kurussi and Waugh, 2012); essentially producing a chaperone molecule for each protein molecule synthesised.

N-terminal solubility promoting fusions are more successful than carboxyl-terminal



(C-) fusions, likely due to these domains being synthesised first and folding while the passenger protein is being synthesised. This proposition is supported by observations that C-terminal, in contrast to N-terminal, MBP and thioredoxin fusions do not usually support over-expression of soluble protein, but still may aid in protein re-folding (Sachdev and Chirgwin, 1998; Dyson *et al.*, 2004).

### 1.2.5 Protein domain evolution

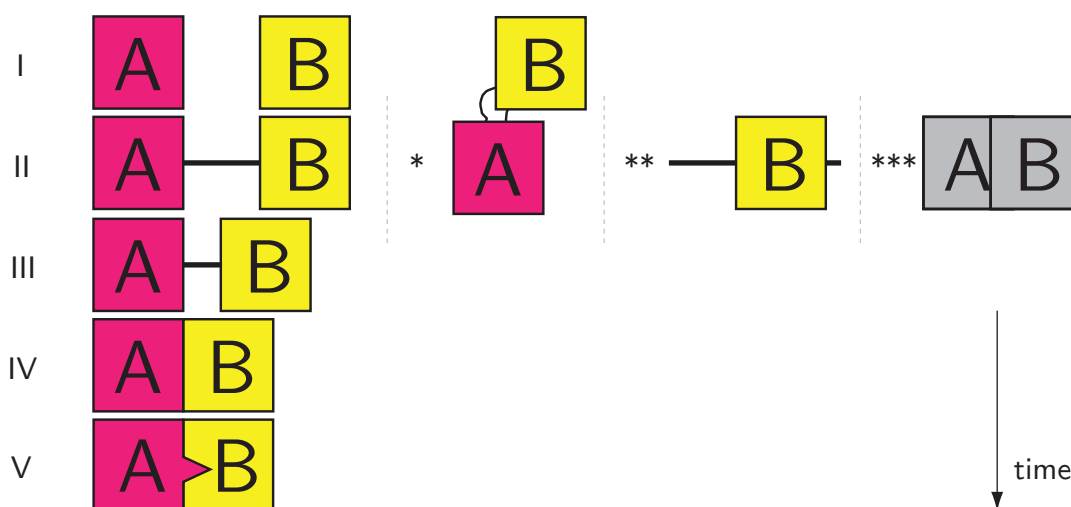
Protein domains are distinct independently folded protein units, and can manifest as distinct proteins, contain flexible extensions, or be joined into multi-domain assemblies. Distinct protein domains often represent modules of function, mutation and/or fusion of which can lead to new protein chemistry and functions (Apic *et al.*, 2001; Gough, 2005; Vogel *et al.*, 2005). Fusion of two protein domains allows new cooperative functions, for example enabling sequential reactions to be promoted (Ostermeier and Benkovic, 2000). *De novo* evolution of a functional polypeptide is many orders of magnitude slower than the rate at which an existing protein domain can differentiate to a new function or specificity, or be joined with other protein domains to create new combinations of specificity and function. In addition, distinct protein domains fold more quickly than multi-domain proteins. As such, protein domain recombination is a process through which many protein functions have developed over time (Apic *et al.*, 2001; Qian *et al.*, 2001; Gerstein and Levitt, 1998).

Recombination of genes encoding protein domains can produce a range of domain

configurations; however, most frequently the result is addition of DNA encoding a new domain at the end of an existing gene (Björklund *et al.*, 2005; Weiner *et al.*, 2006). Exploring the possible fusion events for protein domains shown in Figure 1.4, at step I) A and B are two separate genes. Then at step II) a recombination event produces soluble fusion protein A–B (various other possible configurations are indicated by \*, \*\*, \*\*\*). Insertion of a new domain into the start or end of a protein — where flexible non-structured regions are common and disruption of native contacts are less detrimental to protein folding — is the most likely event to produce a soluble fusion protein. The result of this type of gene fusion also commonly makes fused domains joined by flexible linkers. Other domain fusion orientations are possible if insertion occurs at a non integral position, such as an external loop (\*; Pascarella and Argos, 1992). A recombination event forming a fusion product of a partial protein domain A, can result in production of a fusion protein where a single domain is able to stably fold (B at \*\*), or otherwise a protein incapable of productively folding, resulting in aggregation (\*\*\*).

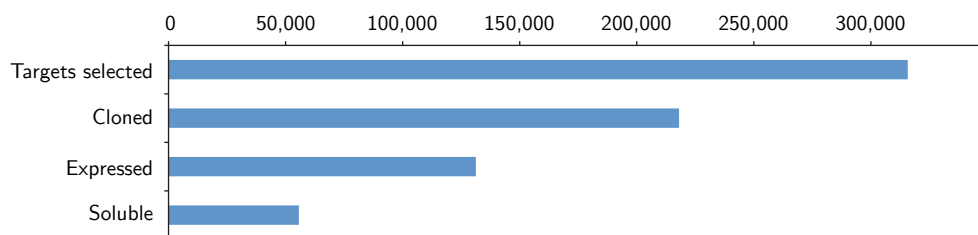
### 1.3 Small proteins are often better suited for three-dimensional structure determination

Determination of three-dimensional shape can help illuminate the properties of a protein of interest. However, many properties of a protein may make it unsuitable for structural determination. TargetTrack (formerly TargetDB; Chen *et al.*, 2004), a database tracking structural genomics sample progress, clearly



**Figure 1.4: Protein domain fusion is a process whereby genes encoding two independent protein domains recombine.** I) Protein domains A and B are independent. II) Insertion of a protein domain produces domains fused end to end, often linked by a flexible polypeptide linker; \* domain fusion can result in domain B inserting into a loop in domain A, producing a soluble multi-domain protein where both domain A and B can productively fold; \*\* insertion of protein domain B disrupts folding of domain A, producing a flexible extension to protein B; \*\*\* protein domains A and B fuse in a position which perturbs productive folding of the multi-domain protein producing a non-functional, aggregating product. III) and IV) Following domain fusion, flexible regions linking domains A and B can atrophy over time, reducing spacing between them. V) Domains A and B evolve a folding pathway which integrates A and B into a distinct folded unit.

shows the high attrition of target proteins through the pipeline to protein structure determination. Following cloning of target protein genes, both the lack of protein expression and solubility greatly reduce the number of useful targets (Figure 1.5). Longer proteins display poorer expressability, and separately, poorer protein solubility when over-expressed (Canaves *et al.*, 2004; Pédelacq *et al.*, 2005; Slabinski *et al.*, 2007). However, even when longer proteins are expressed in soluble form, other properties of larger size, such as increased flexibility and consequently poorer crystallisability, confer poor suitability for structural determination (Christendat *et al.*, 2000; Goh *et al.*, 2004). Currently, structural genomics initiatives try to avoid large proteins due to poor suitability and success rates (Oldfield *et al.*, 2005; Chandonia *et al.*, 2006; Slabinski *et al.*, 2007).



**Figure 1.5: Bottlenecks in structural biology pipelines.** TargetTrack (Chen *et al.*, 2004) lists success statistics for protein structural determination experiments from structural genomics groups. A large number of protein targets are not able to proceed past expression and soluble expression stages. Accessed 14 May 2013; <http://www.sbk.org/tt/>.

### 1.3.1 Nuclear magnetic resonance and protein size

Solution nuclear magnetic resonance (NMR) spectroscopy allows investigation of protein structure in the solution state, and can provide information on protein folding, structure, protein interfaces, binding processes and protein dynamics. Protein NMR has numerous inherent restrictions which bias applicability to small proteins, placing an upper limit on structural determination at 25 kDa for routine NMR (Güntert, 1998; Kwan *et al.*, 2011). Protein NMR experiments require high quality, well resolved NMR spectra, which become increasingly difficult to obtain as protein size increases.

Protein size is inversely correlated with molecular tumbling, where slow molecular tumbling reduces the lifetime of NMR excited states and broadens NMR spectra, giving rise to the size restrictions of NMR (Keeler, 2011; Kwan *et al.*, 2011). Molecular tumbling is further restricted at lowered temperatures where the increased viscosity of water slows the rate of protein tumbling and reduces NMR

signals. These issues restrict useful temperatures at which NMR measurements can be made and this further challenges structure determination of larger proteins that typically have reduced temperature stability. Analysis of proteins up to 50 kDa is plausible for many cases; however, proteins larger than 20–30 kDa require special high field instruments and sample deuteration (Gardner *et al.*, 1998; Kwan *et al.*, 2011). Although some complex state-of-the-art techniques increase the size threshold for protein NMR in solution into the 100s of kDa, they do not efficiently enable structure determination (reviewed in Kwan *et al.*, 2011).

Macromolecular NMR is also limited to well-behaved proteins. NMR produces low signal intensities, requiring high protein concentrations, and long timescales over which to take significant measurements. The requirement for high protein concentration restricts study of proteins of poor solubility, a problem that is exacerbated by the need for low ionic strength buffers for useful signal intensity (Kwan *et al.*, 2011).

Simplification of the protein molecule into smaller, soluble fragments reduces the complexity of both the NMR experiment and sample preparation. A protein of a single, soluble domain (10–20 kDa) is easier to produce, more likely to be soluble, produces a superior spectrum and is easier to analyse following NMR measurement. A more soluble protein will also often have reduced requirement for a high ionic strength buffer, and better temperature stability compared to a low solubility sample.

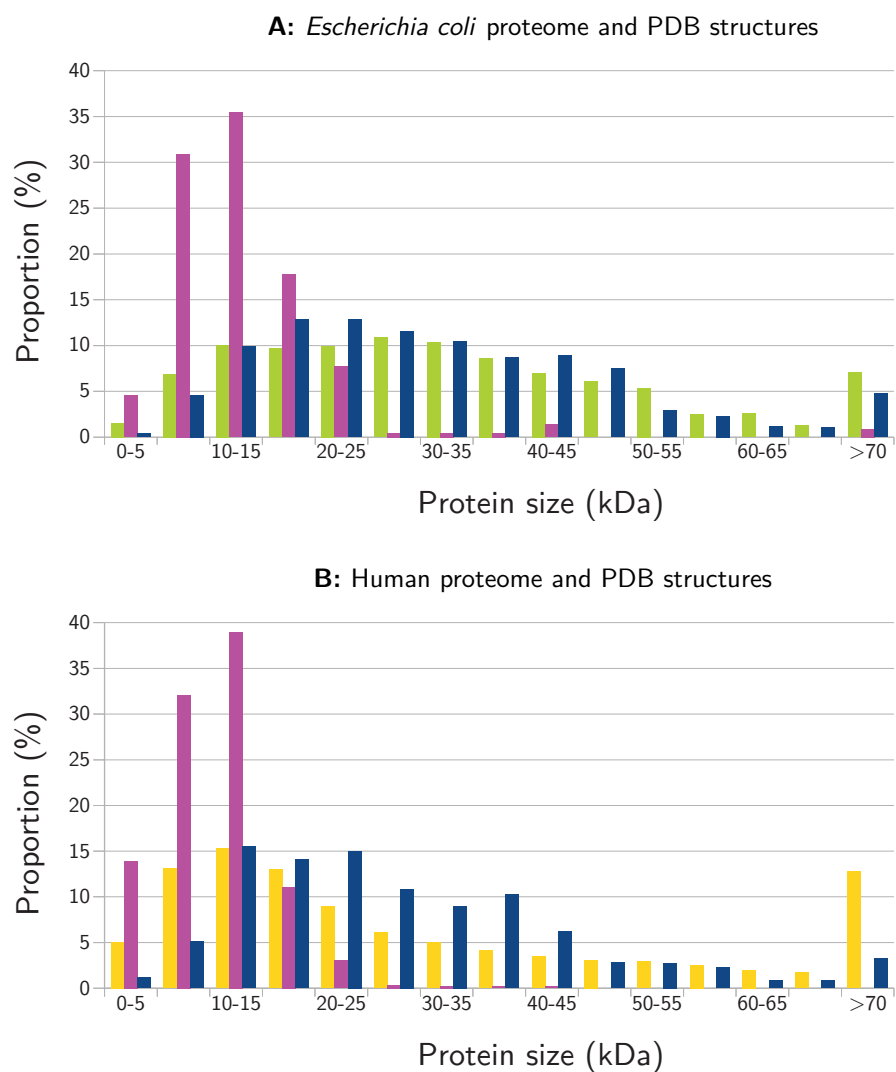
### 1.3.2 X-ray crystallography and protein size

Many characteristics of larger proteins can impede structure determination by X-ray crystallography, separate from protein expressability and solubility. Numerous studies have highlighted protein properties which negatively affect protein crystallisation, including average hydropathy, isoelectric point, presence and length of low-complexity disordered regions, proportion of coil and/or coiled-coil secondary structures, presence of signal peptides and presence of transmembrane helices (Canaves *et al.*, 2004; Goh *et al.*, 2004; Chandonia *et al.*, 2006; Slabinski *et al.*, 2007). These protein features that appear to impede their successful crystallisation are less associated with properties of individual folded domain units, but instead increase with numbers of domains, and the addition of associated flexible inter-domain regions.

Structures in the protein data bank (PDB) show an apparent size bias towards smaller proteins. Comparing the two proteomes most represented (*E. coli* and human) to molecule size in PDB entries reveals that proteins  $\sim 50$  kDa and above are under-represented and this bias may be related to the advantages of smaller proteins for these techniques (Figure 1.6).

## 1.4 Protein domain identification

The techniques used to produce proteins for biochemical study can produce poor yields for larger proteins due to poor expressability or solubility. In many



**Figure 1.6: Size distribution of published protein structures.** Comparison of size for **A** *Escherichia coli* and **B** human protein structures in the protein data bank (PDB), by technique and the respective proteome as of March 2013. Complete reference proteomes were extracted from UniProtKB (<http://www.uniprot.org>) and protein size estimated by converting length in amino acids to molecular weight (1 residue = 110 Da). The PDB structures were extracted based on molecular weight of chain A from all unique structures for each species (proteins with 100% similarity were removed). *E. coli* proteome, ■; human proteome, ■; NMR structures, ■; X-ray structures, ■.

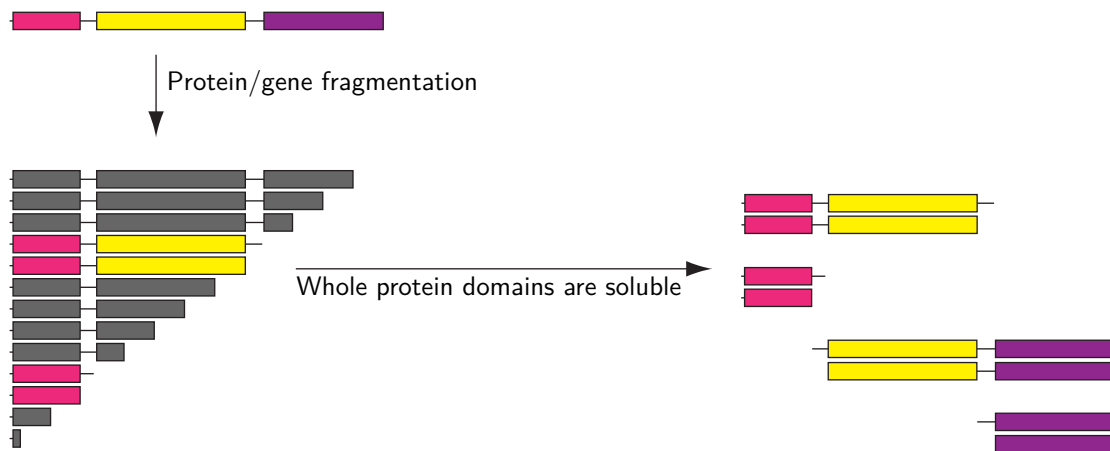
cases, over-expression of a larger, or otherwise problematic protein can overload the innate capacity of expressing cells to fold proteins. A large proportion of known proteomes are composed of large multi-domain proteins, 40% in both prokaryotic domains and 65% in eukaryotes (Ekman *et al.*, 2005). As distinct protein domains often fold more efficiently than larger multi-domain proteins, a greater yield of soluble protein can be achieved by targeting distinct protein domains. The common “domains in series” configuration of multi-domain proteins usually allows for straightforward segregation of large multi-domain proteins into distinct domains in determined structures, but it is often more difficult to predict domain boundaries from primary sequence alone.

In general, truncation of a multi-domain protein to complete folding domains produces protein fragments that are able to independently fold. However, truncation of a protein within a distinct folding region more often produces an aggregating protein species which can not fold appropriately (Figure 1.7). Many proteins can be fragmented effectively into distinct soluble domains, although some proteins prove problematic to fragment successfully. Indeed predicting protein domains can be difficult (Sippl, 2009), as can finding an appropriate position within a protein to form truncations.

#### 1.4.1 Computational methods for protein domain identification

Computational methods are available for identification of the modular pieces of a protein, which is often straightforward for proteins with close homologues





**Figure 1.7: Domain fragmentation.** Cartoon representation of a multi-domain protein and truncated protein domains. Long multi-domain proteins are composed of distinct folded domains. Truncation of individual domains allows biochemical study of each piece in isolation and can be more successful. Truncation of multi-domain proteins to distinct folded domains often results in a soluble protein. However, truncation of a protein to a position within a folded unit commonly produces proteins which do not fold. Coloured boxes represent well folded domains and grey boxes represent incomplete protein domains which can not productively fold.

of known structure or domain boundaries, especially when domains are linked by long flexible regions. Where close homologues are known, protein sequence alignments can provide suggestions on where a folding unit ends and can be useful in determining a suitable protein truncation. However, the ends of protein domains often play a smaller role in folding pathways and structure stabilisation, and are more likely to be obscured over evolutionary time.

#### 1.4.1.1 Hidden Markov models and Pfam for protein domain identification

Statistical methods for predicting protein domains assume that the function, structure and folding properties of a protein are conserved within a family, and that this information must be represented somehow in the primary amino acid

sequence. Thus, when folded protein features are conserved, the responsible components of the amino acid sequence must also be conserved (Krogh *et al.*, 1994). In the vast majority of instances, the structure and folding properties of proteins within a domain family are unknown, so statistical methods for identifying conserved folded features from the primary sequence are required.

Hidden Markov model (HMM) profiles provide the probabilities of the identity of each sequence position in a family of proteins and can even cope with insertions and deletions. These HMM profiles can, with high success, identify related protein sequences (Krogh *et al.*, 1994), and with further examination of a HMM profile can inform on the common structure present in the family.

The HMM methodology is utilised for the Pfam database, a tool for analysis and collection of protein family relationships (Bateman *et al.*, 2000, 2002). For proteins represented by a Pfam entry (which covers about 80% of the UniProt Knowledgebase; Punta *et al.*, 2012), domain annotation can often be straightforward. However, computational methodologies may not be appropriate in all cases and are not useful for proteins with no homologues of known structure. Knowing the likely protein domain boundary does not always help in identifying an engineered protein that can be expressed in a soluble form.

### 1.4.2 Experimental methods for protein domain identification

Varied methodologies are available for experimentally identifying protein domains and these work either by truncating the protein itself or by expressing genes coding for shortened mutants of the protein of interest. Limited proteolysis approaches the identification of protein domains by relying on protease sensitivity of the flexible inter-domain linkers commonly present between protein domains, and has been in use for a long time (Schechter and Berger, 1967; Fontana *et al.*, 2004). Treatment of purified multi-domain proteins with limited amounts of purified protease can result in cleavage of the full-length multi-domain protein preferentially at flexible inter-domain linkers. These flexible linkers are able to adapt into the protease active site, where bulky domains are inaccessible (Herschlag, 1988). It can then be inferred from the positions of proteolysis which regions of a protein are distinct folding units.

For limited proteolysis, useful protease active sites are selective and require peptide binding in an unfolded state, which can often require a flexible extension 8–10 amino acids long that is not obstructed by the folded domains present (Herschlag, 1988). However, limited proteolysis does not select proteins for the complete information to correctly fold or for regions important for maintaining protein solubility, and in some cases, domain fragments liberated by limited proteolysis are not expressable as soluble protein and/or are not soluble, presumably because protease cleavage can often occur at surface loops within folded domains.

Linkers for closely fused domains are poorly accommodated within protease active

sites and these are harder to identify (Tougu *et al.*, 1994), and potentially flexible linkers, which are identified in a binary manner with limited proteolysis, may be in some form, functional. Also, to perform limited proteolysis, one must be able to produce folded full-length protein. Methodologies which do not require the ability to first produce soluble full-length proteins can potentially provide protein domain information where techniques such as limited proteolysis can not.

## 1.5 Protein domain libraries

Gene fragmentation approaches the domain identification problem from a perspective of protein folding; a truncated gene is expressed and assessed for the ability to fold into soluble protein. Gene fragment libraries, where many length variants are produced, contain protein coding genes shortened by some semi-random process (Ostermeier *et al.*, 1999; Ostermeier, 2003), which must then be screened for production of soluble protein.

Methodologies for gene fragmentation can be classified into two modes: I) uni-directional and II) bi-directional fragmentation. Uni-directional gene truncation methodologies commonly utilise directional nuclease enzymes such as exonuclease III (*ExoIII*) to processively remove one-or-other gene end (Ostermeier *et al.*, 1999), while bi-directional fragmentation methodologies fragment the gene at each end simultaneously by digestion with DNase I, utilising endonuclease V hydrolysis at randomly incorporated dUMP sites (Miyazaki, 2002), or simply by cloning sheared DNA. Bi-directional gene truncation has the advantage over

uni-directional truncation that all possible variants of a gene may be produced; however, a complete bi-directional library for a gene  $N$  nucleotides long is comprised of  $\approx \frac{N^2}{2}$  mutants while complete uni-directional libraries contain only  $N$  mutants. Thus, uni-directionally deleted gene libraries are more straightforward to sample to completion and should be sufficient as most protein domains occur in series.

Production of uni-directionally deleted protein-fragment libraries can be targeted for discovery of useful suspected domain boundaries, but may also be performed over the entire gene to illuminate soluble-protein break points that are not otherwise obvious. The methodology also has the advantage over limited proteolysis, in that a number of slightly distinct deletions may be identified, for example including and excluding flexible linkers or disordered regions that may play functional roles.

## 1.6 Selecting for protein solubility

There are two categories of methodologies to screen for over-produced protein solubility: I) direct examination from a soluble cell lysate, or II) implicit examination using a proxy — where a measurable signal can be generated *in vivo* that indicates soluble/insoluble protein has been over-expressed — to deduce expression of soluble protein. Direct measurement of protein solubility from cell lysates is straightforward; however, this methodology is poorly suited for examining many proteins in a timely fashion.

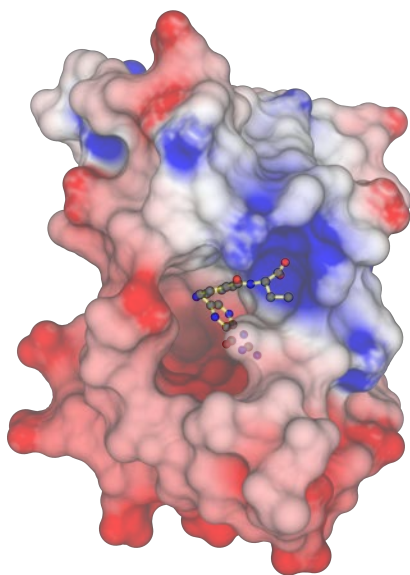
Proxy measurements for protein solubility allow uncomplicated examination of protein solubility during over-expression by examining colony phenotype or growth characteristics of the cell lines involved, and can be carried out at scale. To this end, protein solubility reporting techniques regularly utilise a C-terminal fusion protein, which allows for selection based on truncated gene reading frame, and protein solubility (Waldo, 2003; Dyson *et al.*, 2008). However, proxy measurements for protein solubility require greater care to ensure the protein is folded and useful; small peptides, proteolytic fragments, unintended translation initiation and “passenger solubilised” proteins can produce a soluble proxy protein in the absence of a useful protein of interest (Kawasaki and Inagaki, 2001; Nakayama and Ohara, 2003). Other methodologies for discovering well folded proteins include selection regimes utilising phage display (Sieber *et al.*, 1998; Jung *et al.*, 1999; Christ and Winter, 2006) or use of cellular stress/mis-folding responsive promoters, which drive a reporter gene (Cha *et al.*, 1999; Cortazzo *et al.*, 2002; Lesley *et al.*, 2002).

### 1.6.1 Dihydrofolate reductase and solubility reporting

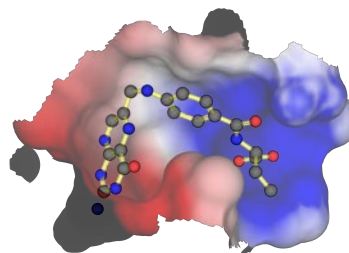
Dihydrofolate reductase (DHFR) is essential in *E. coli* as the principal source of tetrahydrofolate (THF) by NADPH reduction of dihydrofolate (DHF). THF is itself essential due to its involvement as a coenzyme for synthesis of purines, thymidylate and several amino acids (Fierke *et al.*, 1987). DHFR can be inhibited with the antibiotic trimethoprim (TMP) which binds in the DHF binding site

of DHFR in place of DHF (Figure 1.8; Sawaya and Kraut, 1997; Barrow *et al.*, 2004).

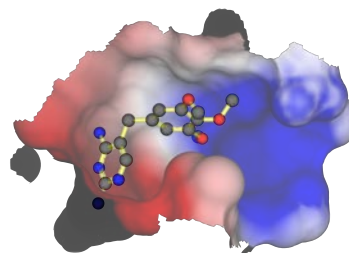
**A:** *Escherichia coli* dihydrofolate reductase occupied by dihydrofolate in its binding pocket



**B:** *E. coli* DHFR: dihydrofolate binding site occupied by dihydrofolate



**C:** *E. coli* DHFR: dihydrofolate binding site occupied by trimethoprim



**Figure 1.8: Trimethoprim blocks the dihydrofolate binding site in dihydrofolate reductase. A:** Front view of *Escherichia coli* DHFR in complex with dihydrofolate (PDB: 1RF7; Sawaya and Kraut, 1997). **B:** Close up of *E. coli* DHFR binding site with resident DHF. **C:** Trimethoprim was modelled into 1RF7 using the structure of *Bacillus anthracis* DHFR with bound TMP (PDB: 3JW3; Barrow *et al.*, 2004) and is shown as a close up of *E. coli* DHFR binding site.

Complementation of *E. coli* growth with TMP resistant murine DHFR in the presence of TMP forms the basis of several bacterial selection systems (Mössner *et al.*, 2001; Koch *et al.*, 2006; Dyson *et al.*, 2008). However, over-expression of soluble, endogenous *E. coli* DHFR can also overcome the TMP sensitivity of *E. coli* (Liu *et al.*, 2006). As such, both murine and *E. coli* DHFR have been used

as solubility reporting fusion partners (Liu *et al.*, 2006; Dyson *et al.*, 2008) where only cells expressing soluble DHFR-fusions can be grown in the presence of TMP. However, this methodology has one *caveat*: fusion proteins that are capable of correctly folding, yet are aggregation prone, are capable of “mopping up” TMP prior to aggregation, thus it may serve more as a reporter of protein folding than of overall protein folding and aggregation resistance.

Liu and colleagues (2006) showed that strains expressing various proteins fused to the N-terminus of DHFR are resistant to TMP in proportion to the solubility of the N-terminal protein. In these experiments two of the proteins were insoluble when expressed alone and were not solubilised when expressed as a DHFR fusion. After showing that TMP resistance was correlated with the solubility of the N-terminal fusion protein, the authors then searched for solubilising mutations of a newly discovered acetyltransferase (ACE) by performing DNA shuffling (Stemmer, 1994). The resulting mutated ACE genes were then expressed as a library of N-terminal fusion proteins to DHFR in *E. coli*, and from these they were able to identify soluble ACE mutants by their resistance to TMP.

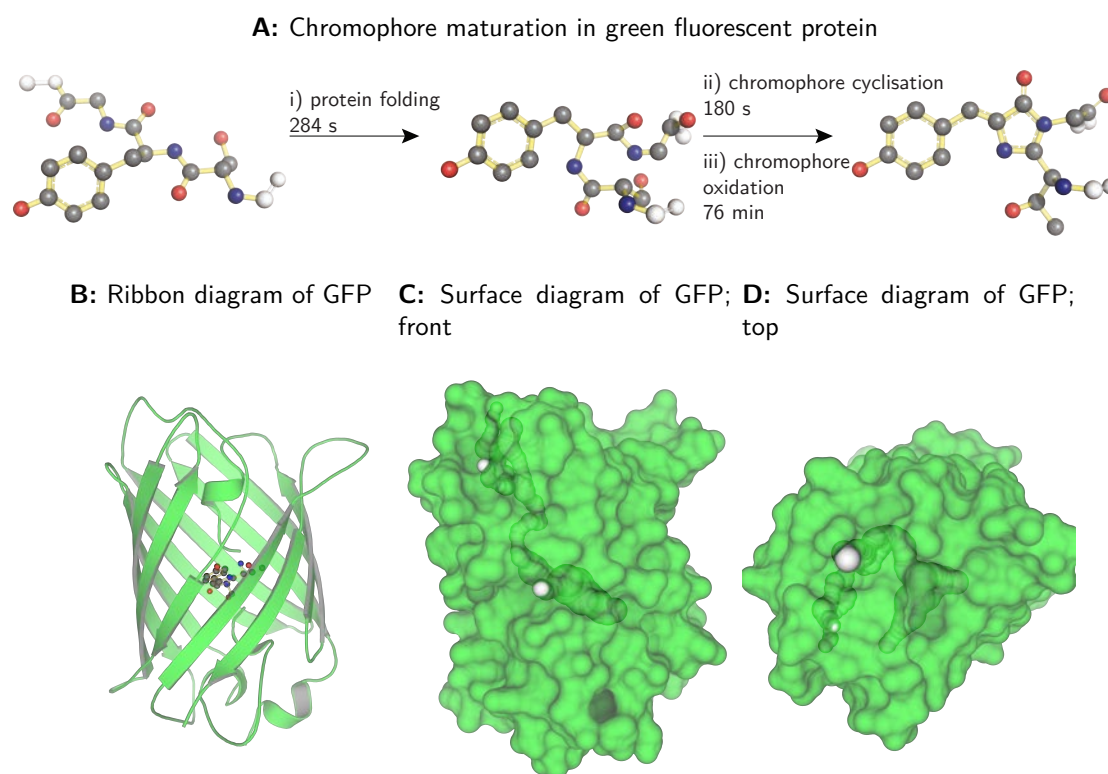
In another study, Dyson and colleagues (2008) were able to isolate soluble protein fragments using murine DHFR as a C-terminal solubility reporting fusion protein. Genes for a transcription factor and cell adhesion protein were randomly fragmented with a nuclease, cloned upstream of DHFR and then expressed as libraries in *E. coli*. From these experiments they were able to identify numerous soluble fragments of both proteins by selecting for strains that could grow in the presence of TMP.



### 1.6.2 Green fluorescent protein and solubility reporting

Green fluorescent protein (GFP) is a well documented protein solubility reporter (Waldo *et al.*, 1999; Pédelacq *et al.*, 2002; Nakayama and Ohara, 2003; Yumerefendi *et al.*, 2010). Maturation of GFP — from folding to assembly of its chromophore — is a slow, multi-step and ordered process that has mostly been studied *in vitro* (Figure 1.9; Reid and Flynn, 1997; Sniegowski *et al.*, 2005b; Pouwels *et al.*, 2008); however, the time-scale for fluorescence development *in vivo* is similar (Reid and Flynn, 1997). The GFP maturation process proceeds through protein folding and then intra-molecular cyclisation at residues Ser65–Tyr66–Gly67, after which reaction of the modified SYG residues with molecular oxygen generates the green fluorescent chromophore (Figure 1.9A Reid and Flynn, 1997; Pouwels *et al.*, 2008). For maturation of the chromophore, each step must occur in sequence. Correct protein folding places the SYG residues in a conformation required for rearrangement and does not occur in incorrectly folded GFP and non-cyclised SYG does not oxidise (Reid and Flynn, 1997; Pouwels *et al.*, 2008). During the final step of chromophore maturation, the chromophore is protected inside the GFP structure and does not have ready access to oxygen, which must instead diffuse through restricted pores within GFP (see Figure 1.9B–D; Heim *et al.*, 1994; Reid and Flynn, 1997; Rosenow *et al.*, 2005; Sniegowski *et al.*, 2005a). The unfavourable oxidation step proceeds with a half-time on the order of one hour.

The slow maturation rate of GFP makes it exceptionally useful as a protein-fusion



**Figure 1.9: Chromophore maturation in green fluorescent protein.** Development of green chromophore fluorescence in green fluorescent protein takes a long time after expression of the protein. GFPs have internal cavities through which molecular oxygen can access the immature chromophore. The crystal structure for GFP (PDB: 1EMA; Ormo *et al.*, 1996) was analysed using Caver 3 (Chovancova *et al.*, 2012), identifying internal channels with internal bottleneck dimensions greater than 0.9 Å. **A:** Ball-and-stick representation of the GFP chromophore maturation pathway (reaction half-times from Reid and Flynn (1997)). **B:** Ribbon diagram of GFP (front view) showing the internal chromophore as a ball-and-stick model. **C:** (front) and **D:** (top) Surface diagram of GFP (green) with internal cavities (white).

solubility reporter. Once GFP has folded correctly, molecular oxygen must diffuse to the solvent restricted protein core to oxidise the chromophore intermediate to give rise to fluorescence and this does not occur when GFP is expressed in inclusion bodies (Reid and Flynn, 1997). A GFP-fusion must fold correctly and be maintained in solution for a long time prior to attainment of green fluorescence.

Numerous studies have used C-terminal GFP-fusions to measure the solubility and stability of fused proteins (e.g., Pédelacq *et al.*, 2002; Nakayama and Ohara, 2003; Pédelacq *et al.*, 2005; Yumerefendi *et al.*, 2010). For example, Waldo and colleagues (1999) compared 20 *Pyrobaculum aerophilum* proteins that were expressed with and without C-terminal GFP fusions in *E. coli*. These authors found a very strong correlation between the measured fluorescence of cells that expressed fusion-GFP and the amount of soluble protein without GFP. Waldo *et al.* (1999) then randomly mutated genes for an insoluble ferritin protein and cloned them to make C-terminal GFP fusions. When they expressed a library of mutated ferritin-GFP genes in *E. coli* they were able to identify soluble and functional ferritin. These results show that GFP is useful for correlating fluorescence and the expression of productively folded and soluble proteins fused with GFP.

## 1.7 Aims of this Thesis

In contrast to full-length proteins, smaller, distinct protein domains are generally over-expressed more efficiently, fold faster and more efficiently, require less folding

chaperone activity, are less likely to aggregate, and are more suited to structural determination and other studies. Yet, protein domains are individual modules of function and can yield much information about the biochemistry of a full-length protein. However, identification of useful truncated protein domains can be difficult. This work aimed to develop a pragmatic methodology to identify soluble protein fragments (Chapter 3). We propose a plasmid system for gene truncation and protein solubility selection using GFP or DHFR as solubility reporting fusion proteins. Neither GFP nor DHFR appear to passenger solubilise protein fusions (Dyson *et al.*, 2004) and have previously been used for similar purposes (Waldo *et al.*, 1999; Pédelacq *et al.*, 2005; Liu *et al.*, 2006; Dyson *et al.*, 2008).

Although bi-directional gene fragmentation can potentially produce isolated examples of every domain present within a protein — and individual protein folding units may be better behaved than a series of domains — this new technique deletes genes uni-directionally. This methodology provides for more straightforward sample manipulation and mutant analysis is more straightforward as the possible variation is much smaller ( $N$  *vs.*  $\approx \frac{N^2}{2}$  for bi-directional truncation). Where the goal is identification of protein domain end points, deletion of a gene uni-directionally and then selecting for soluble truncated proteins should be sufficient, as only truncation end points outside of domains should make soluble proteins. If indeed truncated proteins with more than one domain are less soluble than if they had been first produced as single domain constructs, in many cases they should still be partially soluble. Information about the solubility of mutants truncated at different positions in the protein sequence can be used to infer the proteins domain boundaries and facilitate the design of single domain proteins.

This Thesis further aimed to use this pragmatic domain identification technique to identify soluble domains of *Acinetobacter baylyi* DNA primase and the *Staphylococcus aureus* cell division protein EzrA. In the case of *A. baylyi* DNA primase, the protein has been resistant to identification of soluble constructs for the N-terminal domain by traditional means. Meanwhile no domains are known for *S. aureus* EzrA or its homologues.

# Chapter 2

## General materials and methods

### 2.1 Chemicals, reagents, enzymes and instruments

All chemicals and reagents were of high quality and were purchased from commercial suppliers. Restriction enzymes, mung bean nuclease and T4 deoxyribonucleic acid (DNA) ligase for molecular genetics were obtained from New England Biolabs. DNA polymerase I Klenow fragment, exonuclease III, S1 nuclease and T4 DNA ligase for construction of DNA truncation libraries were from Promega. KOD DNA polymerase was from Novagen. ACCUZYME, AccuSure, VELOCITY and BIOTAQ Red thermostable DNA polymerases were supplied by Bioline (Australia), as were HyperLadder I DNA molecular size marker and HyperPAGE protein molecular weight marker.

MilliQ water (18 M $\Omega$ .cm resistivity ultrapure water) was sourced from a MilliQ system (Merck Millipore) and used in all experiments.

## 2.2 Bacterial strains and transformation

The *E. coli* strains used in this study and their genotypes are listed in Table 2.1.

**Table 2.1:** Bacterial strains used in this work.

Strain	Genotype
AN1459 <sup>1</sup>	<i>E. coli</i> K12 F <sup>-</sup> <i>supE44 thi-l leuB6 thr-l ilvC hsdR recA srlA::Tn10</i>
BL21( $\lambda$ DE3) <i>recA</i> <sup>2</sup>	<i>E. coli</i> B F <sup>-</sup> <i>ompT gal dcm lon hsdS<sub>B</sub>(r<sub>B</sub><sup>-</sup>m<sub>B</sub><sup>-</sup>) gal <math>\lambda</math>(DE3 [<i>lacI lacUV5-T7 genel indl sam7 nin5</i>]) <i>recA srlA::Tn10</i></i>
BL21( $\lambda$ DE3) <i>recA</i> /pLysS <sup>3</sup>	<i>E. coli</i> B F <sup>-</sup> <i>ompT gal dcm lon hsdS<sub>B</sub>(r<sub>B</sub><sup>-</sup>m<sub>B</sub><sup>-</sup>) gal <math>\lambda</math>(DE3 [<i>lacI lacUV5-T7 genel indl sam7 nin5</i>]) <i>recA srlA::Tn10 pLysS(Cm<sup>R</sup>)</i></i>

<sup>1</sup> Elvin *et al.* (1986)

<sup>2</sup> Studier and Moffatt (1986); Williams *et al.* (2002)

<sup>3</sup> Studier and Moffatt (1986); Studier (1991); Williams *et al.* (2002)

### 2.2.1 Routine growth of *Escherichia coli*

Strains of *E. coli* were routinely grown in LB medium (Millers modification; 10 g.L<sup>-1</sup> tryptone, 5 g.L<sup>-1</sup> yeast extract and 10 g.L<sup>-1</sup> NaCl; Miller, 1972) supplemented as required with agar (15 g.L<sup>-1</sup>) and antibiotics (see Table 2.2). Cultures were grown at 37°C (unless otherwise stated) with shaking or in a plate incubator as required.

**Table 2.2:** Antibiotics used in this work.

Antibiotic	Concentration used	Stock solution
ampicillin	100 $\mu\text{g.mL}^{-1}$	100 $\text{mg.mL}^{-1}$
chloramphenicol	34 $\mu\text{g.mL}^{-1}$	34 $\text{mg.mL}^{-1}$
kanamycin	50 $\mu\text{g.mL}^{-1}$	50 $\text{mg.mL}^{-1}$

The standard measure of culture growth was optical density at 600 nm ( $A_{600}$ ) measured using a BioPhotometer (Eppendorf). Culture density was measured in a 1 cm path-length cuvette after dilution in fresh growth medium to cell densities that gave readings below  $A_{600}$  of 1.

### 2.2.2 Clonal isolation of transformants

*E. coli* strains were clonally isolated by inoculation onto LB-agar using a sterile pipette tip. Successive streaks with a sterile inoculating loop provided single, isolated colonies after overnight incubation.

### 2.2.3 Transformation of *Escherichia coli*

Electrocompetent *E. coli* were prepared by inoculating LB with an overnight culture and grown to an  $A_{600}$  of 0.6 AU. Bacterial cells were then collected by centrifugation at  $8,000 \times g$  for 10 min. Cells were washed four times with  $\frac{1}{10}$  culture volume of water by resuspension and centrifugation at  $8,000 \times g$  for 10 min. Cells were resuspended in  $\frac{1}{50}$  culture volume of 10% glycerol, frozen in



liquid nitrogen and stored at  $-80^{\circ}\text{C}$ .

Transformation of electrocompetent *E. coli* cells was carried out using a MicroPulser system (Bio-Rad; Ausubel *et al.*, 1987; Miller and Nickoloff, 1995) as per manufacturer's instructions with a chilled 0.1 cm cuvette (single 1.8 kV pulse; Bio-Rad) with either 1  $\mu\text{L}$  of supercoiled plasmid DNA or 4  $\mu\text{L}$  of ligation mixture. Immediately following transformation, *E. coli* cells were recovered with 1 mL LB and incubated at  $37^{\circ}\text{C}$  for 1 h. Recovered transformed *E. coli* cells were spread on an appropriate LB agar plate and incubated overnight at the appropriate temperature.

Transformation efficiency of ligations was increased when required by heat inactivating T4 DNA ligase by incubation at  $65^{\circ}\text{C}$  for 10 min and/or desalting the ligation mixture using a Nanosep size exclusion spin column (MWCO 300 kDa, Pall).

#### 2.2.4 Long term storage of bacterial strains

For storage, *E. coli* strains were prepared by adjusting 1 mL of an overnight culture to contain 7% dimethyl sulfoxide (DMSO) and stored at  $-80^{\circ}\text{C}$ .

## 2.3 Molecular genetics

### 2.3.1 Preparation of plasmid DNA

#### 2.3.1.1 Plasmid purification by mini-prep

Small scale preparation ( $< 20 \mu\text{g}$  DNA) of cloned plasmids from *E. coli* was performed using QIAprep Spin Miniprep kit (QIAGEN) using the methods supplied, except that cells for plasmid DNA preparation were collected from overnight cultures on LB agar plates containing the appropriate antibiotics (Section 2.2.1) and plasmids were eluted and stored in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0 (TE).

#### 2.3.1.2 Plasmid purification by maxi-prep

Large scale purification ( $< 500 \mu\text{g}$  DNA) of plasmids from *E. coli* was performed using QIAGEN MAXI kits using the protocol supplied for low copy number plasmids. Plasmids were eluted and stored in TE.

### 2.3.2 Restriction digestion of DNA

Restriction endonuclease digests were carried out in buffers supplied and at temperatures recommended by the manufacturer (New England Biolabs; Appendix D). When two or more restriction enzymes with different buffer requirements were used, digestion conditions were as suggested by the manufacturer for double

or sequential digestion. As a general rule, reactions were designed to digest twice the amount of DNA present; the number of units of (each) enzyme was twice the DNA content of the mixture in  $\mu\text{g}$  and the reaction was for 1 h.

### 2.3.3 Preparation of oligonucleotides

Synthetic oligonucleotides used throughout this study were supplied as lyophilised powders at PCR/sequencing grade from GeneWorks. Prior to use oligonucleotides were resuspended and stored in TE at 100 mM ( $-20^{\circ}\text{C}$ ) or 5 mM ( $4^{\circ}\text{C}$ ). Oligonucleotide sequences can be found in Appendix C.

### 2.3.4 Amplification of DNA by polymerase chain reaction

Several thermostable DNA polymerases were used in this work; some of them were used under unique DNA extension conditions. Typical polymerase chain reaction (PCR) cycle parameters were 60 s incubation at  $96^{\circ}\text{C}$  followed by 30 cycles of  $96^{\circ}\text{C}$  for 15 s;  $55^{\circ}\text{C}$  for 15 s; and see Table 2.3 for DNA extension conditions. Reactions were completed with a final cycle at  $72^{\circ}\text{C}$  for 5 min.

A typical PCR mixture contained the appropriate enzyme buffer (Appendix D) with 500  $\mu\text{M}$  each deoxynucleoside 5'-triphosphate (dNTP) and 2 mM  $\text{MgCl}_2$  and 50 mUnits of DNA polymerase per 10  $\mu\text{L}$  total volume. PCR products were analysed by agarose gel electrophoresis (Section 2.3.6).

**Table 2.3:** Extension parameters for thermostable DNA polymerases

DNA polymerase	Extension temperature	Incorporation rate
BIOTAQ Red	72°C	30 s.kb <sup>-1</sup>
ACCUSURE (activated at 95°C for 10 min)	72°C	30 s.kb <sup>-1</sup>
ACCUZYME	72°C	120 s.kb <sup>-1</sup>
VELOCITY	72°C	30 s.kb <sup>-1</sup>
KOD	72°C	30 s.kb <sup>-1</sup>

#### 2.3.4.1 Colony PCR

Colony PCR was conducted on clonally isolated *E. coli* by inoculating a single colony to a 10 µL BIOTAQ Red PCR mixture and an LB agar plate in series. Reaction conditions were as in Section 2.3.4 with an initial denaturation step for three min at 96°C for cell lysis. Cells on the plate were grown overnight and PCR positive mutants were selected from the matching smear.

#### 2.3.5 Colony PCR and storage of truncation libraries

To allow storage of colonies from truncation libraries, the colony PCR methodology (Section 2.3.4.1) was modified to allow for freezer strains to be made at the same time as colony PCR. Prior to inoculation of BIOTAQ PCR mixtures, storage strains of isolated colonies were produced by inoculation of 200 µL LB medium containing 7% DMSO in wells of a 96-well plate. 1 µL aliquots of the colony storage solution were used to inoculate a colony PCR as per Section 2.3.4.1. Truncation library colonies in 96-well plates were stored at −80°C.

### 2.3.6 Electrophoresis of DNA

Agarose gels for electrophoresis of DNA were cast in a Mini-Sub Cell GT apparatus (Bio-Rad) containing agarose concentrations of 0.5–2% (w/v) in TBE (89 mM Tris-borate, 2 mM EDTA; Sambrook *et al.*, 1989) containing 0.5 µg/mL ethidium bromide. DNA samples were mixed with  $\frac{1}{6}$  volume of TriColor DNA loading dye (Bioline) and loaded into agarose wells. Samples were electrophoresed at 50 V, or transferred to the gel matrix at 20 V then separated at 50 V using a PowerPac Basic Power Supply (Bio-Rad). DNA in agarose gels was visualised using a long-wave UV handheld lamp or visualised and imaged using a short-wave UV transilluminator (Molecular Imager Gel Doc XR+ System; Bio-Rad).

### 2.3.7 Isolation of DNA reaction products by agarose gel electrophoresis

Isolation of DNA fragments separated by agarose gel electrophoresis (Section 2.3.6) was done using QIAquick Gel Extraction kits (QIAGEN). DNA bands were visualised under long-wave UV and gel slices extracted using a scalpel. DNA extraction from agarose gel slices followed manufacturer's instructions; DNA was eluted and stored in TE.

### 2.3.8 Isolation of DNA reaction products by silica column

Purification of PCR products and some DNA restriction fragments was carried out using QIAquick DNA purification kits. For purification of products larger than 100 bp, the manufacturer's protocol for purification of PCR products was followed. For purification of 70–100 bp products, a modified gel purification protocol was performed following manufacturer's instructions (substituting the total reaction volume for the value of gel volume). DNA was eluted and stored in TE.

### 2.3.9 Isolation of DNA reaction products by ethanol precipitation

DNA precipitation was performed by addition of 0.1 volume of 3 M sodium acetate (pH 5.3), 0.1 volume of 125 mM EDTA and 2.5 volumes of absolute ethanol. After 15 min at 0°C, the sample was centrifuged for 15 min at  $21,000 \times g$ . The supernatant was replaced with 70% ethanol ( $-20^{\circ}\text{C}$ ), re-centrifuged and then removed. Once the pellets had air dried, the purified DNA was gently resuspended in TE to a suitable volume.

### 2.3.10 Ligation of DNA

Routine DNA ligations were done using T4 DNA ligase (New England Biolabs) overnight at 4°C.

#### 2.3.10.1 Ligation of cohesive DNA termini

Typical subcloning or cloning of PCR products aimed to include 40 ng of plasmid DNA, 200 ng of insert DNA and 200 cohesive end units of T4 DNA ligase per 20  $\mu$ L reaction in New England Biolabs T4 DNA ligase buffer (Appendix D).

#### 2.3.10.2 Ligation of oligonucleotide linkers to plasmid DNA

Routine ligation of oligonucleotide linkers into plasmid DNA aimed to include 40 ng of plasmid DNA, 25  $\mu$ M of each unphosphorylated, unhybridised oligonucleotide and 200 cohesive end units of T4 DNA ligase per 20  $\mu$ L reaction in New England Biolabs T4 DNA ligase buffer.

#### 2.3.11 Dye terminator sequencing of DNA

Plasmid DNA sequences were determined using DNA products made using ABI BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). Cycle parameters were: 60 s incubation at 96°C, followed by 10 cycles of 96°C for 10 s, 50°C for 5 s, 60°C for 75 s, then 5 cycles of 96°C for 10 s, 50°C for 5 s, 60°C for 90 s, then 5 cycles of 96°C for 10 s, 50°C for 5 s, 60°C for 120 s (Platt *et al.*, 2007). Product DNA was recovered by ethanol precipitation (Section 2.3.9). Purified sequencing reaction products were separated and visualised on an ABI 3130xl Genetic Analyzer (Applied Biosystems). Sequencing of the protein coding regions and multiple cloning sites (MCSs) within T7 expression plasmids used primers PET3, PET4, 235 and 549; PET3 primes the sense strand from the T7 promoter,

PET4 primes the anti-sense strand from the T7 terminator and primers 235 or 549 prime the anti-sense strand from the start of the *egfp* gene (Table 2.4).

**Table 2.4:** Sequencing primers.

Primer	Sequence 5'-3'	Position	DNA Strand
PET3 <sup>1</sup>	CGACTCACTATAGGGAGACCACAAC	T7 promoter	sense
PET4 <sup>1</sup>	CCTTTCGGGCTTTGTTAGCAG	T7 terminator	anti-sense
235	CTCGCCCTTGCTCACC	start of <i>egfp</i>	anti-sense
549	CAGGATGGGCACCACC	start of <i>egfp</i>	anti-sense

<sup>1</sup> (Neylon *et al.*, 2000)

## 2.4 Estimation of DNA concentrations

### 2.4.1 Estimation of DNA concentration by agarose gel electrophoresis

Estimation of DNA concentration by agarose gel electrophoresis was conducted with multiple dilutions of a single sample by comparison with known amounts of DNA molecular size standard, following staining with ethidium bromide and visualisation with UV irradiation.

### 2.4.2 Spectrophotometric determination of DNA concentration

Spectrophotometric estimation of DNA concentration was performed by measuring absorbance at 260 nm ( $A_{260}$ ) using a NanoDrop 2000c instrument (Thermo Scientific), using two  $\mu$ L samples referenced against equivalent/identical



buffer. In early experiments, spectrophotometric estimation of concentrations of DNA was made by measuring  $A_{260}$  using a UVette cuvette and BioPhotometer spectrophotometer (Eppendorf).

## 2.5 Protein over-expression and purification

Protein coding DNA sequences were placed downstream of the T7 promoter and the ribosome binding site (RBS) of vector pETMCSI (and derivatives; Neylon *et al.* 2000), by placing the initiation codon within the *Nde*I (5'-CATATG-3') site. Alternatively N-terminal deletion plasmids use an *Nco*I (5'-CCATGG-3') site for initiation following library construction. pETMCSI and related plasmids confer resistance to ampicillin through the included *bla* gene.

### 2.5.1 Protein expression by auto-induction

Over-expression of protein encoded on T7 promoter plasmids was carried out in the *E. coli* BL21( $\lambda$ DE3)*recA* strain by the auto-induction method (Studier, 2005). Culture vessels (2 L baffled Erlenmeyer flasks) contained 350 mL of ZYP-5052 medium (10 g.L<sup>-1</sup> tryptone, 5 g.L<sup>-1</sup> yeast extract, 2 mM MgSO<sub>4</sub>, 50 mM Na<sub>2</sub>HPO<sub>4</sub>, 50 mM KH<sub>2</sub>PO<sub>4</sub>, 25 mM [NH<sub>4</sub>]<sub>2</sub>SO<sub>4</sub>, 5 g.L<sup>-1</sup> glycerol, 0.5 g.L<sup>-1</sup> glucose, 2 g.L<sup>-1</sup>  $\alpha$ -lactose monohydrate) and were incubated at 30°C with shaking at 200 rpm for 24–36 h. Following cell growth, all steps in protein isolation were conducted at 4°C unless otherwise stated.

### 2.5.2 Cell lysis by French press

As required, *E. coli* cells were harvested by centrifugation ( $8,000 \times g$ ), resuspended in a lysis buffer containing 50 mM Tris-HCl, 300 mM NaCl (15 mL per g of cells), and lysed in a French press cell disruptor using a 40 K pressure cell (Thermo Electron Corporation) set to 16,000 psi. Typical samples were lysed in 30 mL aliquots with three passes through the chilled cell.

### 2.5.3 Clarification of bacterial lysates

Following French press lysis, cellular lysate was clarified by centrifugation in a chilled SS34 rotor in a Sorvall RC6+ centrifuge at  $30,000 \times g$  for 30 min. The supernatant containing soluble proteins was either flash frozen with liquid N<sub>2</sub> and stored at  $-80^{\circ}\text{C}$  or directly processed. When required, proteins in the pellet were solubilised in 8 M urea.

### 2.5.4 Protein purification using ÄKTA FPLC systems

Proteins were purified chromatographically using fast protein liquid chromatography (FPLC) with ÄKTApurifier systems (GE Healthcare Life Sciences). Resins, protocols and columns used in protein chromatography are described in the text. Proteins and buffers for chromatography were vacuum filtered through 0.45 µm filters (Merck Millipore).

### 2.5.5 Determination of protein concentration

Protein concentrations were determined by measuring absorbance at 280 nm ( $A_{280}$ ) using either 500  $\mu$ L quartz cuvettes and a BioPhotometer spectrophotometer (Eppendorf) or two  $\mu$ L samples using a NanoDrop 2000c. Predicted molar extinction coefficients of proteins were estimated from their amino acid compositions using ProtParam (Gasteiger *et al.*, 2005).

### 2.5.6 Protein dialysis

Protein buffer exchange was performed by dialysis in Spectra/Por dialysis membrane sacs (Spectrum Laboratories), typically with a molecular weight cut off (MWCO) less than half of the protein's molecular weight. Dialysis was at 4°C using at least 10 volumes of buffer relative to sample, with three buffer changes.

### 2.5.7 Storage of proteins

Partially or fully purified proteins were frozen in liquid N<sub>2</sub> and stored at −80°C.

### 2.5.8 Concentration of protein samples

Purified proteins were concentrated using Amicon Ultra-4 centrifugal filter units (Merck Millipore) with a MWCO less than half of the protein's molecular weight following manufacturer's instructions. Filter units were rinsed repeatedly

with MilliQ water prior to addition of the protein sample. Centrifugation for protein concentration used a Sorvall Super T21 refrigerated centrifuge fitted with an ST-H750 rotor (Thermo-Scientific). To prevent over concentration, protein samples were mixed by inversion between additional centrifugation steps.

### 2.5.9 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis

Electrophoresis of proteins under denaturing conditions used NuPAGE Novex Bis-Tris pre-cast polyacrylamide gels (Invitrogen) or Mini-PROTEAN TGX 4–20% pre-cast polyacrylamide gels (Bio-Rad), with electrophoresis at 200 V for an appropriate duration.

Proteins for sodium dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis (PAGE) were mixed with an appropriate amount of either SDS sample buffer (1X: 50 mM Tris-HCl pH 6.8, 2% w/v SDS, 0.1% bromophenol blue; Sambrook *et al.*, 1989) or NuPAGE LDS sample buffer (1X: 61.75 mM Tris-HCl pH 8.5, 0.5% lithium dodecyl sulphate, 2.5% glycerol, 0.13 mM EDTA, 0.06 mM SERVA Blue G250, 0.44 mM phenol red; Invitrogen) freshly prepared with 50 mM dithiothreitol (DTT), and then heated at 80°C for 10 min prior to loading into an appropriate pre-cast SDS-PAGE gel. Before loading, crude protein samples from SDS-lysed cells were routinely drawn into a 50 µL syringe (Hamilton) to shear DNA.

SDS-PAGE gels were washed with water, then fixed and stained for 20 min in

hot staining solution (40% methanol, 10% acetic acid and 0.2% Coomassie blue) and destained in hot de-staining solution (10% acetic acid and 10% propan-2-ol). De-stained SDS-PAGE gels were imaged with a Molecular Imager Gel Doc XR+ System (Bio-Rad).

#### 2.5.9.1 Invitrogen NuPAGE gels

NuPAGE 4–12% Bis-Tris pre-cast gels (Invitrogen) were prepared in a XCell4 SureLock Midi-Cell tank (Invitrogen) with MES running buffer (50 mM MES, 50 mM Tris-base, 0.1% SDS, 1 mM EDTA, pH 7.3) as per manufacturer's instructions. Samples were loaded into wells and separated by electrophoresis for 45 min at 200 V with a PowerPac Basic Power Supply (Bio-Rad).

#### 2.5.9.2 Bio-Rad gels

Mini-PROTEAN TGX 4–20% pre-cast polyacrylamide gels (Bio-Rad) were prepared in a Mini-PROTEAN System tank (Bio-Rad) with Tris/glycine running buffer (25 mM Tris, 192 mM glycine, 0.1% SDS, pH 8.3). Samples were loaded into wells and separated by electrophoresis as above (Section 2.5.9.1).

### 2.5.10 Mass spectrometry of proteins samples

Prior to mass spectrometric analysis, protein samples were dialysed at 4°C (three buffer changes) in either native (100 mM ammonium acetate pH 7.2

and 1 mM  $\beta$ -mercaptoethanol) or denaturing (0.1% formic acid and 0.5 mM  $\beta$ -mercaptoethanol) buffers.

Mass spectrometry was done at the University of Wollongong Mass Spectrometry User Resource and Research Facility with either a quadrupole time-of-flight (Q-ToF) Ultima mass spectrometer (factory modified for high mass/charge transmission; Waters), or a Synapt Q-ToF/ion mobility mass spectrometer (Waters), each calibrated using a solution of caesium iodide (10 mg.mL<sup>-1</sup> in 70% isopropanol). Mass spectra were collected in positive ion mode using electrospray ionisation and analysed using MassLynx 4.1 software (Waters). Numerous acquisitions were combined and smoothed using a Savitzky-Golay algorithm.

# Chapter 3

## Gene deletion and solubility selection

### 3.1 Introduction

Multiple distinct properties of a protein can lead to a poor soluble protein yield during over-expression, producing protein aggregates that are not generally useful for experimental study. Large multi-domain proteins can contain multiple regions that fold poorly or slowly, causing aggregation of over-expressed protein. In these situations, the contribution of each troublesome protein region to reducing soluble protein yield is multiplicative. As a result, expression of protein domains individually can produce higher soluble protein yields compared to expression of the same domains as part of a multi-domain protein.

By separating a large multi-domain protein — that does make soluble protein when over-expressed — into distinct smaller, folded sections, soluble protein yield is improved and the resulting proteins can be more suitable for experimental

study. However, designing a truncated version of a protein is not necessarily straightforward. Multiple methodologies have been developed for discovering soluble protein truncation variants; however, in the case of some protein variants a difference of a single amino acid can lead to poor yield of an over-expressed protein and make identification of useful truncated protein more difficult.

Uni-directionally deleted gene libraries for a gene of interest can be made using a combination of *ExoIII* and mung bean nuclease (Henikoff, 1984); and can be utilised to provide a random pool of genes encoding proteins of varied sizes. A truncated protein library should contain mostly incomplete domain units — which have been truncated within a folding unit — but in the mix there may also be useful mutants which are not truncated within a domain. Protein mutants which have been conveniently truncated in a unstructured region, in many cases, are likely to represent the boundaries of protein domains. Screening of truncated gene libraries for useful domain truncations can then make use of the expected differences between mutants which contain whole folding units and those that do not. Intact folded protein domains should often be soluble, whereas incomplete domains will often aggregate; but in the end, having any soluble, well-behaved protein to work with is all that is required to enable structural and functional studies to proceed.

By performing domain deletion experiments so that the resulting truncated proteins are expressed as fusions to a second protein with an easily observable phenotype, colonies expressing useful protein mutants can be identified easily by culture on agar plates. Although this methodology may not be useful for



all proteins, protein domain identification using this technique promises to be quick and straightforward for screening lots of unique protein mutants at once. Truncated protein mutants which are folded, soluble and useful for further study can then be expressed separately and used for other experiments.

### 3.1.1 Gene truncation using exonuclease III

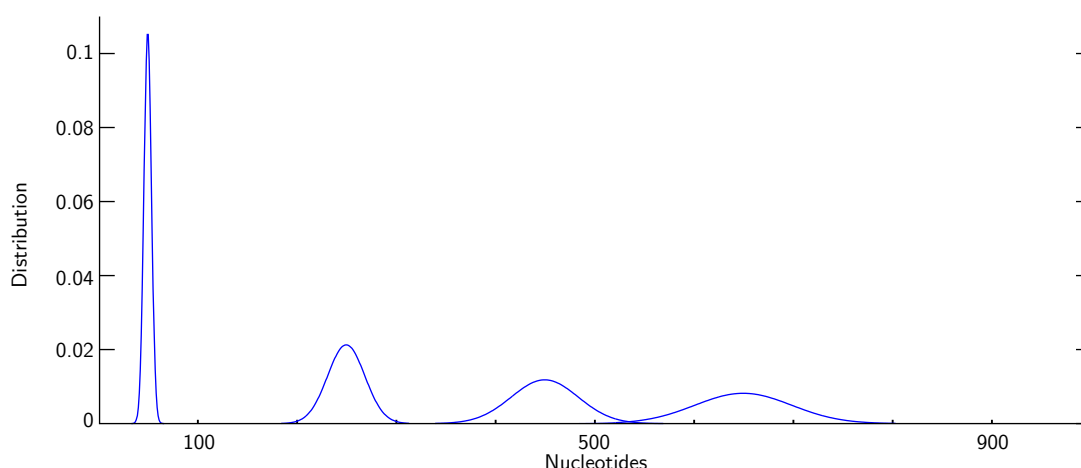
*ExoIII* is a double stranded DNA (dsDNA) specific exonuclease which hydrolyses DNA in the 3' to 5' direction in a non-processive manner to generate products with single-stranded (ss) 5' overhangs (Putney *et al.*, 1981; Henikoff, 1984). It can then be utilised in combination with a ssDNA nuclease to perform uni-directional DNA truncation (Hoheisel, 1993; Ostermeier, 2003). The *ExoIII* reaction is well understood and deletion using *ExoIII* allows tailoring of deletion distributions to interesting regions by careful timing of truncation reactions (Ostermeier, 2003).

*ExoIII* can only digest DNA from 3' ends in dsDNA, and cannot hydrolyse DNA past a modified  $\alpha$ -thio-substituted ( $\alpha$ S) nucleotide (Putney *et al.*, 1981; Henikoff, 1984). To produce uni-directional gene truncations in a plasmid destined to subsequently be recircularised (as here), dissimilar DNA ends are required in a dsDNA fragment, one end that is resistant to digestion with *ExoIII* and one that is not. Often *ExoIII* protected ends are produced by using a DNA polymerase to incorporate  $\alpha$ S nucleotides at 5' DNA overhangs of plasmids that have been linearised by a restriction endonuclease (producing a linearised plasmid with two *ExoIII* resistant ends). Subsequent removal of a single *ExoIII*

protected DNA end with a second endonuclease produces a template suitable for uni-directional truncation. Since uni-directional gene truncation with *ExoIII* removes only one DNA strand, production of plasmids with truncated genes requires removal of ssDNA using mung bean (or S1) nuclease. DNA truncation products made with *ExoIII* and mung bean nuclease often ligate poorly, so DNA end repair using the Klenow fragment of DNA polymerase I is required prior to plasmid ligation. Hoheisel (1993) experimentally identified that the *ExoIII* reaction produces products of lengths that follow a Gaussian distribution (Equation 3.1), where  $L$  = mean truncation length,  $N$  = number of bases truncated and standard deviation  $\sigma = cL$  ( $c = 0.075$ ; however, 9% of truncated gene products are truncated disproportionately; Hoheisel, 1993; Ostermeier, 2003).

$$G(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \quad z = \frac{L - N}{\sigma} \quad (3.1)$$

As the *ExoIII* reaction proceeds, the distribution of truncation lengths expands, so that deleting short sections of a gene produces a pool of different lengths with a small standard deviation compared to longer deletions. As a result, removal of small portions of a gene requires many individual time-point samples to adequately analyse an extended region of interest to compensate for the tight grouping of truncations, while removing longer sections produces a wide spread of sizes; for illustration see Figure 3.1.



**Figure 3.1: Distribution of exonuclease III truncation reaction products.** The predicted distribution of products of an ideal *ExoIII* uni-directional truncation reaction modelled using Equation 3.1 where  $L$  = mean truncation length (50, 250, 450 and 650 nucleotides),  $N$  = truncated gene size and  $c = 0.075$  (Hoheisel, 1993; Ostermeier, 2003). As the *ExoIII* deletion reaction proceeds the distribution of different sized products expands.

## 3.2 Aims

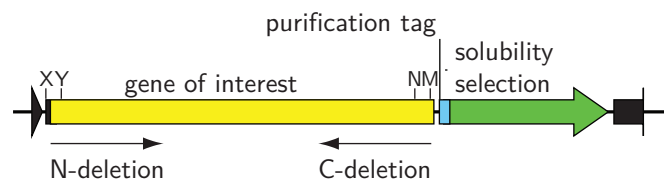
This Chapter presents a plasmid based methodology to I) produce a comprehensive large library of protein mutants of varied lengths and II) use a pragmatic assay of truncated mutants to select for protein solubility (and those in the correct reading frame) to infer successful expression of folded protein segments. The plasmids designed for these experiments also encode appropriately placed protein purification tags for straightforward purification of interesting mutant proteins and are modular to allow easy modification to remove the purification and/or solubility reporting fusion protein depending on experimental need.

### 3.3 A technique for gene truncation and solubility selection

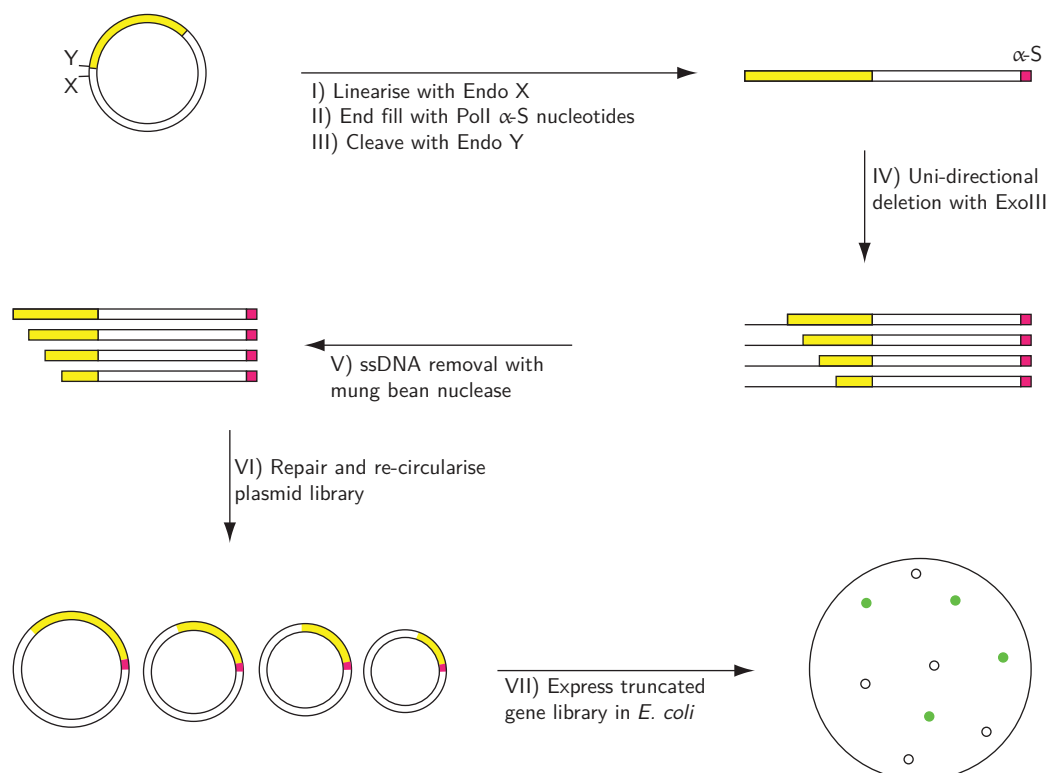
The methodology developed in this work is based around a series of plasmids described in this Chapter. Separate plasmids for N- or C-terminal gene deletion encode downstream purification and solubility reporting protein fusions and contain two restriction sites  $X$  and  $Y$  (or  $M$  and  $N$ ) at the end of the gene to be truncated (Figure 3.2A). Restriction endonuclease treatment of a deletion plasmid at  $X$  (or  $M$ ) generates a 5' overhang that is end filled with DNA polymerase I (Klenow fragment) and dNTP- $\alpha$ S to make a linear DNA that is resistant to *ExoIII* hydrolysis. Then, digestion with  $Y$  (or  $N$ ) generates a new 5' overhang that is susceptible to uni-directional digestion with *ExoIII* (Figure 3.2B). Timed exonuclease reactions generate a library of gene deletions targeted to end in a region of interest. After a truncated gene library has been made, the plasmids are circularised and the pool can then be transformed into an *E. coli* protein over-expression strain and cultured on agar plates where induction of over-expression produces a different truncated protein in each colony; over-expressing colonies that develop the phenotype associated with expression of a folded solubility reporting protein should be expected to contain in-frame, folded and soluble truncated mutants of the protein of interest. Plasmids with genes that direct expression of putative soluble protein fragments can then be easily identified and used for further study.

The technique we developed uses a suite of plasmids derived from the T7 expression plasmid pETMCSI (Neylon *et al.*, 2000) which allow high-level protein over-expression. Separate vectors were made (see Section 3.4.9) for each of: I) gene

**A: Overview of plasmid design for uni-directional gene truncation**



**B: Overview of methodology for generating uni-directional gene truncation library**



**Figure 3.2: Methodology for uni-directional gene truncation.** **A:** Plasmids for domain truncation contain an N-terminal gene of interest and, when plasmids are appropriately truncated, an in-frame sequence encoding a C-terminal purification tag and solubility reporting protein. Restriction sites *X* and *Y* (at the start of the gene or *M* and *N* at the end) are used to produce a DNA template for uni-directional gene deletion. **B:** Methodology for uni-directional gene deletion and solubility selection: I, Linearisation of deletion plasmids with restriction enzyme *X* (or *M*). II, End filling with DNA polymerase I (Klenow) and  $\alpha$ S nucleotides produces linear DNA resistant to *ExoIII*. III, Cleavage with *Y* produces a 5' overhang from which *ExoIII* can produce uni-directional deletions through the gene of interest. IV, Timed *ExoIII* reactions allow the size of truncated genes to be targeted to regions of interest. V, ssDNA extensions are removed from deleted plasmids with mung bean nuclease. VI, Truncated plasmids are repaired and ligated using DNA polymerase, dNTP and T4 DNA ligase. VII, Plasmid pool is transformed into an *E. coli* strain to express truncated proteins and solubility reporting fusion proteins, enabling useful truncated proteins to be identified or selected.

deletion from the 5' or 3' end; II) protein purification with a His<sub>6</sub>, Ktag (Bioline Australia, personal communication) or biotinylation tag (*E. coli* biotin ligase recognition sequence); and III) passenger solubility reporting using enhanced green fluorescent protein (EGFP) or *E. coli* dihydrofolate reductase DHFR.

### 3.3.1 Gene fusion for solubility selection

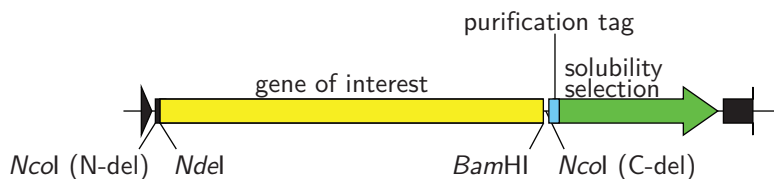
The gene fusion plasmids presented here are designed to encode solubility indicating proteins (EGFP or DHFR), which have been previously shown to indicate folding and solubility when expressed as fusions to other proteins (see Sections 1.6.1 and 1.6.2).

A uni-directional gene truncation library produced with *ExoIII* results in DNA products terminating at all three codon positions. As a result, two out of three resulting genes within a library will contain a non-sense gene of interest. Selection against out-of-frame truncated mutants is carried out by placing the solubility reporting gene downstream of the gene of interest so that genes resulting from truncation with *ExoIII* — when not in the correct reading frame — will either not initiate protein translation (out-of-frame N-deleted mutants), or not be expressed as a fusion with the solubility reporting protein (out-of-frame C-deleted mutants). In addition, to stop non-truncated plasmids from producing solubility reporter in *in vivo* solubility assays, the plasmids have been designed so that I) genes of interest in N-terminal deletion plasmids are not expressed or II) genes of interest in C-terminal deletion plasmids are not in-frame with (and have stop codons

before) the solubility reporters until after truncations have been made with *ExoIII*.

### 3.3.2 Layout of plasmids for gene truncation using exonuclease III

All gene deletion plasmids in this work have been designed so that the start codon of a gene of interest is placed within a unique *NdeI* restriction site (5'-CATATG-3') and the end of the gene is immediately followed by a *BamHI* site (Figure 3.3). The separate plasmids for generating N- or C-terminal deleted genes are designed so that digestion with *NcoI* (5'-CCATGG-3') and subsequent protection using DNA polymerase I (Klenow) to incorporate dGMP $\alpha$ S (and dAMP, dCMP and dTMP) produces *ExoIII* resistant DNA ends.



**Figure 3.3: General layout of gene truncation and solubility reporting plasmids.** The start codon of a gene of interest is placed at a unique *NdeI* restriction site and the gene directly followed by a unique *BamHI* site. N-terminal deletion plasmids have a unique *NcoI* restriction site that ultimately contains the start codon of the gene of interest and C-terminal deletion plasmids have an *NcoI* restriction site downstream of the end of the gene of interest.

#### 3.3.2.1 Plasmids for N-terminal deletion

In an experiment to generate an N-terminal gene deletion library (Figure 3.4), the gene of interest is placed in-frame with the solubility reporter. The original start codon of the gene is deleted by *ExoIII*, so plasmids for gene deletion are designed to introduce a new start codon. The use of a unique *NcoI* site — which





contains the sequence ATG — to produce the *ExoIII* resistant  $\alpha$ S-protected sites in N-terminal deletion plasmids is placed in an appropriate position downstream of the RBS for initiation of protein synthesis. The initiation codon present in the *NcoI* site of N-deletion plasmids is out-of-frame with the gene, ensuring that it can only be over-expressed after truncation by *ExoIII*. Digestion of the *ExoIII* resistant linear plasmids with *NdeI* makes a DNA substrate for uni-directional *ExoIII* truncation.

#### 3.3.2.2 Plasmids for C-terminal deletion

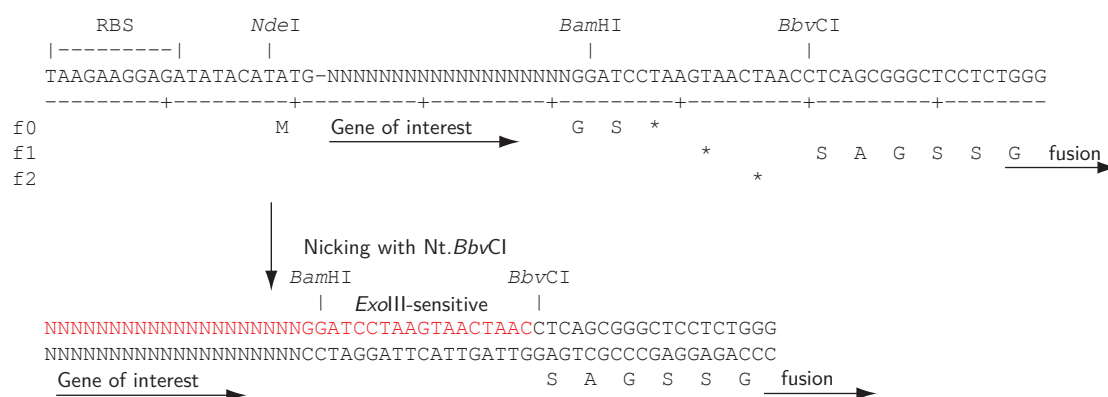
For C-terminal gene deletion (Figure 3.5), the gene of interest is inserted in-frame with the protein translation initiation codon and the end of the gene is followed by stop codons in all three reading frames to prevent expression of the solubility reporter in the event that deletion from the 3' end has not occurred. The unique *NcoI* site in the C-terminal deletion plasmids is placed downstream of the gene of interest and is used to generate the *ExoIII* resistant site by end filling. After gene truncation, the protected *NcoI* site codes for the start of the flexible linker prior to the protein purification tag. Protection using dGTP $\alpha$ S generates a sequence encoding a peptide linker starting with His-Gly while use of dCTP $\alpha$ S makes one that begins with Gly. The *ExoIII* sensitive end is generated by digesting the plasmid with *Bam*HI. A gene of interest cloned into frame 1 cannot be expressed in-frame with the solubility reporter until the plasmid has been digested with *ExoIII*.

A further C-terminal deletion plasmid was designed that uses Nt.*Bbv*CI — a



**Figure 3.5: Cloning site for C-terminal gene truncation and solubility reporting plasmids.** C-terminal deletion DNA sequence for uni-directional gene truncation; f0–2 show the three reading frames where f0 is the reading frame of the initiation codon. Placing a gene of interest so that it ends in f1 (by addition of a single nucleotide between the gene coding sequence and *Bam*HI site) will prevent expression of in-frame solubility reporter unless truncated with *Exo*III. After a gene of interest (represented by NNN...) is placed between the *Nde*I and *Bam*HI restriction sites, cutting the plasmid with *Nco*I, end filling with dGTPαS (red); dATP, dCTP and dTTP (green) and DNA polymerase I (Klenow) makes a linear plasmid resistant to deletion using *Exo*III; then digestion with *Bam*HI removes one protected end to enable C-terminal uni-directional deletion with *Exo*III. Following gene truncation, one in three mutants will be in-frame with the downstream solubility reporting fusion gene for *in vivo* selection for expression and solubility of in-frame proteins.

modified type I restriction endonuclease that produces a single DNA nick — from which *ExoIII* can uni-directionally delete DNA (Figure 3.6). Treatment of these C-terminal deletion plasmids with *Nt.BbvCI* makes a DNA nick between the sequences CC and TCAGC, forming a single 3' end that *ExoIII* can attack. After initiation of gene truncation, *ExoIII* must remove 19 nucleotides before deletions are generated in the gene of interest. C-terminal gene truncation plasmids for use with *Nt.BbvCI* are similar in design to the other C-terminal deletion plasmids, except that the C-terminal peptide linker starts with Ser-Ala-Gly and a gene of interest is translated in frame 0 (not 1), and cannot be expressed in-frame with the solubility reporter until the plasmid has been digested with *ExoIII*.



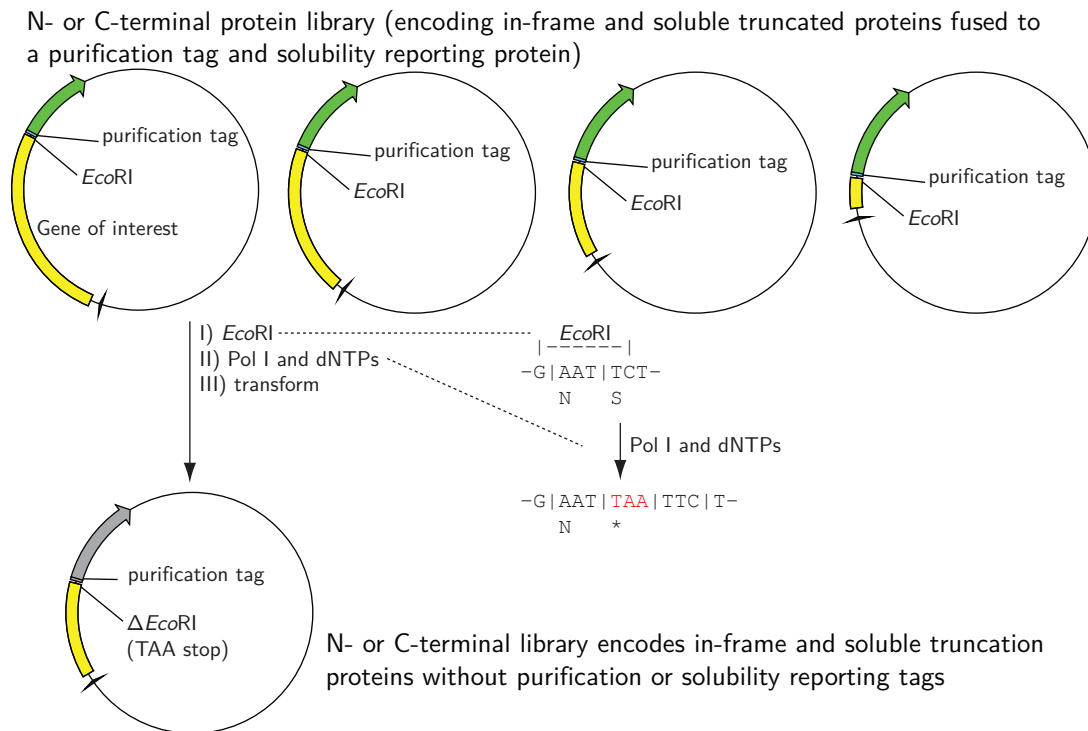
**Figure 3.6: Alternate cloning site for C-terminal gene truncation and solubility reporting plasmids.** Uni-directional C-terminal deletion DNA sequence; f0–2 show the three reading frames where f0 is the reading frame of the gene of interest and f1 the solubility reporter. A gene cloned into f0 will not be in-frame with the solubility reporter unless truncated with *ExoIII*. Uni-directional gene deletion using this series of plasmids is accomplished by treatment with *Nt.BbvCI*, which generates a DNA nick — a single strand break — making a single 3' end available for digestion with *ExoIII*. Where a gene of interest has been cloned with the start codon at the *NdeI* site — and after uni-directional gene truncation — one in three mutants will be in-frame with the downstream purification and solubility reporting fusion gene. Deleted genes can be screened *in vivo* for expression of in-frame, expressed and soluble protein.

### 3.3.2.3 Genetic removal of fusion tags from gene truncation plasmids

The series of T7 expression plasmids are capable of inducible high-level protein over-expression and make production of useful protein truncation mutants straightforward. However, following identification of putative soluble protein truncations, further experiments using fusions to the solubility reporting protein are not necessarily useful. To facilitate easy removal of the solubility reporting and/or protein purification tags, restriction sites were incorporated such that simple genetic modifications could be performed to appropriately introduce stop codons with minimal sample handling or purification.

Separating the truncated gene sequence from the purification fusion tag is an *EcoRI* site. The restriction site was carefully placed such that digestion with *EcoRI* and then filling the ssDNA overhangs converts the original GAATTC sequence to GAATTAATTC making an in-frame stop codon and resulting in a C-terminal protein sequence which no longer has the purification tag or solubility reporting fusions (Figure 3.7). Following fusion removal from truncated genes, the unnatural C-terminal sequences are: N-terminally deleted proteins, GSSGN\*; C-terminally deleted proteins, HGSSGN\* and C-terminally deleted proteins (made with *Nt.BbvCI*), SAGSSGN\*.

In addition, the sequences encoding the purification tags are separated from the downstream solubility reporting fusion gene by an *MfeI* site (Figure 3.8A). The restriction sequence was placed so that treatment with *MfeI* and then ligation of a 5'-unphosphorylated palindromic DNA oligonucleotide (Figure 3.8C)



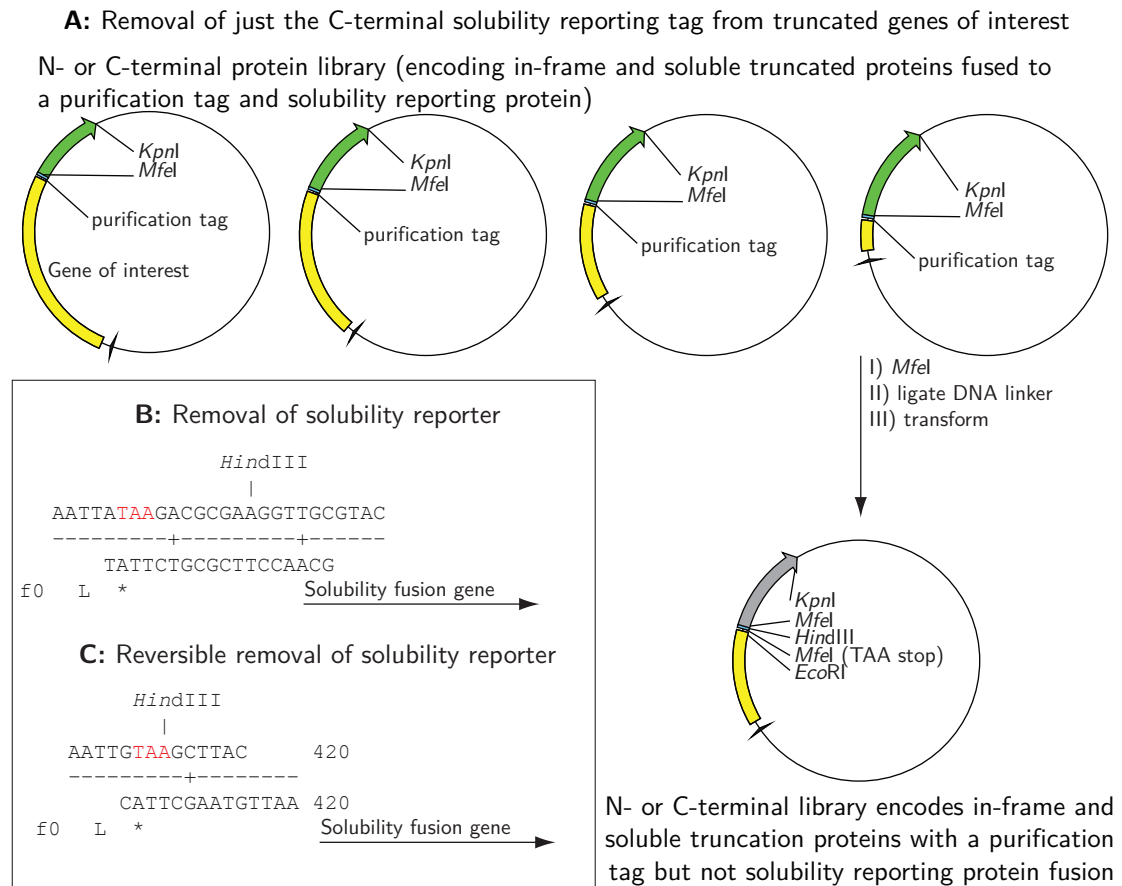
**Figure 3.7: Removal of purification and solubility reporter tags from truncated genes of interest.** Methodology for converting gene libraries so that they no longer encode C-terminal fusions to the truncated gene of interest, yellow; in-frame solubility reporter gene, green; out-of-frame solubility reporter gene, grey. Both the purification and solubility reporting genes are downstream of the *EcoRI* site in these gene truncation plasmids. Plasmids are converted by linearising the library with *EcoRI*, end filling with DNA polymerase I (Klenow fragment) and dNTPs, and then transformation into an *E. coli* BL21( $\lambda$ DE3)*recA* expression strain. Plasmids are efficiently recircularised by ligation *in vivo*. After treatment with *EcoRI* and DNA polymerase I (Klenow fragment), recovered plasmids no longer encode C-terminal fusions to the truncated gene of interest.

would produce a C-terminal sequence of QL\* directly after the purification tag. Plasmids modified by this process can be screened for successful recombinants by the loss of solubility reporter phenotype when expressed in BL21( $\lambda$ DE3)*recA* cells. Otherwise entire selected or unselected truncated gene libraries can be digested with *MfeI* and *KpnI* and the tag removed by replacing the solubility reporting gene with a (unphosphorylated) DNA linker (Figure 3.8B). C-terminal His<sub>6</sub> fusion tags after removal of solubility reporting fusions are; N-terminally deleted, GSSGNSHHHHHHQL\*; C-terminally deleted, HGSSGNSHHHHHHQL\* and C-terminally deleted (nicked with *Nt.BbvCI*) SAGSSGNSHHHHHHQL\*.

## 3.4 Assembly of plasmids for gene truncation and solubility selection

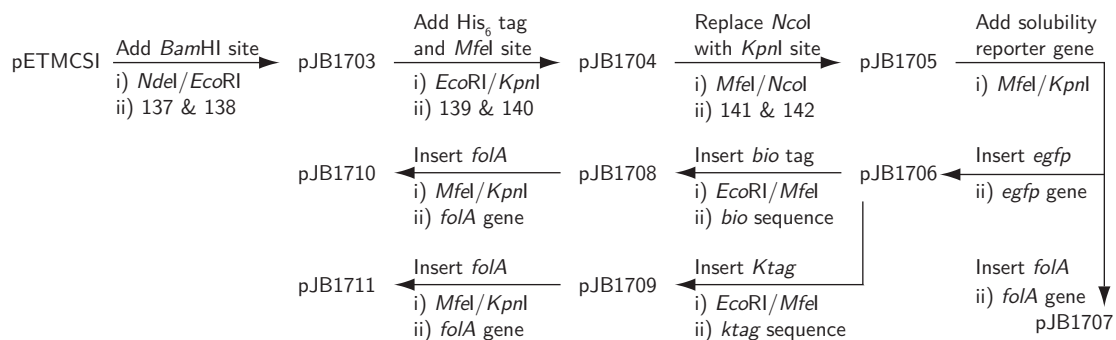
### 3.4.1 Strategy for construction of solubility selection plasmids

The strategy for making the different plasmids with protein purification (His<sub>6</sub> tag, biotinylation sequence and ktag) and solubility selection (EGFP and DHFR) protein tags is outlined in Figure 3.9. First a new plasmid with a multiple cloning site (pJB1705) was made by three successive linker ligation cloning steps (making pJB1703 and pJB1704 in the process). From pJB1705 a protein purification and solubility selection plasmid (pJB1706) was assembled that encodes a C-terminal His<sub>6</sub> and EGFP fusion protein. From pJB1706, biotinylation and Ktag EGFP variants were made by replacing the His<sub>6</sub> tag between the *EcoRI* and *MfeI* sites with DNA encoding either the *E. coli* biotin ligase recognition sequence or the



**Figure 3.8: Removal of solubility reporting gene fusion from truncated genes of interest.** **A**, Removal of solubility reporter from the reading frame of the gene of interest and protein purification tag is carried out by treating the plasmids with *MfeI*, and ligating an oligonucleotide linker to make a stop codon upstream of the solubility reporter gene. **B**, Oligonucleotides 5'-AATTATAAGACGCGAAGGTTGCGTAC and 5'-GCAACCTTCGCGTCTTAT can be inserted between the *MfeI* and *KpnI* sites of truncated plasmids. **C**, the self complementary oligonucleotide 420 can be inserted into the *MfeI* site of truncated plasmids. Gene of interest, yellow; in-frame solubility reporting gene, green; out-of-frame gene for solubility reporting protein, grey.

Ktag peptide sequence. Finally, biotinylation and Ktag versions with a DHFR solubility selection fusion protein were made by replacing the *egfp* sequence in pJB1708 and Ktag-EGFP with *folA*.



**Figure 3.9: Strategy for construction of protein purification and solubility selection plasmids.** Plasmids with varied protein purification (His<sub>6</sub> tag, biotinylation sequence and ktag) and solubility reporting (EGFP and DHFR) protein tags were constructed from pETMCSI. Detailed steps are explained in the text. Numbers refer to 5'-unphosphorylated oligonucleotide linkers with appropriate overhanging ends.

### 3.4.2 Construction of pJB1703

The creation of a new T7 based multiple cloning site for solubility selection was done in three sections using oligonucleotide linkers (summarised in Figure 3.10). pJB1703 was constructed by digesting pETMCSI (Neylon *et al.*, 2000) with the restriction enzymes *NdeI* and *EcoRI* (Section 2.3.2), from which linearised plasmid was purified by agarose gel electrophoresis (Section 2.3.7). Complementary oligonucleotides 137 and 138, with complementary overhangs to *NdeI/EcoRI* digested pETMCSI, were mixed with it and ligated (Section 2.3.10.2; Figure 3.10). Ligated plasmids were transformed into *E. coli* AN1459 (Section 2.2). Transformant colonies were clonally isolated, grown on an LB agar plate and



used for plasmid isolation (Section 2.3.1.1). Selected plasmids that tested positive for the introduction of a *Bam*HI restriction site were sequenced to confirm linker insertion (Section 2.3.11).

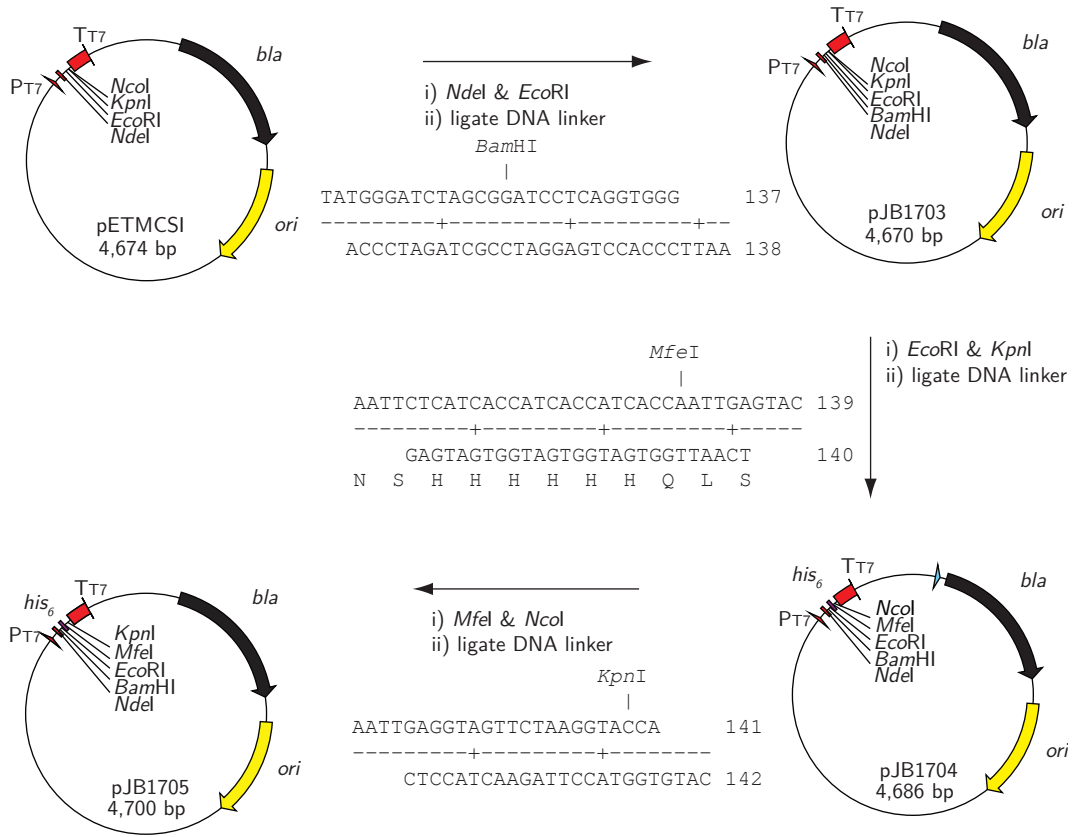
### 3.4.3 Construction of pJB1704

To make pJB1704, which introduces a six-histidine encoding sequence, complementary oligonucleotides 139 and 140 were inserted between the *Eco*RI and *Kpn*I sites of gel purified pJB1703 (Figure 3.10). Ligated plasmids were transformed into AN1459 and then transformants were streak-diluted on LB agar plates to obtain single colonies, smeared on an LB agar plate, grown and plasmids were isolated. Successful cleavage by *Mfe*I indicated the presence of the new linker. Plasmids which tested positive for *Mfe*I digestion were sequenced to confirm the presence of the insert.

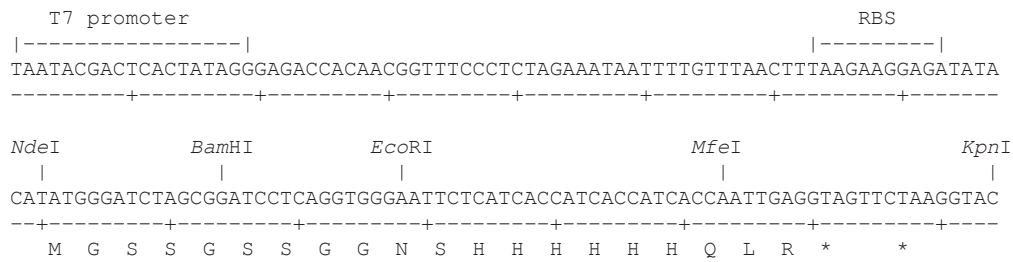
### 3.4.4 Construction of pJB1705

pJB1705 was made by replacing the small *Mfe*I and *Nco*I restriction fragment of pJB1704 with complementary oligonucleotides 141 and 142 (Figure 3.10). Ligated plasmids were transformed into AN1459, and transformants were clonally isolated. Plasmids from selected transformants which tested positive for *Kpn*I digestion were sequenced to confirm that the insert was present. Note that the linker sequence is such that the *Nco*I site is not regenerated.

### A: Construction of pJB1705



### B: Multiple cloning site of pJB1705



**Figure 3.10: Construction of pJB1705.** Successive DNA linkers replaced the multiple cloning site of pETMCSI. **A:** pETMCSI was digested with *NdeI* and *EcoRI* and ligated with oligonucleotide linkers 137 and 138 to make pJB1703. pJB1704 was made by inserting oligonucleotide linkers 139 and 140 between the *EcoRI* and *KpnI* sites of pJB1703. pJB1705 was made by inserting oligonucleotide linkers 141 and 142 between the *MfeI* and *NcoI* sites of pJB1704. **B:** Multiple cloning site of pJB1705 showing phage T7 promoter (*P<sub>T7</sub>*), RBS and the reading frame from the start codon. pJB1705 encodes a His<sub>6</sub> purification tag.

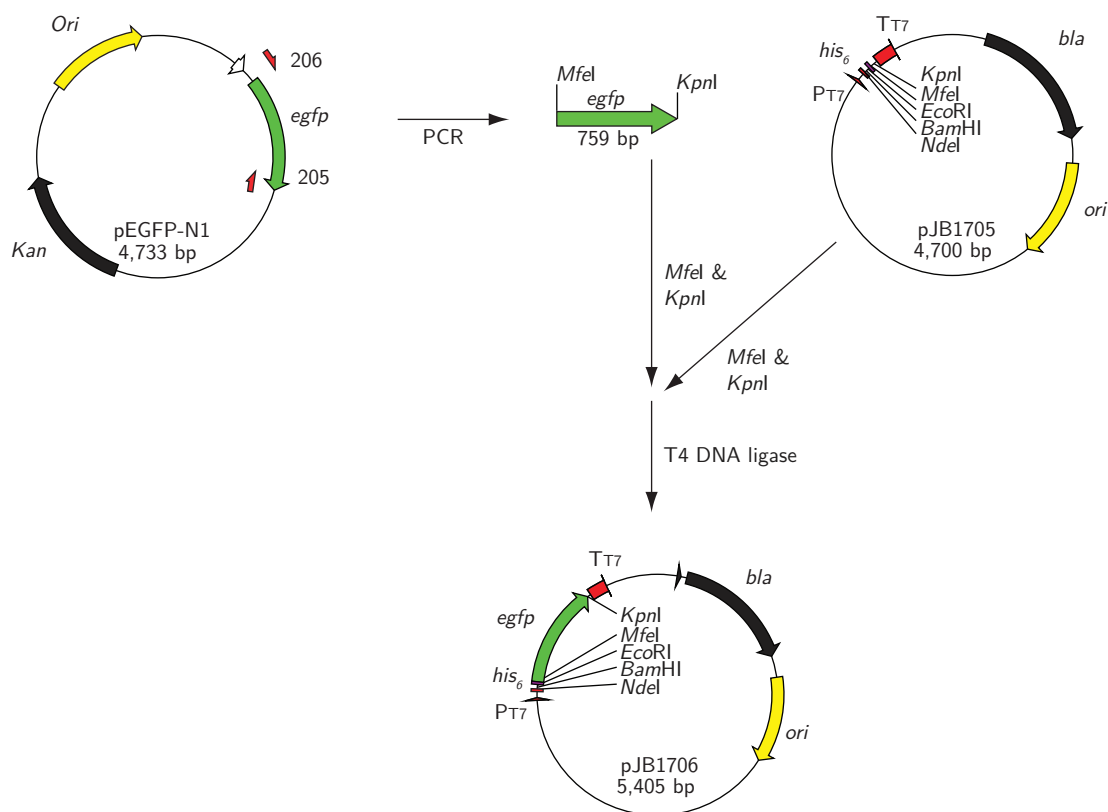
### 3.4.5 Construction of pJB1706

To make a His<sub>6</sub>-EGFP fusion plasmid, the *egfp* gene was amplified from pEGFP-N1 (Clontech) with oligonucleotides 205 and 206 (Appendix C) using Accuzyme DNA polymerase. Primer 206 introduces an *Mfe*I restriction site at the start of *egfp* and 205 a *Kpn*I site after it. The PCR product was purified by using a silica column (Section 2.3.8) and digested with restriction enzymes *Mfe*I and *Kpn*I for insertion between the *Mfe*I and *Kpn*I sites of the expression plasmid pJB1705 (Section 3.4.4). AN1459 transformants were clonally isolated and plasmids were extracted. They were digested using *Eco*RI and *Kpn*I to identify plasmids containing the 759 bp *egfp* gene fragment. Plasmids were sequenced to confirm the inserted sequence, and one was retained as pJB1706 (Figure 3.11).

Transformants of BL21(λDE3)*recA* containing pJB1706 were observed to have a green fluorescent phenotype on LB agar plates. Although expression of EGFP was not intentionally induced, the *lacUV5* promoter controlling expression of T7 RNA polymerase, which in turn transcribes mRNA from T7 promoters, is not fully repressed in BL21(λDE3)*recA* grown in LB medium (Mertens *et al.*, 1995).

### 3.4.6 Construction of pJB1707

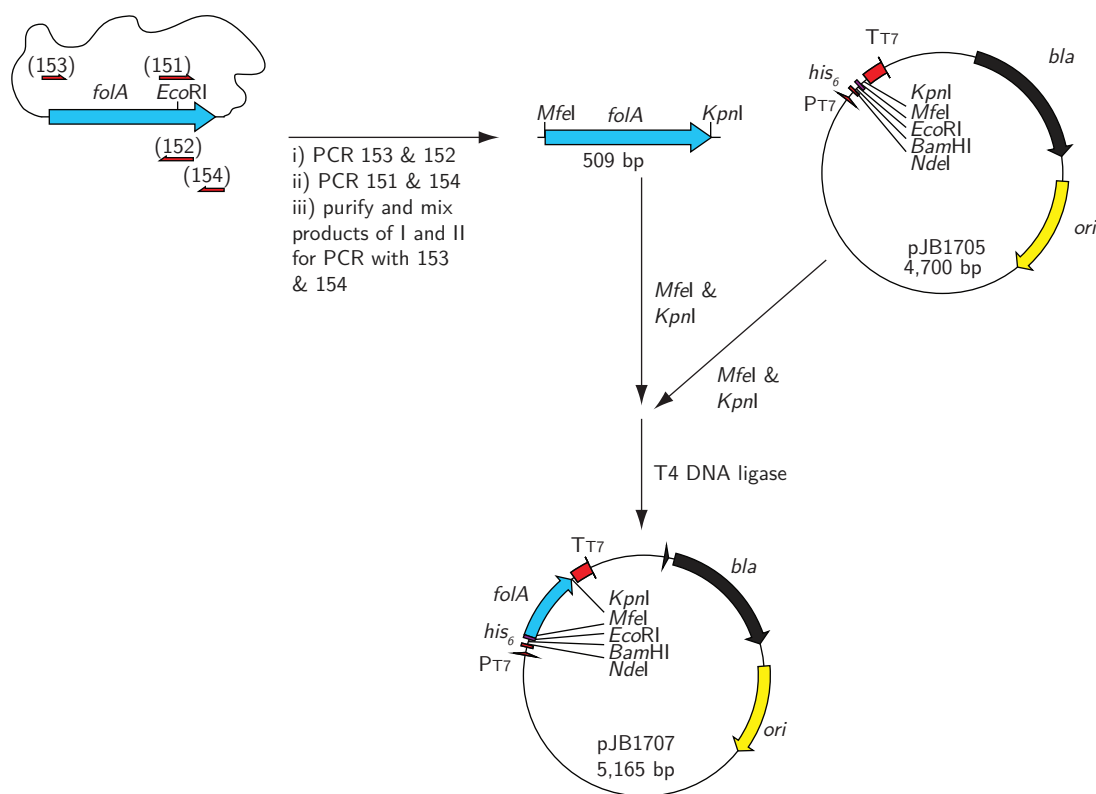
The *E. coli folA* gene, which codes for DHFR was prepared for cloning into the solubility selection vector by PCR. The *E. coli folA* gene contains an *Eco*RI site which was silently mutated by strand overlap-extension PCR. Oligonucleotides 152, 153 (Appendix C) and KOD DNA polymerase were used to amplify the



**Figure 3.11: Construction of pJB1706: a C-terminal His<sub>6</sub>-EGFP fusion plasmid for protein purification and solubility reporting.** The *egfp* gene was amplified from pEGFP-N1 (Clontech) by PCR using oligonucleotides 205 and 206, digested with *MfeI* and *KpnI* and inserted between the *MfeI* and *KpnI* sites of pJB1705 (Figure 3.10). pJB1706 directs over-expression of EGFP with an additional N-terminal sequence of MGSSGSSGGNSHHHHHHQL.

5' segment of *folA* and oligonucleotides 151, 154 (Appendix C) and Accuzyme DNA polymerase were used to amplify the 3' fragment of the gene (Figure 3.12). Oligonucleotide primers 151 and 152 are complementary to each other and encode a silent mutation removing the *EcoRI* site. The two *folA* PCR products were purified using QIAQuick columns, and assembled by PCR with Accuzyme DNA polymerase using the outer primers 153 and 154.

The purified *folA* PCR product was digested with *MfeI* and *KpnI* for insertion



**Figure 3.12: Construction of pJB1707: a His<sub>6</sub>-DHFR fusion plasmid for protein purification and solubility reporting.** The *folA* gene was amplified from *E. coli* genomic DNA by PCR in two overlapping sections using oligonucleotide pairs 153–152 and 151–154 and then recombined in a further PCR using 153 and 154. The product was digested with *MfeI* and *KpnI* and inserted between the *MfeI* and *KpnI* sites of pJB1705 (Figure 3.10). pJB1707 directs over-expression of DHFR with an additional N-terminal sequence of MGSSGSSGGNSHHHHHHQL.

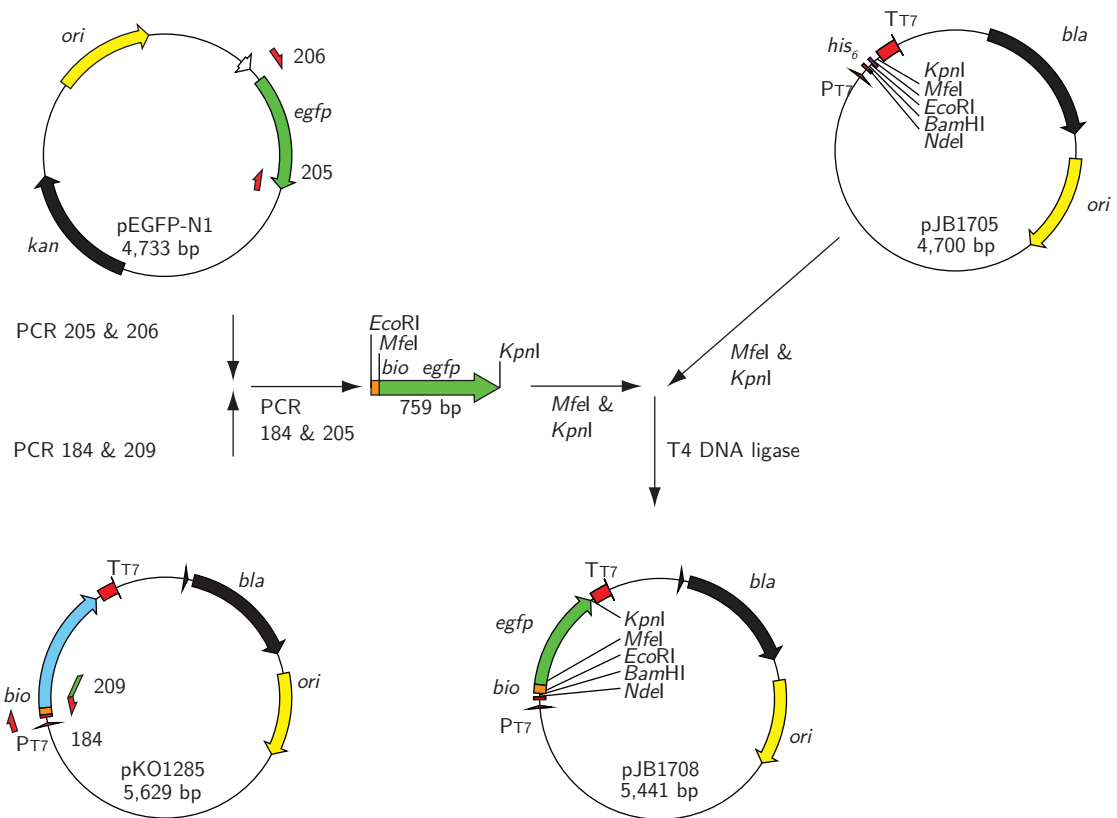
between the *MfeI* and *KpnI* sites of the expression plasmid pJB1705 (Section 3.10). Transformants of AN1459 were clonally isolated, and extracted plasmids were digested with *EcoRI* and *KpnI*; successful restriction digestion of pJB1707 gave a product of 509 bp. Selected plasmids were sequenced to confirm the insert, and one was stored as pJB1707.

### 3.4.7 Construction of enhanced green fluorescent protein expression plasmids with biotinylation or Ktag purification tags

Alternate versions of the EGFP expression plasmid pJB1706 were designed with an *E. coli* biotinylation (Beckett *et al.*, 1999) or a Ktag (Bioline, Aust.) purification sequence. The cloning of these purification tags was such that an initial PCR reaction amplified the protein purification tag sequence from an existing plasmid, and the primers were designed with a complement to the 5' end of the *egfp* PCR product generated previously (Section 3.4.5). A nested PCR containing the purification tag PCR product and the *egfp* PCR product created a purification tag-*egfp* gene fusion which was then cloned between the *Mfe*I and *Kpn*I sites of pJB1705 (Figure 3.13).

The biotinylation tag insert for pJB1708 was produced by amplifying the DNA encoding the small *E. coli* biotinylation sequence MAGLNDIFEAQK<sup>bio</sup>IEWHEH from pKO1285 (a derivative of pKO1274 which has the *E. coli tus* gene inserted; Jergic *et al.*, 2007) using oligonucleotides 184 and 209 (Appendix C) and Accuzyme DNA polymerase. The two amplification products were then combined by PCR using oligonucleotides 184 and 205.

The gene insert for pJB1710 was constructed by amplifying the small Ktag region from pQIS214 (gift from Bioline Aust.) using oligonucleotides 207 and 208 (Appendix C) and Accuzyme DNA polymerase, and the two amplification products were then assembled by PCR using oligonucleotides 208 and 205.



**Figure 3.13: Construction of pJB1708: a bio-EGFP fusion plasmid for protein purification and solubility reporting.** The *E. coli* biotin ligase recognition sequence MAGLNDIFEAQK<sup>bio</sup>IEWHEH was PCR amplified from pKO1285 using primers 184 and 209 and the *egfp* gene was amplified from pEGFP-N1 by PCR with 205 and 206. The *E. coli* biotinylation sequence and *egfp* gene were combined by PCR and cloned into the *MfeI* and *KpnI* sites of pJB1705 (Figure 3.10). The same methodology was used to produce a plasmid with a Ktag purification sequence (pJB1710). pJB1708 and pJB1710 direct over-expression of EGFP with an additional N-terminal sequence of MGSSGSSGGNS**MAGLNDIFEAQKIEWHEH**QL or MGSSGSSGGNS**EDVDECSENMSAQLCQL** respectively (purification tag sequence in bold).

pJB1705 that had been digested with *EcoRI* and *KpnI* was ligated with *EcoRI* and *KpnI* digested and gel purified *bio-egfp* and *Ktag-egfp* amplification products. Ligated plasmids were selected in BL21( $\lambda$ DE3)*recA*. Leaky expression from the T7 promoter (Mertens *et al.*, 1995) allowed transformants to be identified by their green fluorescent phenotype. Plasmids were prepared from green fluorescent cells and sequenced to confirm the small insert in pJB1708 (*bio-egfp*) and pJB1710 (*Ktag-egfp*; Figure 3.13).

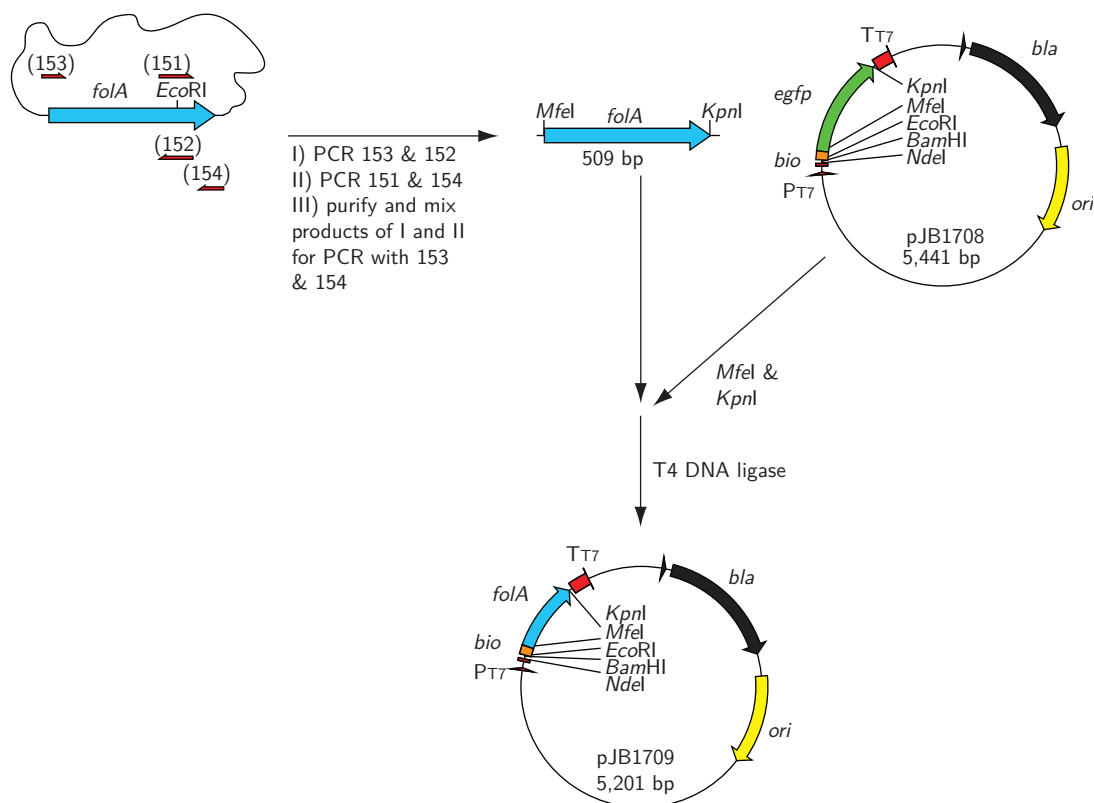
#### 3.4.8 Construction of dihydrofolate reductase expression plasmids with biotinylation or Ktag purification tag

Alternate versions of the DHFR expression plasmid (pJB1707) with an *E. coli* biotinylation sequence (pJB1709) or a Ktag (pJB1711; Bioline, Aust.) were made from pJB1708 (bio-EGFP) and pJB1710 (Ktag-EGFP), respectively, by replacement of the *egfp* gene (Figure 3.14). pJB1708 and pJB1710 were digested with *MfeI* and *KpnI*, gel purified and ligated with the *folA* PCR product used for cloning the His<sub>6</sub>-DHFR plasmid (pJB1707; Figure 3.12). The sequences of the *folA* inserts in pJB1709 and pJB1711 were confirmed by DNA sequencing.

#### 3.4.9 Construction of gene-truncation plasmids

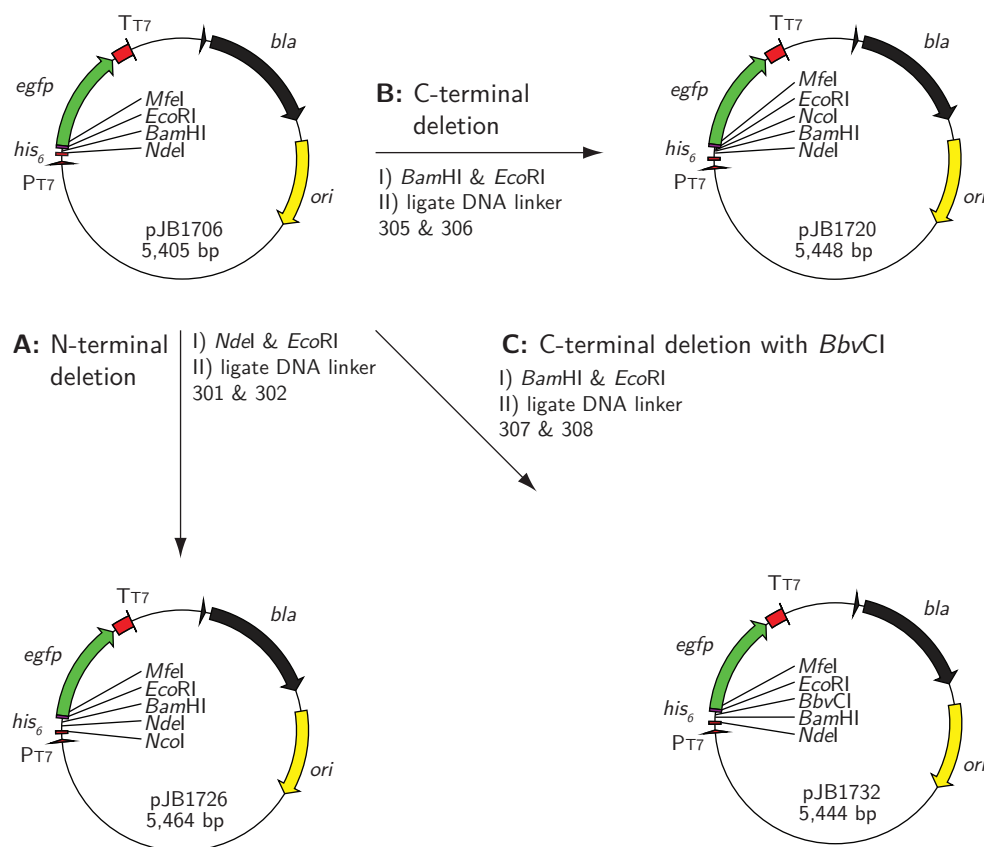
The suite of N-terminal deletion plasmids was produced by separately digesting pJB1706–1711 (Figure 3.9) with *NdeI* and *EcoRI*. The linear plasmids were gel purified and ligated with oligonucleotides 301 and 302 (Appendix C) to make





**Figure 3.14: Construction of pJB1709: a bio-DHFR fusion plasmid for protein purification and solubility reporting.** A DHFR expression plasmid with an N-terminal *E. coli* biotinylation sequence was made by replacing the *egfp* gene from pJB1708 with the *E. coli folA* gene. The *folA* gene was amplified from genomic DNA by PCR in two overlapping sections using oligonucleotide pairs 153–152 and 151–154 and then recombined in a further PCR using primers 153 and 154, digested with *MfeI* and *KpnI* and inserted between the *MfeI* and *KpnI* sites in pJB1708. An identical methodology was used to produce a plasmid with a Ktag purification sequence (pJB1711). pJB1709 and pJB1711 direct over-expression of DHFR with an additional N-terminal sequence of MGSSGSSGGNS**MAGLNDIFEAQKIEWHEHQL** or MGSSGSSGGNS**EDVDECSENMSAQLCQL**, respectively (purification tag sequence in bold).

N-deletion plasmids pJB1726–pJB1731 (Figure 3.15A; Table 3.1). Note that the linker ligation destroys the original *NdeI* site, replacing it with an *NcoI* site, and creating a new *NdeI* site downstream. A new *BamHI* site is installed just before the *EcoRI* site (see sequence in Figure 3.4).



**Figure 3.15: Gene truncation plasmids.** Plasmids for gene truncation were constructed from purification-solubility fusion expression plasmids pJB1706, pJB1707, pJB1708, pJB1709, pJB1710 and pJB1711 (see Figure 3.9).

C-terminal deletion plasmids were produced by separately digesting pJB1706–1711 with *BamHI* and *EcoRI*. Each linear plasmid was gel purified and ligated with

**Table 3.1: Gene deletion and solubility selection plasmids.**

Purification tag and solubility selection gene	Plasmid
N-terminal deletion plasmids (Figure 3.15A)	
<i>His<sub>6</sub>-egfp</i>	pJB1726
<i>His<sub>6</sub>-folA</i>	pJB1727
<i>bio-egfp</i>	pJB1728
<i>bio-folA</i>	pJB1729
<i>Ktag-egfp</i>	pJB1730
<i>Ktag-folA</i>	pJB1731
C-terminal deletion plasmids (Figure 3.15B)	
<i>His<sub>6</sub>-egfp</i>	pJB1720
<i>His<sub>6</sub>-folA</i>	pJB1721
<i>bio-egfp</i>	pJB1722
<i>bio-folA</i>	pJB1723
<i>Ktag-egfp</i>	pJB1724
<i>Ktag-folA</i>	pJB1725
C-terminal deletion plasmid for use with <i>Bbv</i> CI (Figure 3.15C)	
<i>His<sub>6</sub>-egfp</i>	pJB1732

oligonucleotide linkers 305 and 306 (Appendix C) to make C-deletion plasmids pJB1720–pJB1725 (Figure 3.15B; Table 3.1). This strategy moves the *Bam*HI site a further 21 bp downstream of the *Nde*I site (see Figure 3.5) to ensure efficient restriction endonuclease cleavage (during insertion of the gene of interest) and installs an *Nco*I site preceded by TAA stop codons in all three reading frames (see Figure 3.5).

A similar strategy was used to make a further C-terminal deletion plasmid which provides a Nt.*Bbv*CI nicking endonuclease cleavage site. pJB1706 was digested

with *Bam*HI and *Eco*RI, gel purified and ligated to oligonucleotide linkers 307 and 308 (Appendix C) to make pJB1732 (Figure 3.15C; Table 3.1; see also Figure 3.6).

### 3.4.10 Truncation apparatus

An apparatus for making highly sampled *Exo*III truncated libraries was constructed but was not required for the work performed in this Thesis. See Appendix A.

## 3.5 Conclusion

Multi-domain proteins are very often arranged with each folded protein domain in series, and may be linked by non-structured flexible regions. These protein domains have structures and functions which can be studied in isolation and provide information on their contribution to the properties of the full-length protein. This can be particularly useful when the full length protein cannot be studied. Yet, it can sometimes be problematic to produce soluble protein domains. We have developed a methodology that makes many truncated versions of a protein of interest by uni-directional deletion into a gene — in an expression plasmid — using *Exo*III. As complete protein domains are more likely to fold well and not aggregate — in contrast to incomplete domains — genes that direct expression of soluble proteins are probably directing production of proteins that do not contain incomplete segments of a folded domain. Therefore, genes for

soluble shortened proteins are those that have been truncated to regions between folding units — a domain boundary. The presence of a C-terminal fusion protein (EGFP or DHFR) means that once an expression plasmid library for many truncated versions of a protein is produced, and is transformed into an *E. coli* expression strain, protein over-expression in unique colonies can allow screening for putative soluble proteins by looking for the distinct phenotypes related to correctly folded and soluble EGFP or DHFR.

The plasmids made for this new technique, after gene truncation with *ExoIII*, form expression plasmids that direct expression of truncated proteins with, in order, a C-terminal purification tag and a solubility reporter protein. This arrangement of fusion protein genes in these plasmids, while initially allowing selection of soluble truncated mutants, can be modified in a straightforward manner to direct expression of truncated proteins alone, or just with a useful protein purification tag.

To generate truncations of a protein of interest, its gene can be placed between the *NdeI* (containing the gene start codon) and *BamHI* (directly following the last codon) restriction sites of any of the plasmids. In general the same gene sequence can be inserted into plasmids for either N- or C-terminal deletion. However, for C-terminal deleted genes, an additional nucleotide following the gene of interest ensures that the library template must be truncated by *ExoIII* before the gene of interest can be in-frame with the solubility reporter to prevent recovery of non-truncated genes in libraries. The plasmids presented here were used to produce and select for soluble truncated proteins, as described in the following Chapters of this Thesis.

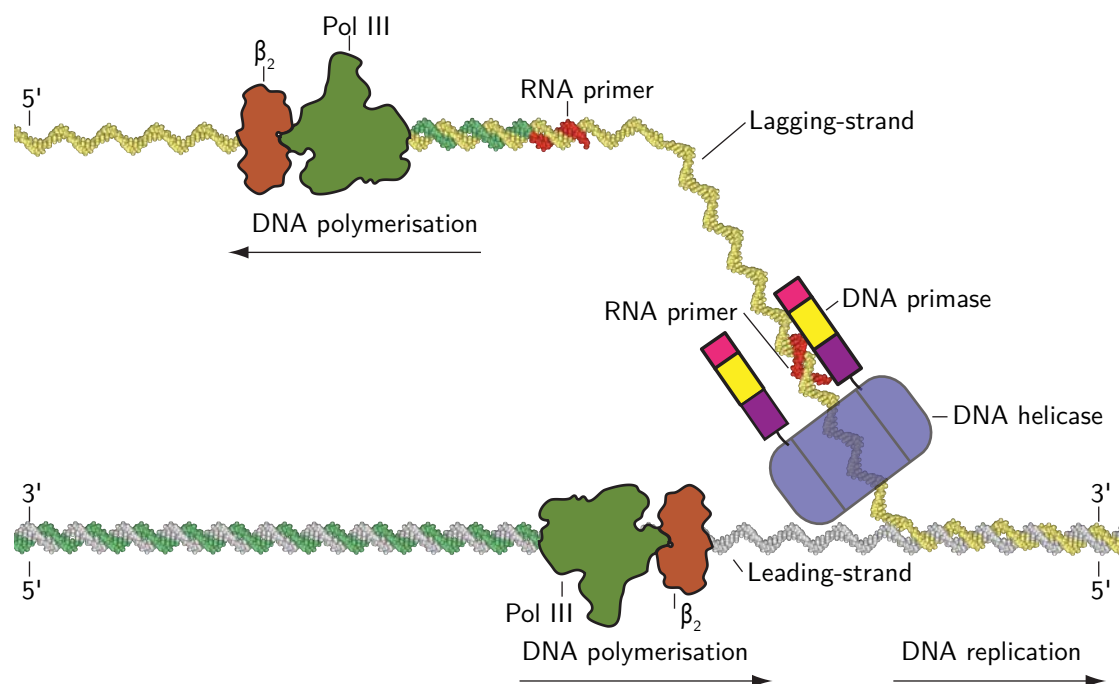
## Chapter 4

# Domain identification in *Acinetobacter baylyi* DNA primase

### 4.1 DNA primase and DNA replication

Chromosome replication in bacteria requires a large number of proteins to work in a concerted manner (Benkovic *et al.*, 2001). To provide DNA polymerase III — the chromosomal DNA replicase — with access to ssDNA, the hexameric helicase DnaB translocates along DNA in the 5' to 3' direction, unwinding dsDNA in its path. Once the parental DNA strands have been separated, each is duplicated by individual DNA polymerase III molecules tethered to the helicase.

DNA synthesis by DNA polymerases occurs only in the 5' to 3' direction, dictating that replication of antiparallel dsDNA occurs in distinct modes on the two complementary strands (Figure 4.1; Ogawa and Okazaki, 1980). On the 3'



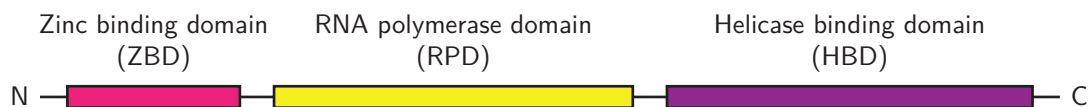
**Figure 4.1: DNA replication fork.** Chromosome replication occurs by two different modes, depending on the orientation of the template DNA strand and occurs after DNA helicase has separated the parental DNA strands. Leading-strand replication is straightforward, where DNA polymerase III can synthesise the leading-strand duplicate in the 5' to 3' direction. As DNA polymerase III can only synthesise DNA with 5' to 3' polarity, the lagging-strand needs to be duplicated in the opposite orientation to the lagging-strand. DNA primase synthesises an RNA primer on the lagging strand so that DNA polymerase III can begin DNA synthesis.

to 5' leading strand template, replication is relatively straightforward and can progress continuously, but on the complementary 5' to 3' lagging strand template, DNA polymerisation must still occur with 5' to 3' polarity. The cellular DNA replication machinery facilitates replication fork movement in one direction by copying the lagging strand template in many short 5' to 3' segments known as Okazaki fragments. Subsequent ligation of Okazaki fragments results in faithful replication of the lagging strand template (Ogawa and Okazaki, 1980; Wu *et al.*, 1992). Lagging strand DNA synthesis is further complicated because the genomic DNA polymerase requires the 3'-OH of a pre-existing 5' nucleotide to incorporate

a new 3' residue; these 5' residues are not present on unwound, lagging strands. To prime a lagging strand for DNA polymerisation, DNA primase (DnaG) places a short RNA primer on the nascent lagging strand, and then DNA polymerase III is able to copy the lagging strand template (Bouché *et al.*, 1978; Wu *et al.*, 1992).

#### 4.1.0.1 DNA primase (DnaG)

The importance of DNA replication makes DnaG an essential protein (Katayama *et al.*, 1989; Mushegian and Koonin, 1996; Hutchison *et al.*, 1999; Forsyth *et al.*, 2002; Gerdes *et al.*, 2003; Kobayashi *et al.*, 2003; Gil *et al.*, 2004; Glass *et al.*, 2006) that performs three distinct functions (Frick and Richardson, 2001). The three functional domains of DNA primase perform, in order from amino- to carboxyl-terminus (Figure 4.2), single stranded DNA recognition/binding (zinc-binding domain, or ZBD; Mendelman *et al.*, 1994; Griep and Lokey, 1996; Biswas and Weller, 1999), RNA synthesis (RNA polymerase domain, RPD; Frick and Richardson, 2001), and helicase binding (helicase binding domain, HBD; Tougu and Marians, 1996; Bird *et al.*, 2000; Mitkova *et al.*, 2003; Oakley *et al.*, 2005).



**Figure 4.2: The domains of DNA primase.** DNA primase has three domains. The zinc-binding domain binds the lagging strand DNA, the RNA polymerase synthesises an RNA primer and the helicase-binding domain attaches DNA primase to DNA helicase.



#### 4.1.1 Preliminary work on *Acinetobacter baylyi* DNA primase

Species in the *Acinetobacter* genus have emerged as clinically acquired pathogens over the last 20 years. Many *Acinetobacter* strains have developed resistance to some, or all current classes of antibiotic, and members of this genus appear to be especially adept at acquiring new resistance mechanisms (Bergogne-Berezin and Towner, 1996; Gales *et al.*, 2006; Perez *et al.*, 2007), resulting in a significant number of pathogenic outbreaks, predominantly in hospitals (Villegas and Hartstein, 2003; Dijkshoorn *et al.*, 2007).

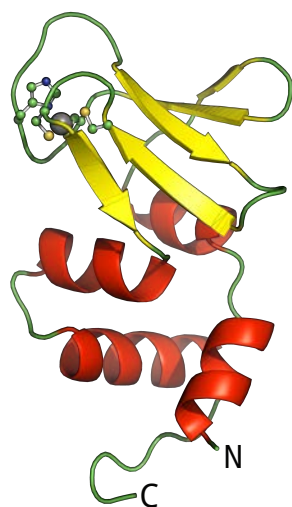
The limited availability of effective antibiotics, and increasing observations of highly antibiotic resistant *Acinetobacter spp.* is troubling. The emergence of antibiotic resistant and pathogenic *Acinetobacter spp.* warrants investigation of important biochemical pathways in the search for novel antibiotic targets. Towards the goal of identifying new antibiotic targets against *Acinetobacter spp.*, the Dixon group has started projects to investigate components of DNA replication in *A. baylyi* ADP1 (Robinson *et al.*, 2010). One such project involves DnaG primase. Members of the Dixon research group have in the past set out to make single and double domain *A. baylyi* DnaG mutants with the goal of identifying new antibiotic targets. These experiments used sequence alignments to predict domain boundaries, and successfully produced both C-terminal HBD and ZBD-RPD domain constructs. However, these rational design experiments did not find suitable breakpoints between the ZBD and RPDs able to produce soluble RPD-HBD or ZBD protein constructs (Andrew Robinson and Stephanie Ruiz; unpublished).

### 4.1.2 The DNA primase zinc-binding domain

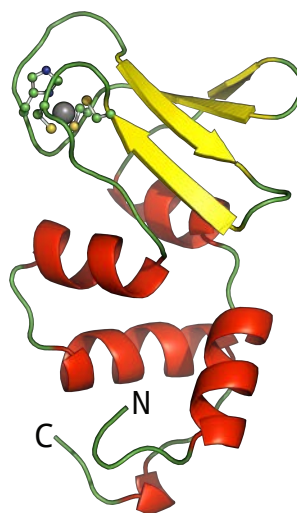
During Okazaki fragment synthesis, the bacterial primase ZBD is believed to be involved with RNA primer initiation site recognition, lagging strand binding and in limiting RNA primer length (Kuchta and Stengel, 2010). Multiple primase molecules are tethered to the helicase during DNA replication, where they can act on the lagging strand after the double helix has been separated. A model proposed for phage T7 DNA gp4 helicase-primase suggests that neighbouring primase domains — each bound to the same helicase — interact in *trans*, so that the ZBD of one primase unit binds to the template strand just ahead of the path of a neighbouring primase, blocking further synthesis of the RNA primer (Qimron *et al.*, 2006). Mutations in the *E. coli* ZBD increase the length of RNA primers (and reduce primase activity; Corn *et al.*, 2005), supporting that the ZBD in *E. coli* is responsible for limiting RNA primer size, similar to the suggested manner for T7 DNA primase (Kuchta and Stengel, 2010).

The structure of the ZBD appears to be highly conserved throughout bacteria; the two known ZBD structures from the distantly related Aquificae and Firmicute phyla are highly similar (C $\alpha$  RMSD: 1.16 Å over 88 residues; Figure 4.3; Shindyalov and Bourne, 1998; Pan and Wigley, 2000; Jia *et al.*, 2004; Corn *et al.*, 2005). The ZBD structures are folded (Pan and Wigley, 2000) and resistant to proteolysis (Tougu *et al.*, 1994; Bird *et al.*, 2000).

**A:** The zinc-binding domain of *Aquifex aeolicus* DnaG



**B:** The zinc-binding domain of *Geobacillus stearothermophilus* DnaG

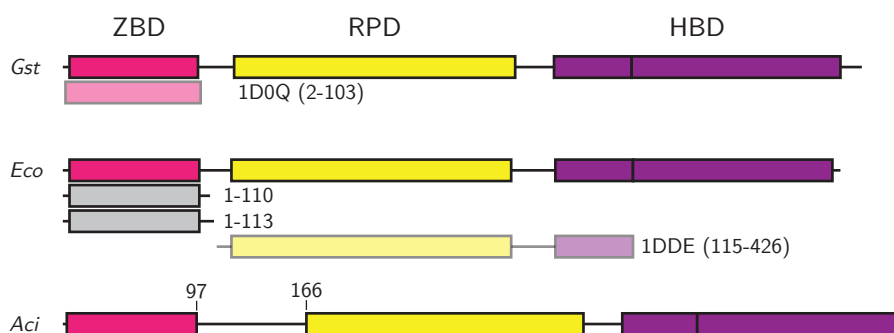


**Figure 4.3: Bacterial primase zinc-binding domains.** Cartoon diagram of **A**, ZBD of *Aquifex aeolicus* (PDB: 2AU3; residues 3–96; Corn *et al.*, 2005) and **B**, ZBD of *Geobacillus stearothermophilus* primase (PDB: 1D0Q; residues 2–101; Pan and Wigley, 2000). Red,  $\alpha$ -helix; yellow,  $\beta$ -sheet; green, loop. Zinc-binding residues are shown in ball-and-stick representation with the zinc atom in grey. The ZBDs of *A. aeolicus* and *G. stearothermophilus* primases are highly conserved ( $C\alpha$  RMSD: 1.16 Å).

#### 4.1.3 Few soluble constructs for bacterial zinc-binding domains are known

Although DnaG has three distinct folded units, for most bacterial species only functional domains for the RPD and HBD have been produced. With the exception of the ZBD of the thermophile *Geobacillus stearothermophilus* (Pan *et al.*, 1999) — from which the ZBD was identified by limited proteolysis and then cloned for soluble over-expression (Pan *et al.*, 1999; Bird *et al.*, 2000) — no bacterial ZBD has been isolated as a folded, soluble protein.

A boundary for the N-terminal ZBD has been known for some time thanks to proteolysis experiments. Limited trypsin digests of *E. coli* DnaG initially produce two fragments, a 16 kDa C-terminal fragment, and a 49 kDa N-terminal fragment which contains the RPD and the ZBD (Tougu *et al.*, 1994); extended proteolysis produces distinct ZBD fragments that have been cleaved after residues Arg110 and Arg113 (Figure 4.4). Although two protein structures for ZBDs are known and proteolytic fragments have been produced for *E. coli* and *G. stearothermophilus* DnaG, our group has had no success producing soluble *A. baylyi* or *E. coli* ZBDs (Nicholas Dixon and Andrew Robinson, personal communication).



**Figure 4.4: Domain location and known soluble constructs of DNA primase.** DnaG proteins from *G. stearothermophilus* (Gst), *E. coli* (Eco) and *A. baylyi* (Aci). Each protein is shown according to annotations in Pfam (<http://pfam.sanger.ac.uk>; Bateman *et al.*, 2000, 2002): (■), N-terminal zinc-binding; (■), RNA polymerase and (■), C-terminal helicase binding domains. Also shown are soluble DnaG domain mutants with breakpoints near the ZBD (pale-coloured); *G. stearothermophilus* 2–103, Pan *et al.* (1999) and *E. coli* 115–426, Keck *et al.* (2000). Shown in grey are ZBD fragments produced by limited proteolysis (Tougu *et al.*, 1994).

The *Aquifex aeolicus* multi-domain ZBD-RPD structure shows the ZBD sitting snugly against the surface of the RPD (PDB: 2AU3; Corn *et al.*, 2005). Potentially the DnaG ZBD is often closely associated with the RPD, and this is supported by the delayed liberation of the ZBD during limited proteolysis (Bird *et al.*, 2000). The close association of the ZBD and RPD may explain the difficulty in producing

appropriate truncations for soluble ZBD other than in *G. stearothermophilus*, as the region linking the domains may be short and hard to identify.

#### 4.1.4 The zinc-binding domain of *Acinetobacter baylyi* DNA primase is followed by a novel sequence insertion

Interestingly, the *A. baylyi* DnaG protein contains a large 69 amino acid extension to the conventional C-terminus of the ZBD and a similar sequence appears in all sequenced members of the *Moraxellaceae* family (of which *Acinetobacter* species are members; Robinson *et al.*, 2010). The sequence insertion, which ranges from 50–100 amino acids, appears to be restricted to the *Moraxellaceae*, where the sequence has very poor conservation (Figure 4.5), and has no homology to any sequence present in the nucleotide data banks.

## 4.2 Aims

We were interested in dissecting the ZBD from the RPD of *A. baylyi* DnaG as a testing platform for the pragmatic protein domain identification methodology proposed in Chapter 3. We aimed to perform deletion of the *dnaG* gene to express libraries of deletions of the DnaG protein from both the I) C-terminus and II) N-terminus surrounding the ZBD sequence extension and RPD boundaries; and III) to examine any putative soluble DnaG truncated proteins for *bona fide* protein solubility and foldedness.

Abu	IRFLMDIDNR	NFIDVMKELS	GNTGVLEPKD	NTDNKKLSY	---	TROVT-K	PSTPPKTVA	----	E--PT	125
Aby	IRFLMDIDNR	NFIDVMKELS	SKSGVELPKD	NFEOKKLSY	---	KRNTQ-K	PEPKPVVNT	----	EKSAP	127
Aca	IRFLMDIDNR	NFIDVMKELS	SSSTGIELPKD	NTENKKLSY	---	TROVT-K	PPVTKTNA	----	E--PV	125
Aol	IRFLMDIDNR	NFIDVMKELS	SSAGVELPKD	NTENKKLSY	---	TROVT-K	PPVAKA-NTA	----	E--PV	125
Par	ISFLRDYENL	TFIEAVNELS	KOTGIEVPKE	EQONVSYQR	---	AKPKPK	SAIKPATSAT	HLQKND	SQPV	133
Pcr	ISFLRDYENL	TFIEAVNELS	KOTGIEVPKE	EQONVSYQR	---	APVKPKSKPK	SAVKPATSAT	HLRKND	SQPV	137
P.sp	IKFLREVENQ	TMEAVCELS	ROTGIEIPKE	DNKDLRYKR	---	SAKPTPT	APAPSWTNG	KRLN-D	THTA	133
Mca	ITFLKEFERM	SFIESVKELS	EOTGIELPKD	DDQKKRKY	---	KKTVKTQG	AKLPTQSP	----	DPSSR	129
Pae	LFVMDHDQL	EFQAVEELA	KRAGMDVPRE	ERGRHETPR	---	---	---	---	---	110
Pbr	LFVMDHDNL	DFQAVEELA	KRAGMEIPRE	ESGR-CHKPR	---	---	---	---	---	109
Avi	LFVMDHDHL	EFQAVEELA	RRAGLEVPRE	ERGG-PHOPR	---	---	---	---	---	109
Cja	LFVMDYERU	SFQAVEOLA	RITGLEVPRE	VQSEAEKRE	---	---	---	---	---	110
Abu	TQ---API	----	----	EDQ	YNTFEPVY	----	F	DD	----	144
Aby	SHHAAS	----	----	ESSEQI	DQLQAPSY	----	F	DD	----	151
Aca	IPQOPS	----	----	DES	YNTLEPVF	----	F	DD	----	145
Aol	TPTPOOPT	----	----	DES	YNTLEPVF	----	F	DD	----	147
Par	TNNK-SSMSM	PA	----	DDT	QPTY	----	---	DD	----	161
Pcr	SNNQ-SSIIM	PA	----	DDT	Q	----	---	DD	----	165
P.sp	ATTPHSGQTN	GLATNTTSVA	DAPPAYDED	SYFNLD	SAPP	SDWDM-DTPG	DNLGYS	SGFT	DMSSSNFGGQ	203
Mca	QNFIIHQ	----	----	DTR	PSPATDWVND	LSAY	----	DTY	----	171
Pae	----	----	----	----	----	----	----	----	SGLA	110
Pbr	----	----	----	----	----	----	----	----	----	109
Avi	----	----	----	----	----	----	----	----	----	109
Cja	----	----	----	----	----	----	----	----	----	110
Abu	----	P	FAQFEQPF	FD	MPV	----	----	QEG	----	169
Aby	----	P	FAQFDQSYMG	FEDAP	----	----	----	QEG	----	177
Aca	----	P	FAQFEQPF	FDEPV	----	----	----	QEG	----	170
Aol	----	P	FAQFEQPF	FDEPV	----	----	----	QEG	----	172
Par	PPLDAYDAVP	YAMDGYDAH	QDDNYPPAW	LAGGDMAGLH	GSDINHSFDN	NNDSDNEDG	----	----	NYDLLE	226
Pcr	PPLDAYDAVP	YVMDGYDAH	QDDNGYPPAW	LTGDDNTALY	----	N-DN	----	SISDKDG	----	222
P.sp	ATI	----	----	----	----	----	----	EDG	----	217
Mca	ELYCTRAPIL	TQSPNAAQCD	----	----	----	----	----	----	LYSLIT	197
Pae	----	----	----	----	----	----	----	QP	TDSPLYPLIS	122
Pbr	----	----	----	----	----	----	----	QP	TDSPLYPLIT	121
Avi	----	----	----	----	----	----	----	QS	ADSPYPLIA	121
Cja	----	----	----	----	----	----	----	QE	KKS-LYSLLE	121
Abu	NVAQFYEHQL	PTSQKAKN	----	YFKRGISD	QTIQFWRLGY	APEDMOHL	----	EKAFFPY	----	227
Aby	NVAQFYEHQL	PNSNKAQQ	----	YFKRGISA	ETIQFWRLGY	APEDMOHL	----	EKAFFPY	----	235
Aca	NVAQFYEHQL	PHSQKAKN	----	YFKRGITN	QTIQFWRLGY	APEDMOHL	----	EKAFFPY	----	228
Aol	NVAQFYEHQL	PNSQKAKN	----	YFKRGITN	QTIQFWRLGY	APEDMOHL	----	EKAFFPY	----	230
Par	KIQQFYQNL	SIHPHAKH	----	YFLSRGISD	EIFETFGIGY	APFGMOHL	----	EHQFPQ	----	284
Pcr	KIQQFYQNL	SIHPHAKH	----	YFLSRGISD	EIFETFGIGY	APFGMOHL	----	EHQFPQ	----	280
P.sp	KIQQFYQNL	RNNLHAMA	----	YETERGLTD	ATINEFGIGY	APTGMOHL	----	EEAFPO	----	275
Mca	AVHDYYQLML	NNFTFAKQ	----	YFLDRGISE	ETIOTFGIGY	APDGMOHL	----	EQVFPO	----	255
Pae	AAAEFYKQAL	KSHPARKAAV	NYLKRGLTG	ETARDFGLGF	APPGWNLK	HLGGDNLQK	AMLD-AGL	----	----	189
Pbr	AAADFYRQAL	KSHPARKAAV	DYLKRGLTG	ETARDFGLGF	APPGWNLK	HLSSDTLQOR	AMID-AGL	----	----	188
Avi	TAADYYRQAL	KSHPARQAV	DYLKRGLTG	VIARDFALGF	APPGWNLK	HLGGDALQK	AMIE-AGL	----	----	188
Cja	KADDFYQHQL	ROHPSKHLAV	NYLKRGLTG	KIAKTYGVGF	APPGWNLK	TLGQDDDDKH	LLIQ-GM	----	----	188

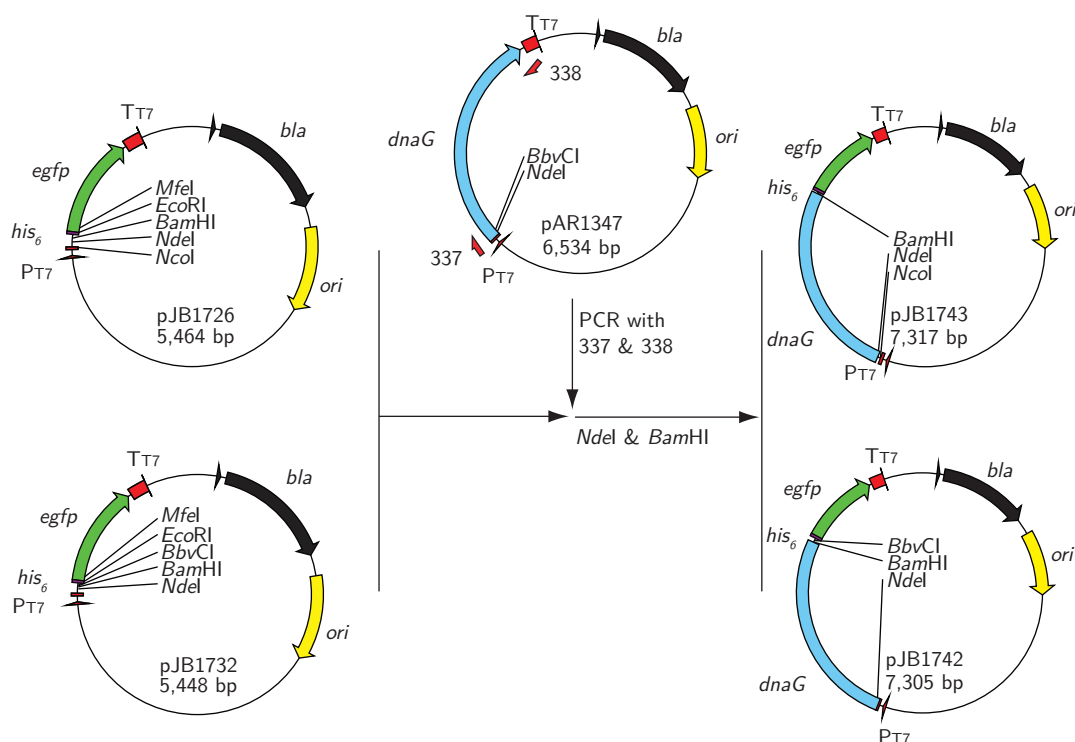
**Figure 4.5: Zinc-binding domain extension in *Moraxellaceae*.** Protein sequence alignment showing the novel ZBD sequence insertion of the *Moraxellaceae* compared to the closest relatives from *Pseudomonadales*. Alignments were generated using ClustalW (Thompson *et al.*, 2002) and were manually adjusted. *Moraxellaceae* species: Abu, *Acinetobacter baumannii*; Aby, *Acinetobacter sp.* ADP1 (baylii); Aca, *Acinetobacter calcoaceticus*; Aol, *Acinetobacter oleivorans*; Par, *Psychrobacter arcticus*; Pcr, *Psychrobacter cryohalolentis*; P.sp, *Psychrobacter sp.* PRwf-1; Mca, *Moraxella catarrhalis*. *Pseudomonadulus* species: Pae, *Pseudomonas aeruginosa*; Pbr, *Pseudomonas brassicacearum*; Avi, *Azotobacter vinelandii*; Cja, *Cellvibrio japonicus*. Amino acid residues are shown in single-letter abbreviations and are coloured: red, acidic; blue, basic; orange, non-polar; green, polar; yellow, cysteine.

## 4.3 Materials and methods

### 4.3.1 Plasmids for gene truncation of *Acinetobacter baylyi* DNA

#### primase

Plasmids were constructed for gene truncation of *A. baylyi* DnaG by inserting the *dnaG* gene from pAR1347 (provided by Andrew Robinson) into pJB1726 and pJB1732 (Section 3.4.9), to produce pJB1743 and pJB1742 for N- and C-terminal deletions, respectively (Figure 4.6). pJB1743 allows 5'-terminal gene deletion and pJB1742 3'-terminal gene deletion with each plasmid providing DNA encoding a C-terminal His<sub>6</sub> tag followed by the EGFP fusion protein for protein purification and solubility selection. The open reading frame (ORF) for *A. baylyi* DnaG was PCR amplified from pAR1347 using oligonucleotide primers 337 and 338 (Appendix C). The PCR primers produced an amplification product with an appropriate *Nde*I restriction site at the start of the *dnaG* ORF and a *Bam*HI restriction site at the end of the ORF (Section 3.3.2.2). To facilitate truncation generation using *Nt.Bbv*CI, PCR primer 337 also caused a silent mutation in the *A. baylyi* DnaG sequence such that the *Bbv*CI restriction site present at the start of DnaG was removed. Plasmids pJB1726, pJB1732 and the *dnaG* PCR product were digested with *Nde*I and *Bam*HI, gel purified and ligated. Successful transformants were selected and extracted plasmids were sequenced to confirm their identities.



**Figure 4.6: Plasmids for gene truncation of *dnaG*.** pJB1743 (for N-terminal deletions) and pJB1742 (C-terminal deletions) were made by placing the *A. baylyi dnaG* gene from pAR1347 between the *NdeI* and *BamHI* sites of pJB1726 and pJB1732 (Section 3.4.9). The *dnaG* gene was amplified by PCR using oligonucleotide primers 337 and 338 (Appendix C), digested with *NdeI* and *BamHI* and gel purified. pJB1726 and pJB1732 direct expression of C-terminal His<sub>6</sub> and EGFP fusion proteins for protein purification and solubility selection.

## 4.3.2 Library preparation

### 4.3.2.1 Preparation of plasmid DNA for uni-directional truncation

Gene deletion plasmids containing *dnaG* were prepared for uni-directional deletion using exonuclease III. For C-terminal deletion of *dnaG*, 30 µg of pJB1742 was digested with 30 U of the nicking endonuclease Nt.*BbvCI* in NEB restriction enzyme buffer 4 with 100 µg.mL<sup>-1</sup> BSA for 2 h at 37°C and the enzyme was heat inactivated by incubation at 80°C for 20 min.



For N-terminal deletion of *dnaG* 20 µg of purified pJB1743 was digested with 20 U of *NcoI* in NEB restriction enzyme buffer 3 with 100 µg.mL<sup>-1</sup> BSA for 2 h at 37°C, and the enzyme was heat inactivated by incubation at 80°C for 20 min. The *NcoI*-digested plasmid was then end protected by adding 40 µM each of dGTPαS (Glen Research), dATP, dCTP, dTTP and 20 U DNA polymerase I (Klenow fragment), and the mixture was incubated at 30°C for 15 min; DNA polymerase I was then heat inactivated by incubation at 65°C for 20 min. Linearised pJB1743 was made susceptible to uni-directional *ExoIII* digestion using 20 U *NdeI*, incubated for 2 h at 37°C and *NdeI* was heat inactivated by incubation at 80°C for 20 min.

#### 4.3.2.2 Uni-directional truncation using exonuclease III

Exonuclease III reactions were performed in 60 mM Tris-HCl pH 7.6, 5 mM MgCl<sub>2</sub>, 1 mM DTT and 100 µg.mL<sup>-1</sup> BSA. Reaction mixtures were prepared with 1 µg of target plasmid DNA and 100 U of *ExoIII* in 20 µL per time point sample. Exonuclease reactions were performed in a heat block and initiated by introducing *ExoIII* to warmed DNA in reaction buffer followed by mixing by pipette.

Exonuclease reactions were terminated by chelation of magnesium and increased acidity. Each 20 µL time point sample (3' deleted; 15 s intervals between 6 min 30 s and 8 min 45 s and 5' deleted: 15 s intervals between 1 min 25 s and 2 min 40 s) was introduced into a separate pre-prepared tube with independent 2 µL droplets of 3 M sodium acetate pH 5.2 and 125 mM EDTA,

such that addition of the exonuclease treated DNA sample mixed the other two solutions. DNA was immediately precipitated from terminated exonuclease reactions by addition of 50  $\mu$ L of 100% ethanol (see Section 2.3.9). After the DNA pellet had been washed with 70% ethanol and dried, TE was added to resuspend the truncated DNA. The truncated plasmid DNA from multiple *ExoIII* reactions were pooled to make separately a 3' deleted and a 5' deleted gene library.

#### 4.3.2.3 Removal of ssDNA from exonuclease truncated DNA

Resuspended exonuclease truncated plasmid DNA was incubated in a total volume of 10  $\mu$ L of 50 mM sodium acetate pH 5, 30 mM NaCl, 1 mM ZnSO<sub>4</sub> and 1 U of mung bean nuclease per 1  $\mu$ g DNA. Mung bean nuclease reactions were incubated at 30°C for 30 min and then terminated by placing on ice and addition of SDS to a final concentration of 0.01%. Truncated plasmid DNA was precipitated using sodium acetate, EDTA and ethanol (Section 2.3.9), washed with 70% ethanol and then resuspended in TE.

#### 4.3.2.4 Ligation of the truncated plasmid pool

Mung bean nuclease digested *A. baylyi* N- and C-terminally deleted plasmid pools were repaired using DNA polymerase I (Klenow fragment) and then ligated using T4 DNA ligase. Precipitated DNA pellets were resuspended in TE and then diluted in a dual purpose DNA polymerase I/T4 DNA ligase reaction buffer (50 mM Tris-HCl pH 7.6, 10 mM MgCl<sub>2</sub>, 10 mM DTT and 1 mM ATP) to a DNA concentration of 6 ng. $\mu$ L<sup>-1</sup>.

The truncated DNA in dual purpose buffer was warmed to 37°C and then DNA polymerase I (Klenow fragment; 1 U per µg of DNA) was added and the mixture incubated for 5 min, after which 40 µM each dATP, dCTP, dGTP and dTTP was added and incubation continued for a further 30 min. The mixtures were then cooled to 4°C and T4 DNA ligase (1 U per 300 ng of DNA) was added before incubation overnight at 4°C.

To increase transformation efficiency, salt was removed from the ligated gene truncation plasmid pools. Plasmid pools were diluted to 500 µL in TE and concentrated by centrifugal ultrafiltration, re-diluted in TE and concentrated a second time.

#### 4.3.2.5 Library transformation and expression

Ligated gene truncation libraries were transformed into *E. coli* BL21(λDE3)*recA* and cultured on agar plates to induce expression of truncated DnaG-EGFP fusion proteins (Section 2.2.3). After one h recovery in LB, transformed library pools were plated onto LB agar or Selection agar (15 g.L<sup>-1</sup> agar in 5 g.L<sup>-1</sup> yeast extract, 2 mM MgSO<sub>4</sub>, 50 mM Na<sub>2</sub>HPO<sub>4</sub>, 50 mM KH<sub>2</sub>PO<sub>4</sub>, 25 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 0.1 g.L<sup>-1</sup> ampicillin, 0.5 g.L<sup>-1</sup> glucose, 2 g.L<sup>-1</sup> α-lactose and 5 g.L<sup>-1</sup> glycerol). Plated gene truncation libraries were incubated at 30°C overnight and stored at 4°C thereafter.

### 4.3.3 Truncation mutation identification

Successful transformants from truncated DnaG-EGFP fusion libraries were observed under a long-wave UV handheld lamp to excite EGFP and produce fluorescence. Selected bacterial colonies were clonally isolated on LB agar plates (Section 2.2.2), with incubation overnight at 30°C.

Truncated DnaG-EGFP fusion library transformants were screened for insert size by colony PCR and simultaneously stored as recoverable cultures at  $-80^{\circ}\text{C}$  as described in Section 2.3.5. Briefly, clonally isolated bacterial colonies were picked and inoculated into 200  $\mu\text{L}$  LB-Miller with 7% DMSO from which a 1  $\mu\text{L}$  sample was used for colony PCR. A sample of each identified colony was also inoculated onto an agar plate grid prior to the storage of bacterial cultures at  $-80^{\circ}\text{C}$ . Samples of each bacterial culture smeared onto an agar plate grid were incubated overnight at 30°C and observed for degree of EGFP fluorescence phenotype by comparison to an identical culture solely expressing EGFP or not (*i.e.*, containing pJB1706 or non-truncated deletion plasmid, respectively). Nucleotide sequences were determined for selected plasmids (Section 2.3.11).

### 4.3.4 Removal of EGFP sequence from fusion genes

Following library selection for expression of soluble N- or C-terminal deleted *A. baylyi* DnaG-EGFP fusion proteins, selected sequenced plasmids were chosen for expression of truncated DnaG proteins without EGFP. To facilitate straightforward removal of EGFP from the protein sequence, a stop codon

was introduced following the His<sub>6</sub> sequence for each plasmid (see Section 3.3.2.3). Selected plasmids encoding *A. baylyi* DnaG truncation mutants were digested with *Mfe*I, and ligated with the self-complementary oligonucleotide 420 (AATTGTAAGCTTAC; Section 2.3.10). Ligation products were transformed into BL21(λDE3)*recA* and selected on expression-inducing Selection agar plates. Transformants were selected based on the loss of their green fluorescent phenotype due to the successful introduction of a stop codon preceding the *egfp* ORF. The plasmids were then purified and sequenced for confirmation.

#### 4.3.5 Removal of His<sub>6</sub>-tag and EGFP sequence from fusion proteins

In a similar fashion to removal of the EGFP fusion sequence (Section 4.3.4), some *A. baylyi* DnaG truncation mutants were selected for expression without the His<sub>6</sub> or *egfp* fusion tags (Section 3.3.2.3). The selected plasmids were linearised with *Eco*RI and then the sticky ends were filled in using DNA polymerase I (Klenow fragment) and 40 μM each dNTP in the *Eco*RI restriction enzyme buffer for 30 min at 30°C. The end-filled linear plasmid was transformed directly into BL21(λDE3)*recA* and white, non-fluorescent transformants were selected as above. The plasmids were then purified and sequenced for confirmation.

### 4.3.6 Expression, solubility examination and purification of truncation mutants

To examine the solubility of DnaG truncation mutants identified by solubility library selection experiments, selected proteins were expressed with a C-terminal His<sub>6</sub>-tag in *E. coli* BL21( $\lambda$ DE3)*recA*/pLysS by auto-induction (see Section 2.5.1). Individual 400 mL cultures were incubated with shaking for 24 h at 30°C. To compare the solubility of the selected truncation mutants to each other, a sample of each over-expressed BL21( $\lambda$ DE3)*recA* culture was then recovered by centrifugation at  $8,000 \times g$  and the bacterial pellets were resuspended to an  $A_{600}$  of 10 AU in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole.

To facilitate efficient lysis of the cultures that overproduced DnaG truncation mutants and T7 lysozyme, the resuspended bacterial cells were lysed by repeated freeze-thaw cycles between  $-80^{\circ}\text{C}$  and  $4^{\circ}\text{C}$  or processed through a French press cell three times; cell lysate from freeze-thaw lysis was then passed through a syringe needle to shear DNA. Following cell lysis, the soluble and insoluble cellular fractions were separated by centrifugation at  $30,000 \times g$  for 30 min. Equal samples of whole cell lysate and soluble cellular fractions for each truncation mutant were analysed directly by SDS-PAGE (Section 2.5.9), while insoluble cellular fractions were resuspended in a small volume of 8 M urea and then diluted in SDS-loading buffer to be in proportion to the total cell lysate, then analysed by SDS-PAGE.

#### 4.3.6.1 Purification of over-expressed His<sub>6</sub>-tagged truncated proteins

Cells from 400 mL cultures of isolates that over-expressed DnaG truncation mutants with C-terminal His<sub>6</sub>-tags (see Section 4.3.6) were resuspended in 30 mL of immobilised metal ion chromatography (IMAC) binding buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole) and lysed by French press. The soluble fraction was recovered by centrifugation at  $30,000 \times g$  for 30 min.

Once the cleared bacterial lysate had been filtered through a 0.45  $\mu$ m cassette (Merck Millipore), it was applied to a 1 mL HisTrap HP column (GE Healthcare Life Sciences) using an ÄKTApurifier (GE Healthcare Life Sciences). The column was washed with IMAC binding buffer. Once the  $A_{280}$  reached baseline, the bound protein was eluted using IMAC binding buffer with a step gradient of imidazole from 20 mM through 500 mM. Eluted fractions were pooled based on purity observed by SDS-PAGE; the purified proteins eluted at concentrations suitable for later experiments. Mass spectrometry analysis (Section 2.5.10) of His<sub>6</sub>-tagged DnaG truncated mutants in 0.1% formic acid produced electrospray ionization (ESI)-mass spectra in excellent agreement with the sequence predicted masses (results not shown).

#### 4.3.6.2 Improved purification of His<sub>6</sub>-tagged truncated proteins

In later purifications of His<sub>6</sub>-tagged proteins, the protocol was modified to include a wash step for the protein bound column using 10 column volumes (CV) of IMAC binding buffer with 1 M NaCl to remove non-specifically bound species,

then 10 CV of IMAC binding buffer, prior to a step elution using 50 mM Tris-HCl pH 8.0, 300 mM NaCl and 90 mM imidazole. Eluted protein was pooled and dialysed into 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA and 1 mM DTT.

#### 4.3.7 Circular dichroism of truncated protein

*A. baylyi* His<sub>6</sub>-tagged DnaG ZBD truncation mutants purified by the initial method (Section 4.3.6.1) were dialysed into circular dichroism (CD) buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl) and filtered using a 0.22 µm cassette (Merck Millipore) for CD spectroscopy. Far-UV CD spectra were acquired using a Jasco J-810 spectropolarimeter (Jasco, Victoria, Canada) with a Jasco circulating water bath at 25°C. Samples for spectra were diluted in CD buffer to produce useful spectra (approximately 0.2 mg.mL<sup>-1</sup>) in a 0.05 cm quartz cell and the following parameters were used: range, 300–190 nm; recording speed, 100 nm.min<sup>-1</sup>; accumulating 4 scans.

##### 4.3.7.1 Purification of DnaG<sup>1–165</sup> and DnaG<sup>1–170</sup>

Untagged *A. baylyi* DnaG ZBD truncated proteins were over-expressed by auto-induction (see Section 2.5.1). Cultures of BL21(λDE3)*recA*/pLysS containing plasmids that direct expression of DnaG truncations were prepared by shaking at 30°C until the cultures were saturated. The bacterial cells containing expressed DnaG<sup>1–165</sup> and DnaG<sup>1–170</sup> were harvested by centrifugation at 8,000 × *g* for 10 min. Pelleted bacterial cells (10 g) were resuspended in TBS<sub>150</sub> (50 mM Tris-HCl



pH 8.0, 150 mM NaCl, 1 mM EDTA and 1 mM DTT; 5 mL per g cells) and lysed using a French press. Soluble proteins were recovered by centrifugation at  $30,000 \times g$  for 30 min. The soluble protein fraction was treated with  $0.2 \text{ g.mL}^{-1}$   $(\text{NH}_4)_2\text{SO}_4$  and the precipitate collected by centrifugation at  $30,000 \times g$  for 30 min. The protein pellets containing the DnaG truncation proteins were collected and resuspended in TBS<sub>150</sub>, and then dialysed against the same buffer overnight.

The ammonium sulphate isolated protein fractions were then purified in two aliquots by flowing through a DEAE chromatography column (35 mL Toyopearl DEAE-650M) equilibrated with TBS<sub>150</sub> and connected to an ÄKTApurifier system; under these conditions contaminants with a very strong negative charge such as DNA are bound to the DEAE column and the protein of interest flows through. The protein fractions which did not bind to the DEAE column were pooled and proteins were precipitated by addition of  $0.5 \text{ g.mL}^{-1}$   $(\text{NH}_4)_2\text{SO}_4$  and collected by centrifugation at  $30,000 \times g$  for 30 min. The protein pellet containing the DnaG truncation protein was collected and resuspended in TBS<sub>50</sub> (50 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA) and then dialysed overnight in the same buffer.

The DEAE flow-through purified protein sample was then clarified by centrifugation at  $30,000 \times g$  for 30 min and then, in two separate purifications, loaded onto a SuperQ column (30 mL Toyopearl SuperQ-650M) connected to an ÄKTApurifier system, and the column washed with TBS<sub>50</sub>. Once the  $A_{280}$  of the column eluant reached baseline, a linear gradient elution was applied from 50–500 mM NaCl in TBS over 400 mL. Partially purified DnaG<sup>1–165</sup> and DnaG<sup>1–170</sup> eluted

at a NaCl concentration of  $\sim 160$  mM.

Pooled SuperQ fractions were dialysed overnight with TBS<sub>50</sub> and then loaded, in two separate purifications, on to a MonoQ column (8 mL; GE Healthcare Life Sciences). Pure DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> were each eluted as a sharp peak in an elution gradient of 50–500 mM NaCl over 200 mL. Mass spectrometry analysis of DnaG DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> in 0.1% formic acid produced ESI-mass spectra in excellent agreement with their predicted masses based on the gene sequence (data not shown).

Purified *A. baylyi* DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> truncated proteins were dialysed extensively against 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA. Protein samples were then concentrated using Amicon Ultra-4 Centrifugal Filter Units (Merck Millipore) with a MWCO of 4,000 Da.

#### 4.3.8 Nuclear magnetic resonance analysis of *Acinetobacter baylyi*

##### DNA primase zinc-binding domain

An aliquot of concentrated *A. baylyi* ZBD protein was dialysed extensively against 20 mM Tris-HCl pH 7.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA; dialysis was completed without EDTA present in the dialysis buffer. Truncated DnaG<sup>1-165</sup> was stored and shipped in liquid nitrogen to Dr Xun-Cheng Su, State Key Laboratory of Elemento-organic Chemistry, Nankai University, Peoples Republic of China.

Otherwise, aliquots of DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> proteins were dialysed extensively against 50 mM Tris-HCl pH 7.5 (or pH 7.0), 150 mM NaCl, 1 mM DTT and 1 mM EDTA; dialysis was completed without EDTA present in the dialysis buffer. Protein aliquots were stored and shipped on dry ice to Dr Kiyoshi Ozawa at the Australian National University, Canberra, Australia for analysis by TOCSY NMR.

NMR experiments recorded spectra using Bruker Avance 600 MHz NMR spectrometers with cryoprobes at 25°C; 80 ms mixing time was applied for 2D TOCSY spectra.

#### 4.3.9 Protein crystallography of DnaG<sup>1-165</sup> and DnaG<sup>1-165</sup>

Purified DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> were subjected to crystallisation trials using QIAGEN NeXtal JCSG+, PEGs I and PEGs II suites in sitting drop configuration with Corning 96 well crystallisation plates (reservoir volume 50  $\mu$ L). Purified, concentrated protein (1  $\mu$ L ; 10 mg.mL<sup>-1</sup>) was mixed with crystallisation solution (1  $\mu$ L), and the crystallisation plate sealed and then stored at 4°C or 16°C and occasionally inspected. Later crystallisation trails used DnaG<sup>1-165</sup> at 200 mg.mL<sup>-1</sup> and DnaG<sup>1-170</sup> at 44 mg.mL<sup>-1</sup> and the QIAGEN NeXtal JCSG+ suite.

#### 4.3.10 Homology modelling of the *Acinetobacter baylyi* DNA primase zinc-binding and RNA polymerase domains

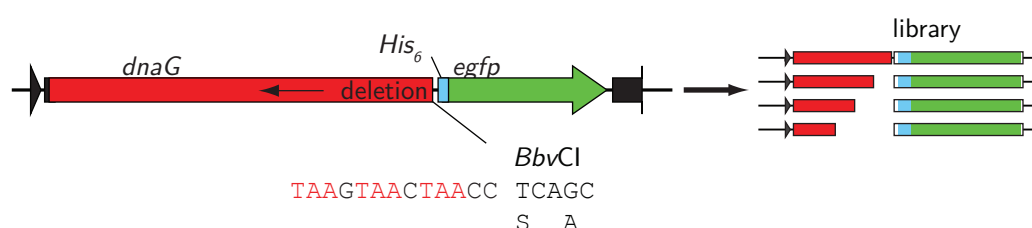
Protein structures were modelled using ModWeb (Eswar *et al.*, 2003), an online server for protein structure modelling using the MODELLER program (Sali and Blundell, 1993; Fiser *et al.*, 2000; Martí-Renom *et al.*, 2000; Eswar *et al.*, 2006) to calculate three-dimensional protein structures against the best available homologues in the PDB. The ModWeb service outputs model quality assessments of the predicted protein structure including the GA341 score which accounts for sequence identity, structural compactness and residue level statistical measures of contacts,  $\phi/\psi$  dihedral angles and accessible surfaces (Melo *et al.*, 2002; Eramian *et al.*, 2008). The GA341 score ranges from 0 to 1 with a score  $\geq 0.7$  suggesting a reliable structure with a probability of the correct fold larger than 95%. QMEAN scores were calculated using the SWISS-MODEL server (<http://swissmodel.expasy.org/qmean/cgi/index.cgi>; Benkert *et al.*, 2008) and protein structures were visualised using the PyMOL Molecular Graphics System, Version 1.5.7 ([www.pymol.org](http://www.pymol.org); Schrödinger, LLC, 2012).

## 4.4 Results

### 4.4.1 C-terminal truncation of *Acinetobacter baylyi* DNA primase

To search for soluble C-terminal truncations representing the *A. baylyi* ZBD, an *ExoIII* gene truncation library was produced from pJB1742 as described in

Section 4.3.2. Treatment of pJB1742 with Nt.*Bbv*CI produces a single DNA nick just upstream of the end of *dnaG* and is separated from *dnaG* by stop codons in all three reading frames. *Exo*III will uni-directionally truncate from the 3' nick, proceed through the three-frame stop codons and then through *dnaG*. Ligation of limited *Exo*III digestions produces plasmids in which truncated genes for *dnaG* are fused to the downstream *egfp* sequence; some of which will be in-frame (Figure 4.7).



**Figure 4.7: Methodology for C-terminal truncation of *dnaG*.** Digestion of pJB1742 with Nt.*Bbv*CI produces a single 3' end (between the sequence CC and TCAGC) for uni-directional deletion using *Exo*III. Initial digestion through three stop codons (red) must occur to allow gene fusion to the downstream *His*<sub>6</sub> and *egfp* sequences. Following uni-directional gene deletion and plasmid repair, one-in-three truncated genes are in-frame with the downstream *egfp* gene, producing a truncated DnaG-EGFP fusion protein when expressed in *E. coli*.

Deletion reactions were carried out at 30°C for between 6 min 30 s and 8 min 45 s. Under these reaction conditions, *Exo*III progresses with a rate of 200 bp.min<sup>-1</sup> and resulted in a library of genes for the 5'-terminus that span about a third of *dnaG* (see Appendix B.1). The samples from all time points were pooled, ligated and transformed into BL21(λDE3)*recA* as described in Section 4.3.2. The recovered transformed BL21(λDE3)*recA* cells were spread onto 10 Selection agar plates containing ampicillin. After overnight incubation at 30°C approximately 2,100 colonies developed, forty-nine of which exhibited a green fluorescent phenotype.

To determine the coverage of the gene library, 15 randomly selected plasmids

were sequenced, but of these eight mutants were completely un-truncated while the seven truncated genes ranged in size from 181 to 830 bp. Sequenced green fluorescent colonies contained truncated genes ranging from 200 to 700 bp (discussed later). To reduce the number of non-truncated mutants in the *A. baylyi* C-terminal deletion library, the library pool was digested with *Bam*HI, a site for which is present in the template plasmid pJB1742 but should not appear in any truncated plasmid. The plasmid library was de-salted and BL21( $\lambda$ DE3)*recA* was transformed with the new *Bam*HI treated library and transformants recovered on 11 agar plates. This resulted in approximately 1,300 colonies, 35 of which were green fluorescent. Twenty randomly selected non-green-fluorescent mutants were sequenced to determine the coverage of truncated *A. baylyi dnaG* genes (Table 4.1). Of the twenty mutants sequenced, two had truncated past the initiation codon at the start of the gene and one mutant did not appear to have been truncated. The remaining 17 plasmids contained genes ranging in size from 115 through 872 bp, representing genes in all three reading frames relative to the C-terminal fusion gene. Three mutants were also slightly truncated through the C-terminal fusion gene, missing between one and four nucleotides, indicating that these had been deleted through the action of *Exo*III, mung bean nuclease or DNA polymerase I (Klenow fragment). The gene deletion library was considered sufficient to identify those expressing soluble proteins that comprise the N-terminal ZBD of DnaG.

Thirty-five plasmids expressing in-frame and putatively soluble DnaG-EGFP fusions — indicated by their green fluorescent colony phenotype — were first

**Table 4.1: Randomly sequenced *dnaG* C-terminally truncated plasmids.** Gene sequences for deleted DnaG proteins (shown in the correct reading frame for *dnaG*) are fused to downstream purification and solubility reporting proteins. Mutants with over-truncated C-terminal fusion tags are displayed in red.

End of truncated <i>dnaG</i> mutant	Out of frame residues	Fusion sequence	Protein length (amino acids)
Over truncated	−14	TCAGCGGGCTCCTCT	−4.7
Over truncated	−8	TCAGCGGGCTCCTCT	−2.7
ATTCGGGTTGCTGTC	CA	TCAGCGGGCTCCTCT	38.3
CATTGCTTTGGCTGT		---GCGGGCTCCTCT	62
TCCTAAAGATAACTT	T	TCAGCGGGCTCCTCT	99.7
ACTTTGAACAAAAGA	AA	TCAGCGGGCTCCTCT	103.3
TTTCCTATAAGCGCA	AT	TCAGCGGGCTCCTCT	109.3
ACCATCTCACCATGC	T	TCAGCGGGCTCCTCT	130.7
GCAATATCCGAATCA		TCAGCGGGCTCCTCT	136
CCTCTTACTTTGACG	AT	TCAGCGGGCTCCTCT	150.3
ACAGTTTGATCAAAG	C	TCAGCGGGCTCCTCT	158.7
ATCAAAGCTATATGG	GT	TCAGCGGGCTCCTCT	161.3
GGTTTTGAAGATGCC		TCAGCGGGCTCCTCT	166
GACCTATTGAAAAAT		TCAGCGGGCTCCTCT	178
TAGCTCAGTTTTATG	AA	TCAGCGGGCTCCTCT	183.3
CAGTTTTATGAAAAA		TCAGCGGGCTCCTCT	185
ACAACGTGGCTTGAG	T	----CGGGCTCCTCT	202.67
AAAGGGCGTGTGGTC		TCAGCGGGCTCCTCT	267
AGACTCCGAGATTTT	C	-----GCTCCTCT	290.67
CCTAAGTAACTAACC		TCAGCGGGCTCCTCT	not truncated

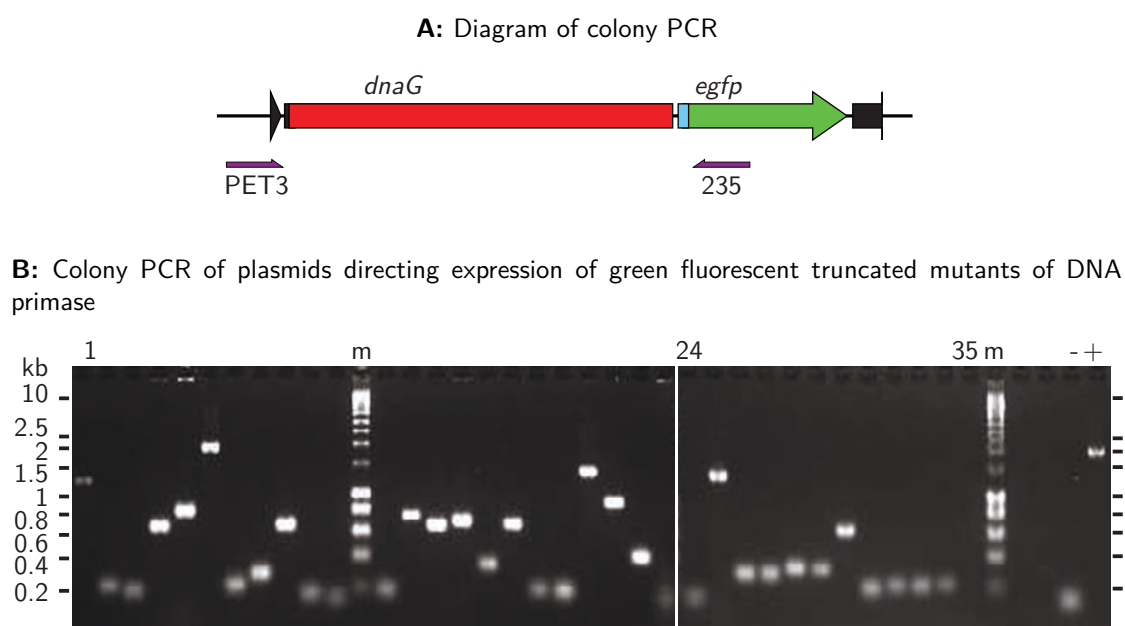
examined by colony PCR to determine the size of the genes (Figure 4.8). One mutant failed to produce a PCR amplification product indicating that one of the two primer sites had been removed. The majority of putative truncated *dnaG* genes ranged in size between 50 and 800 bp.

The recovered green fluorescent colonies had varied green fluorescence intensity, which was easily visible by patching green mutants onto an agar grid (for example, see Figure 4.9); some strains appeared to lose their green fluorescent phenotype on successive culturing (data not shown). The intensity of the green fluorescence phenotype was recorded and aggregated from multiple culture plates and assigned a score from 0–4 (see Table 4.2); a score of zero was assigned when green fluorescence could not be visually determined and a score of four when a colony has the same brightness as His<sub>6</sub>-EGFP expressed from pJB1706 on the same agar plate.

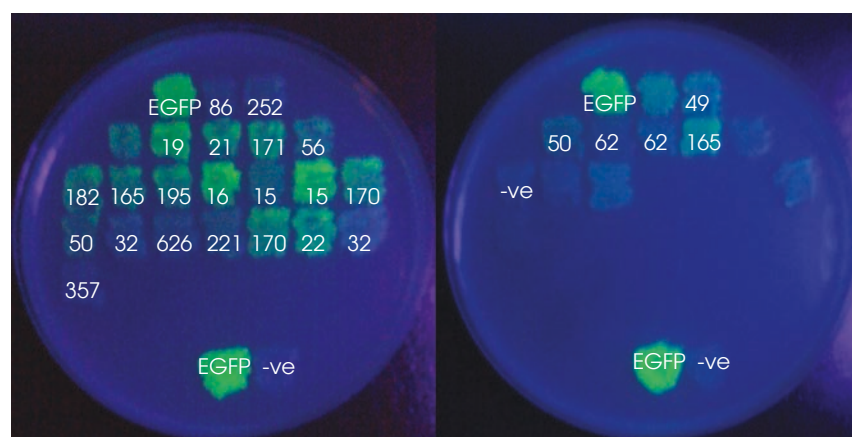
Precise sites of truncation of the putative soluble DnaG proteins, all 35 mutants from the *Bam*HI treated library and 11 from the initial library (pre-*Bam*HI treated), were then determined by dye terminator sequencing of plasmids designated pJB1765–pJB1791 (Table 4.2). Soluble truncation products comprised small peptides, or had C-terminal residues close to the end of the unusual *Moraxellaceae* sequence insertion/domain linker. No putative soluble truncated proteins were observed within the sequence insertion region of DnaG, suggesting that it is important for expressing soluble ZBD constructs.

Multiple sequence alignment of the *Moraxellaceae* insertion sequence (Figure 4.5) shows that around residue 169 (*A. baylyi*), the DnaG sequences begins to





**Figure 4.8: DNA primase C-terminally deleted green fluorescent mutants.** Colony PCR was performed using primers PET3 and 235 to identify the size of genes present in *dnaG* truncated mutants. **A:** Diagram of PCR primer binding sites in gene truncation plasmids; PCR product contain 138 bp additional to the truncated gene. **B:** Agarose gel electrophoresis (1%) of truncated *dnaG* mutants (1–35), non-truncated pJB1742 (+), EGFP expression plasmid pJB1706(–), and HyperLadder I molecular size markers (m).



**Figure 4.9: Green fluorescence phenotype of *Acinetobacter baylyi* DnaG ZBD truncations.** Green fluorescent *E. coli* colonies expressing C-terminally truncated *A. baylyi* DnaG-EGFP fusions were smeared on Selection media plates and incubated overnight at 30°C. The number identifies the C-terminal residue of the truncation. Green fluorescent phenotype was examined under a long-wave UV lamp and assigned a score from 0–4; 0, green fluorescence not visible or 4, expression of His<sub>6</sub>-EGFP from pJB1706.

**Table 4.2: Sequenced C-terminally deleted *A. baylyi* DnaG constructs.** Green fluorescent colonies expressing truncated DnaG-EGFP fusion proteins were sequenced to identify the mutations present. Shown at the **top** are the DNA and protein sequence of untruncated pJB1742 (wt/WT) that was used to generate truncated DnaG genes fused to EGFP. f0–2 indicate the reading frames present in pJB1742; red, stop codons in each reading frame which must be removed for truncated proteins to co-express with EGFP; bold, *Bam*HI site which follows the *A. baylyi* *dnaG* sequence, adding the residues Gly-Ser to the protein; blue, *Bbv*CI site used to generate the 3' nick. Shown in the **body** of the Table are DNA and protein sequences of truncated mutants. Scores for green fluorescence brightness are on a scale from 0–4 where 0 indicates that green fluorescence could not be visually detected and 4, green fluorescence was similar to colonies expressing His<sub>6</sub>-EGFP with no N-terminal fusion. Red, indicates regions of the flexible linker that have been deleted. One *dnaG* truncated mutant was out of frame with *egfp*, one mutant did not produce readable sequencing and four contained no *dnaG* gene (not shown).

<i>wt</i>	TTAAGATTATTATCT <b>GGATCC</b> TAAGTA <b>ACTAACC</b> TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	WT
f0	L R L L S G S *	
f1	+1 * S A G S S G N S H H H H H	
f2	+2 *	
DNA sequence		Plasmid
Score	Protein sequence	Mutation
(4)	CCTCAGCATACCATTGATCAAATTCTAGATCGGACT P Q H T I D Q I L D R T	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–15
(4)	CAGCATACCATTGATCAAATTCTAGATCGGACTGAT Q H T I D Q I L D R T D	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–16
(4)	ATTGATCAAATTCTAGATCGGACTGATCTGGTTGAG I D Q I L D R T D L V E	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–19
(3)	CAAATTCTAGATCGGACTGATCTGGTTGAGTTAATA Q I L D R T D L V E L I	GCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–21
(3)	CAAATTCTAGATCGGACTGATCTGGTTGAGTTAATA Q I L D R T D L V E L I	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–21
(2)	ATTCTAGATCGGACTGATCTGGTTGAGTTAATAGGC I L D R T D L V E L I G	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC 1–22

Continued on next page...

**Table 4.2** – continued from previous page

<i>wt</i>	TTAAGATTATTATCTGGATCCTAAGTAACTAACC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	WT
<i>f1</i>	+1 *	S A G S S G N S H H H H H	
DNA sequence			Plasmid
Score	Protein sequence	Mutation	
	CGGACTGATCTGGTTGAGTTAATAGGCCAGAGAGTC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1791
(1)	R T D L V E L I G Q R V	S A G S S G N S H H H H H	1–25
	CGGACTGATCTGGTTGAGTTAATAGGCCAGAGAGTC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1754
(1)	R T D L V E L I G Q R V	S A G S S G N S H H H H H	1–25
	ATAGGCCAGAGAGTCAAACATAAAAAAACAGGTAGA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1778
(0)	I G Q R V K L K K T G R	S A G S S G N S H H H H H	1–32
	ATAGGCCAGAGAGTCAAACATAAAAAAACAGGTAGA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1783
(0)	I G Q R V K L K K T G R	S A G S S G N S H H H H H	1–32
	TGTCCATTTTCATCAAGAAAAAAGCCCTTCTTTCCAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1786
(1)	C P F H Q E K S P S F H	S A G S S G N S H H H H H	1–49
	TGTCCATTTTCATCAAGAAAAAAGCCCTTCTTTCCAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1755
(1)	C P F H Q E K S P S F H	S A G S S G N S H H H H H	1–49
	TGTCCATTTTCATCAAGAAAAAAGCCCTTCTTTCCAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1756
(1)	C P F H Q E K S P S F H	S A G S S G N S H H H H H	1–49
	CCATTTTCATCAAGAAAAAAGCCCTTCTTTCCATGTA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1777
(1)	P F H Q E K S P S F H V	S A G S S G N S H H H H H	1–50
	CCATTTTCATCAAGAAAAAAGCCCTTCTTTCCATGTA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1787
(1)	P F H Q E K S P S F H V	S A G S S G N S H H H H H	1–50
	AAAAGCCCTTCTTTCCATGTATATCGAGACAAGGGC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1757
(2)	K S P S F H V Y R D K G	S A G S S G N S H H H H H	1–55
	AGCCCTTCTTTCCATGTATATCGAGACAAGGGCTAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1770
(2)	S P S F H V Y R D K G Y	S A G S S G N S H H H H H	1–56

Continued on next page...

**Table 4.2** – continued from previous page

<i>wt</i>	TTAAGATTATTATCTGGATCCTAAGTAACTAACC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	WT
<i>f1</i>	+1 *	S A G S S G N S H H H H H	
DNA sequence			Plasmid
Score	Protein sequence	Mutation	
	CATGTATATCGAGACAAGGGCTATTACCATTGCTTT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1758
(1)	H V Y R D K G Y Y H C F	S A G S S G N S H H H H H	1–60
	CATGTATATCGAGACAAGGGCTATTACCATTGCTTT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1759
(1)	H V Y R D K G Y Y H C F	S A G S S G N S H H H H H	1–60
	TATCGAGACAAGGGCTATTACCATTGCTTTGGCTGT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1788
(0)	Y R D K G Y Y H C F G C	S A G S S G N S H H H H H	1–62
	TATCGAGACAAGGGCTATTACCATTGCTTTGGCTGT	CAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1789
(0)	Y R D K G Y Y H C F G C	P A G S S G N S H H H H H	1–62
	ATTGATGGGCGTAACTTTATTGATGTCATGCAGGAA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1765
(0)	I D G R N F I D V M Q E	S A G S S G N S H H H H H	1–86
	GCACAGTTTGATCAAAGCTATATGGGTTTTGAAGAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1772
(3)	A Q F D Q S Y M G F E D	S A G S S G N S H H H H H	1–165
	GCACAGTTTGATCAAAGCTATATGGGTTTTGAAGAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1790
(3)	A Q F D Q S Y M G F E D	S A G S S G N S H H H H H	1–165
	GCACAGTTTGATCAAAGCTATATGGGTTTTGAAGAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1760
(3)	A Q F D Q S Y M G F E D	S A G S S G N S H H H H H	1–165
	GCACAGTTTGATCAAAGCTATATGGGTTTTGAAGAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1761
(3)	A Q F D Q S Y M G F E D	S A G S S G N S H H H H H	1–165
	AGCTATATGGGTTTTGAAGATGCCCCTCAAGAAGGC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1776
(3)	S Y M G F E D A P Q E G	S A G S S G N S H H H H H	1–170
	AGCTATATGGGTTTTGAAGATGCCCCTCAAGAAGGC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1781
(3)	S Y M G F E D A P Q E G	S A G S S G N S H H H H H	1–170

Continued on next page...

**Table 4.2** – continued from previous page

<i>wt</i>	TTAAGATTATTATCTGGATCCTAAGTAACTAACC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	WT
<i>f1</i>	+1 *	S A G S S G N S H H H H H	
DNA sequence			Plasmid
Score	Protein sequence		Mutation
	TATATGGGTTTTGAAGATGCCCCTCAAGAAGGCAAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1769
(3)	Y M G F E D A P Q E G N	S A G S S G N S H H H H H	1-171
	TATATGGGTTTTGAAGATGCCCCTCAAGAAGGCAAT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1762
(3)	Y M G F E D A P Q E G N	S A G S S G N S H H H H H	1-171
	AATCTTTATGACCTATTGGAAAATGTAGCTCAGTTT	-----GGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1771
(2)	N L Y D L L E N V A Q F	- - G S S G N S H H H H H	1-182
	AATCTTTATGACCTATTGGAAAATGTAGCTCAGTTT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1763
(2)	N L Y D L L E N V A Q F	S A G S S G N S H H H H H	1-182
	AAACAGCTGCCAAATAGTAATAAGGCACAGCAATAT	---GCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1773
(2)	K Q L P N S N K A Q Q Y	- A G S S G N S H H H H H	1-195
	GAAAAACAGCTGCCAAATAGTAATAAGGCACAGCAA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1764
(2)	E K Q L P N S N K A Q Q	S A G S S G N S H H H H H	1-195
	TGGCGTTTGGGTTATGCACCCGAAGACTGGCAGCAC	CAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1780
(1)	W R L G Y A P E D W Q H	P A G S S G N S H H H H H	1-221
	TCAAGTGATAGTGGACGTGACTTTGACCTGCTACGT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1766
(2)	S S D S G R D F D L L R	S A G S S G N S H H H H H	1-252
	TCAAGTGATAGTGGACGTGACTTTGACCTGCTACGT	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1785
(2)	S S D S G R D F D L L R	S A G S S G N S H H H H H	1-252
	ACATTATTTAAACAGAACTCTAGAATCACGATTGCC	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1784
(0)	T L F K Q N S R I T I A	S A G S S G N S H H H H H	1-357
	CAGCTCAATGAATTAAGCAAACAGATCAACTTAAGA	TCAGCGGGCTCCTCTGGGAATTCTCATCACCATCAC	pJB1779
(1)	Q L N E L S K Q I N L R	S A G S S G N S H H H H H	1-626

be conserved, suggesting that the RPD begins around residue 169, which is 12 residues preceding the Pfam assignment for homology to the RPD. Putative soluble N-terminal fragments 1–165, 1–170 and 1–175 had a strong green fluorescent phenotype and deletion end-points were clustered in this region, while slightly longer mutants containing short portions of the RPD were also identified as putatively soluble (see Figure 4.21A in Discussion).

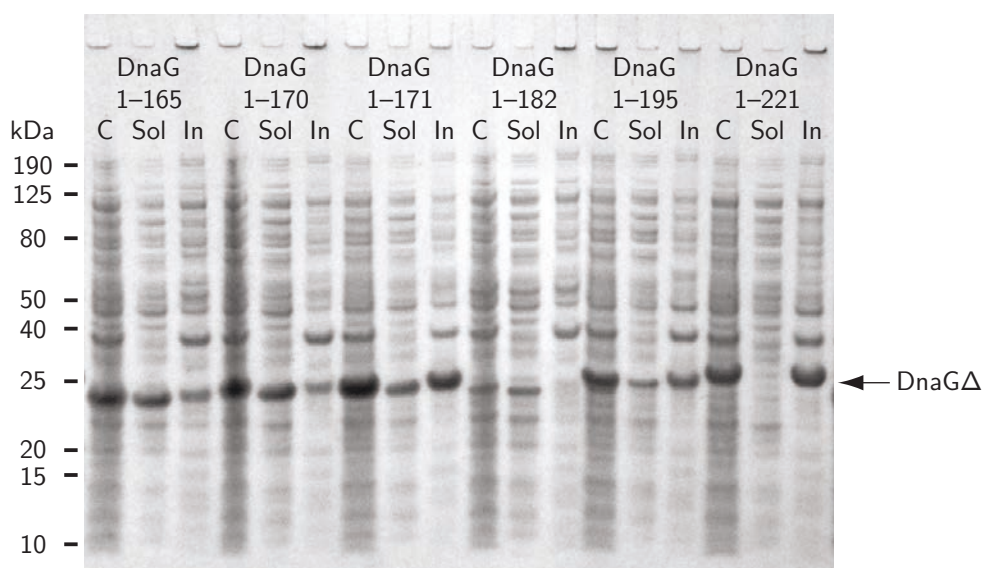
#### 4.4.2 Examination of soluble N-terminal DNA primase mutants

To investigate if the truncated proteins identified by the EGFP solubility selection screen were in fact soluble, those proteins which might represent a useful ZBD protein were modified to no longer express an EGFP C-terminal fusion, but to still possess a C-terminal His<sub>6</sub>-tag (pJB1793–pJB1790; Table 4.3; Section 4.3.4). The genes were over-expressed by auto-induction in BL21(λDE3)*recA*/pLysS. Since no proteins corresponding to just the predicted ZBD (~ 100 residues; Figures 4.4, 4.5 and 4.21A) were obtained by screening for a green fluorescent phenotype, mutants were selected which might represent the smallest soluble construct containing the ZBD and the following insertion sequence.

The T7 lysozyme containing, mutant DnaG over-expressed cells were collected by centrifugation and resuspended to equal concentration of cells, lysed by repeated freeze-thaw cycles and then centrifuged to separate soluble and insoluble components, which were analysed by SDS-PAGE (Figure 4.10). Each plasmid directed over-expression of a protein of the expected size and the amount of

**Table 4.3: Plasmids for expression of putatively soluble DNA primase N-terminal fragments.** Selected *A. baylyi* DnaG ZBD mutant plasmids were modified to direct expression of C-terminal His<sub>6</sub>-tagged protein without EGFP fusion.

Deletion mutant	Molecular weight (Da)	Plasmid
DnaG <sup>1-165</sup>	20,539.8	pJB1793
DnaG <sup>1-170</sup>	21,022.3	pJB1794
DnaG <sup>1-171</sup>	21,136.4	pJB1795
DnaG <sup>1-182</sup>	22,442.9	pJB1796
DnaG <sup>1-195</sup>	23,972.5	pJB1797
DnaG <sup>1-221</sup>	27,192.2	pJB1798



**Figure 4.10: Soluble over-expression of DNA primase C-terminal deletion mutants.** N-terminal fragments of *A. baylyi* DnaG were over-expressed in BL21(λDE3)*recA* by auto-induction at 30°C. Following cell lysis, the cellular, C; soluble, Sol and insoluble, In fractions were each analysed by SDS-PAGE. HyperPage molecular size markers are indicated. The arrow indicates the over-expressed protein.

protein expression varied significantly among the mutants. In general, the yield of soluble over-expressed truncated DnaG corresponded to the green fluorescence intensity of cells expressing these proteins as fusions to EGFP; where a brighter green fluorescent colony coincided with higher soluble protein yield. Of the

mutants examined, DnaG<sup>1-165</sup>, DnaG<sup>1-170</sup> and DnaG<sup>1-171</sup> produced a substantial proportion of soluble over-expressed protein and showed bright green fluorescence when expressed as a fusion to EGFP. Of these three proteins, DnaG<sup>1-170</sup> showed a lower level of over-expression, yet produced similar yields of soluble protein.

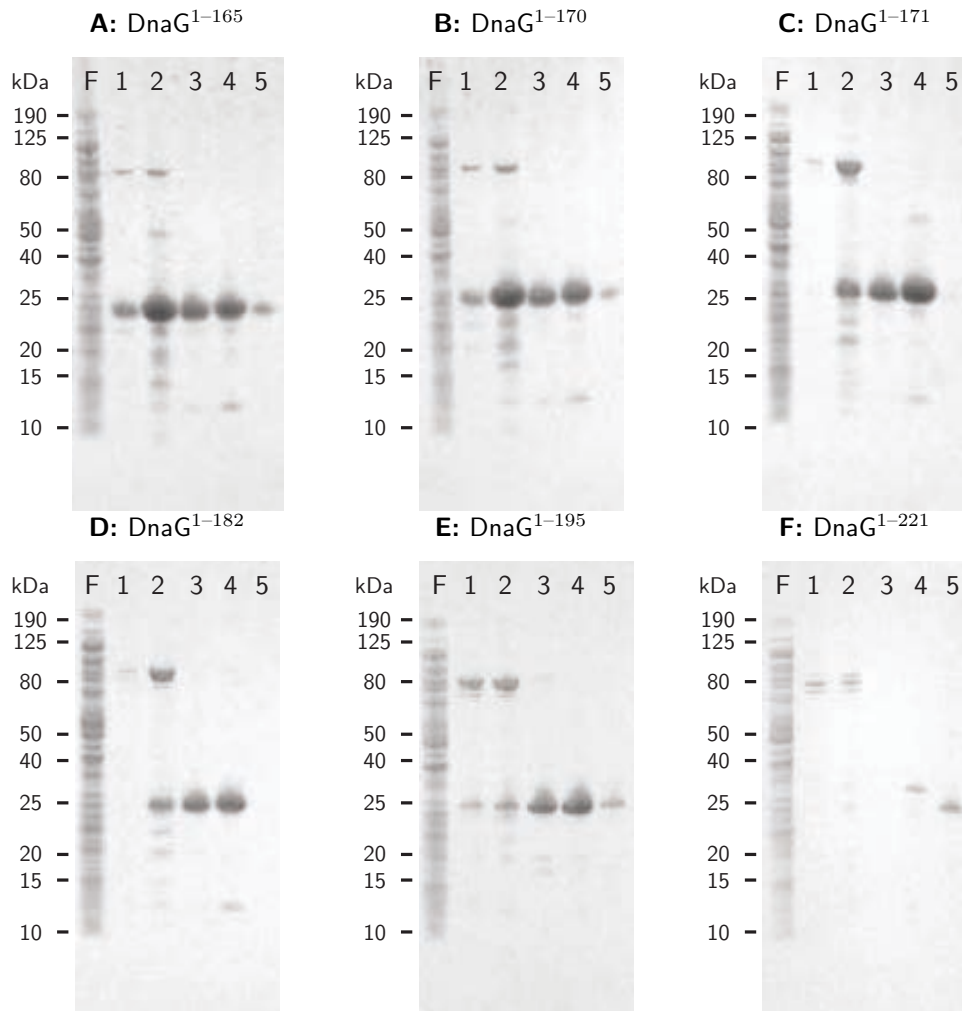
DnaG<sup>1-182</sup> and DnaG<sup>1-195</sup> produced less soluble protein than the preceding mutants and had correspondingly less bright green fluorescence when fused to EGFP. DnaG<sup>1-182</sup> over-expressed poorly but appeared to be very soluble at this level of protein expression, whereas, DnaG<sup>1-195</sup> over-expressed well but was poorly soluble. DnaG<sup>1-221</sup> produced no detectable soluble protein when over-expressed; this mutant had a very weakly fluorescent EGFP fusion. Most of the soluble proteins were also evident in the insoluble cellular fraction, but this might be expected when a protein is highly over-expressed, or cell lysis may not have been complete.

#### 4.4.3 Purification of soluble N-terminal fragments of DNA primase

To further confirm the solubility and foldedness of the truncation mutants identified, each was purified by one step IMAC (Figure 4.11). Each of DnaG<sup>1-165</sup>, DnaG<sup>1-170</sup>, DnaG<sup>1-171</sup>, DnaG<sup>1-182</sup> and DnaG<sup>1-195</sup>, which appear in the soluble cellular fraction were purified by IMAC and remained in solution. On the other hand, DnaG<sup>1-221</sup>, which does not appear in the soluble cell fraction was not enriched using IMAC. These results highlight that truncated DnaG mutants identified by expression of the corresponding EGFP fusion protein are actually



soluble.



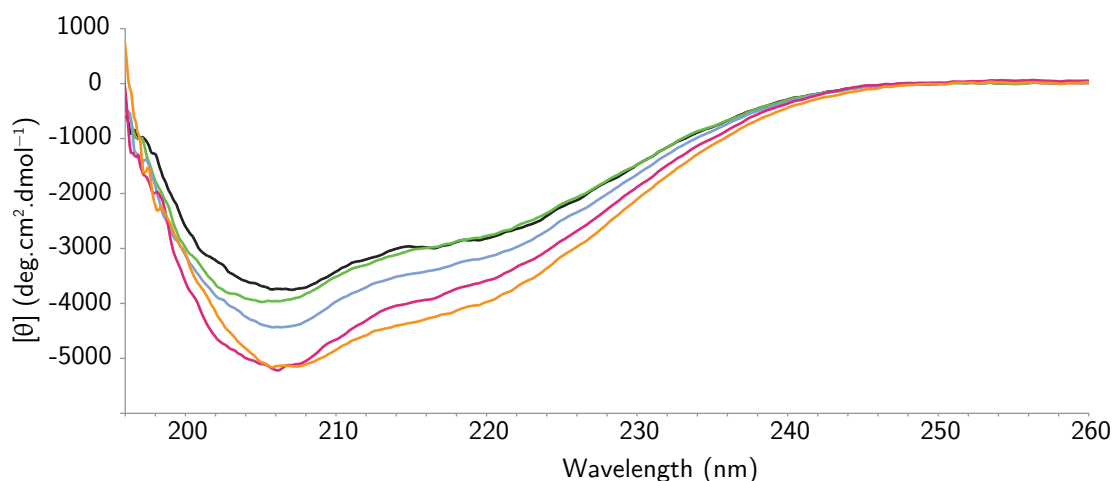
**Figure 4.11: Purification of DNA primase C-terminally truncated proteins.** C-terminal-His<sub>6</sub> tagged fragments of DnaG: **A**; DnaG<sup>1-165</sup>, **B**; DnaG<sup>1-170</sup>, **C**; DnaG<sup>1-171</sup>, **D**; DnaG<sup>1-182</sup>, **E**; DnaG<sup>1-195</sup>, **F**; DnaG<sup>1-221</sup>, were over-expressed in BL21(λDE3)*recA* by auto-induction at 30°C. The soluble cellular fraction was purified by one-step IMAC and analysed by SDS-PAGE. Shown on gels are F, IMAC unbound fraction and 1-5, successively eluted IMAC fractions. Migration of HyperPage molecular size markers are indicated.

#### 4.4.4 Circular dichroism of the DNA primase zinc-binding domain

We wished to examine whether the N-terminal domain fragments of DnaG were folded. To help assess the foldedness of the ZBD-extension proteins, low resolution structural information was obtained by recording circular dichroism (CD) spectra for each soluble truncated DnaG mutant. CD is a technique to acquire information on the secondary structures present in a protein. Each secondary structural element differentially absorbs left- and right-handed circularly polarised light and gives rise to characteristic CD spectra (Kelly *et al.*, 2005). The CD spectra produced for all mutants (Figure 4.12) were comparable, consistent with them being folded and indicating that the proteins share secondary structural elements. In particular, the weak negative ellipticity at wavelengths  $< 210$  nm indicated the presence of  $\alpha$ -helix and  $\beta$ -sheet structures in the ZBDs (*cf.* Figure 4.3).

#### 4.4.5 Examination of the foldedness of C-terminally truncated DNA primase by nuclear magnetic resonance spectroscopy

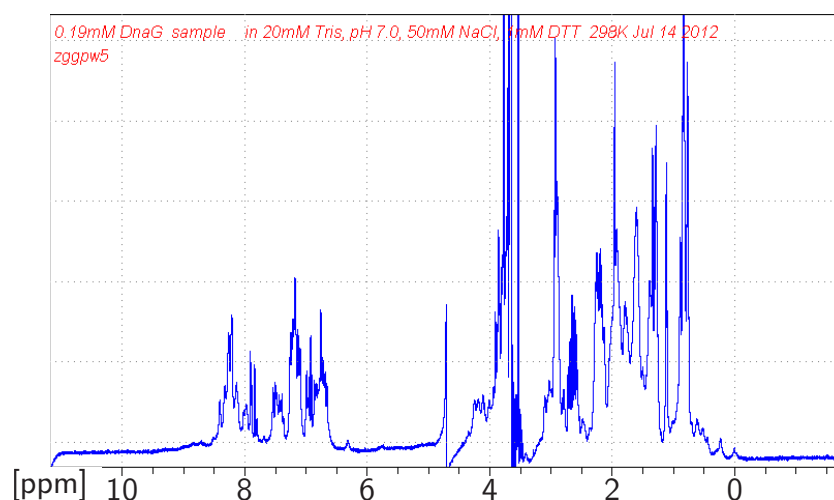
Further characterisation of protein foldedness in truncated *A. baylyi* DnaG proteins was performed using NMR spectroscopy. 1D proton NMR can infer if a protein is in a folded state. In an unfolded protein, methyl protons are solvent exposed and all will produce similar NMR signals. However, when a protein is folded, some of these methyl groups are sequestered in the hydrophobic protein core, where they will produce unique (sharp) NMR signals related to their unique



**Figure 4.12: Circular dichroism of DNA primase zinc-binding domain-extension proteins.** Circular dichroism spectra were recorded for purified His<sub>6</sub>-tagged *A. baylyi* ZBD-extension proteins in 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA. Circular dichroism spectra were converted to molar ellipticity ([θ]). CD spectra: (■), DnaG<sup>1-165</sup>; (■), DnaG<sup>1-170</sup>; (■), DnaG<sup>1-171</sup>; (■), DnaG<sup>1-182</sup>; (■), DnaG<sup>1-195</sup>.

chemical environment (−0.5 to 1.5 ppm; McDonald and Phillips, 1967; Christendat *et al.*, 2000). Further information on protein conformation can be inferred from the presence of Cα (5–6 ppm) proton chemical shifts which are indicative of β-sheet hydrogen bonds (McDonald and Phillips, 1967; Christendat *et al.*, 2000; Rehm *et al.*, 2002). The 1D NMR spectrum of non-His<sub>6</sub>-tagged DnaG<sup>1-165</sup> (Figure 4.13) shows well dispersed signals in the methyl and amide regions, but not Cα, consistent with the protein possessing a folded and predominantly α-helical structure, as predicted from the known ZBD structures (Figure 4.3). In addition to indicating that DnaG<sup>1-165</sup> contains a folded α-helical region, the 1D NMR spectrum showed sharp amide resonances (7–9 ppm) suggestive of a folded structure.

2D T<sub>O</sub>Tal C<sub>O</sub>rr<sub>O</sub>lation S<sub>P</sub>ectroscop<sub>Y</sub> (TOCSY) NMR introduces a secondary spectral axis, whereby under the right conditions, the chemical shifts for each



**Figure 4.13: One dimensional nuclear magnetic resonance spectrum of DnaG<sup>1–165</sup>.** To see if purified DnaG<sup>1–165</sup> was folded, 1D NMR was recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT by the research group of Dr Xun-Cheng Su at the State Key Laboratory of Elemento-organic Chemistry, Nankai University, Peoples Republic of China.

proton in a spin-system can be observed within distinct amino acids. These spin-system chemical shifts can be seen perpendicular to each other in both spectral axes and can be used to identify responsible residues. As proteins increase in size, the number of proton resonances increases, resulting in chemical shift overlap in NMR spectra. Further, as the protein molecules tumble more slowly as the proteins increase in size, chemical shift signals broaden (Jardetzky and Roberts, 1981). The result of these two effects is that, in 2D experiments, chemical shifts from protons within large molecules broaden to the point that they are indistinguishable from the background (Lecroisey *et al.*, 1997). However, flexible regions within a protein move more quickly, sharpening their resonances to detectable levels. Thus, an appropriate 2D NMR approach can produce a “flexible” residue spectrum, whereby chemical shifts are visible only for mobile residues in larger molecules. In addition to defining flexible residues, this 2D

NMR experiment can offer insight into the conformation of these flexible, NMR-observable residues. When in a disordered conformation, amino acids show resonances at characteristic random-coil values (Wishart *et al.*, 1995; Wishart, 2011) and the contribution of random-coil chemical shifts in “flexible” NMR spectra then indicate if these regions of a protein are disordered or not.

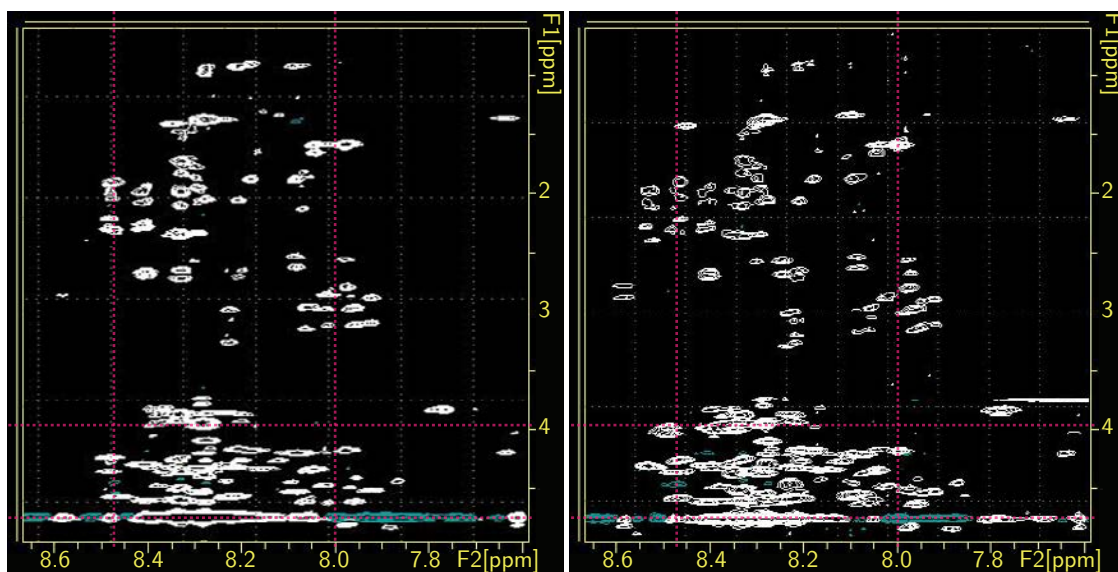
2D TOCSY NMR spectra were recorded for His<sub>6</sub>-tagged DnaG<sup>1-165</sup> (pH 7.5) and DnaG<sup>1-170</sup> (pH 7.0; Figure 4.14) and these “flexible” NMR spectra show a significant number of residues with random-coil chemical shifts. Comparison of spectra for the two similar proteins indicates the appearance of an additional glutamic acid random-coil chemical shift in the spectrum of DnaG<sup>1-170</sup>. The additional C-terminal residues of DnaG<sup>1-170</sup>-His<sub>6</sub>, compared to DnaG<sup>1-165</sup>-His<sub>6</sub>, contain this additional Glu, strongly suggesting that the flexible residues are present at the C-terminus of these proteins (see Figure 4.5).

#### 4.4.6 N-terminal fragments of DNA primase bind zinc

One of the defining features of ZBDs in primase is that they bind tightly to zinc. To test the ability of the DnaG N-terminal proteins to bind zinc, *A. baylyi* ZBD-extension mutants were analysed by ESI-mass spectrometry under both denaturing and native conditions. Proteins maintained in acidic buffers, which promote protein unfolding (denaturation), lose non-covalently bound species as the folded regions to which they associate are no longer present. However, proteins maintained in ammonium acetate buffers can often retain their natively folded

**A:** *A. baylyi* DnaG: DnaG<sup>1-165</sup>-His<sub>6</sub> TOCSY NMR spectra; pH 7.5.

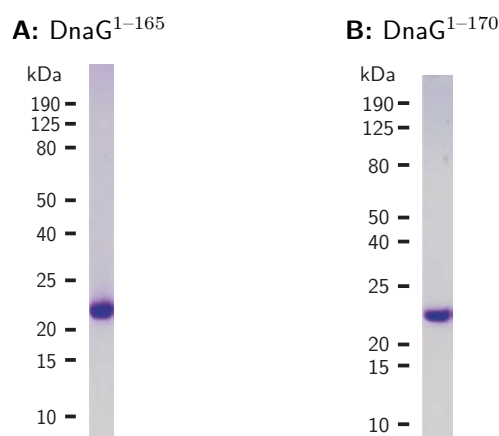
**B:** *A. baylyi* DnaG: DnaG<sup>1-170</sup>-His<sub>6</sub> TOCSY NMR spectra; pH7.0.



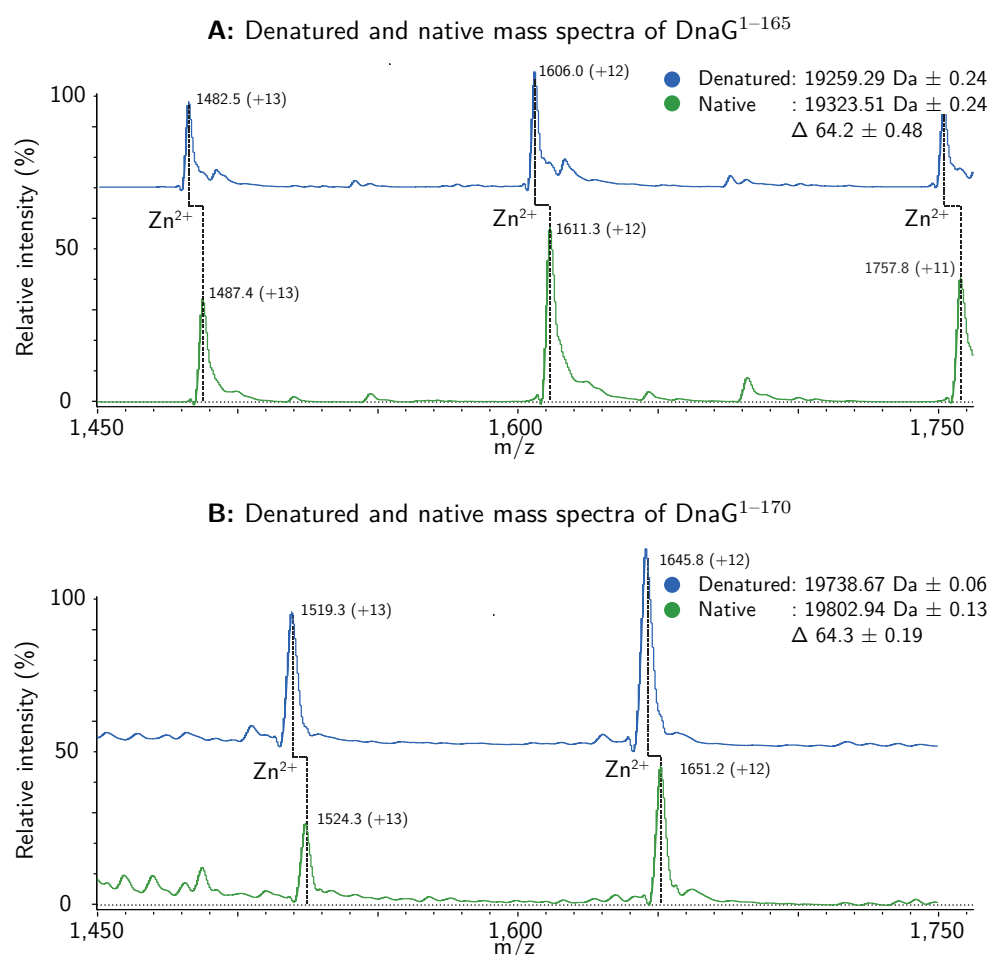
**Figure 4.14: Two dimensional TOCSY NMR spectra of DnaG<sup>1-165</sup>-His<sub>6</sub> and DnaG<sup>1-170</sup>-His<sub>6</sub>.** To see if purified ZBD-extension proteins were folded, TOCSY NMR spectra were recorded by Dr Kiyoshi Ozawa at the Australian National University. 2D TOCSY NMR spectra (80 ms mixing time) were recorded for: **A**, DnaG<sup>1-165</sup>-His<sub>6</sub> (pH 7.0; 50 mM Tris, 150 mM NaCl, 1 mM DTT, 1 mM EDTA) resulting in some precipitation of the protein over the course of the experiment; and **B**, DnaG<sup>1-170</sup>-His<sub>6</sub> (pH 7.5; 50 mM Tris, 150 mM NaCl, 1 mM DTT). Red dashed lines indicate the range of random-coil values for: vertical, amide proton and horizontal, C $\alpha$  proton chemical shifts (Wishart *et al.*, 1995; Wishart, 2011)

conformations (Felitsyn *et al.*, 2002; Hernández and Robinson, 2007), and so can be observed in mass spectrometry with ligands still intact.

To determine if *A. baylyi* ZBD-extension mutants contain a folded ZBD, non-His<sub>6</sub>-tagged DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> were purified (Sections 4.3.5 and 4.3.7.1; Figure 4.15) and the proteins were prepared for mass spectrometry. Contaminating metals were removed from the samples by dialysis against 1 mM EDTA before analysis. The electrospray mass spectra of denatured DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> in 0.1% formic acid showed that the proteins were complete, each



**Figure 4.15: Purified untagged DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup>.** Two ZBD-extension proteins were expressed without His<sub>6</sub> or EGFP fusions (Sections 4.3.5 and 4.3.7.1). Purified: **A**, DnaG<sup>1-165</sup> and **B**, DnaG<sup>1-170</sup> were analysed by SDS-PAGE and stained with Coomassie blue.



**Figure 4.16: Positive ion electrospray mass spectrum of denatured and native DNA primase zinc-binding domain-extension mutants.** Denatured mass analysis (blue) was performed in 0.1% formic acid and native mass analysis (green) was performed in 100 mM ammonium acetate pH 7.2 for: **A**, DnaG<sup>1-165</sup> and **B**, DnaG<sup>1-170</sup>. The mass obtained is consistent with the predicted mass the respective proteins. Native protein samples contain additional mass (64.2 and 64.3 Da) consistent with a single bound zinc ion.

giving a mass in close agreement with the predicted mass without N-terminal methionine (DnaG<sup>1-165</sup>, 19,257.4 Da and DnaG<sup>1-170</sup>, 19,739.9 Da; Figure 4.16). However, under native mass spectrometry conditions both DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> displayed a mass higher by 64 Da than under denaturing conditions, consistent with a single bound zinc (atomic mass 65.4 Da), supporting that *A. baylyi* ZBD-extension mutants contain a folded ZBD.

#### 4.4.7 Crystallisation of DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup>

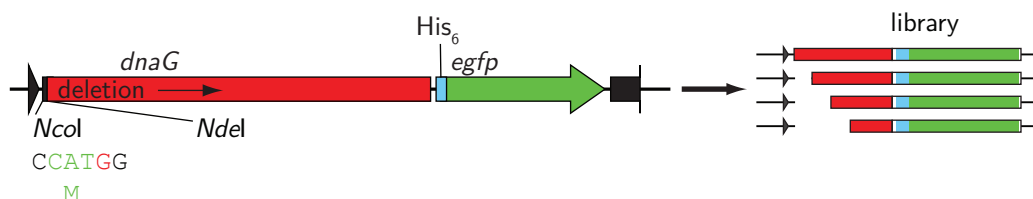
To attempt to determine the three-dimensional structures of the *A. baylyi* ZBD-extension proteins, numerous attempts to crystallise both DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> were made, but were unsuccessful, even using highly concentrated protein samples (up to 10 mM in the case of DnaG<sup>1-165</sup>). Many protein properties can contribute to failed protein crystallisation. In the case of the *A. baylyi* ZBD-extension proteins, the C-terminal extensions may be flexible. The *Moraxellaceae* sequence insertion shows poor conservation (Figure 4.5) suggesting that it may be unstructured. This is supported by NMR data that show significant portions of the proteins are in a random-coil conformation (Figure 4.14).

#### 4.4.8 N-terminal truncation of *Acinetobacter baylyi* DNA primase

To search for soluble N-terminal truncations complementing those identified in Section 4.4.1, an N-deleted gene truncation library was produced using pJB1743 as described in Section 4.3.2. *NcoI* linearised pJB1743 was end filled using



dGTP $\alpha$ S, dATP, dCTP, dTTP and the Klenow fragment of DNA polymerase I. Digestion of  $\alpha$ -S protected pJB1742 with *Nde*I removed the dGTP $\alpha$ S end adjacent to the start codon of *dnaG* and allowed uni-directional deletion using *Exo*III (Figure 4.17). The  $\alpha$ -S protected *Nco*I site initiates translation in pJB1743 and following ligation of limit *Exo*III truncations directs over-expression of N-terminally truncated DnaG-EGFP fusion proteins when a mutant is in the appropriate reading frame. Truncating a protein composed of globular domains from both the N- and C-termini would be expected to produce similar protein break-points as we expect these to occur at or near domain boundaries.



**Figure 4.17: Methodology for N-terminal truncation of DNA primase.** *Nco*I linearised pJB1743 was protected with dGTP $\alpha$ S (red) and dATP, dCTP and dTTP (green) and then digested with *Nde*I to allow uni-directional deletion through the N-terminus of *A. baylyi dnaG*. The filled *Nco*I site provides the start codon for truncated genes. Following limit digests of pJB1743 using *Exo*III and plasmid repair, one-in-three truncated genes are in-frame with the start codon and so direct over-expression of N-terminally deleted DnaG proteins fused with the downstream EGFP when expressed in *E. coli*.

Deletion reactions were carried out at 30°C for between 1 min 25 s and 2 min 40 s. These conditions should have resulted in a library of genes with the N-terminal domain removed and sampled through the ZBD-extension (the library was focussed to provide deletion end-points approximately between nucleotides 300 and 600 of *dnaG*; see Appendix B.2). Truncated library samples were pooled, ligated and transformed into BL21( $\lambda$ DE3)*recA* as described in Section 4.3.2. Library transformed BL21( $\lambda$ DE3)*recA* cells were spread onto 12 Selection

agar plates containing ampicillin. After overnight incubation at 30°C a total of approximately 4,000 colonies developed, 63 of which showed green fluorescence. After the agar plates had been stored at 4°C for 24 h, a further 39 green fluorescent colonies became apparent.

Twenty randomly selected non-fluorescent colonies were chosen and the mutant plasmids purified and sequenced (Table 4.4). The genes present for N-terminally deleted DnaG were truncated by between 246 and 720 nucleotides, represented all three reading frames and did not appear to have truncated at the  $\alpha$ -S protected end. The N-terminal deletion library likely provided sufficient coverage of the region of interest. The size of the genes present in the 102 green fluorescent N-terminally deleted DnaG mutants was determined by colony PCR and mostly revealed genes ranging between 350 and 1,900 bp (data not shown; full-length *dnaG* is 1,890 bp). Twenty mutants appeared to contain small < 200 bp genes and these were not investigated further.

The identities of the remaining 82 putative soluble EGFP fusion proteins were then determined by DNA sequencing and the mutants were scored based on the brightness of their green fluorescent phenotype (Table 4.5). Firstly, multiple genes for proteins starting at residues 168–170 of DnaG had a bright green fluorescent phenotype, suggesting that they represent soluble proteins. These DnaG mutants have terminal residues very close to those identified by C-terminal deletion of DnaG and appear to be at the start of homology for the RPD (Figure 4.5). However, in contrast to the C-terminal deletion library, N-terminal deletions were putatively soluble when truncation occurred at the end of the ZBD and

throughout the *Moraxellaceae* sequence insertion/domain linker residues 101 through 168 (of *A. baylyi* DnaG; Figure 4.21B in Discussion).

**Table 4.4: Randomly sequenced N-terminally truncated DNA primase plasmids.** The fusion sequence, which provides a start codon containing  $\alpha$ S protected dGMP (bold) is fused to truncated *dnaG* genes (shown in the correct reading frame for *dnaG*).

Fusion sequence	Out of frame residues	End of truncated <i>dnaG</i> mutant	Protein length (amino acids)
<b>CCATG</b>	C	GCTCAAGTGATAGTG	389.7
<b>CCATG</b>	AG	TGCAGAAACTATTCA	426.3
<b>CCATG</b>	GC	ACAGCAATATTTTAA	436.3
<b>CCATG</b>	GT	AGCTCAGTTTTATGA	450.3
<b>CCATG</b>	G	ACCTATTGGAAAATG	455.7
<b>Deleted</b>		ATGGGTTTTGAAGAT	469
<b>CCATG</b>		TATATGGGTTTTGAA	470
<b>CCATG</b>		GCACAGTTTGATCAA	476
<b>CCATG</b>	GA	CGATCCATTCGCACA	479.3
<b>CCATG</b>		GCACCCTCTTACTTT	485
<b>CCATG</b>	GG	TTTACAAGCACCCCTC	487.3
<b>CCATG</b>		GAACAGATTGACGGT	492
<b>CCATG</b>	CA	TGCTGCAATATCCGA	499.3
<b>CCATG</b>	CA	CCATGCTGCAATATC	500.3
<b>CCATG</b>	A	CTGAAAAATCTGCAC	507.7
<b>CCATG</b>		ACTGAAAAATCTGCA	508
<b>CCATG</b>		GAACAAAAGAAACTT	529
<b>CCATG</b>	AA	AGATAACTTTGAACA	532.3
<b>CCATG</b>	TC	GGGCGTTGAACTTCC	538.3
<b>CCATG</b>	G	ATGTCATGCAGGAAC	547.7

As previously observed, the recovered colonies had varied green fluorescence intensity for the genes of various sizes. Several mutants with very poor green fluorescence brightness were produced that had truncations within the ZBD of *A. baylyi* DnaG, where production of only part of the ZBD folding unit should

**Table 4.5: Sequenced N-terminally deleted *A. baylyi* DnaG.** Plasmids from green fluorescent colonies expressing truncated DnaG-EGFP fusion proteins were sequenced to identify the mutations present. Displayed at the top are the DNA and protein sequences of untruncated pJB1743 (*wt*/WT) that was used to generate truncated *dnaG* genes fused to *egfp*; shown in red are the nucleotides removed during linearisation for uni-directional truncation. f0–1 indicate the reading frames present in pJB1743; f0; reading frame of initiation codon, f1; reading frame of non-truncated *dnaG* in pJB1743. Scores for green fluorescence brightness are on a scale from 0–4 where 0 indicates that green fluorescence could not be visually detected and 4, green fluorescence was similar to colonies expression His<sub>6</sub>-EGFP with no N-terminal fusion. One mutant did not provide useful DNA sequencing data (not shown).

wt	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTCCTCAGCATACCATTGATCAA	WT
f0	M A T S *		
f1		M A I P Q H T I D Q	
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(2)	CCAT <b>G</b>	ATGGCTATTCCTCAGCATACCATTGATCAAATTCTA	pJB1866
	M	M A I P Q H T I D Q I L	1-629
(2)	CCAT <b>G</b>	ATGGCTATTCCTCAGCATACCATTGATCAAATTCTA	pJB1869
	M	M A I P Q H T I D Q I L	1-629
(1)	CCAT <b>G</b>	CCTCAGCATACCATTGATCAAATTCTAGATCGGACT	pJB1822
	M	P Q H T I D Q I L D R T	4-629
(0)	CCAT <b>G</b>	GGTAGAACCTATTCGGGTTGCTGTCCATTTTCATCAA	pJB1810
	M	G R T Y S G C C P F H Q	31-629
(1)	CCAT <b>G</b>	CGAGACAAGGGCTATTACCATTGCTTTGGCTGTCAG	pJB1858
	M	R D K G Y Y H C F G C Q	52-629
(3)	CCAT <b>G</b>	GATATTGATGGGCGTAACTTTATTGATGTCATGCAG	pJB1823
	M	D I D G R N F I D V M Q	74-629
(2)	CCAT <b>G</b>	GGGCGTAACTTTATTGATGTCATGCAGGAAGTGTCT	pJB1845
	M	G R N F I D V M Q E L S	77-629
(2)	CCAT <b>G</b>	GATGTCATGCAGGAAGTGTCTAGTAAATCGGGCGTT	pJB1857
	M	D V M Q E L S S K S G V	82-629

Continued on next page. . .

Table 4.5 – continued from previous page

<i>wt</i>	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTCCTCAGCATACCATTGATCAA	WT
<i>f0</i>	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(2)	CCAT <b>G</b>	CAGGAACTGTCTAGTAAATCGGGCGTTGAACTTCCT	pJB1837
	M	Q E L S S K S G V E L P	85–629
(3)	CCAT <b>G</b>	CTGTCTAGTAAATCGGGCGTTGAACTTCCTAAAGAT	pJB1805
	M	L S S K S G V E L P K D	87–629
(2)	CCAT <b>G</b>	TCGGGCGTTGAACTTCCTAAAGATAAAGTTTGAACAA	pJB1841
	M	S G V E L P K D N F E Q	91–629
(2)	CCAT <b>G</b>	CTTCCTAAAGATAAAGTTTGAACAAAAGAACTTTCC	pJB1827
	M	L P K D N F E Q K K L S	95–629
(1)	CCAT <b>G</b>	GAACAAAAGAACTTTCTCTATAAGCGCAATACACAA	pJB1826
	M	E Q K K L S Y K R N T Q	101–629
(1)	CCAT <b>G</b>	CAAAAGAACTTTCTCTATAAGCGCAATACACAAAAA	pJB1876
	M	Q K K L S Y K R N T Q K	102–629
(1)	CCAT <b>G</b>	CTTTCCTATAAGCGCAATACACAAAAACCAGAACCG	pJB1816
	M	L S Y K R N T Q K P E P	105–629
(2)	CCAT <b>G</b>	TATAAGCGCAATACACAAAAACCAGAACCGAAACCT	pJB1853
	M	Y K R N T Q K P E P K P	107–629
(2)	CCAT <b>G</b>	TATAAGCGCAATACACAAAAACCAGAACCGAAACCT	pJB1854
	M	Y K R N T Q K P E P K P	107–629
(2)	CCAT <b>G</b>	AAGCGCAATACACAAAAACCAGAACCGAAACCTGTT	pJB1815
	M	K R N T Q K P E P K P V	108–629
(1)	CCAT <b>G</b>	CAAAACCAGAACCGAAACCTGTTGTAAATACTGAA	pJB1846
	M	Q K P E P K P V V N T E	112–629

Continued on next page. . .

**Table 4.5** – continued from previous page

wt	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTCTCAGCATACCATTGATCAA	WT
f0	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(3)	CCAT <b>G</b>	CCTGTTGTAAATACTGAAAAATCTGCACCATCTCAC	pJB1821
	M	P V V N T E K S A P S H	118–629
(2)	CCAT <b>G</b>	GTTGTAAATACTGAAAAATCTGCACCATCTCACCAT	pJB1803
	M	V V N T E K S A P S H H	119–629
(1)	CCAT <b>G</b>	AATACTGAAAAATCTGCACCATCTCACCATGCTGCA	pJB1880
	M	N T E K S A P S H H A A	121–629
(2)	CCAT <b>G</b>	AAATCTGCACCATCTCACCATGCTGCAATATCCGAA	pJB1850
	M	K S A P S H H A A I S E	124–629
(2)	CCAT <b>G</b>	TCTACCATGCTGCAATATCCGAATCATCGGAACAG	pJB1832
	M	S H H A A I S E S S E Q	128–629
(2)	CCAT <b>G</b>	CACCATGCTGCAATATCCGAATCATCGGAACAGATT	pJB1807
	M	H H A A I S E S S E Q I	129–629
(2)	CCAT <b>G</b>	CACCATGCTGCAATATCCGAATCATCGGAACAGATT	pJB1808
	M	H H A A I S E S S E Q I	129–629
(2)	CCAT <b>G</b>	CACCATGCTGCAATATCCGAATCATCGGAACAGATT	pJB1813
	M	H H A A I S E S S E Q I	129–629
(2)	CCAT <b>G</b>	CACCATGCTGCAATATCCGAATCATCGGAACAGATT	pJB1829
	M	H H A A I S E S S E Q I	129–629
(2)	CCAT <b>G</b>	CACCATGCTGCAATATCCGAATCATCGGAACAGATT	pJB1856
	M	H H A A I S E S S E Q I	129–629
(2)	CCAT <b>G</b>	GCTGCAATATCCGAATCATCGGAACAGATTGACGGT	pJB1831
	M	A A I S E S S E Q I D G	131–629

Continued on next page. . .

Table 4.5 – continued from previous page

wt	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTTCCTCAGCATACCATTGATCAA	WT
f0	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(2)	CCAT <b>G</b>	GCAATATCCGAATCATCGGAACAGATTGACGGTTTA	pJB1872
	M	A I S E S S E Q I D G L	132-629
(1)	CCAT <b>G</b>	TCCGAATCATCGGAACAGATTGACGGTTTACAAGCA	pJB1849
	M	S E S S E Q I D G L Q A	134-629
(2)	CCAT <b>G</b>	TCATCGGAACAGATTGACGGTTTACAAGCACCCCTCT	pJB1852
	M	S S E Q I D G L Q A P S	136-629
(2)	CCAT <b>G</b>	TCATCGGAACAGATTGACGGTTTACAAGCACCCCTCT	pJB1873
	M	S S E Q I D G L Q A P S	136-629
(3)	CCAT <b>G</b>	TCGGAACAGATTGACGGTTTACAAGCACCCCTCTTAC	pJB1817
	M	S E Q I D G L Q A P S Y	137-629
(2)	CCAT <b>G</b>	TTACAAGCACCCCTCTTACTTTGACGATCCATTTCGCA	pJB1811
	M	L Q A P S Y F D D P F A	143-629
(2)	CCAT <b>G</b>	TTACAAGCACCCCTCTTACTTTGACGATCCATTTCGCA	pJB1833
	M	L Q A P S Y F D D P F A	143-629
(2)	CCAT <b>G</b>	CCCTCTTACTTTGACGATCCATTTCGCACAGTTTGAT	pJB1818
	M	P S Y F D D P F A Q F D	146-629
(2)	CCAT <b>G</b>	CCCTCTTACTTTGACGATCCATTTCGCACAGTTTGAT	pJB1870
	M	P S Y F D D P F A Q F D	146-629
(2)	CCAT <b>G</b>	TACTTTGACGATCCATTTCGCACAGTTTGATCAAAGC	pJB1809
	M	Y F D D P F A Q F D Q S	148-629
(2)	CCAT <b>G</b>	TACTTTGACGATCCATTTCGCACAGTTTGATCAAAGC	pJB1878
	M	Y F D D P F A Q F D Q S	148-629

Continued on next page...

**Table 4.5** – continued from previous page

<i>wt</i>	CCAT <b>GCGACCTCGTGA</b> –/19 bp	ATGGCTATTCCTCAGCATACCATTGATCAA	WT
<i>f0</i>	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(2)	CCAT <b>G</b>	CAGTTTGATCAAAGCTATATGGGTTTTGAAGATGCC	pJB1839
	M	Q F D Q S Y M G F E D A	155–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1806
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1812
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1814
	M	Q E G N L Y D L L E N V	168–629
(2)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1838
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1848
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1855
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1861
	M	Q E G N L Y D L L E N V	168–629
(3)	CCAT <b>G</b>	CAAGAAGGCAATCTTTATGACCTATTGGAAAATGTA	pJB1877
	M	Q E G N L Y D L L E N V	168–629
(2)	CCAT <b>G</b>	GAAGGCAATCTTTATGACCTATTGGAAAATGTAGCT	pJB1819
	M	E G N L Y D L L E N V A	169–629
(2)	CCAT <b>G</b>	GAAGGCAATCTTTATGACCTATTGGAAAATGTAGCT	pJB1844
	M	E G N L Y D L L E N V A	169–629

Continued on next page. . .



**Table 4.5** – continued from previous page

wt	CCAT <b>GCGACCTCGTGA</b> –/19 bp	ATGGCTATTTCCTCAGCATACCATTGATCAA	WT
f0	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(3)	CCAT <b>G</b>	GGCAATCTTTATGACCTATTGGAAAATGTAGCTCAG	pJB1804
	M	G N L Y D L L E N V A Q	170–629
(3)	CCAT <b>G</b>	GGCAATCTTTATGACCTATTGGAAAATGTAGCTCAG	pJB1843
	M	G N L Y D L L E N V A Q	170–629
(3)	CCAT <b>G</b>	GGCAATCTTTATGACCTATTGGAAAATGTAGCTCAG	pJB1851
	M	G N L Y D L L E N V A Q	170–629
(3)	CCAT <b>G</b>	GGCAATCTTTATGACCTATTGGAAAATGTAGCTCAG	pJB1875
	M	G N L Y D L L E N V A Q	170–629
(1)	CCAT <b>G</b>	TTGGAAAATGTAGCTCAGTTTTATGAAAAACAGCTG	pJB1825
	M	L E N V A Q F Y E K Q L	176–629
(0)	CCAT <b>G</b>	GAAAAACAGCTGCCAAATAGTAATAAGGCACAGCAA	pJB1820
	M	E K Q L P N S N K A Q Q	184–629
(1)	CCAT <b>G</b>	AAACAGCTGCCAAATAGTAATAAGGCACAGCAATAT	pJB1842
	M	K Q L P N S N K A Q Q Y	185–629
(2)	CCAT <b>G</b>	AAGGCACAGCAATATTTTAAACAACGTGGCTTGAGT	pJB1834
	M	K A Q Q Y F K Q R G L S	192–629
(2)	CCAT <b>G</b>	AAGGCACAGCAATATTTTAAACAACGTGGCTTGAGT	pJB1836
	M	K A Q Q Y F K Q R G L S	192–629
(2)	CCAT <b>G</b>	AAACAACGTGGCTTGAGTGCAGAACTATTCAATTC	pJB1828
	M	K Q R G L S A E T I Q F	198–629
(0)	CCAT <b>G</b>	CAACGTGGCTTGAGTGCAGAACTATTCAATTCTGG	pJB1824
	M	Q R G L S A E T I Q F W	199–629

Continued on next page...

Table 4.5 – continued from previous page

<i>wt</i>	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTCCTCAGCATACCATGATCAA	WT
<i>f0</i>	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(2)	CCAT <b>G</b> ATTCAATTCTGGCGTTTGGGTTATGCACCCGAAGAC	pJB1868	
	M I Q F W R L G Y A P E D	207–629	
(2)	CCAT <b>G</b> GCACCCGAAGACTGGCAGCACCTTGAAAAAGCCTTT	pJB1859	
	M A P E D W Q H L E K A F	215–629	
(1)	CCAT <b>G</b> TTAAAGCAAGTGGGACTCATCCGCTCAAGTGATAGT	pJB1847	
	M L K Q V G L I R S S D S	233–629	
(1)	CCAT <b>G</b> AGTGATAGTGGACGTGACTTTGACCTGCTACGTGAG	pJB1864	
	M S D S G R D F D L L R E	242–629	
(2)	CCAT <b>G</b> CGTGAGCGTGTGCATCTTCCCGATTTCGTGATCATAAA	pJB1862	
	M R E R V I F P I R D H K	252–629	
(2)	CCAT <b>G</b> CGTGAGCGTGTGCATCTTCCCGATTTCGTGATCATAAA	pJB1865	
	M R E R V I F P I R D H K	252–629	
(2)	CCAT <b>G</b> CGTGAGCGTGTGCATCTTCCCGATTTCGTGATCATAAA	pJB1867	
	M R E R V I F P I R D H K	252–629	
(2)	CCAT <b>G</b> CGTGAGCGTGTGCATCTTCCCGATTTCGTGATCATAAA	pJB1871	
	M R E R V I F P I R D H K	252–629	
(3)	CCAT <b>G</b> CGTGAGCGTGTGCATCTTCCCGATTTCGTGATCATAAA	pJB1879	
	M R E R V I F P I R D H K	252–629	
(2)	CCAT <b>G</b> AAACCTCAAGGCAAAAGACTGGTTAATGGTAGAAGGC	pJB1874	
	M K L K A K D W L M V E G	305–629	
(2)	CCAT <b>G</b> GCAACCTTAGGCACAGCCAGTAATGCTGATCACTTA	pJB1863	
	M A T L G T A S N A D H L	333–629	

Continued on next page. . .

**Table 4.5** – continued from previous page

<i>wt</i>	CCAT <b>G</b> CGACCTCGTGA- /19 bp	ATGGCTATTCCTCAGCATACCATTGATCAA	WT
<i>f0</i>	M A T S *		
DNA sequence			Plasmid
Score	Protein sequence		Mutation
(0)	CCAT <b>G</b>	CAGAACTCTAGAATCACGATTGCCTTTGATGGCGAT	pJB1830
	M	Q N S R I T I A F D G D	350–629
(0)	CCAT <b>G</b>	ATTGCCTTTGATGGCGATGCAGCAGGGCAAAAAGCG	pJB1860
	M	I A F D G D A A G Q K A	356–629
(2)	CCAT <b>G</b>	GACATCTCTACGCCTGAGGGCAAAAGTCAGGTCATG	pJB1835
	M	D I S T P E G K S Q V M	431–629
(0)	CCAT <b>G</b>	CTTTATATTCAC TTCGAGGCTTTGCGGGCTTTTATT	pJB1840
	M	L Y I H F E A L R A F I	508–629

lead to protein misfolding and/or aggregation. Truncation at residue 74 (which is near the end of the ZBD) produced a mutant colony phenotype suggesting a well expressed and soluble protein. Compared to the *G. stearothermophilus* ZBD structure (Figure 4.3B), truncation at residue 74 leaves only the complete final twelve residue  $\alpha$ -helix of the ZBD; presence of the corresponding region in the *A. baylyi* DnaG truncation mutant appears to not perturb protein folding or produce significant aggregation. Mutants were also identified with truncation end-points in the RPD, where they appear to dissect this domain in loop regions, discussion of which can be found in Section 4.4.10.

The size and high quality of the library of transformants suggests that each mutant within the region where the *ExoIII* deletions were focussed (i.e. between residues 100 and 250; Appendix B.2) should be observed more than once. The mutants with the strongest green fluorescence brightness were over-represented in the mutants identified, likely due to the ease with which these mutants are observed on agar plates. On the other hand, of all the green fluorescent mutants sequenced, 39 were represented only once, and of these, most had a very poor green fluorescence phenotype and/or were outside of the focussed region, making them less likely to be visually identified. However, two mutants with strong green fluorescent phenotypes (score of  $3/4$ ; Table 4.5) and which were expected to be represented more than once in the library were recovered only once (DnaG<sup>118–629</sup> and DnaG<sup>137–629</sup>). Aside from these two, the recovery of identical putatively soluble DnaG mutants, and the gene sizes identified by sequencing randomly selected mutants, supports that the library recovered most of the possible truncated proteins detectable by this method.

#### 4.4.9 Solubility of N-terminally deleted DNA primase mutants

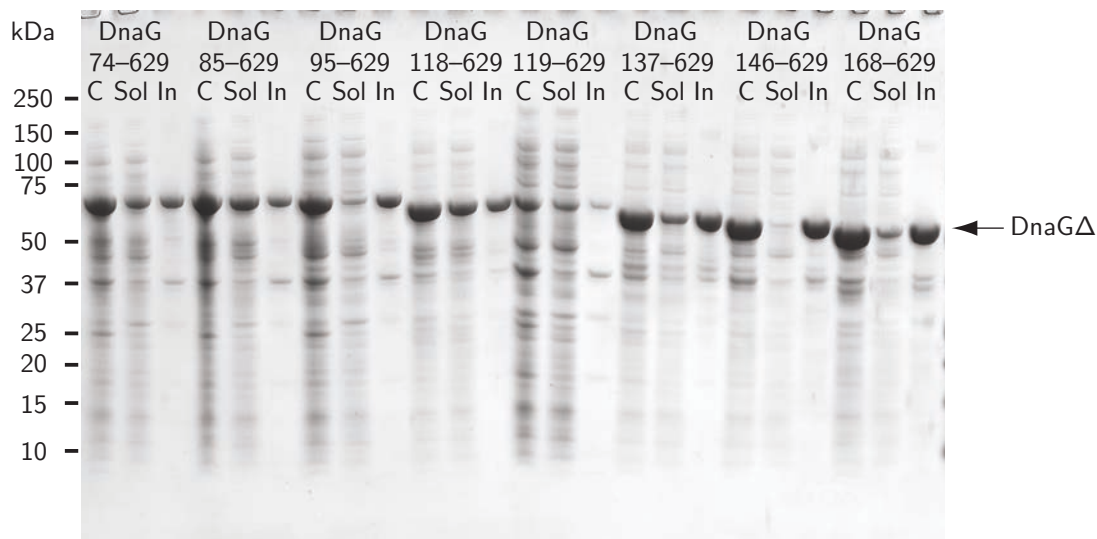
To confirm the solubility of the mutants identified by N-terminal deletion of DnaG, a range of putatively soluble mutants was selected and they were modified to no longer express an EGFP C-terminal fusion, but still possess a C-terminal His<sub>6</sub>-tag (pJB1951–1960; Table 4.6). These proteins were over-expressed by auto-induction in BL21(λDE3)*recA*.

**Table 4.6: Plasmids for expression of putative soluble DNA primase C-terminal fragments.** Plasmids for putative soluble mutants of *A. baylyi* DnaG deleted from the N-terminus were modified to express them as C-terminal His<sub>6</sub>-tagged proteins without EGFP fusion.

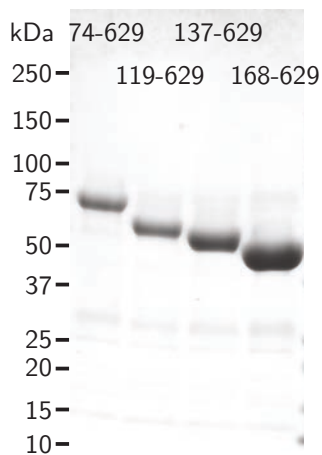
Deletion mutant	Molecular weight (Da)	Plasmid
DnaG <sup>74–629</sup>	65,370.7	pJB1951
DnaG <sup>85–629</sup>	64,094.3	pJB1953
DnaG <sup>95–629</sup>	63,049.2	pJB1955
DnaG <sup>118–629</sup>	60,250.0	pJB1956
DnaG <sup>119–629</sup>	60,152.9	pJB1957
DnaG <sup>137–629</sup>	58,306.9	pJB1958
DnaG <sup>146–629</sup>	57,364.9	pJB1959
DnaG <sup>168–629</sup>	54,808.2	pJB1960

Cells from cultures of strains predicted to express N-terminally deleted *A. baylyi* DnaG fragments were collected by centrifugation and resuspended to equal concentrations of cells, and lysed by being passed through a French press three times; the cell lysate was then centrifuged to separate the soluble and insoluble fractions, then analysed by SDS-PAGE (Figure 4.18A). Each mutant

**A: Solubility of N-terminally deleted DNA primase**



**B: Purified N-terminally deleted DNA primase**



**Figure 4.18: Over-expression and protein solubility of DNA primase N-terminally deleted mutants.** All proteins were over-expressed in BL21( $\lambda$ DE3)*recA* by auto-induction at 30°C. **A:** Following cell lysis the C, cellular; Sol, soluble and In, insoluble fractions were each analysed by SDS-PAGE. Molecular size markers are indicated. **B:** Four soluble expressed N-terminally truncated DnaG proteins were purified by one-step IMAC and pooled, eluted protein analysed by SDS-PAGE.

plasmid directed over-expression of a protein of the expected size, but with varied levels of protein over-expression. Of the mutants examined, strains containing DnaG<sup>74-629</sup>, DnaG<sup>85-629</sup>, DnaG<sup>119-629</sup> and DnaG<sup>118-629</sup> showed a high proportion of soluble over-expressed protein and produced a strong green fluorescence phenotype when expressed as EGFP fusion proteins in *E. coli*. Strains producing DnaG<sup>95-629</sup>, DnaG<sup>137-629</sup>, DnaG<sup>168-629</sup> produced a substantial proportion of soluble over-expressed protein but also a significant proportion of over-expressed protein was collected in the insoluble cellular fraction, while DnaG<sup>146-629</sup> appeared almost exclusively in the insoluble fraction. These observations are consistent with the green fluorescent phenotype of these mutants when expressed as fusions to EGFP.

Each mutant strain identified as producing putatively soluble DnaG truncations using EGFP as a solubility reporter produced soluble over-expressed protein in proportion to its EGFP brightness phenotype. To further confirm the solubility of these proteins, four truncated proteins, DnaG<sup>74-629</sup>, DnaG<sup>119-629</sup>, DnaG<sup>137-629</sup> and DnaG<sup>168-629</sup> were purified by one-step IMAC and were enriched and soluble (Figure 4.18B).

#### 4.4.10 Modelled protein structures of the *Acinetobacter baylyi*

##### zinc-binding and RNA polymerase domains

As crystallisation trials of N-terminal ZBD-extension mutants did not yield useful protein crystals, hypothetical protein structures were modelled for the *A. baylyi* DnaG ZBD and also the RPD. In a general sense, protein homology models are

produced by aligning the sequence of the protein of interest against the known structure of a related protein (Sali and Blundell, 1993; Fiser *et al.*, 2000; Haas *et al.*, 2013). The backbone of the protein of interest is then reproduced from the known structure, which then allows insertion of side-chain atoms and model optimisation. Optimisation of a model can allow improvement of protein models by adjusting atom positions to minimise collisions. The most robust information provided by modelled protein structures is the position of backbone atoms while side-chain positions are less certain (Schwede *et al.*, 2009), and of course comparative protein modelling is unreliable for regions of proteins with no known structural homologue.

MODELLER is a comparative protein structure modelling program (Sali and Blundell, 1993; Fiser *et al.*, 2000) that produces homology models of protein structures by satisfying/optimising spatial restraints of structures based on template structures and *de novo* modelling of protein loops (Eswar *et al.*, 2006; Martí-Renom *et al.*, 2000), resulting in a calculated protein structure for the position of all non-hydrogen protein atoms. A web interface for comparative protein structure modelling — MobWeb — allows automated generation of structural templates from structures in the PDB that have sequence based similarity to the protein of interest (ModPipe v2.2.0; Eswar *et al.*, 2003). Models are then comparatively built for each conserved sequence template with MODELLER to satisfy spatial restraints (MODELLER v9.11, released Sep. 6, 2012; Fiser *et al.*, 2000; Sali and Blundell, 1993). ModWeb then evaluates the resulting models using several model scoring methods to generate an aggregate measure of model quality.



Having generated a protein structure model, one needs to obtain an indication of the likelihood that the structure is realistic. To meet this goal, many methods have been developed to evaluate model protein structures (Melo and Feytmans, 1998; Samudrala and Moulton, 1998; Martí-Renom *et al.*, 2000; Zhou and Zhou, 2002; Wallner and Elofsson, 2003; Pettitt *et al.*, 2005; Tosatto, 2005; Benkert *et al.*, 2008; Eramian *et al.*, 2008; Randall and Baldi, 2008). One such homology structure evaluation algorithm, QMEAN, compares modelled secondary structures with secondary structure predicted from the primary amino acid sequence, long range atomic interactions, local backbone geometry and the likelihood of residue burial, ultimately to arrive at a reliability score for fold likelihood (Benkert *et al.*, 2008, 2011). QMEAN returns a value for the likelihood of the fold ranging from 0 to 1 where higher values indicate a better quality structure.

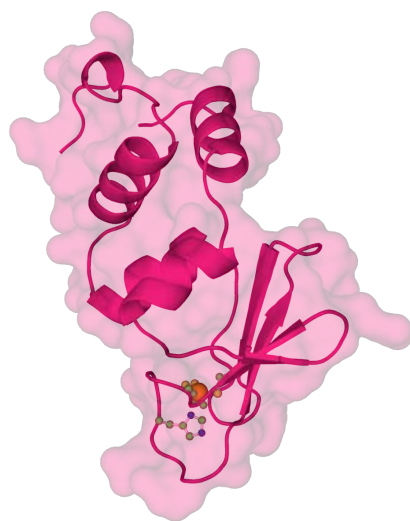
The distantly related *A. aeolicus* and *G. stearothermophilus* ZBD structures are highly conserved (Figure 4.3; Pan and Wigley, 2000; Corn *et al.*, 2005) suggesting that the bacterial ZBD should be well suited to comparative protein structure modelling. The larger RPD is also well conserved among bacteria, especially so at the two N-terminal sub-domains (*A. aeolicus* and *E. coli* C $\alpha$  RMSD: whole RPD, 4.05 Å over 288 residues; N-terminal sub-domain, 2.14 Å over 112 residues; topim sub-domain, 1.90 Å over 120 residues; and C-terminal sub-domain, 3.46 Å over 48 residues; Keck *et al.*, 2000; Corn *et al.*, 2005).

#### 4.4.10.1 Modelled protein structure of the *Acinetobacter baylyi* zinc-binding domain

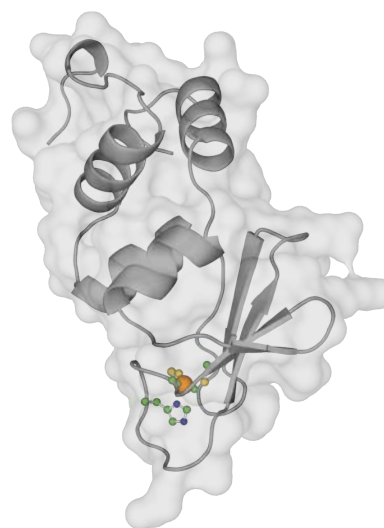
The most reliable structural model for the DnaG ZBD (Figure 4.19A) was built using the *G. stearotheophilus* ZBD (C $\alpha$  RMSD: 0.19 Å; 1D0Q chain A; Figure 4.19B; Pan and Wigley, 2000). However, a model based on the ZBD of *A. aeolicus* also produced a high probability structure (C $\alpha$  RMSD: 0.59 Å; 2AU3; Corn *et al.*, 2005). The model structures of the ZBD had GA341 (confidence) scores of 1, indicating that the protein conformations were structurally favourable and well supported. Neither model produced predictions for the poorly conserved and likely flexible ZBD-extension sequence. In addition to the GA341 scores produced by ModWeb, a QMEAN score of 0.716 was also calculated for the modelled *A. baylyi* ZBD structure, indicating a significantly reliable model structure. The *A. baylyi* DnaG ZBD modelled structure presented here represents residues 2–101, which correspond to residues 4–103 in the *G. stearotheophilus* ZBD structure (48% sequence identity).

ModWeb does not incorporate ligands into the models it produces. To include the zinc molecule which should be present in the *A. baylyi* ZBD structure, a single zinc atom was modelled into the predicted ZBD structure by aligning the four zinc coordinating side-chain atoms from the *G. stearotheophilus* ZBD (PDB: 1D0Q) with the corresponding atoms in the predicted *A. baylyi* structure. A zinc atom was then introduced and bonds added as in 1D0Q. The distances for the modelled Cys (sulphur)-Zn and His (nitrogen)-Zn bonds are within the ranges observed in crystal structures for other Cys (sulphur)-Zn and His (nitrogen)-Zn bonds (Hsin *et al.*, 2008) and the zinc coordination geometry is tetrahedral.

**A:** *Acinetobacter baylyi* zinc-binding domain model



**B:** *Geobacillus stearothermophilus* zinc-binding domain structure



**Figure 4.19: Structural model of DNA primase zinc-binding domain.** Protein structure models for the DnaG ZBD were produced using ModWeb (Eswar *et al.*, 2003). A statistically significant structural model was produced using the ZBD of *G. stearothermophilus* as the template (1D0Q; Pan and Wigley, 2000). **A:** *A. baylyi* ZBD structure, and **B:** *G. stearothermophilus* ZBD structure. ZBD structures are represented as ribbons under transparent surface with the C-terminus at the top-left and zinc coordinating residues in ball-and-stick representation and zinc atom in orange.

The predicted structure of the *A. baylyi* DnaG ZBD — in close agreement with the *G. stearothermophilus* structure against which it was modelled — contains a central four-stranded antiparallel  $\beta$ -sheet region from Lys28 to Cys59, where the zinc coordinating residues Cys38, His41, Cys59 and Cys62 are present in loops extending from the ends of the second and fourth  $\beta$ -strands. Flanking the antiparallel  $\beta$ -sheet region are two  $\alpha$ -helices at the N-terminus and three  $\alpha$ -helices at the C-terminus.

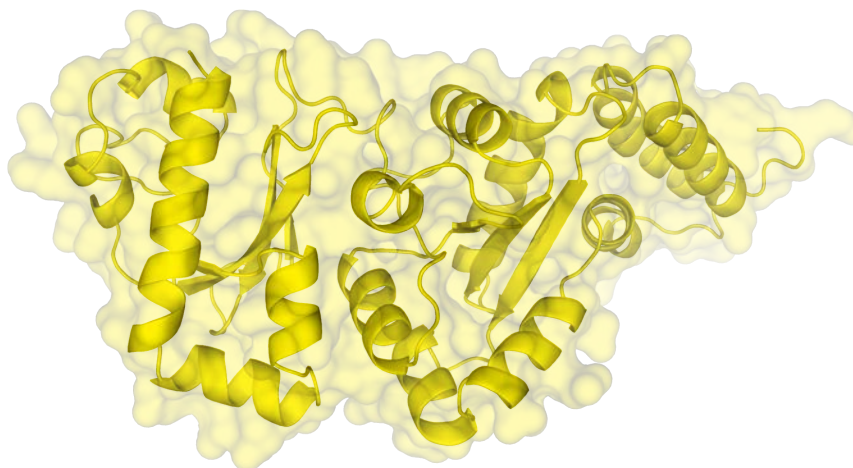
#### 4.4.10.2 Modelled protein structure of *Acinetobacter baylyi* RNA polymerase domain

Modelling of the RPD of *A. baylyi* DnaG returned a protein structure (Figure 4.20A) based on the *E. coli* RPD structure (C $\alpha$  RMSD: 0.88 Å; 1DD9 chain A; Figure 4.20B; Keck *et al.*, 2000). The *A. baylyi* model contains residues 171 to 475 modelled against *E. coli* DnaG residues 115–424 (40% sequence identity). The significance of the model is further supported by a GA341 of 1 and QMEAN score of 0.785. The modelled *A. baylyi* RPD structure contains the three sub-domains characteristic of bacterial primase RPDs (Keck *et al.*, 2000). At its centre, the metal binding toprim fold is flanked by an N-terminal mixed  $\alpha/\beta$  fold and a C-terminal fold made from three anti-parallel  $\alpha$ -helices.

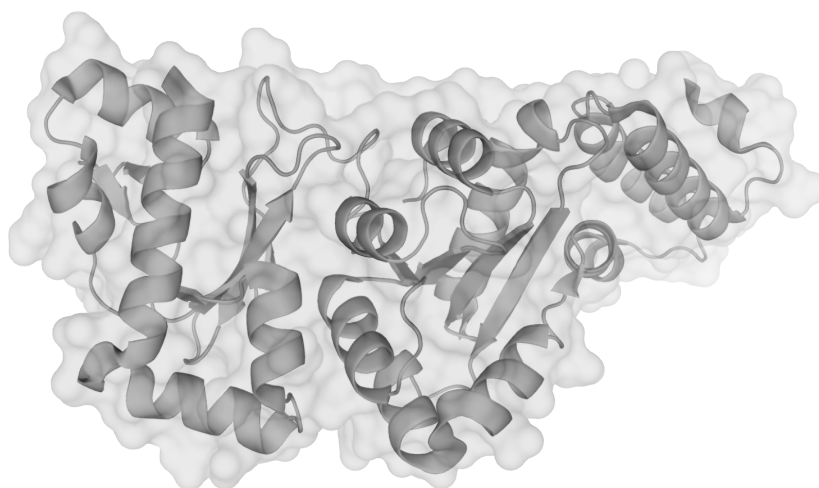
#### 4.4.11 N-terminal DNA primase deletion mutants and predicted protein structure

Even though truncation within the RPD — with its three distinct sub-domains — was not intended, some N-terminally deleted mutants were over truncated, identifying intra-RPD mutants that make soluble proteins. When comparing these intra-RPD mutants to the predicted structure of the *A. baylyi* domain, several of these appear at rational positions. Pfam prediction of the RPD of *A. baylyi* DnaG begins at residue Phe182, yet alignment of DnaG sequences closely related to *A. baylyi* show high similarity from around residue 169 (Figure 4.5); the first residue in the modelled structure is residue 171. Soluble N-terminal deletion mutants beginning at the start of the RPD (Table 4.7) remove the first

**A:** *Acinetobacter baylyi* RNA polymerase domain model



**B:** *Escherichia coli* RNA polymerase domain structure

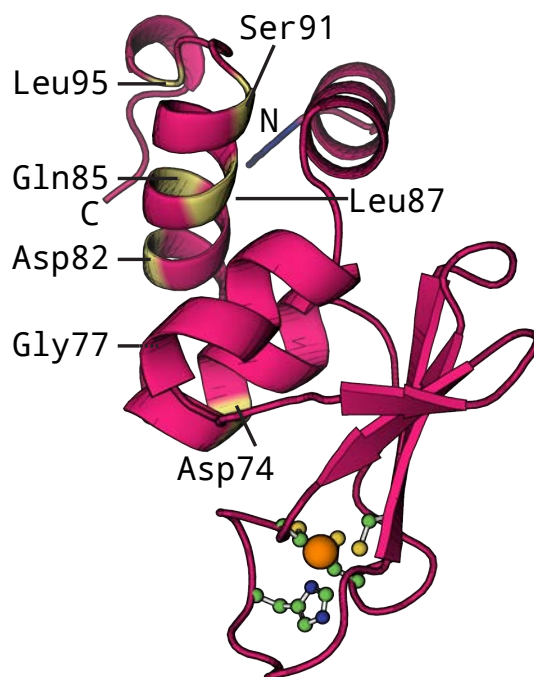


**Figure 4.20: Structural model of DNA primase RNA polymerase domain.** The protein structure was modelled using ModWeb (Eswar *et al.*, 2003). A highly supported structural model was produced using the *E. coli* RPD as the template (1DD9; Keck *et al.*, 2000). **A:** *A. baylyi* RPD structure, and **B:** *E. coli* RPD structure are represented as ribbons under transparent surface with the N-terminus at the left-hand side.

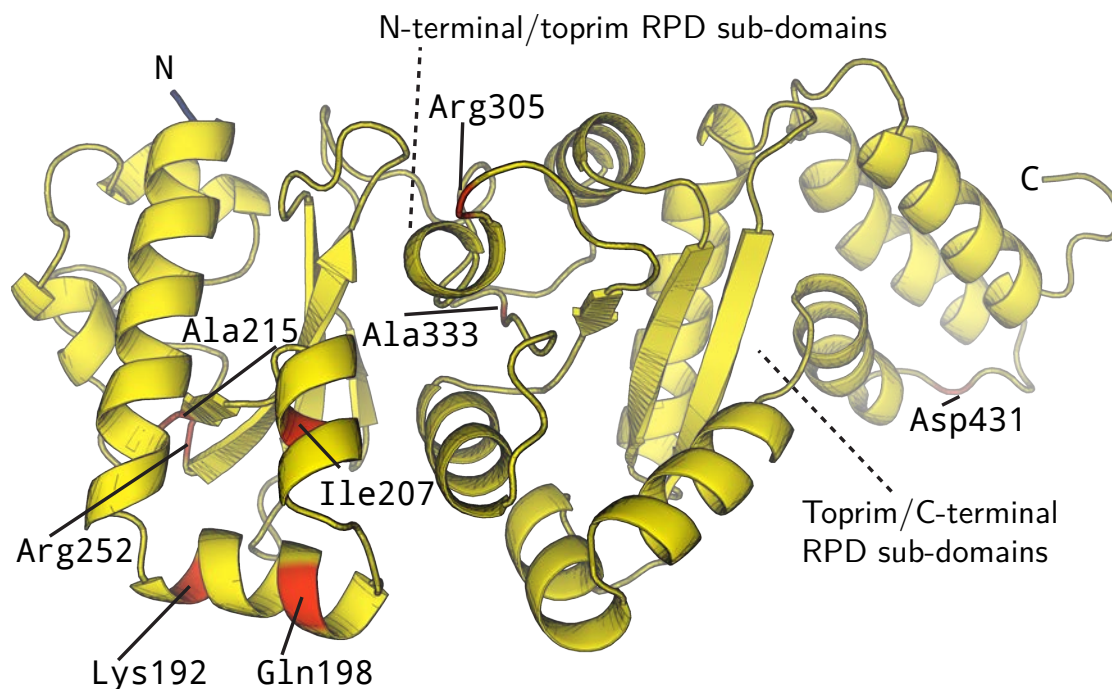
**Table 4.7: Putatively soluble N-terminally deleted mutants of DNA primase.** Identity of putative soluble protein mutants that are truncated within regions of modelled structures (Section 4.4.8).

Mutant
ZBD
74–629
77–629
82–629
85–629
87–629
91–629
RPD
192–629
198–629
207–629
215–629
252–629
305–629
333–629
431–629

**C: *Acinetobacter baylyi* zinc-binding domain**



**D: *Acinetobacter baylyi* RNA polymerase domain**



**Figure 4.20: N-terminal breakpoints of truncated DNA primase.** *A. baylyi* models of **A:** ZBD and **B:** RPD. Terminal residues of putative soluble mutants are coloured yellow (ZBD) or red (RPD) and the N-terminal residues are coloured blue and labelled N. ZBD structures show zinc coordinating residues in ball-and-stick representation with the zinc atom in orange. Dashed lines indicate the boundaries of distinct RPD sub-domains.

(DnaG<sup>192-629</sup>), second (DnaG<sup>198-629</sup>) and third (DnaG<sup>207-629</sup>)  $\alpha$ -helices (Figure 4.20D). Truncation at residue 252 separates the three C-terminal  $\beta$ -strands from the N-terminal sub-domain and apparently allows production of soluble protein. Further truncation through DnaG reveals protein end-points near the ends of the RPD sub-domains; DnaG<sup>305-629</sup> and DnaG<sup>333-629</sup> appear to separate, roughly, the central toprim fold from the N-terminal fold, and DnaG<sup>431-629</sup> the toprim fold from the C-terminal fold.

Soluble N-terminally deleted DnaG mutants that truncate within the ZBD occur at the end of the domain and are not very interesting. DnaG<sup>74-629</sup>, DnaG<sup>77-629</sup>, DnaG<sup>82-629</sup>, DnaG<sup>85-629</sup>, DnaG<sup>87-629</sup> and DnaG<sup>91-629</sup>, incorporate residues of the last  $\alpha$ -helix and short loop of the ZBD into the expressed protein (Figure 4.20C).

#### 4.4.11.1 C-terminal DNA primase deletion mutants and predicted protein structure

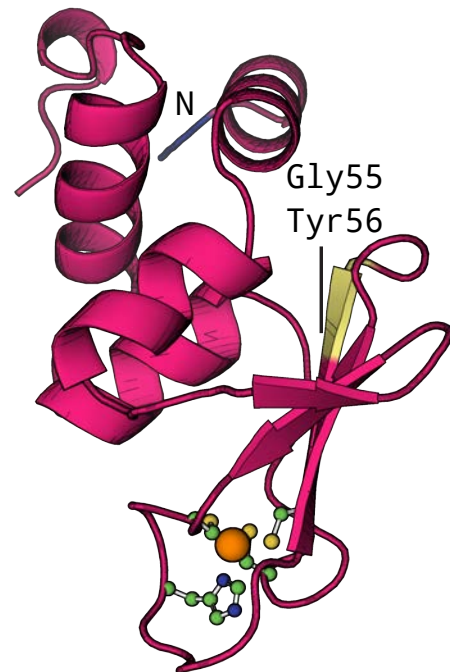
Few soluble C-terminally deleted DnaG mutants have truncations that map to regions of the modelled structures. Of the truncations terminating inside the ZBD, only two, short, putatively soluble mutants were identified, DnaG<sup>1-55</sup> and DnaG<sup>1-56</sup> (Table 4.9), which contain the three zinc-binding anti-parallel  $\beta$ -sheets and two short N-terminal  $\alpha$ -helices (Figure 4.20E). Each of these proteins contain the important zinc-binding region but may be too small to fold. The next shortest soluble C-terminally deleted DnaG mutants contain the whole ZBD and inter-domain linker, which is likely unstructured.

For C-terminally truncated proteins terminating in the RPD, DnaG<sup>1-195</sup> contains

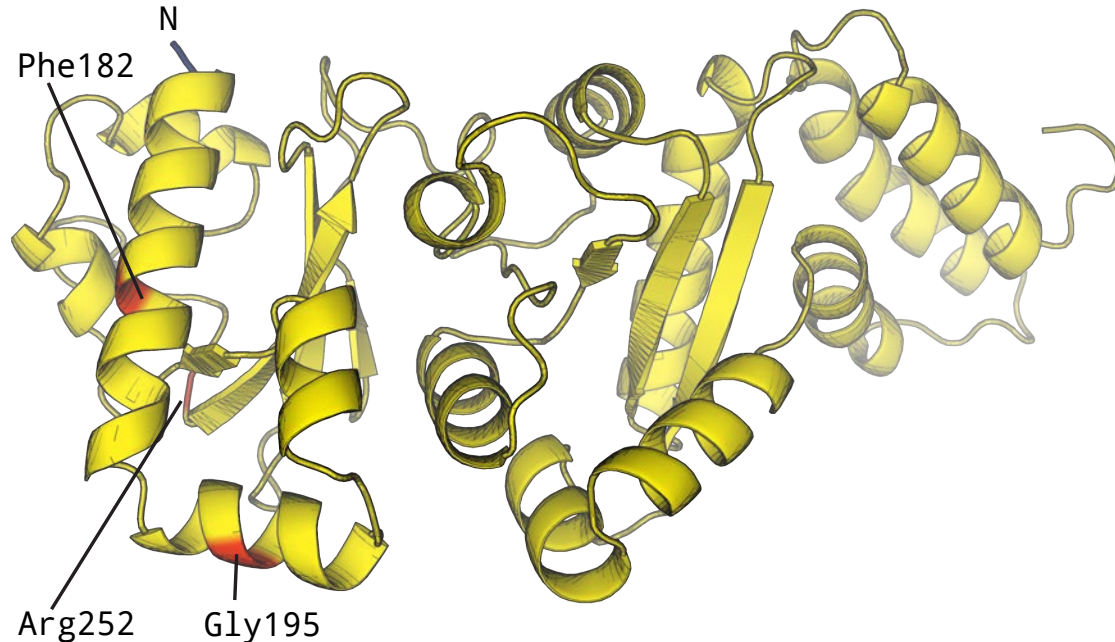
**Table 4.9: Putatively soluble C-terminally deleted mutants of DNA primase.** Identity of putatively soluble protein mutants that are truncated within regions of modelled structure (Section 4.4.1).

Mutant
ZBD
1-55
1-56
RPD
1-182
1-195
1-252

**E:** *Acinetobacter baylyi* zinc-binding domain



**B:** *Acinetobacter baylyi* RNA polymerase domain



**Figure 4.20: C-terminal breakpoints of truncated DNA primase.** A. *baylyi* models of **A:** ZBD and **B:** RPD. Terminal residues of putative soluble mutants coloured yellow (ZBD) or red (RPD) and N-terminal of structure are coloured blue and labelled N. ZBD structures show zinc coordinating residues in ball-and-stick representation with the zinc atom in orange.



the complete first N-terminal  $\alpha$ -helix (Figure 4.20B) and DnaG<sup>1-182</sup> contains the first 12 residues of this helix. Interestingly, the C-terminus of DnaG<sup>1-252</sup> — that ends at the same position as the N-terminally deleted mutant DnaG<sup>252-629</sup> — separates completely the three C-terminal  $\beta$ -strands from the N-terminal fold of the RPD.

## 4.5 Discussion

The aim of the work in this Chapter was to use the new technique presented in Chapter 3 to identify the domain boundary of the *A. baylyi* DnaG ZBD and to produce soluble domain constructs for the N-terminal ZBD and RPD-HBD of DNA primase, which we had otherwise not been able to produce.

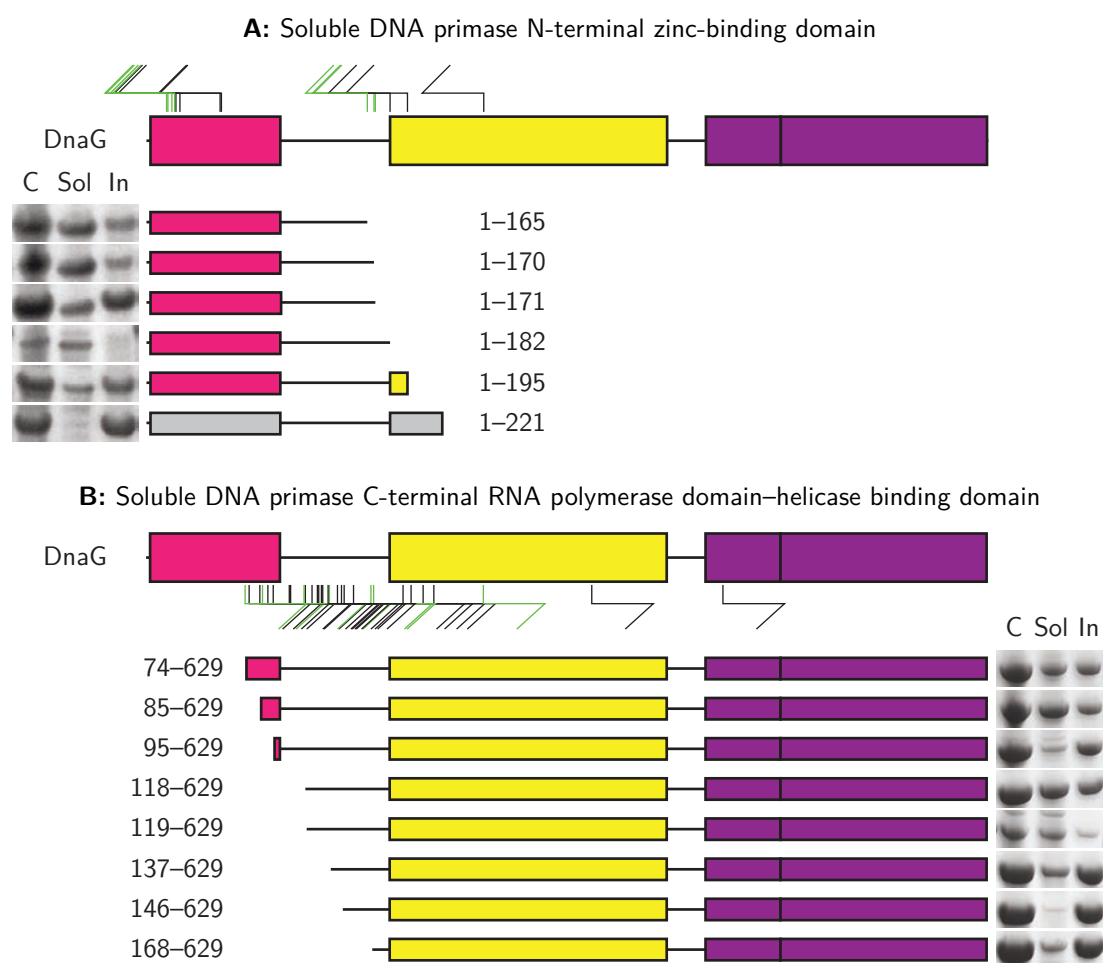
To achieve these aims, libraries of N- or C-terminally deleted *A. baylyi dnaG* genes were prepared using *ExoIII* and mung bean nuclease. Once expression plasmid pools for truncated *dnaG* were prepared, transformation and culture of these plasmids in the *E. coli* expression strain BL21( $\lambda$ DE3)*recA* allowed co-expression of the truncated genes with a 3'-*egfp* fusion for indicating solubility of the truncated proteins. EGFP is not fluorescent when first synthesised; protein folding must first occur, and then slow chemical steps generate the internal chromophore required for fluorescence (Section 1.6.2). These processes that mature non-fluorescent EGFP are slow and sensitive to protein solubility, making EGFP useful as a reporter of solubility of proteins to which it is fused. The EGFP solubility screening methodology used here is pragmatic and makes the

assumption that intact protein domains are well structured and therefore are more likely to be soluble. The other side of this coin is that truncated proteins which contain incompletely folded domains are likely to not fold properly and perturb maturation of EGFP fluorescence, producing different colony phenotypes for cells expressing soluble and insoluble EGFP-fusions. Shortened protein mutants that, when expressed, allow the C-terminal EGFP to become fluorescent were proposed to be truncated between domain boundaries. Still, some complete protein domains are not necessarily soluble when expressed alone and although this methodology is not of use in producing these proteins in soluble form, insoluble domains are of little experimental use. In this sense, our selection procedure is pragmatic.

#### 4.5.1 C-terminal deletions of DNA primase

This work was unsuccessful in producing genes encoding canonical, soluble ZBD for *A. baylyi* DnaG primase. However, in *A. baylyi*, DnaG contains an unusual 69 residue insertion at the end of the predicted ZBD boundary (Figure 4.5). *A. baylyi dnaG* ZBD genes which incorporate this linker region can direct expression of soluble proteins that remain soluble when purified (Figure 4.21A). This work reports experiments confirming that these proteins are indeed soluble, incorporate zinc and that at least a portion of the proteins are structured.

When the ZBD-extension proteins were first produced they contained a C-terminal EGFP fusion, which is not necessarily useful in further experiments and may in fact act to help solubilise the proteins. When EGFP was genetically removed,



**Figure 4.21: Soluble truncated DNA primase. A:** Soluble N-terminal (ZBD) fragments of DnaG, **B:** soluble C-terminal (RPD-HBD) fragments of DnaG. Coloured rectangles are Pfam domain annotations for *A. baylyi* DnaG: (■), ZBD; (■), RPD and (■), HBD. Arrows above (C-terminally deleted DnaG) and below (N-terminally deleted DnaG) indicate termini of truncation mutants identified in this study; green arrows indicate mutants with strong green fluorescence phenotype. Thinner representations below show data for truncated proteins for which protein solubility was determined without fusion to EGFP; included are SDS-PAGE gels showing whole cell, soluble and insoluble fractions. Grey cartoon indicates proteins for which no soluble protein was produced.

protein solubility was retained and the proteins could be purified. Purified DnaG<sup>1-165</sup>, DnaG<sup>1-170</sup>, DnaG<sup>1-171</sup>, DnaG<sup>1-182</sup> and DnaG<sup>1-195</sup> displayed similar CD spectra and these spectra were consistent with the proteins being at least partially structured. Further examination of two of the ZBD-extension proteins

using 1D and TOCSY NMR supports the conclusion that the proteins are at least partially folded and composed of mostly  $\alpha$ -helices, but also contain unfolded regions. Attempts to crystallise DnaG<sup>1–165</sup> and DnaG<sup>1–170</sup> were unsuccessful, which may be due to the presence of unstructured regions in the ZBD extension.

Examination of the identities of putatively soluble C-terminally deleted mutants of DnaG (Figure 4.21A) suggest that truncations within the *Moraxellaceae* ZBD extension do not produce soluble ZBD proteins. The sequence insertion within *A. baylyi* DnaG appears, by sequence alignment to related species (Figure 4.5), to end at residues 168, where interestingly, DnaG<sup>1–165</sup>, DnaG<sup>1–170</sup> are the shortest soluble proteins identified. The Pfam annotation of *A. baylyi* DnaG places residue 182 at the start of the RPD, although all 14 residues from 169–182 are well conserved in the *Moraxellaceae* species. The shortest of these DnaG mutants — DnaG<sup>1–165</sup> — terminates just before the end of the *Moraxellaceae* extension, and two more very soluble proteins DnaG<sup>1–170</sup> and DnaG<sup>1–171</sup> terminate at the very start of the RPD. Mutants longer than these proteins have reduced solubility (or over-express poorly), which would seem to be in agreement with the start of the folded RPD being around residue 169, since incomplete folding segments of the RPD should reduce solubility of these proteins. The discrepancy between Pfam assignment of the RPD and the extra 14 conserved *Moraxellaceae* residues might help explain why previous work to produce truncated proteins at this domain boundary were unsuccessful.

Investigation of DnaG<sup>1–165</sup> and DnaG<sup>1–170</sup> by mass spectrometry revealed that these two proteins bind zinc tightly. Analysis of these mutants, after extensive

dialysis against EDTA, revealed an additional mass consistent with bound zinc in buffer which supports natively folded protein conformations but the additional mass was not found when DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> were prepared in formic acid. Previous studies have shown that the complete primase ZBD structure is required for strong zinc-binding; a short 35 residue, partially folded zinc-binding peptide (K34-G69) from *E. coli* DNA primase binds zinc very weakly compared to the full length primase (K34-G69,  $K_D \sim 10-100$  nM and primase,  $K_D < 1$  pM; Griep and Lokey, 1996; Griep *et al.*, 1997). The weaker zinc-binding by the partially folded K34-G69 peptide allows removal of zinc from the peptide by chelation with EDTA, whereas it is not removed from full length primase. The fact that DnaG<sup>1-165</sup> and DnaG<sup>1-170</sup> can tightly hold onto zinc during dialysis against EDTA suggests that these ZBD-extension mutants have a natively-folded ZBD, and that therefore, it is the 69 residue extension sequence that contains the unstructured regions observed by NMR.

Although structures of DNA primase ZBDs are known, and limited proteolysis liberates the *E. coli* ZBD from primase (Tougu *et al.*, 1994), previous members of our research group have unsuccessfully attempted to make ZBD truncations of both the *E. coli* and *A. baylyi* proteins (Nicholas Dixon and Andrew Robinson; personal communication). In both *E. coli* and *G. stearothermophilus* DNA primases, the ZBD is liberated more slowly by limited proteolysis than the other domains of DnaG (Tougu *et al.*, 1994; Bird *et al.*, 2000), suggesting that the ZBD may be closely associated with the RPD. The ZBD is in fact observed on the surface of the primase RPD in the *A. aeolicus* structure (Corn *et al.*, 2005). The close association of the ZBD and RPD may, in many species, have either I)

resulted in stabilisation of the ZBD by association with the RPD, or II) aided folding of the ZBD, which may explain the difficulties in producing soluble ZBD constructs. The thermophilic nature of *G. stearothermophilus* may have provided selection pressure for increased stability/folding efficiency of its ZBD.

In that case, the absence of distinct and soluble ZBD mutants of *A. baylyi* DnaG from recovered shortened genes would also suggest that the ZBD of *A. baylyi* DnaG is either not soluble, or is unable to efficiently fold, in isolation. Yet, soluble ZBD mutants containing also the *Moraxellaceae* insertion sequence can be expressed as folded, soluble proteins. Thus, in *A. baylyi* — and likely also in the other *Moraxellaceae* species — the extension following the ZBD may have adapted to add stability to the ZBD, or take part in the folding pathway of the ZBD, allowing over-expression of folded and soluble ZBD. The fact that when the ZBD is proteolytically removed from *E. coli* DnaG, it is not soluble, would, if the relationship is general for many bacteria, suggest that ZBD stabilisation, rather than folding, is important for isolated ZBD solubility. In this case, the extension to the ZBD of *A. baylyi* — and maybe the rest of the *Moraxellaceae* — has been selected over time to stabilise the ZBD. As the original insertion into the ancestral DNA primase must have been tolerated, the closer proximity of the extension sequence to the ZBD — at least when not associated with the RPD — may have provided the selective pressure for reassignment of the stabilising interaction from the RPD to the extension.

### 4.5.2 N-terminal deletions of DNA primase

An N-terminal deletion library, complementary to the C-terminal library, was produced to identify soluble RPD-HBD proteins (Figure 4.21B). Very soluble N-terminally deleted *A. baylyi* DnaG proteins were identified starting at residues 168 and 170, which is directly where the *Moraxellaceae* sequences begin to be well conserved (Figure 4.5), and this likely marks the start of the RPD. Several truncated, putatively soluble mutants began slightly into the N-terminus of the RPD (truncated at residues 192, 198 and 207) and three were spaced within the RPD (residues 252, 305 and 333). On the whole, few putatively soluble truncation mutants were observed encroaching into the start of the RPD of primase, which we should expect if the methodology selects for complete domains which fold well. However, in contrast to the C-terminal deletion, N-terminal deletion produced soluble mutants that start in the ZBD-extension and even in the ZBD (Figure 4.21A). The large number of soluble mutants beginning in the ZBD-extension may support that this region is unstructured. In the end, comparison of mutants obtained by N-terminal and C-terminal deletion strongly suggests that the N-terminal domain boundary of the RPDs is at residues 168–170, and this will be the region targeted in further work.

### 4.5.3 DNA primase truncation mutations and modelled protein structures

This Chapter provides well supported homology models for the ZBD and RPD of *A. baylyi* DnaG. Mapping the end-points of soluble truncated DnaG mutants onto modelled ZBD and RPD structures suggests that truncation breakpoints of soluble proteins can occur in the first few secondary structural elements of the domain (at specific positions). Identification of soluble proteins using a solubility reporting fusion protein selects proteins that over-express well and are soluble. Based on these selection criteria, proteins with truncated secondary structures — that may or may not be folded — and portions of natively unfolded flexible extensions do not necessarily negatively affect selection.

Soluble N-terminally deleted DnaG mutants were also identified that truncate the known sub-domains of the RPD. This was interesting as these truncations were unexpected. Unlike at the start of the RPD, where several mutants were identified, sub-domain mutants did not have varied end-points, but since these regions were not targeted by the deletion study, it stands to reason that these were likely under represented in the library. It might be expected that a deletion library focussed at these regions would reveal more soluble intra-RPD mutants. Nonetheless, these adventitious non-targeted soluble mutants are an added bonus, providing further insight into the available soluble variants of the protein.



#### 4.5.4 General discussion

Attempting to reconcile the divergence of solubility between C-terminally and N-terminally shortened DnaG proteins, where truncation occurs in the sequence insertion following the ZBD, is not straightforward. C-terminally shortened proteins containing the whole ZBD are not expressed solubly without the whole extension, but many different truncations within the extension from the complementary N-terminally deleted library were soluble. It appears that the ZBD of *A. baylyi* requires the full ZBD-extension for solubility, but that this region is unstructured and unimportant for the solubility of the RPD.

In the experiments presented here, the green fluorescence phenotype of *E. coli* cells expressing truncated DnaG-EGFP fusion proteins is shown to correlate strongly with protein solubility when the EGFP is removed; these data are thus consistent with previous reports (Waldo *et al.*, 1999; Pédelacq *et al.*, 2005). Examining all mutants identified using the EGFP solubility reporter shows that many of the possible protein constructs are soluble, the majority of which appear to truncate to regions outside of known structural elements. Whether or not all of these proteins are fully folded is unknown, but seems unlikely. It is probable that some contain unfolded extensions and were identified because these extensions do not lead to significant levels of protein aggregation. Proteins which produce the brightest green fluorescent phenotype when over-expressed as fusions to EGFP, however, appeared at fewer truncation positions and paint a clearer picture, where they strongly correlate with the domain boundary at the N-terminus of the RPD. This technique for protein truncation and solubility selection has proven useful

for identifying soluble protein constructs, which previously were not produced by means of rational design.

From a pragmatic viewpoint, having a way to produce soluble versions of small domains enables study of their structures and functions that would otherwise be impossible. For example, the extent of unstructured regions at the termini can readily be identified by NMR, but this can only be done if samples of soluble proteins can be produced in the first place.

## Chapter 5

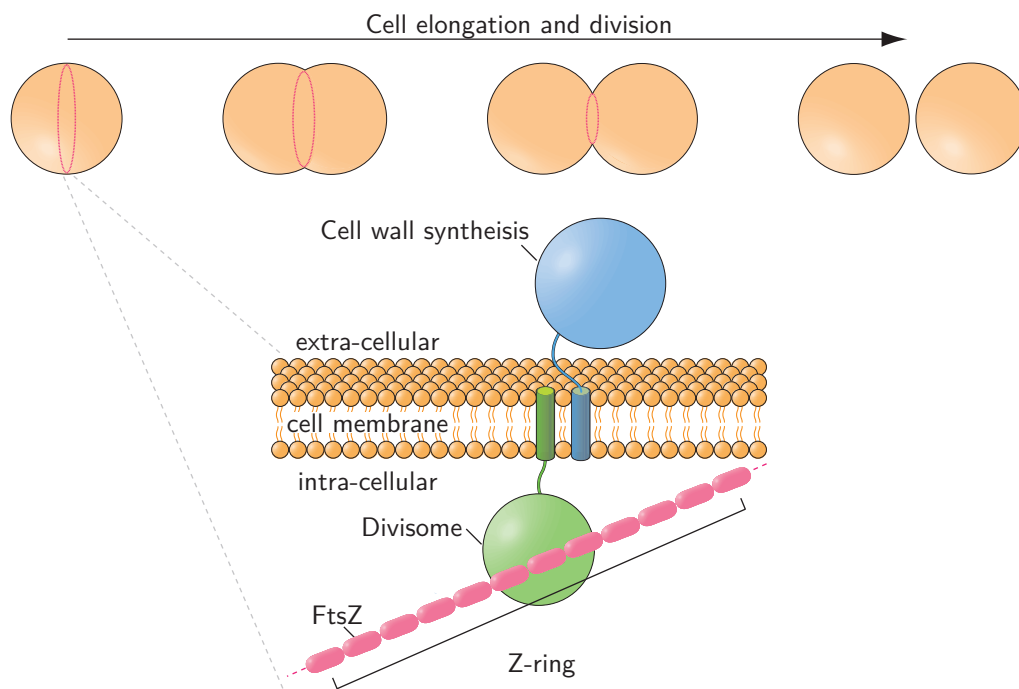
# Soluble domain constructs of *Staphylococcus aureus* septation ring formation regulator (EzrA)

### 5.1 Introduction

#### 5.1.1 Cell division and Z-rings

Cell division in bacteria is a tightly controlled process, in which many proteins work in concert to segregate the required cellular components for daughter cell production, and to facilitate separation of the new cells (Adams and Errington, 2009). Integral to division of the cell, FtsZ polymerisation forms a ring structure around the perimeter of the cell centre known as the Z-ring (Bi and Lutkenhaus, 1991; Adams and Errington, 2009). With the Z-ring in place, complementary

proteins are mobilised to the mid-cell to form the divisome which coordinates cell membrane synthesis, cell wall synthesis and cell division (Figure 5.1; Goehring and Beckwith, 2005; Adams and Errington, 2009).



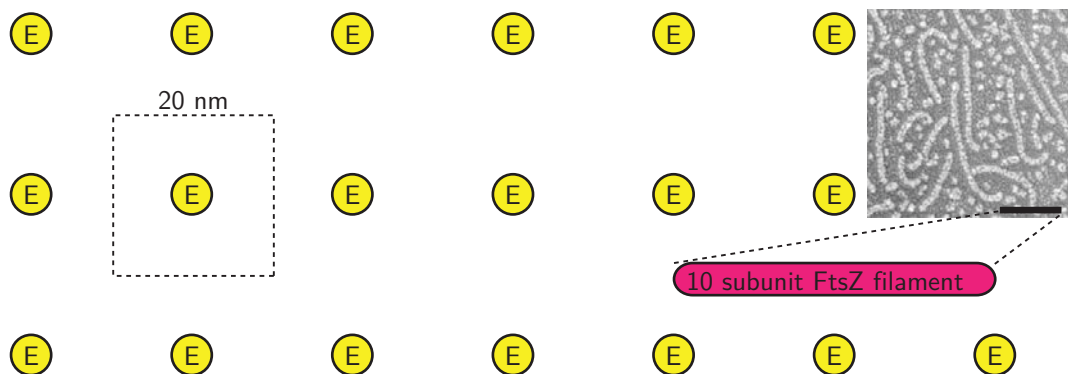
**Figure 5.1: Cell division and the divisome.** Bacterial cell division occurs by cell elongation at a structure known as the Z-ring, which is a polymer of the FtsZ protein. On the inside of a cell, a group of proteins known as the divisome facilitates the process of cell division. On the outside of the cell, a second group of proteins synthesise and cross link peptidoglycans for synthesis of new cell wall. The divisome and cell wall synthesis proteins drive cell elongation and over time constriction of the Z-ring separates the daughter cells. Image inspired by Adams and Errington (2009).

In both *Bacillus subtilis* and *E. coli*, Z-ring formation is observed to be independent of FtsZ concentration. Throughout the cell cycle FtsZ maintains a constant concentration, and in fact experimentally increasing the concentration of FtsZ does not alter the regulation of Z-ring formation (Weart and Levin, 2003). The tight regulation of Z-ring formation in bacterial cell division therefore indicates the action of other participants.

### 5.1.2 EzrA has two Z-ring regulating roles

One candidate for regulation of Z-ring formation in the medically relevant Firmicutes bacteria is the membrane anchored protein EzrA (extra Z-rings; Levin *et al.*, 1999). *Bacillus subtilis* EzrA contains an N-terminal membrane anchor and occurs throughout the cell membrane where it appears to guard against FtsZ polymerisation. In *ezrA* null mutants, Z-rings lose septal specificity, with Z-rings then occurring anywhere from the poles to mid-cell (Levin *et al.*, 1999), whereas mild over-expression of EzrA inhibits FtsZ assembly (Haeusser *et al.*, 2004). Removal of the N-terminal membrane anchor abolishes the ability of EzrA to negatively control Z-ring formation (Haeusser *et al.*, 2004). Although *B. subtilis ezrA* null strains lose spatial control of Z-ring formation, the proportion of growing populations containing Z-rings is unchanged, indicating that EzrA is not involved in the cell-cycle switch for Z-ring formation (Levin *et al.*, 1999).

EzrA is well suited for negative regulation of Z-ring formation, with cellular concentrations around 10,000–20,000 molecules per cell (Haeusser *et al.*, 2004). It associates tightly with the cell membrane ( $\frac{2}{3}$  of EzrA is extracted with membranes; Bhavsar *et al.*, 2005). With a cell membrane area of  $3.9 \times 10^6 \text{ nm}^2$  (rod with two hemispheres of  $r = 0.415 \text{ }\mu\text{m}$  and  $h = 2.45 \text{ }\mu\text{m}$ ; Maass *et al.*, 2011), EzrA may be spaced as closely as one EzrA molecule per 300–700  $\text{nm}^2$  of cell membrane, thus providing a dense packing to block FtsZ polymerisation (Figure 5.2).



**Figure 5.2: EzrA spacing and FtsZ polymer size.** EzrA is densely packed on the internal face of the cell membrane in *B. subtilis*. Inside the average *B. subtilis* cell, the internal surface area is up to  $3.9 \times 10^6 \text{ nm}^2$  (Maass *et al.*, 2011), with between 10,000 and 20,000 molecules of EzrA (Haeusser *et al.*, 2004) associating tightly with the cell membrane (over  $\frac{2}{3}$  of EzrA is membrane associated; Bhavsar *et al.*, 2005). A grid of EzrA monomers (yellow balls), each occupying an area of  $400 \text{ nm}^2$  is shown to scale, along with a 10 subunit FtsZ polymer of  $4 \times 40 \text{ nm}$  (pink rod; Romberg *et al.*, 2001). EzrA monomers (62 kDa) are depicted to scale as a spherical and globular protein (diameter of  $\approx 5 \text{ nm}$ ). Inset top right: electron micrograph of *B. subtilis* FtsZ polymers with scale bar of 40 nm (Adapted from Haeusser *et al.*, 2004).

However, as somewhat of a contradiction to the apparent Z-ring inhibition role of EzrA, it has also been shown to co-localise with Z-rings and this occurs even when the N-terminal membrane anchor is deleted (Haeusser *et al.*, 2004). This mid-cell localisation of EzrA has been suggested to occur through a short conserved C-terminal region (the QNR patch) that, when removed, eliminates its ability to localise to mid-cell. *ezrA*<sup>ΔQNR</sup> mutants experience delayed — but not absent — cell division, and still negatively regulate FtsZ polymerisation (Haeusser *et al.*, 2007).

### 5.1.3 Cell wall synthesis

EzrA is also involved in coordinating cell wall synthesis. The bacterial cell wall, enlightened mostly from experiments in *B. subtilis*, is composed of a single massive molecule of cross-linked peptidoglycan polymers and is important for maintaining cellular integrity. For successful cell division, this massive structure must be remodelled to allow daughter cell production, requiring careful choreography of the many proteins associated with the divisome.

Cell wall synthesis in the Firmicutes occurs in two distinct modes, mediated by two spatially distinct pathways: I) cell division (septum formation), common to both coccoid and rod-shaped species, and II) cell elongation (lateral wall synthesis), occurring only in, and responsible for the unique morphology of rod-shaped cells (Jones *et al.*, 2001; Daniel and Errington, 2003; Scheffers and Pinho, 2005; Carballido-López, 2006). Cell septum peptidoglycan synthesis — depletion of which can produce defects in daughter cell production — is more important than lateral wall peptidoglycan synthesis, where deletion of genes encoding important cell elongation proteins can convert rod-shaped species to have viable coccoid phenotypes (Begg and Donachie, 1985; Henriques *et al.*, 1998).

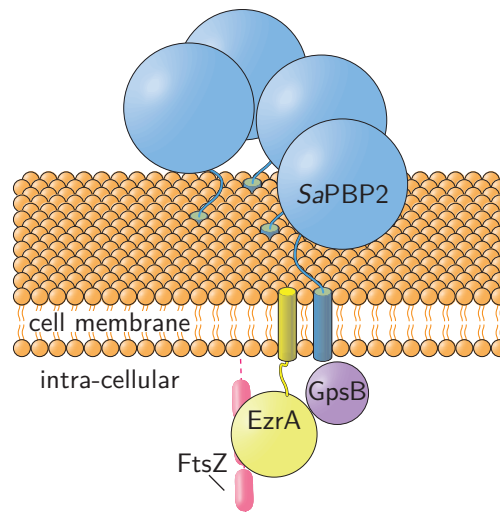
Septal cell wall synthesis occurs mainly through the action of the septum specific peptidoglycan synthesising penicillin-binding proteins (PBPs), located on the outer surface of the cytoplasmic membrane (Scheffers and Pinho, 2005). Class A PBPs facilitate both crosslinking between and elongation of peptidoglycans while class B PBPs catalyse cross-linking only (Ghuysen, 1991). Localisation of cell

division PBPs is dependent on the presence of Z-rings as Z-ring depletion results in dispersal of septal PBPs. The septal localisation of division PBPs suggests communication from the intra-cellular divisome proteins arranged at the Z-ring to the PBPs positioned on the outside of the cell membrane, but this relationship is poorly understood (Scheffers and Pinho, 2005).

#### 5.1.4 EzrA helps coordinate cell wall synthesis at the site of cell division

Bacterial-two-hybrid analysis in *B. subtilis* indicates two cell division proteins which interact with the essential septal PBP1 (*ponA*; also called PBP1A, not to be confused with a related protein known as PBP1a), EzrA and GpsB (guiding PBP1 shuttling), which also appear to interact with each other. Neither EzrA nor GpsB is individually essential in *B. subtilis* but gene mutations are synthetically almost inviable in combination (Kobayashi *et al.*, 2003; Claessen *et al.*, 2008). PBP1 takes part in both lateral cell wall and septal peptidoglycan synthesis, localising to mid-cell at the onset of cell division, but otherwise is distributed throughout the cell wall; this spatial pattern of localisation is shared by GpsB. Depletion of EzrA removes GpsB and PBP1 localisation to mid-cell (Figure 5.3) and results in a thinner cell phenotype, suggesting a role for EzrA in coordinating these components of cell wall synthesis to the divisome.





**Figure 5.3: Coordination of cell wall synthesis at the divisome.** Currently PBP2 is thought to be the enzyme required for cell wall synthesis in *Staphylococcus aureus*. The *B. subtilis* homologue of PBP2, PBP1, has been shown by bacterial-two-hybrid experiments to associate with EzrA and GpsB at mid-cell during cell division (Claessen *et al.*, 2008). Depletion of EzrA stops GpsB and PBP1 from localising at the mid-cell.

### 5.1.5 The importance of EzrA in cell division in *Staphylococcus aureus*

Cell wall synthesis in the major human pathogen *Staphylococcus aureus* occurs at the division septum only (Pinho and Errington, 2003; Scheffers and Pinho, 2005). Here, the cell wall synthesis apparatus localises and is coordinated by the cell division apparatus, that alternates, in turn, among the three perpendicular axes of the cell.

Due to the simplified cell wall synthesis methodology applied by *S. aureus*, the proteome contains a minimal complement of PBPs, where PBP2 (*ponA*; homologous function to PBP1 in *B. subtilis*) localises to the cell septum (Pinho and Errington, 2003), and is the only PBP present that can both elongate and

cross-link peptidoglycans (Pinho and Errington, 2003; Scheffers and Pinho, 2005). Localisation of the divisome and coordination of cell wall synthesis is imperative in *S. aureus*, as eliminating Z-rings through FtsZ depletion delocalises PBP2, and upon loss of this septal localisation, peptidoglycan synthesis continues in a distributed fashion. This results in a sickly cell phenotype where the cells continue to enlarge until they are too large and lyse (Pinho and Errington, 2003). A similar defect occurs through application of methicillin, which delocalises *S. aureus* PBP2 and leads to cell death (Pinho *et al.*, 2001). Dissection of the mechanism for coordinating the internal cell division machinery with external cell wall synthesis might suggest novel antibiotic targets. Previous evidence that EzrA plays a role in localising the *B. subtilis* PBP2 homologue to the divisome makes EzrA an intriguing target for study.

Indeed, *ezrA* has been established by multiple studies to be an essential gene in *S. aureus*. Transposon mutagenesis has identified *ezrA* (SAV1717 in the Mu50 reference genome; Bae *et al.*, 2004) as an essential gene in both the Newman and SH1000 strains (Chaudhuri *et al.*, 2009) and this conclusion was supported by an antisense RNA screen in a wild-type but otherwise unidentified *S. aureus* strain (named *ytwP* in this study; Ji *et al.*, 2001). Steele and colleagues (2011) showed evidence for the essentiality of *ezrA* in *S. aureus* SH1000 by deleting the 291 C-terminal codons from chromosomal *ezrA* in a strain complemented with plasmid encoded full-length EzrA under stringent control; this strain could not grow in the absence of expression of full length EzrA.

However, controversy exists as to whether *ezrA* is indeed an essential gene in *S.*

*aureus*, as not all systematic studies to identify essential genes have identified *ezrA*. Moreover, the work of Jorge and colleagues (2011) directly contradicts the previous studies that identified EzrA as being an essential protein. They present evidence that *S. aureus* can survive when EzrA is I) depleted using antisense RNA to inhibit *ezrA* expression, II) not expressed through withholding inducer from a strain that expresses *ezrA* under the strictly controlled p<sub>spac</sub> promoter, and III) eliminated in null mutants of the COL, NCTC8325-4, Newman, RN4220 and SH1000 strains. The data presented for both the antisense RNA and p<sub>spac</sub> experiments indicate that a small proportion of residual EzrA remains *in vivo*, which might help explain the lack of lethality. However, the inviability of null mutants is hard to reconcile. Two further genome wide studies of genes essential in *S. aureus* failed to identify *ezrA*; however, this was because *ezrA* was not identified specifically due to low homology with *ezrA* from other species (< 30% identity at the protein level *c.f.* *B. subtilis*, *Enterococcus faecalis*, *E. coli*, *Streptococcus pneumoniae* and *Pseudomonas aeruginosa*; Ji *et al.*, 2001; Ko *et al.*, 2006).

The conflicting results of Jorge *et al.* (2011), compared with those of other groups (Ji *et al.*, 2001; Bae *et al.*, 2004; Chaudhuri *et al.*, 2009; Steele *et al.*, 2011), are hard to reconcile, and no obvious strain or methodological difference is identifiable. Nonetheless, Jorge *et al.* (2011) do report sickly cell phenotypes when EzrA is not expressed, including larger cells, delocalised cell wall synthesis and delocalised Z-rings.

### 5.1.6 EzrA as a potential protein–protein interaction hub

Recent evidence implicates EzrA as an interaction hub in *S. aureus*, where bacterial two-hybrid experiments have indicated that it interacts directly with 12 cell division/peptidoglycan synthesis proteins, of which at least 10 are encoded by essential genes (Table 5.1; Bae *et al.*, 2004; Chaudhuri *et al.*, 2009; Steele *et al.*, 2011). The implication of these many interactions is that EzrA may play an important role in coordination of many functions of the divisome in addition to regulating Z-ring formation and localisation of cell wall synthesis.

Using a *S. aureus* strain expressing an EzrA-GFP fusion from the native *ezrA* promoter and inducible FtsZ, Steele *et al.* (2011) were able to show that EzrA co-localises with Z-rings. Further experiments with an inducible *ezrA* strain showed that EzrA is necessary for localisation of GspB-GFP and PBP2-GFP to the septum and in fact, for active peptidoglycan synthesis. These results provide strong evidence that de-coupling of cell wall synthesis from cell division is lethal in *S. aureus* and that divisome construction — initiated by Z-ring formation — is coordinated by EzrA through the cell membrane to the peptidoglycan synthesising PBP2. Regardless of the conflicting reports of *ezrA* essentiality in *S. aureus*, clearly cells depleted of EzrA are sickly, if not inviable. Further examination of protein interactions with EzrA should provide insight into the coordination of bacterial cell division, and has the potential to yield novel antibiotic targets. Of initial interest are the EzrA interactions with FtsZ and PBP2, where disruption of the EzrA–FtsZ interaction should banish PBP2 (and potentially other proteins) from the divisome and lead to cell lysis by delocalisation of peptidoglycan

**Table 5.1: Essentiality of proteins suspected of interacting with EzrA.** List of proteins that were identified as interacting with *S. aureus* EzrA by bacterial-two-hybrid assay (Steele *et al.*, 2011). Proteins are named as in *S. aureus*, and alternate names are in footnotes. Protein functions are described by as per Lutkenhaus and Addinall (1997); Henriques *et al.* (1998); Mercer and Weiss (2002); Scheffers and Pinho (2005); Adams and Errington (2009).

Protein	Known role(s)	Essential	
		<i>S. aureus</i> <sup>1</sup>	<i>B. subtilis</i> <sup>2</sup>
EzrA <sup>3</sup>	Negative regulator of Z-ring formation away from septum. Regulates Z-ring dynamics when localised to mid-cell.	√ <sup>4,5</sup>	X, √ (GpsB <sup>-</sup> / FtsL <sup>-</sup> / SepF <sup>-</sup> )
DivIB	Unclear role in septal peptidoglycan synthesis	√ <sup>4</sup>	√
DivIC	Unclear role in septal peptidoglycan synthesis	√ <sup>4</sup>	√
FtsA	FtsZ membrane anchor. Required for Z-ring assembly, organisation and ancillary protein recruitment	√ <sup>4</sup>	√
FtsL	Unclear role in septal peptidoglycan synthesis	√ <sup>4</sup>	√
FtsW	Unclear role in stabilisation of Z-rings and recruitment of septal peptidoglycan synthesis enzymes	√ <sup>4</sup>	√
FtsZ	Structural unit of Z-ring	√ <sup>4</sup>	√
GpsB <sup>6</sup>	Recruitment of PBP2 ( <i>B. subtilis</i> PBP1) to mid-cell during cell division (and non-division lateral cell wall synthesis in rod shaped species)	√ <sup>4</sup>	X, √ (EzrA <sup>-</sup> ) <sup>7</sup>
PBP1 <sup>8</sup>	Cell division specific peptidoglycan synthesis	√ <sup>4</sup>	√
PBP2 <sup>9</sup>	Peptidoglycan synthesis	√ <sup>4</sup>	X
PBP3	Cell division specific peptidoglycan synthesis	X	NA
RodA	Required for rod cell morphology in <i>B. subtilis</i> , unknown role in <i>S. aureus</i>	X (√) <sup>4,5</sup>	√
SepF <sup>10</sup>	Required for proper septal morphology	√ <sup>4</sup>	X, √ (EzrA <sup>-</sup> ) <sup>11</sup>

<sup>1</sup> Chaudhuri *et al.* (2009)

<sup>3</sup> EzrA was previously named YtwP and SAV1717

<sup>5</sup> Ji *et al.* (2001)

<sup>7</sup> Claessen *et al.* (2008)

<sup>9</sup> Named PBP1 in *B. subtilis*

<sup>11</sup> Hamoen *et al.* (2005)

<sup>2</sup> Kobayashi *et al.* (2003)

<sup>4</sup> Bae *et al.* (2004)

<sup>6</sup> GpsB gene name is *ypsB*

<sup>8</sup> Named PBP2B in *B. subtilis*

<sup>10</sup> SepF gene name is *ylmF*

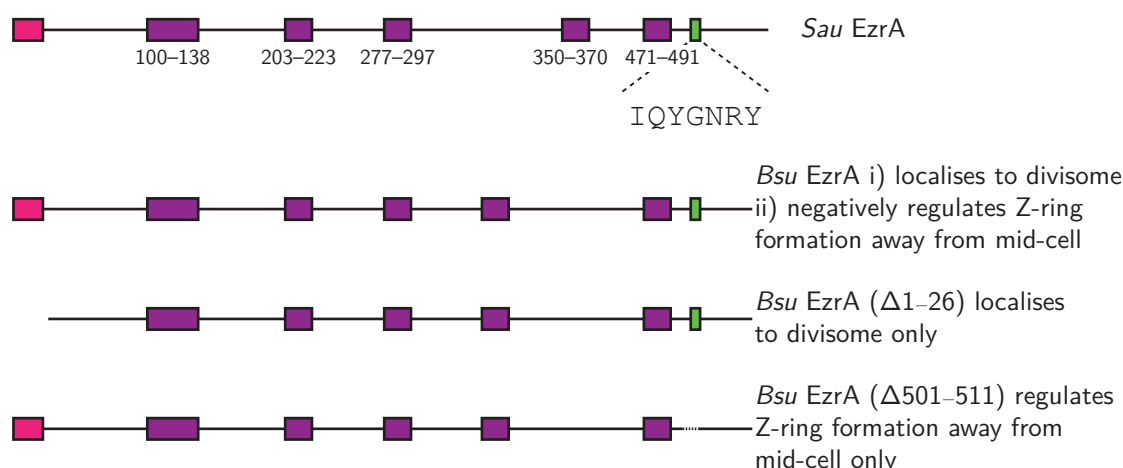
synthesis, while disruption of the EzrA–PBP2 interaction should disable cell wall growth.

Of these two interactions, that between EzrA and FtsZ is most intriguing as an antibacterial target. The number of different putative interactions of EzrA suggest that it may localise other proteins to mid-cell, and the cell lysis phenotype would be promising from the standpoint of recruiting the immune system in infected patients.

### 5.1.7 Little is currently known about the structure of EzrA

EzrA is a relatively recently identified participant in cell division (Levin *et al.*, 1999) where *ezrA* mutants of *B. subtilis* were observed to possess extra Z-rings. Since the discovery of EzrA, there have been very few *in vitro* biochemical studies of it. After beginning this work, soluble N-terminally deleted EzrA mutants have been identified by limited proteolysis (EzrA<sup>164–564</sup>, EzrA<sup>250–564</sup>; Son and Lee, 2013), but other domain boundaries are unknown and its interactions are in question. No three-dimensional structural information is currently published for any EzrA protein and no structurally known domains occur within EzrA according to Pfam (Finn *et al.*, 2010), but five coiled-coil domains are predicted to be present (Figure 5.4).

The N-terminal membrane anchor of EzrA is dispensable for the interaction with the divisome and PBP2; instead, the membrane anchor serves the purpose



**Figure 5.4: Known phenotypes of *ezrA* mutants.** Displayed are *S. aureus* and *B. subtilis* EzrA proteins with known *in vivo* *B. subtilis* *ezrA* phenotypes (Haeusser *et al.*, 2007). Predicted coil-coil domains as annotated in Pfam are shown as purple boxes, the N-terminal membrane anchor is shown as a pink box and the QNR patch is shown in green. The full length *B. subtilis* protein is 562 amino acids long and *S. aureus* 564.

of embedding EzrA in the cell membrane to guard against aberrant Z-ring polymerisation (Figure 5.4). The well conserved near-C-terminal seven amino acid QNR patch (residues 506–512 in *S. aureus*) has been shown to be important in mid-cell localisation but not inhibition of FtsZ assembly by *in vivo* deletion studies (Haeusser *et al.*, 2007).

## 5.2 Aims

EzrA is an interesting target for antibacterial development as two protein–protein interactions in which it is involved are known and many more are suspected, and most of these interaction partners are essential for growth and/or cell survival in *S. aureus* (Table 5.1). Since EzrA is likely to be composed of several protein domains but is poorly conserved and has no homologues with solved

protein structures, we were interested in searching for soluble protein truncations using the methodology described in Chapter 3. In addition to interest in the proposed FtsZ-interacting C-terminal EzrA domain, the numerous suspected protein–protein interactions may occur elsewhere in the protein, and use of a range of soluble protein constructs would facilitate study of the sites of these interactions.

The aims of this Chapter were to perform both I) 5'-terminal and II) 3'-terminal deletion of *S. aureus ezrA* to produce libraries of truncated genes with truncation points distributed throughout. This strategy should allow identification of soluble proteins that are useful for direct study and determination of protein break points that suggest domain boundaries throughout the protein. After identification of soluble truncated EzrA proteins, the further aims were to III) examine if the truncated proteins are folded, and IV) identify EzrA fragments that can still interact with FtsZ.

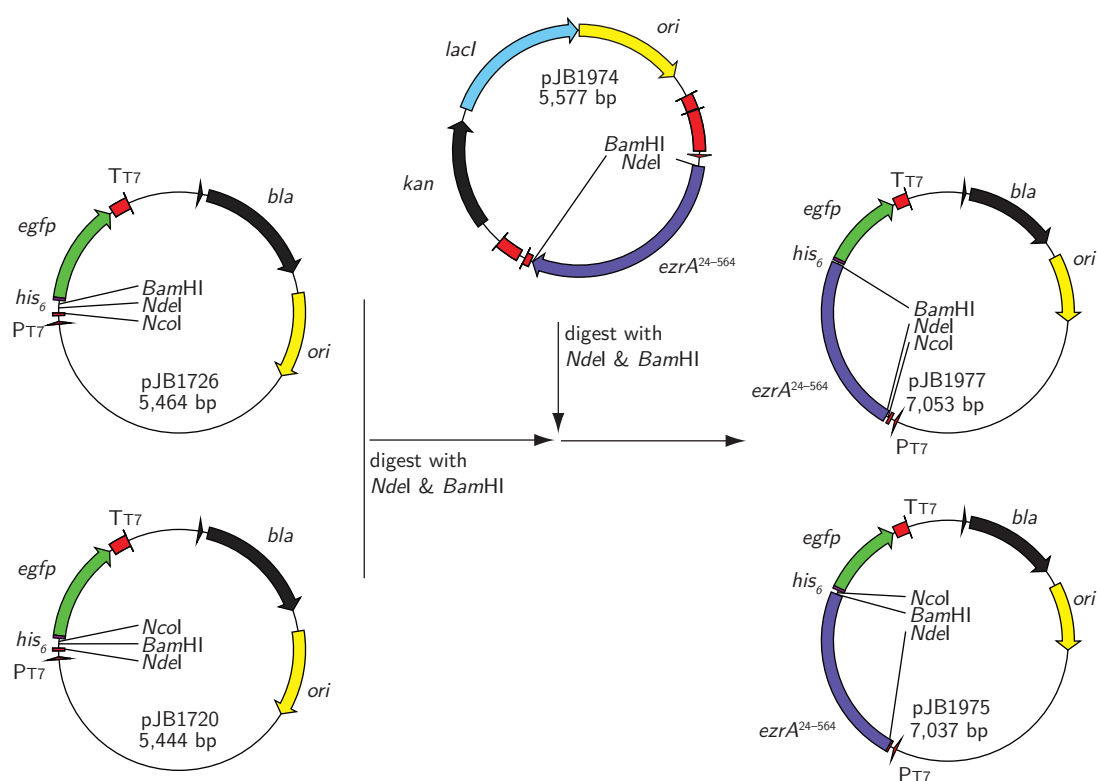
## 5.3 Materials and methods

### 5.3.1 Preparation of *Staphylococcus aureus ezrA* truncation libraries

*E. coli* codon optimised *S. aureus ezrA* was ordered from DNA2.0 (USA) without the 23 residue N-terminal membrane anchor (EzrA<sup>24–564</sup>) in plasmid pJExpress411 (pJB1974). The initiation codon occurs at the unique *Nde*I restriction site and the TAA stop codon is immediately preceded by a unique *Bam*HI site which adds two amino acids (Gly-Ser) to the C-terminus of the expressed EzrA<sup>24–564</sup> protein.



The gene for *EzrA*<sup>24–564</sup> was inserted between the *NdeI* and *BamHI* sites of the C- and N-terminal deletion plasmids pJB1720 and pJB1726 (Figures 3.15B and 3.15A), making pJB1975 and pJB1977 respectively(Figure 5.5).



**Figure 5.5: Plasmids for gene truncation of *Staphylococcus aureus* *ezrA*.** pJB1977 (for N-terminal deletion) and pJB1975 (for C-terminal deletion) were made by placing the *S. aureus* *ezrA*<sup>24–564</sup> gene from pJB1974 into the *NdeI* and *BamHI* sites of pJB1726 and pJB1720. pJB1974 was digested with *NdeI* and *BamHI* and the *ezrA* fragment was gel purified. pJB1726 and pJB1720 express C-terminal His<sub>6</sub> tags and EGFP fusion proteins for protein purification and solubility selection.

To make N- and C- terminal deletion libraries of *ezrA*, 20 µg of purified pJB1977 or pJB1975 were digested with 20 U of *NcoI* in NEB restriction enzyme buffer 3 with 100 µg.mL<sup>-1</sup> BSA for 2 h at 37°C and then the enzyme was heat inactivated by incubation at 80°C for 20 min. Linear plasmids were then end protected with

40  $\mu$ M each of dGTP $\alpha$ S (Glen Research), dATP, dCTP, and dTTP, 20 U DNA polymerase I (Klenow fragment) at 30°C for 15 min. After heat inactivation of DNA polymerase I at 65°C for 20 min, pJB1977 and pJB1975 were prepared for uni-directional *ExoIII* digestion using 20 U *NdeI* or *BamHI*, by incubation for 2 h at 37°C and then the enzymes were heat inactivated by incubation at 80°C. Plasmids containing *ezrA* were then truncated with *ExoIII* (by taking samples of the reaction at 5.0 and 6.5 min), repaired, transformed into *E. coli* and identified as described in Sections 4.3.2 and 4.3.3.

### 5.3.2 Preparation of EzrA truncation-His<sub>6</sub> tagged expression plasmids

Plasmids for expression of truncated EzrA without EGFP were prepared by linearising the selected plasmids with *MfeI* and then ligation of the self-complementary oligonucleotide 420 (AATTGTAAGCTTAC) to form an in-frame stop codon. The resulting plasmids expressed a C-terminal fusion sequence to EzrA of GSSGNSHHHHHHQL\* for N-terminal deletion mutants and HGSSGNSHHHHHHQL\* for C-terminal mutants.

### 5.3.3 Expression, examination of protein solubility and purification of truncation mutants

#### 5.3.3.1 Protein expression and examination of protein solubility

Truncated EzrA mutants identified from library experiments were expressed with a C-terminal His<sub>6</sub>-tag in *E. coli* BL21(λDE3)*recA*/pLysS by auto-induction (Section 2.5.1). Individual 400 mL cultures were incubated with shaking for 24 h at 30°C, after which cells were recovered by centrifugation at  $8,000 \times g$  and portions of the bacterial pellets were resuspended to an  $A_{600}$  of 10 AU in IMAC buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM imidazole) and lysed by passing three times through a chilled French press (Section 2.5.2). Following cell lysis, the soluble and insoluble cellular fractions were separated by centrifugation at  $30,000 \times g$  for 30 min.

Samples of whole cell lysate and soluble protein for each mutant were analysed directly by SDS-PAGE (Section 2.5.9) and insoluble cellular protein was resolubilised in a small volume of 8 M urea then diluted with SDS-loading buffer in proportion to cell lysate and analysed by SDS-PAGE.

#### 5.3.3.2 Purification of truncated EzrA proteins

EzrA mutants were purified by resuspension of cells ( $\sim 4$  g) from over-expressed cultures in IMAC buffer (15 mL per g of cells) and then lysed by passage three times through a chilled French press. Cellular extract was clarified by

centrifugation at  $30,000 \times g$  for 30 min, filtered through a 45  $\mu\text{m}$  syringe filtration device (Merck Millipore) and loaded onto a 1 mL HisTrap HP column (GE Healthcare Life Sciences) connected to an ÄKTApurifier system (GE Healthcare Life Sciences). The column was washed with 10 mL of IMAC buffer, then 10 mL of IMAC buffer with 1 M NaCl and 10 mL further IMAC buffer. Purified truncated EzrA proteins were then eluted immediately using IMAC buffer with 500 mM imidazole, and fractions were pooled (typically 2–5 mL) and diluted to 40 mL in ion exchange buffer TBS<sub>0</sub> (50 mM Tris-HCl pH 8.0, 1 mM DTT and 1 mM EDTA). EzrA proteins were then applied to an 8 mL MonoQ column (GE Healthcare Life Sciences) connected to an ÄKTApurifier and equilibrated with TBS<sub>0</sub>, then eluted with a gradient of 0–400 mM NaCl in TBS<sub>0</sub> over 20 column volumes.

#### 5.3.3.3 Concentration and storage of truncated EzrA proteins

Purified EzrA truncation mutants were dialysed extensively against 20 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA. Protein samples were then concentrated using Amicon Ultra-4 Centrifugal Filter Units (Merck Millipore) with a MWCO of 4,000 Da.

#### 5.3.3.4 Purification of pooled libraries of truncated EzrA

EzrA protein library pools were prepared by mixing plasmids for various His<sub>6</sub>-tagged truncated EzrA proteins. The EzrA plasmid pools were then transformed into BL21( $\lambda$ DE3)*recA* by electroporation and the transformation

pools directly inoculated into 400 mL auto-induction medium (Section 2.5.1). The cultures were grown for 24 h at 30°C and the pool of proteins purified as described in Section 5.3.3.2.

#### 5.3.4 Nuclear magnetic resonance analysis of truncated EzrA

Some EzrA mutants, when purified by anion-exchange chromatography, eluted in two distinct peaks at  $A_{280}$ , containing proteins of identical mass; for these proteins the early eluting species were prepared for NMR analysis. Aliquots of concentrated truncated EzrA mutants (Table 5.2) were dialysed extensively against 20 mM Tris-HCl pH 7.0, 50 mM NaCl, 1 mM DTT and 1 mM EDTA, with EDTA absent from the final dialysis buffer. Truncated EzrA protein aliquots were stored and shipped at liquid nitrogen temperature to Dr Xun-Cheng Su, State Key Laboratory of Elemento-organic Chemistry, Nankai University, Peoples Republic of China. NMR experiments analysed 1D and TOCSY spectra using a Bruker Avance 600 MHz NMR spectrometer with 100 ms mixing times for TOCSY spectra.

#### 5.3.5 Plasmids for over-expression of wild type and biotinylated

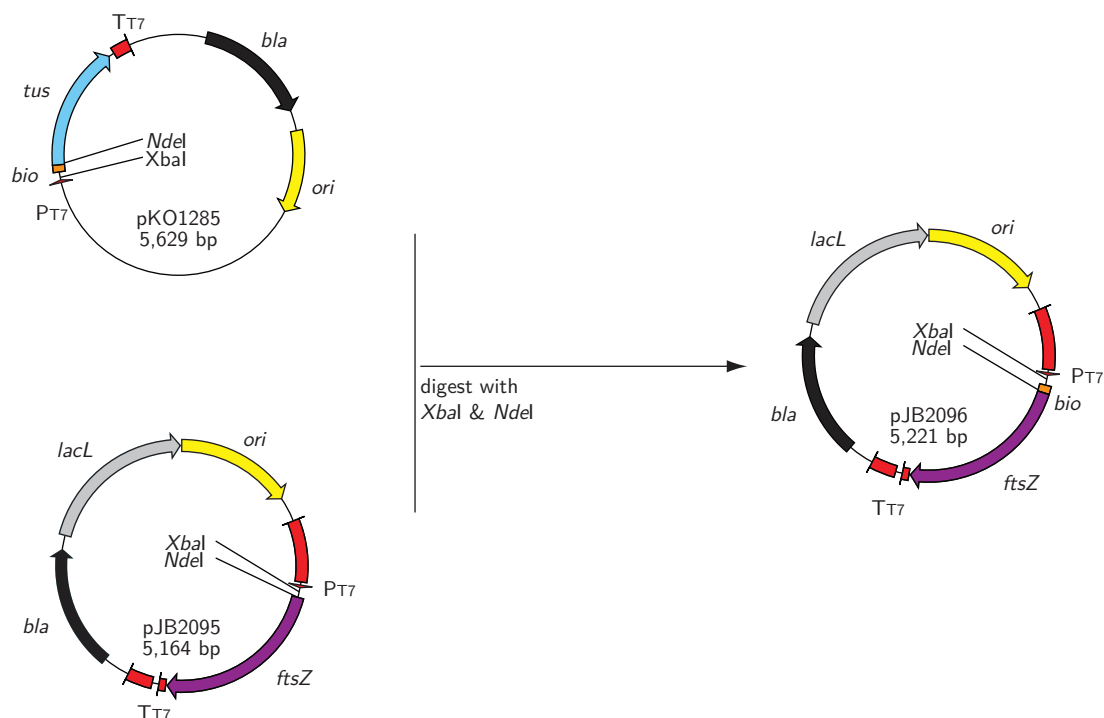
##### *Staphylococcus aureus* cell division protein FtsZ

A synthetic plasmid containing the *E. coli* codon optimised protein coding sequence for *S. aureus* FtsZ was ordered from DNA2.0 in plasmid pJExpress414 (pJB2095); this vector directs expression of FtsZ from a T7 phage promoter, and

**Table 5.2: Concentration of EzrA mutants for analysis by nuclear magnetic resonance.**

Mutant	Concentration
EzrA <sup>24–564</sup>	0.17 mM
EzrA <sup>24–97</sup>	0.3 mM
EzrA <sup>24–126</sup>	0.04 mM
EzrA <sup>24–128</sup>	0.22 mM
EzrA <sup>24–129</sup>	0.3 mM
EzrA <sup>24–139</sup>	0.3 mM
EzrA <sup>24–214</sup>	0.3 mM
EzrA <sup>24–238</sup>	0.3 mM
EzrA <sup>24–476</sup>	0.34 mM
EzrA <sup>277–564</sup>	0.3 mM
EzrA <sup>280–564</sup>	0.24 mM
EzrA <sup>302–564</sup>	0.23 mM
EzrA <sup>381–564</sup>	0.3 mM
EzrA <sup>425–564</sup>	0.16 mM
EzrA <sup>443–564</sup>	0.3 mM
EzrA <sup>484–564</sup>	0.3 mM

the *ftsZ* gene has an *NdeI* site at the start codon. To incorporate DNA encoding the *E. coli* biotinylation sequence (MAGLNDIFEAQK<sup>bio</sup>IEWHEH; Beckett *et al.*, 1999) from pKO1285 (a derivative of pKO1274 which has the *E. coli tus* gene inserted; Jergic *et al.*, 2007) at the start of the coding sequence for FtsZ, the small *XbaI*-*NdeI* restriction fragment of pKO1285 — encoding ribosome-binding site, initiation codon and biotin ligase recognition sequence — was gel purified and substituted for the equivalent fragment from pJB2095, producing pJB2096 (Figure 5.6). The resulting fusion protein, FtsZ<sup>bio</sup>, when biotinylated, incorporates one biotin molecule per molecule of protein.



**Figure 5.6: Plasmid for over-expression of N-terminally biotinylated *Staphylococcus aureus* FtsZ.** The ribosome binding site, initiation codon and *E. coli* biotin ligase recognition sequence from pKO1285 were liberated by digestion with *XbaI* and *NdeI* and ligated between the *XbaI* and *NdeI* sites of pJB2095. pJB2096 directs over-expression of full-length *S. aureus* FtsZ with an N-terminal *E. coli* biotin ligase recognition sequence.

### 5.3.6 Purification of biotinylated *Staphylococcus aureus* FtsZ

pJB2096 was transformed into BL21( $\lambda$ DE3)*recA* and a selected transformant was inoculated into 2.1 L of auto-induction medium and cultured, with shaking, for 24 h at 30°C (Section 2.5.1). FtsZ<sup>bio</sup> containing cells were harvested by centrifugation at  $8,000 \times g$  and resuspended in TBS<sub>150</sub> (50 mM Tris-HCl pH 8.0, 150 mM NaCl and 1 mM EDTA; 15 mL per g of cells), lysed using a chilled French press (Section 2.5.2) and the soluble cell fraction cleared by centrifugation at  $30,000 \times g$  for 30 min at 4°C.

A previous study identified two FtsZ populations when *E. coli* FtsZ was over-expressed in a similar manner to that used here (Lu *et al.*, 1998). The two populations have differing GTPase activity and were discovered as they precipitate at differing ammonium sulphate concentrations (low activity, 0.11 g.mL<sup>-1</sup>; high activity, 0.14 g.mL<sup>-1</sup>). The source of the difference between the two populations of FtsZ is unknown as the two populations gave identical mass by ESI mass spectrometry, and exhibited identical polymerisation characteristics (Lu *et al.*, 1998). Our intent for the FtsZ protein was to identify protein-protein interactions, so the FtsZ purification procedure did not intend to separate the two species (if they exist for the *S. aureus* protein). FtsZ<sup>bio</sup> was precipitated from the clarified cell lysate by addition of 0.2 g.mL<sup>-1</sup> ammonium sulphate, which precipitates the majority of over-expressed FtsZ<sup>bio</sup>. Precipitated protein was separated by centrifugation at 30,000 × *g* for 30 min at 4°C and then resuspended in 160 mL of TBS<sub>150</sub>.

DNA was removed from 40 mL portions of the recovered FtsZ<sup>bio</sup> containing solution by flowing it through a 35 mL DEAE chromatography column (Toyopearl DEAE-650M) in TBS<sub>150</sub> using an ÄKTApurifier and the FtsZ<sup>bio</sup> containing unbound fraction was collected. The relatively pure FtsZ<sup>bio</sup> solution was then pooled, concentrated by precipitation with 0.5 g.mL<sup>-1</sup> ammonium sulphate and then collected by centrifugation at 30,000 × *g* for 30 min. Ammonium sulphate precipitated FtsZ<sup>bio</sup> was resuspended in TBS<sub>150</sub> with 1 mM DTT and 1 mM EDTA and then dialysed against the same buffer before freezing in aliquots in liquid nitrogen and storage at -80°C.



#### 5.3.6.1 Purification of N-terminally His<sub>6</sub> tagged *Escherichia coli* biotin ligase

N-terminally His<sub>6</sub> tagged *E. coli* biotin ligase (BirA) was expressed from pKO1298 (Jergic *et al.*, 2007) in BL21(λDE3)*recA* by auto-induction, with shaking, for 24 h at 30°C in a 400 mL culture (Section 2.5.1). Cells were collected by centrifugation at  $8,000 \times g$  and resuspended in IMAC buffer (50 mM Tris-HCl, 300 mM NaCl and 20 mM imidazole; pH 8.0; 15 mL per g of cells) and lysed using a chilled French press (Section 2.5.2). Insoluble components were separated from the soluble cell lysate by centrifugation at  $30,000 \times g$  for 30 min at 4°C.

Clarified cell lysate (30 mL) containing BirA was filtered through a 45 µm syringe filtration device and applied to a 1 mL HisTrap HP column, washed with 10 column volumes of IMAC buffer, then washed with 10 column volumes of IMAC buffer with 1M NaCl, a further 5 column volumes of IMAC buffer and then eluted using IMAC buffer with 500 mM imidazole. Fractions containing His<sub>6</sub>-BirA were pooled, immediately adjusted to 1 mM EDTA and 1 mM DTT and then dialysed against 50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA and 1 mM DTT (pH 8.0). Purified His<sub>6</sub>-BirA aliquots were frozen in liquid nitrogen and stored at −80°C.

#### 5.3.6.2 *In vitro* biotinylation of *Staphylococcus aureus* FtsZ

FtsZ<sup>bio</sup> was dialysed to 50 mM Bicine pH 8.3 for conjugation to biotin using *E. coli* biotin ligase. *In vitro* biotinylation of 40 µM FtsZ<sup>bio</sup> was performed in 10 mL of 50 mM Bicine pH 8.3, 10 mM ATP, 10 mM Mg(OAc)<sub>2</sub>, 50 µM D-biotin with 100

µg BirA for 30 min at 30°C (Kay *et al.*, 2009). Following *in vitro* biotinylation, FtsZ<sup>bio</sup> was dialysed to 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM DTT and 1 mM EDTA.

#### 5.3.6.3 *In vitro* EzrA pull-down using biotinylated-FtsZ bait

To identify EzrA fragments that are capable of interacting with full length FtsZ, *in vitro* biotinylated FtsZ<sup>bio</sup> was attached to streptavidin-coated resin and incubated with various EzrA mutants (Section 5.3.3.4). Pull-down assays using FtsZ-coated resin were performed in 600 µL sample tubes and buffers exchanged by centrifugation of the samples in a small bench-top centrifuge at  $2,000 \times g$  for 5 s, then removing the supernatant with a pipette. FtsZ<sup>bio</sup> was incubated with 7.5 µL Pierce High-capacity streptavidin resin (maximal binding capacity up to 37.5 µg biotinylated FtsZ) and incubated on ice in TBS<sub>150</sub> (50 mM Tris-HCl pH 8.0, 150 mM NaCl) for 30 min, then thoroughly washed with TBS<sub>150</sub>. To reduce non-specific protein interactions with the streptavidin resin, FtsZ<sup>bio</sup>-coated beads were then incubated with 4 nmol D-biotin and 40 mg BSA for 30 min on ice, then thoroughly washed with TBS<sub>150</sub>. FtsZ<sup>bio</sup>-coated resin was then exchanged to ice-cold TBS<sub>50</sub> (50 mM Tris-HCl pH 8.0, 50 mM NaCl). Samples of EzrA proteins were added in excess and incubated in TBS<sub>50</sub> for 30 min on ice, after which resin was collected by centrifugation at  $2,000 \times g$  for 5 s, and the buffer exchanged for fresh ice-cold TBS<sub>50</sub> twice. Once the FtsZ-coated resin had been sufficiently washed, 15 µL SDS-loading dye was added and the mixture incubated at 80°C for 10 min prior to analysis by SDS-PAGE.

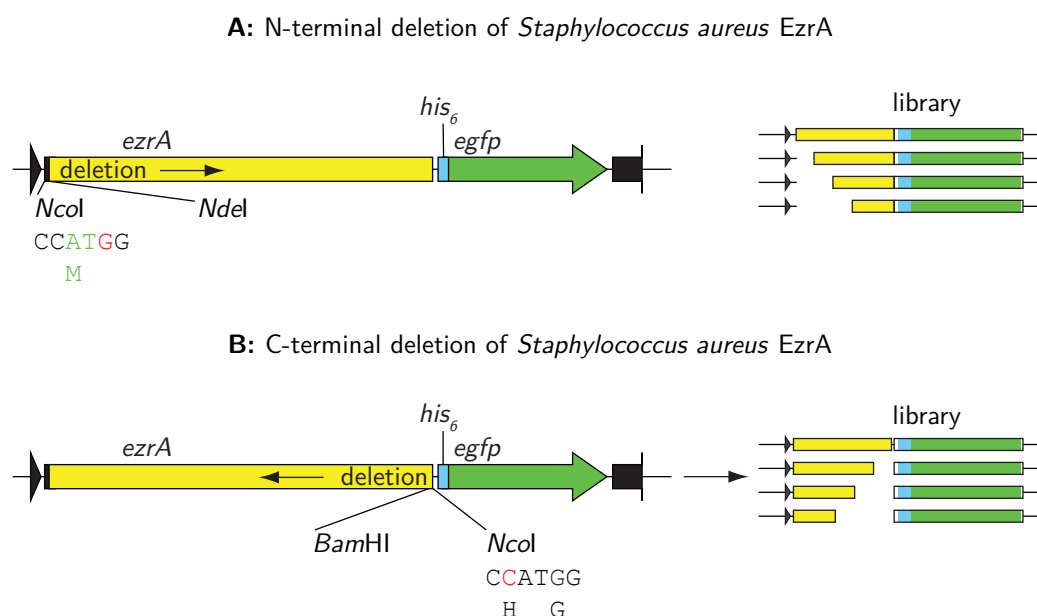
## 5.4 Results

### 5.4.1 Genetic truncation of *Staphylococcus aureus* EzrA

In order to gain insight into the domain organisation of EzrA, both N- and C-terminal truncation was performed. EzrA is annotated in Pfam as containing five coiled-coil domains (Figure 5.4), hinting that it contains up to six distinctly folded regions. In the protein truncation libraries characterised in Chapter 4, truncations targeting removal of the N-terminal domain of DnaG unexpectedly provided soluble proteins outside of the region targeted by the limited *ExoIII* digest; gene deletion using *ExoIII* is known to produce mutants that are disproportionately over and under truncated (Hoheisel, 1993; Ostermeier, 2003). To elucidate soluble EzrA protein domains, a truncation methodology was chosen such that the majority of gene products of truncation would be for protein sizes likely to comprise one or two terminal folded domains. This methodology alone should provide domain break points for up to four EzrA domains. With sufficient mutant sampling, identification of under-truncated soluble protein should then hint at soluble truncations for the remaining protein domain(s), if they exist.

N- and C-terminal EzrA deletion libraries were produced from limited-*ExoIII* digests of linearised pJB1977 and pJB1975 (5.0 and 6.5 min with a rate of 200 bp.min<sup>-1</sup>) as described in Section 5.3.1. *NcoI* digested pJB1977 and pJB1975 were protected from *ExoIII* hydrolysis at both ends by incorporation of dGTP $\alpha$ S and made sensitive to *ExoIII* by removing a single protected end using *NdeI* or *BamHI* (Figure 5.7). For N-terminal deletion of EzrA, the remaining  $\alpha$ -S

protected *Nco*I site initiates translation in pJB1977 and following ligation of limit *Exo*III truncation (pooled 5.0 and 6.5 min samples), directs over-expression of N-terminally truncated EzrA-EGFP fusion proteins (Figure 5.7A). For C-terminal deletion of EzrA, removal of the *Nco*I–*Bam*HI DNA fragment deletes stop codons from all three reading frames as well as the  $\alpha$ -S protected *Nco*I site of pJB1975; ligation of products of limit *Exo*III truncation produces plasmids where 3'-terminally truncated *ezrA* genes are fused in-frame to the downstream *egfp* sequence (Figure 5.7B).



**Figure 5.7: Methodology for 5'- and 3'-truncation of *Staphylococcus aureus* *ezrA*.** **A**, 5'-terminally and **B**, 3'-terminally deleted genes were prepared from pJB1977 or pJB1975 (respectively) digested with *Nco*I and then protected with dGTP $\alpha$ S (red) and dATP, dCTP and dTTP (green). These linearised plasmids were then digested with *Nde*I (N-terminal deletion) or *Bam*HI (C-terminal deletion) to remove the protected end most adjacent to the *ezrA* gene to allow uni-directional deletion. The remaining filled *Nco*I site provides the start codon for 5'-terminally truncated genes or the start of the 3'-terminal gene fusion for 3'-terminally truncated genes. After limit digests using *Exo*III and plasmid repair, one-in-three truncated gene open-reading-frames are in-frame with the downstream *egfp* gene, producing a truncated EzrA protein fused to EGFP when expressed in *E. coli*.

Separately, limit *ExoIII* digests of pJB1977 and pJB1975 were ligated and transformed into BL21( $\lambda$ DE3)*recA*, recovered in LB, then spread onto 10 Selection agar plates (Section 4.3.2.5) and incubated overnight at 30°C. The transformed cultures were stored at 4°C for 48 h as previous evidence suggests that colonies further develop their green fluorescent phenotype during this time. The N-terminally deleted library made from pJB1977 produced 4,381 transformants, of which 140 had a green fluorescent phenotype, while the C-terminally deleted library produced from pJB1975 gave 2,955 colonies, of which 73 produced a green fluorescent phenotype. Previous protein truncation libraries produced in this work (see Chapter 4), showed a correlation between green fluorescent phenotype and protein solubility, and this is supported by previous reported work (Waldo *et al.*, 1999; Pédelacq *et al.*, 2005). Relying on the principle that interrupting protein folding units should reduce protein solubility, sequencing of truncated *ezrA* mutants with the strongest green fluorescent phenotypes was expected to provide insight into the domain architecture of EzrA. As a result 49 5'-terminally deleted (Table 5.3) and 33 3'-terminally deleted (Table 5.4) *ezrA* mutants with a bright green fluorescent phenotype were sequenced to identify the truncated proteins they express.

**Table 5.3: N-terminally deleted *S. aureus* EzrA mutants.** Plasmids from green colonies expressing truncated EzrA-EGFP fusion proteins were sequenced to identify the mutations present. Displayed at the **top** are the DNA and protein sequences of untruncated pJB1977 (*wt*/WT) that was used to generate truncated *ezrA* genes fused to *egfp*; shown in red are the nucleotides removed during linearisation for uni-directional truncation. f0–1 indicate the reading frames present in pJB1977; f0; reading frame of initiation codon, f1; reading frame of non-truncated *ezrA* in pJB1977. Shown in the **body** are DNA and protein sequences of truncated EzrA mutants. Six mutants did not provide useful DNA sequence and 12 contained no or short *ezrA* sequences (not shown).

<i>wt</i>	CCATG	GCGACCTCGTGA	–/19 bp	ATGGCTATTCTCAGCATACCATTTGATCAA	WT
f0	M	A	T	S	*
f1				M R S N K R Q I I E	
DNA sequence					Plasmid
Protein sequence					Mutation
	CCATG	CAGATCATCGAAAAAGCAATAGAGAGAAAGAATGAAATT			pJB2052
	M	Q I I E K A I E R K N E			29–564
	CCATG	GTGGAAGGTTACGATCTGGATCATGTGAAAGTTGACAGC			pJB2029
	M	V E G Y D L D H V K V D			244–564
	CCATG	CTGGAAGAGGCGAACGATAAACTGGCGAACATCAACGAC			pJB2046
	M	L E E A N D K L A N I N			277–564
	CCATG	GCGAACGATAAACTGGCGAACATCAACGACAAGCTGGAC			pJB2045
	M	A N D K L A N I N D K L			280–564
	CCATG	GTGAAAGCCAAGAACGACGTCGAAGAAACCAAGGATATC			pJB2050
	M	V K A K N D V E E T K D			302–564
	CCATG	GTTCAGGACAACTTGCAGTACCTGGAGGATCACGTTACC			pJB2024
	M	V Q D N L Q Y L E D H V			381–564
	CCATG	ACCGTAATCAACGATAAGCAGGAGAAGTTGCAAAACCAT			pJB2042
	M	T V I N D K Q E K L Q N			393–564
	CCATG	TCGAAGAAAGAAGAAGTGTATCGTCGCTTGCTGGCTTCC			pJB2044
	M	S K K E E V Y R R L L A			425–564

Continued on next page...

**Table 5.3** – continued from previous page.

<i>wt</i>	CCATG	GCGACCTCGTGA- /19 bp	ATGGCTATTCTCAGCATACCATTGATCAA	WT
<i>f0</i>	M	A	T	S *
DNA sequence				Plasmid
Protein sequence				Mutation
CCATG	CCAGAGCGCTTCATTATCATGAAAAACGAGATTGATCAT			pJB2031
M	P E R F I I M K N E I D			443–564
CCATG	AAAAACGAGATTGATCATGAGGTTCGTGACGTTAATGAG			pJB2036
M	K N E I D H E V R D V N			450–564
CCATG	ATTGATCATGAGGTTCGTGACGTTAATGAGCAGTTTTCT			pJB2021
M	I D H E V R D V N E Q F			453–564
CCATG	ATTCACGTCAAGCAGCTGAAGGATAAAGTCTCTAAGATT			pJB2051
M	I H V K Q L K D K V S K			469–564
CCATG	GATAAAGTCTCTAAGATTGTCATTCAAATGAACACCTTC			pJB2025
M	D K V S K I V I Q M N T			476–564
CCATG	GTCTCTAAGATTGTCATTCAAATGAACACCTTCGAGGAC			pJB2038
M	V S K I V I Q M N T F E			478–564
CCATG	CAAATGAACACCTTCGAGGACGAAGCGAACGACGTGCTG			pJB2039
M	Q M N T F E D E A N D V			484–564
CCATG	AGCAACGTTGACAAGTCTTTGAATGAGGCGGAGCGTCTG			pJB2040
M	S N V D K S L N E A E R			517–564
CCATG	GTTGACAAGTCTTTGAATGAGGCGGAGCGTCTGTTTAAG			pJB2028
M	V D K S L N E A E R L F			519–564
CCATG	GACAAGTCTTTGAATGAGGCGGAGCGTCTGTTTAAGAAC			pJB2035
M	D K S L N E A E R L F K			520–564
CCATG	TCTTTGAATGAGGCGGAGCGTCTGTTTAAGAACAATCGT			pJB2022
M	S L N E A E R L F K N N			522–564

Continued on next page...

**Table 5.3** – continued from previous page.

<i>wt</i>	CCATG	GCGACCTCGTGA- /19 bp	ATGGCTATTCTCAGCATACCATTGATCAA	WT
<i>f0</i>	M	A	T	S *
DNA sequence				Plasmid
Protein sequence				Mutation
CCATG	AAGAACAATCGTTACAAACGCGCCATTGAAATCGCTGAA			pJB2030
M	K N N R Y K R A I E I A			531–564
CCATG	AAGAACAATCGTTACAAACGCGCCATTGAAATCGCTGAA			pJB2041
M	K N N R Y K R A I E I A			531–564
CCATG	AAGAACAATCGTTACAAACGCGCCATTGAAATCGCTGAA			pJB2047
M	K N N R Y K R A I E I A			531–564
CCATG	AATCGTTACAAACGCGCCATTGAAATCGCTGAACAGGTG			pJB2026
M	N R Y K R A I E I A E Q			533–564
CCATG	AATCGTTACAAACGCGCCATTGAAATCGCTGAACAGGTG			pJB2032
M	N R Y K R A I E I A E Q			533–564
CCATG	CGCGCCATTGAAATCGCTGAACAGGTGCTGGAGAGCGTT			pJB2027
M	R A I E I A E Q V L E S			537–564
CCATG	GAAATCGCTGAACAGGTGCTGGAGAGCGTTGAACCGGGC			pJB2020
M	E I A E Q V L E S V E P			540–564
CCATG	GAAATCGCTGAACAGGTGCTGGAGAGCGTTGAACCGGGC			pJB2043
M	E I A E Q V L E S V E P			540–564
CCATG	GTGCTGGAGAGCGTTGAACCGGGCGTCACGAAACACATC			pJB2023
M	V L E S V E P G V T K H			545–564
CCATG	AGCGTTGAACCGGGCGTCACGAAACACATCGAAGAAGAG			pJB2048
M	S V E P G V T K H I E E			548–564
CCATG	AGCGTTGAACCGGGCGTCACGAAACACATCGAAGAAGAG			pJB2049
M	S V E P G V T K H I E E			548–564

Continued on next page. . .



**Table 5.3** – continued from previous page.

<i>wt</i>	CCAT <b>G</b> GCGACCTCGTGA-/19 bp	ATGGCTATTCCTCAGCATACCATTTGATCAA	WT
<i>f0</i>	M A T S *		
DNA sequence			Plasmid
Protein sequence			Mutation
	CCAT <b>G</b>	AAACACATCGAAGAAGAGGTGATCAAGCAA	pJB2037
	M	K H I E E E V I K Q	555–564

**Table 5.4: C-terminally deleted *S. aureus* EzrA.** Plasmids from green colonies expressing truncated EzrA-EGFP fusion proteins were sequenced to identify the mutations present. Shown at the **top** are the DNA and protein sequence of untruncated pJB1975 (*wt*/WT) that was used to generate truncated *ezrA* genes fused to *egfp*. f0–2 indicates the reading frames present in pJB1975. Shown in red are the nucleotides removed during linearisation for uni-directional truncation. Shown in the **body** are DNA and protein sequences of truncated EzrA mutants. Seven mutants did not produce readable DNA sequence and two contained non-truncated EzrA (not shown).

<i>wt</i>	GAGGTGATCAAGCAAGGATCCTAAGTAACTAAC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	WT
f0	E V I K Q G S * V T N	H G S S G N S H H H H H	
f1		+1 *	
f2		+2 *	
DNA sequence			Plasmid
Protein sequence			Mutation
AGCAACAAGAGACAGATCATCGAAAAAGCAATAGAG	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2001	
S N K R Q I I E K A I E	H G S S G N S H H H H H	24–36	
GAAAAAGCAATAGAGAGAAAGAATGAAATTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2004	
E K A I E R K N E I E T	H G S S G N S H H H H H	24–43	
GCGCAACTGAGCAAAGTGAATTTGAAAGGTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2013	
A Q L S K L N L K G E T	H G S S G N S H H H H H	24–62	
GCGCAACTGAGCAAAGTGAATTTGAAAGGTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1995	
A Q L S K L N L K G E T	H G S S G N S H H H H H	24–62	
GCGCAACTGAGCAAAGTGAATTTGAAAGGTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1998	
A Q L S K L N L K G E T	H G S S G N S H H H H H	24–62	
GCGCAACTGAGCAAAGTGAATTTGAAAGGTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2002	
A Q L S K L N L K G E T	H G S S G N S H H H H H	24–62	
GCGCAACTGAGCAAAGTGAATTTGAAAGGTGAAACC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2003	
A Q L S K L N L K G E T	H G S S G N S H H H H H	24–62	
CTGAGCAAAGTGAATTTGAAAGGTGAAACCAAACT	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1994	
L S K L N L K G E T K T	H G S S G N S H H H H H	24–64	

Continued on next page...

**Table 5.4** – continued from previous page.

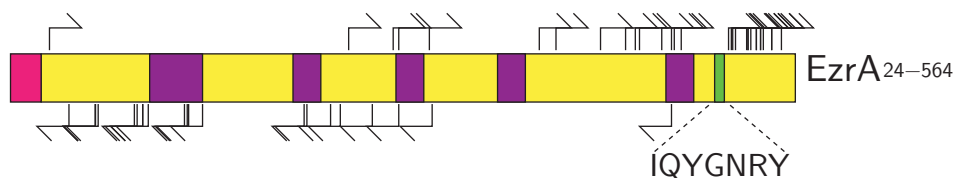
<i>wt</i>	GAGGTGATCAAGCAAGGATCCTAAGTAAC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	WT
<i>f0</i>	E V I K Q G S * V T N	H G S S G N S H H H H H	
DNA sequence			Plasmid
Protein sequence			Mutation
AAGTATCTGGCACCTGTTGAAGAGAAAATTCACAAT	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1993	
K Y L A P V E E K I H N	H G S S G N S H H H H H	24–90	
GCACCTGTTGAAGAGAAAATTCACAATGCCGAAGCG	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1982	
A P V E E K I H N A E A	H G S S G N S H H H H H	24–93	
GAGAAAATTCACAATGCCGAAGCGCTGCTGGATAAG	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1990	
E K I H N A E A L L D K	H G S S G N S H H H H H	24–97	
CACAATGCCGAAGCGCTGCTGGATAAGTTTAGCTTC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2006	
H N A E A L L D K F S F	H G S S G N S H H H H H	24–100	
GACAGCTATGAGCAGAGCTACCAGCAGCAACTGGAG	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1986	
D S Y E Q S Y Q Q Q L E	H G S S G N S H H H H H	24–126	
TATGAGCAGAGCTACCAGCAGCAACTGGAGGACGTT	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2010	
Y E Q S Y Q Q Q L E D V	H G S S G N S H H H H H	24–128	
GAGCAGAGCTACCAGCAGCAACTGGAGGACGTTAAC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2005	
E Q S Y Q Q Q L E D V N	H G S S G N S H H H H H	24–129	
GTTAACGAGATCATCGCACTGTACAAAGATAACGAC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1988	
V N E I I A L Y K D N D	H G S S G N S H H H H H	24–139	
GCGGCACTGAATGAGCAAATGAAACAGCTGCGTAGC	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2012	
A A L N E Q M K Q L R S	H G S S G N S H H H H H	24–212	
CTGAATGAGCAAATGAAACAGCTGCGTAGCTATATG	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB1989	
L N E Q M K Q L R S Y M	H G S S G N S H H H H H	24–214	
CTGATTCGTGAAACCCAAAAAGAACTGCCGGGTCAA	CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT	pJB2011	
L I R E T Q K E L P G Q	H G S S G N S H H H H H	24–231	

Continued on next page. . .

Table 5.4 – continued from previous page.

<i>wt</i>	GAGGTGATCAAGCAAGGATCCTAAGTAACTAAC												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT										WT		
<i>f0</i>	E	V	I	K	Q	G	S	*	V	T	N		H	G	S	S	G	N	S	H	H	H	H	H	
DNA sequence																								Plasmid	
Protein sequence																								Mutation	
GAACTGCCGGGTCAATTCCAAGACCTGAAGTATGGT												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT												pJB1985	
E L P G Q F Q D L K Y G												H G S S G N S H H H H H H												24–238	
GATCATGTGAAAGTTGACAGCACCTTGCAGAGCCTG												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT												pJB1984	
D H V K V D S T L Q S L												H G S S G N S H H H H H H												24–261	
GAGCCGCTGATTAGCCGTCTGGAAGTGAAGAGGCG												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT												pJB1981	
E P L I S R L E L E E A												H G S S G N S H H H H H H												24–280	
GATATGTATGATTTGATTGAGCACGAGGTGAAAGCC												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT												pJB2009	
D M Y D L I E H E V K A												H G S S G N S H H H H H H												24–304	
TCTGAACGCCCCGATTCACGTCAAGCAGCTGAAGGAT												CATGGCTCCTCTGGGAATTCTCATCACCATCACCAT												pJB1992	
S E R P I H V K Q L K D												H G S S G N S H H H H H H												24–476	

As expected, many of the truncated EzrA mutants producing a green fluorescent phenotype when expressed as a fusion to EGFP are clustered at the ends of the protein where the libraries were focussed (Figure 5.8). Strikingly, many truncations within EzrA produce putatively soluble proteins, and with few exceptions, fusion proteins of putatively soluble mutants are within 30 amino acids of one-another. These results indicate that many possible truncation mutations in *S. aureus* EzrA can be tolerated, which may suggest that it is actually a poorly-structured protein throughout much of its length.



**Figure 5.8: EzrA truncations.** Truncation end points of EzrA truncation libraries with a C-terminal EGFP protein-solubility reporter. Arrows indicate: **above**, fluorescence-selected N-terminal deletions and **below**, fluorescence-selected C-terminal deletions of EzrA. Also shown: N-terminal membrane anchor in pink, predicted coiled-coil-folds in purple and QNR patch in green.

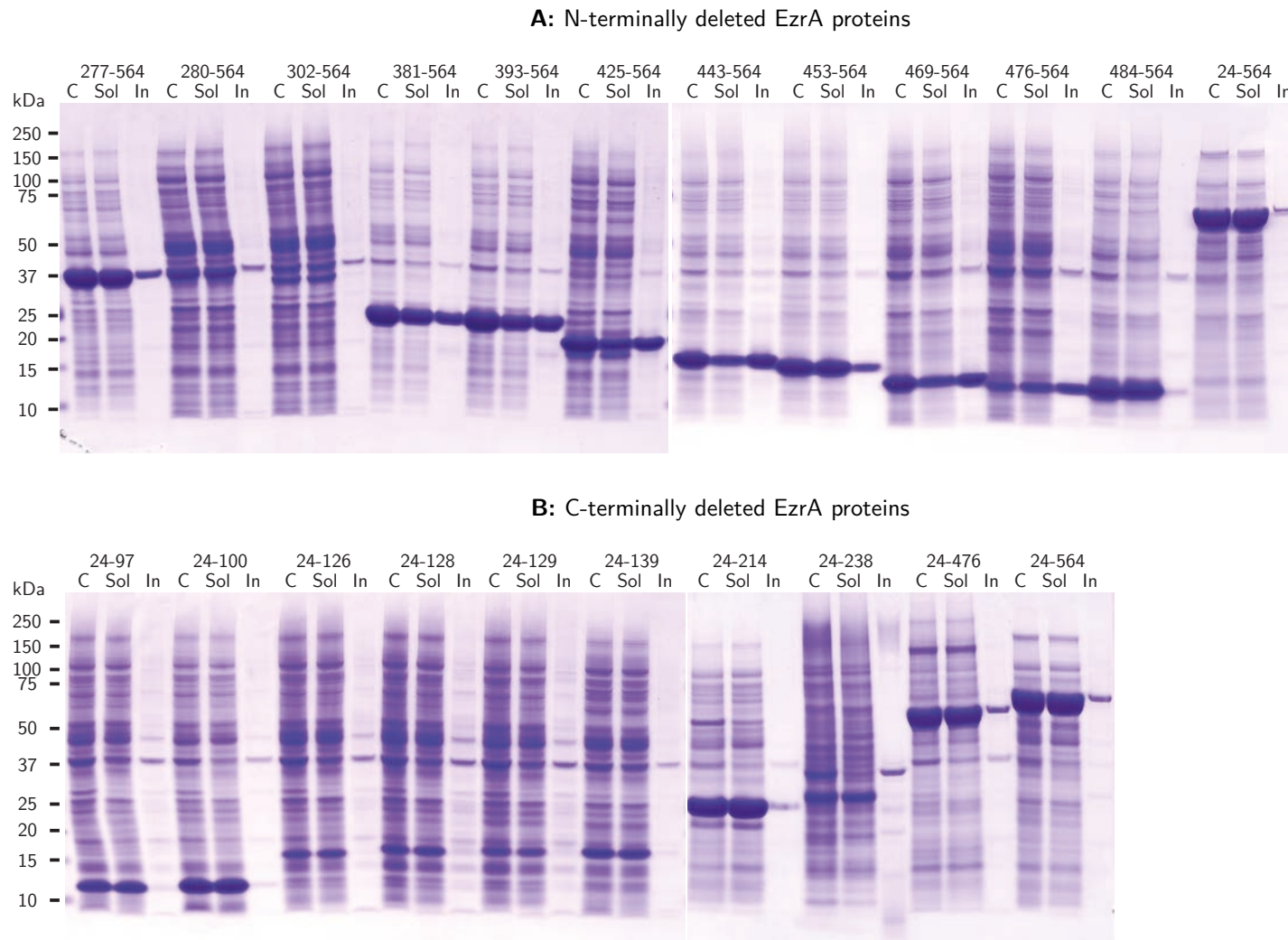
#### 5.4.2 Examination of solubility of truncated EzrA mutants

To explore the solubility of expressed, truncated EzrA mutants, a variety of the truncated *ezrA* plasmids were modified to no longer express the C-terminal EGFP fusion, but still contain a C-terminal His<sub>6</sub>-tag (N-terminal deletions of EzrA, pJB2082–pJB2093; C-terminal deletions of EzrA, pJB2070–pJB2081; Table 5.5). These proteins were over-expressed by auto-induction in BL21(λDE3)*recA*.

**Table 5.5: Identity of putatively soluble truncated *Staphylococcus aureus* *ezrA* plasmids.** Plasmids directing expression of selected *S. aureus* EzrA mutants were modified to express a C-terminal His<sub>6</sub>-tag without an EGFP fusion. C-terminal protein sequences are: for EzrA<sup>24–564</sup> and N-terminally deleted EzrA, GSSGNSHHHHHHQL\* and for C-terminally deleted EzrA HGSSGNSHHHHHHQL\*. EzrA mutants are numbered as in the full-length *S. aureus* protein.

Deletion mutant	Molecular weight (kDa)	Plasmid
Full-length		
EzrA <sup>24–564</sup>		pJB2094
N-terminal deletions		
EzrA <sup>277–564</sup>		pJB2082
EzrA <sup>280–564</sup>		pJB2083
EzrA <sup>302–564</sup>		pJB2084
EzrA <sup>381–564</sup>		pJB2085
EzrA <sup>393–564</sup>		pJB2086
EzrA <sup>425–564</sup>		pJB2087
EzrA <sup>443–564</sup>		pJB2088
EzrA <sup>453–564</sup>		pJB2089
EzrA <sup>469–564</sup>		pJB2090
EzrA <sup>476–564</sup>		pJB2091
EzrA <sup>484–564</sup>		pJB2093
C-terminal deletions		
EzrA <sup>24–97</sup>		pJB2070
EzrA <sup>24–100</sup>		pJB2071
EzrA <sup>24–126</sup>		pJB2072
EzrA <sup>24–128</sup>		pJB2073
EzrA <sup>24–129</sup>		pJB2074
EzrA <sup>24–139</sup>		pJB2075
EzrA <sup>24–214</sup>		pJB2076
EzrA <sup>24–231</sup>		pJB2077
EzrA <sup>24–238</sup>		pJB2078
EzrA <sup>24–476</sup>		pJB2081

Individual cultures of strains that over-expressed truncated EzrA mutants were collected by centrifugation and lysed using a French press. The cell lysates were separated into soluble and insoluble components by centrifugation and then examined by SDS-PAGE (Figure 5.9). All mutants directed expression of



**Figure 5.9: Over-expression and protein solubility of EzrA N- and C-terminally deleted mutants.** **A**, N-deleted and **B**, C-deleted EzrA mutants were over-expressed in BL21( $\lambda$ DE3)*recA* by auto-induction at 30°C. After cell lysis by French press, the: C; cellular; Sol; soluble and In; insoluble fractions were each analysed by SDS-PAGE. Migration of Precision Plus Protein Dual Color molecular size markers are indicated.

soluble protein of the expected sizes except for EzrA<sup>24-231</sup>, which did not produce detectable protein over-expression (not shown).

Full-length EzrA<sup>24-564</sup> was over-expressed well and appeared almost exclusively in the soluble protein fraction. Of the N-terminally truncated EzrA mutants, the largest fragments examined — EzrA<sup>277-564</sup> and EzrA<sup>280-564</sup> — are  $\sim 35$  kDa in size, and wholly contain the three predicted C-terminal coiled-coil domains and the QNR patch; EzrA<sup>302-564</sup> is truncated slightly into the first predicted coiled-coil fold of these mutants. EzrA<sup>277-564</sup> was over-expressed very well with good yields of soluble protein, whereas EzrA<sup>280-564</sup> was expressed poorly in comparison and had a reduced proportion of soluble EzrA. Expression of EzrA<sup>302-564</sup> was poorer still. EzrA<sup>381-564</sup> and EzrA<sup>393-564</sup> are missing two predicted coiled-coil domains with respect to EzrA<sup>277-564</sup>; these proteins were over-expressed well and significant amounts of soluble protein were present in the cell extracts, with about 50% of each present in the insoluble protein fraction.

EzrA<sup>425-564</sup> (18 kDa) was over-expressed at high levels, with an apparent proteolytic fragment (1–2 kDa smaller) also apparent in the soluble fraction, but not in the insoluble fraction. The proteolytic fragment of EzrA<sup>425-564</sup> was similar in size to EzrA<sup>443-564</sup> (16 kDa) and EzrA<sup>453-564</sup> (15 kDa), which were also over-expressed well and were soluble. The presence solely in the soluble fraction of the proteolytic fragment of EzrA<sup>425-564</sup> would imply that the insoluble EzrA<sup>425-564</sup> produced by over-expression is aggregated and therefore inaccessible to cellular proteases. Cleavage of the soluble fraction at a flexible region produces the more compact proteolysed fragment, and this fragment remains soluble. The remaining



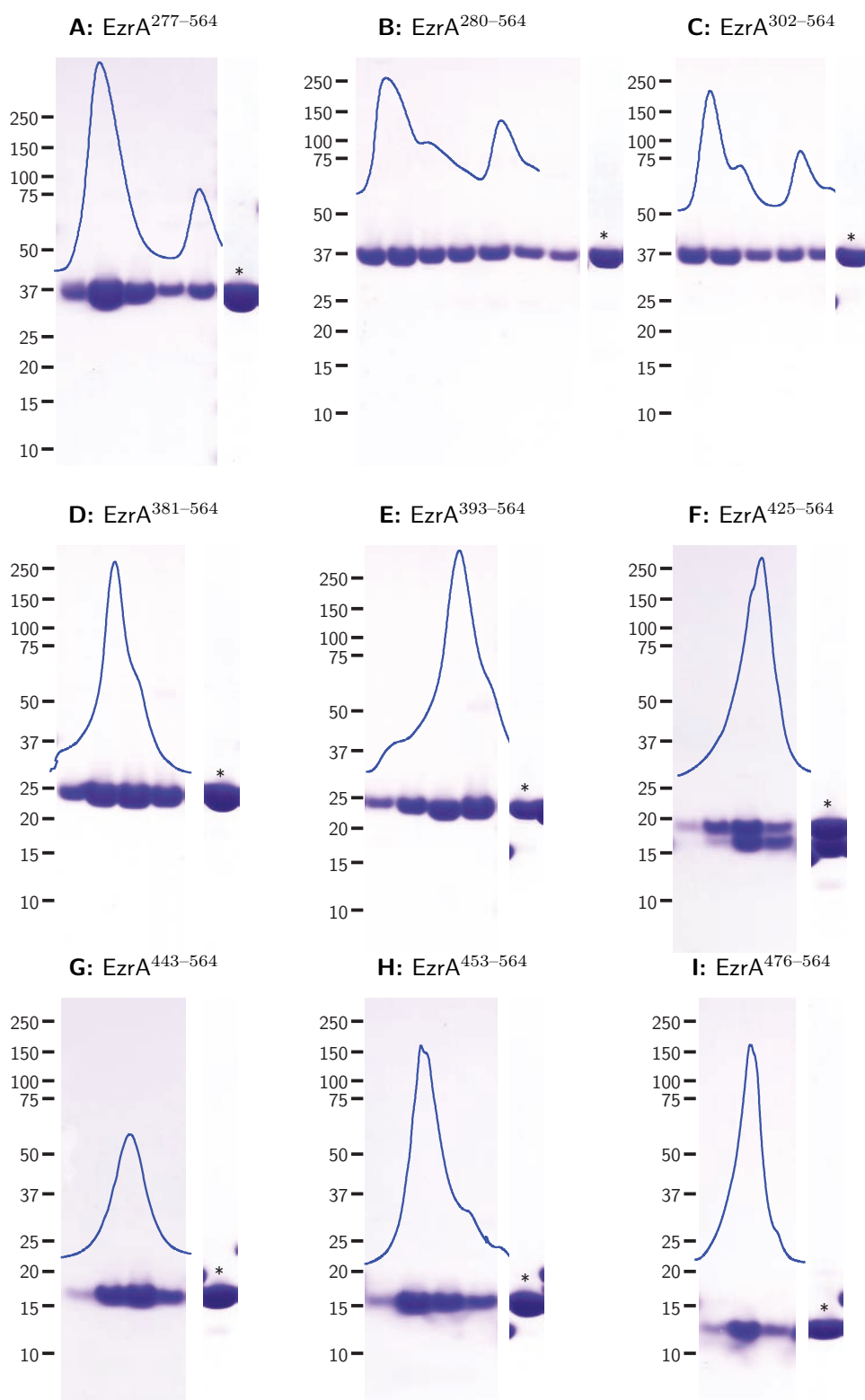
C-terminal EzrA fragments EzrA<sup>453–564</sup>, EzrA<sup>469–564</sup>, EzrA<sup>476–564</sup> and EzrA<sup>484–564</sup> were over-expressed in soluble form but their N-termini are not clustered near distinct regions. EzrA<sup>484–564</sup> also showed evidence of proteolysis.

Of the C-terminally deleted EzrA proteins examined, the small 10–15 kDa mutants EzrA<sup>24–97</sup>, EzrA<sup>24–100</sup>, EzrA<sup>24–126</sup>, EzrA<sup>24–128</sup>, EzrA<sup>24–129</sup> and EzrA<sup>24–139</sup> were over-expressed to moderate levels, and were predominantly found in the soluble cellular fraction. The larger EzrA<sup>24–214</sup> (24 kDa), EzrA<sup>24–238</sup> (27 kDa) and EzrA<sup>24–476</sup> (55 kDa) proteins were also expressed well and were mostly produced as soluble proteins.

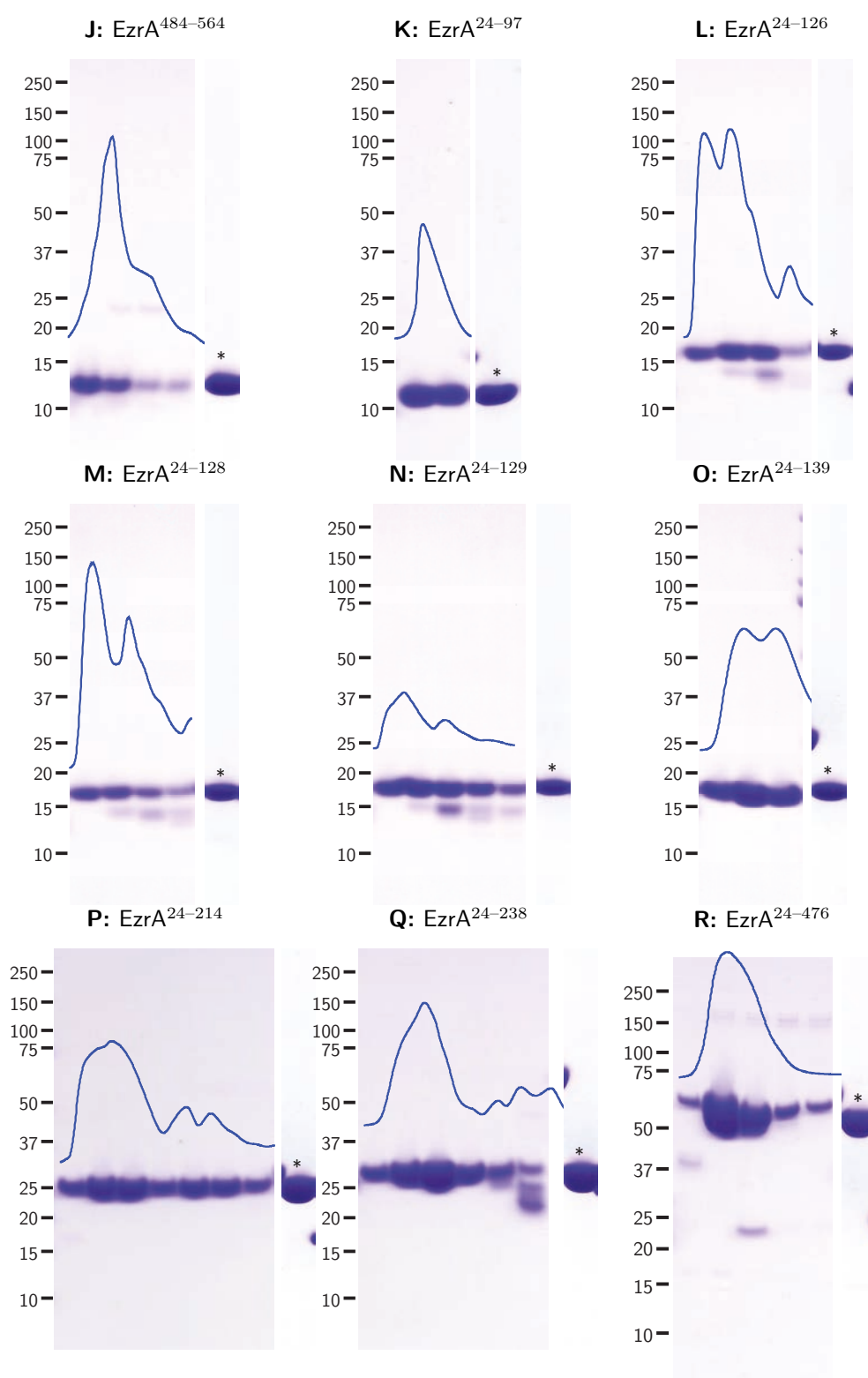
#### 5.4.3 Purification of soluble N- and C-terminal fragments of EzrA

To study the properties of the truncated EzrA proteins, 19 mutants were purified by one step IMAC and 18 were further fractionated by high resolution ion-exchange chromatography (Figure 5.10: A–R; mass spectra of purified proteins are shown in Appendix E 1–17; EzrA<sup>476–564</sup> precipitated in high concentration imidazole buffer when eluted from IMAC). All mutants presumed to be soluble by using EGFP fusion to report solubility remained soluble following purification except EzrA<sup>476–564</sup>, supporting the hypothesis that truncated EzrA mutants identified as soluble EGFP fusion proteins are soluble by themselves.

For unknown reasons, EzrA<sup>277–564</sup>, EzrA<sup>280–564</sup> and EzrA<sup>302–564</sup> eluted from the anion-exchange column in two distinct peaks, yet mass spectra of both the



**Figure 5.10: Purification of N- and C-terminally deleted EzrA mutants.** Soluble N-terminally truncated EzrA proteins were purified by IMAC and then by anion-exchange chromatography (**A–J**). C-terminally truncated EzrA proteins were purified by IMAC and then **K**; cation-exchange or **L–R**; anion-exchange chromatography. SDS-PAGE gels were stained with Coomassie blue. Shown for each protein are chromatographic fractions in order of elution with sample  $A_{280}$  (mAU) overlaid in blue and pooled purified proteins marked \* (**at the right of each panel**). Continued on next page...

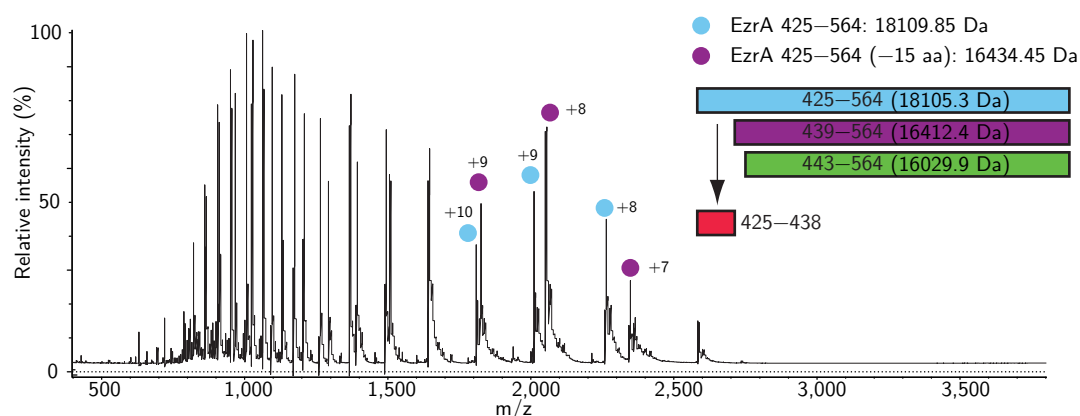


**Figure 5.10: Purification of N- or C-terminally deleted EzrA mutants**— continued from previous page.

early and late eluting peaks for EzrA<sup>277-564</sup> (Appendix E.1A and E.1B) and EzrA<sup>302-564</sup> (Appendix E.3A and E.3B) were identical. This suggests some structural variability between the distinct populations rather than differences in sequence; the remaining, shorter, N-terminally deleted EzrA mutants EzrA<sup>381-564</sup>, EzrA<sup>393-564</sup>, EzrA<sup>425-564</sup>, EzrA<sup>443-564</sup>, EzrA<sup>453-564</sup>, EzrA<sup>476-564</sup> and EzrA<sup>484-564</sup> eluted as single populations when purified by anion-exchange.

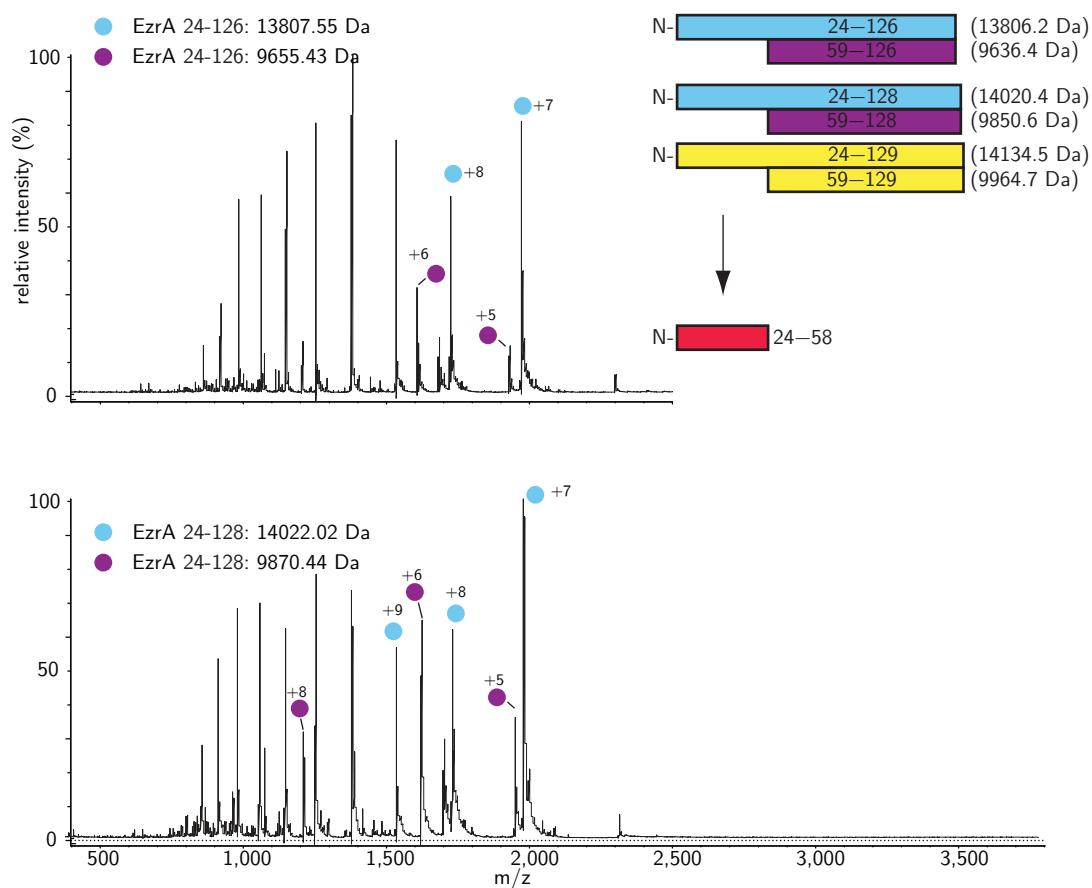
Purification of EzrA<sup>425-564</sup> by sequential IMAC and anion-exchange chromatography recovered both full length EzrA<sup>425-564</sup> and the 1–2 kDa smaller proteolysis fragment (Figure 5.10F). As all the truncated EzrA proteins were expressed with a C-terminal His<sub>6</sub>-tag and purified initially by IMAC, the recovered proteolytic fragment must be missing the N-terminus of EzrA<sup>425-564</sup>. Analysis by mass spectrometry of denatured EzrA<sup>425-564</sup> and its large proteolysis product revealed a mass difference of 1.8 kDa that is consistent with the loss of the 15 N-terminal residues (Figure 5.11). The large proteolytic fragment of EzrA<sup>425-564</sup> is therefore just four amino acids larger than EzrA<sup>443-564</sup>, which was also identified as being soluble (Figure 5.9A). The fact that both domain truncation using our methodology and *in vivo* proteolysis recover proteins that begin in the same region suggests a folding unit boundary at this position in the sequence of EzrA. The larger N-terminally deleted EzrA mutants EzrA<sup>277-564</sup>, EzrA<sup>280-564</sup> and EzrA<sup>302-564</sup> do not proteolyse in this manner, suggesting that EzrA<sup>425-564</sup> might be partially unfolded at its N-terminus, allowing access to a proteolytic enzyme.

Purification of some C-terminally deleted EzrA mutants also revealed evidence of



**Figure 5.11: N-terminal proteolysis of EzrA<sup>425-564</sup>.** Positive ion electrospray mass spectrum of denatured, EzrA<sup>425-564</sup> and a co-purified, *in vivo* proteolysis product. The C-terminus of proteolytic fragments are retained as EzrA<sup>425-564</sup> contains a C-terminal His<sub>6</sub>-tag. Species observed in mass spectra reveal residual fragments consistent with proteolysis between residues 338–339, liberating the N-terminal fragment MSKKEEVYRLLASN. Full length proteins are coloured blue, residual proteolysed EzrA<sup>425-564</sup> purple and a separately identified soluble EzrA truncation mutant EzrA<sup>393-564</sup> green. Calculated peptide masses are indicated in brackets.

proteolysis; *i.e.*, in EzrA<sup>24-126</sup>, EzrA<sup>24-128</sup> and EzrA<sup>24-129</sup>. These proteins contain proteolytic fragments that elute in a later, second peak during anion-exchange chromatography (Figure 5.10L, M, N). Mass spectrometry under denaturing conditions showed the mass of the full length mutants and a proteolytic fragment consistent with loss of the 35 N-terminal amino acids from each mutant (Figure 5.12). This proteolysis event did not occur for EzrA<sup>24-139</sup>, EzrA<sup>24-214</sup>, EzrA<sup>24-238</sup> or EzrA<sup>24-476</sup> although EzrA<sup>24-238</sup> suffered minor proteolysis of the N-terminal 48 and 49 residues (Appendix E.16B).



**Figure 5.12: N-terminal proteolysis of EzrA<sup>24-126</sup>, EzrA<sup>24-128</sup> and EzrA<sup>24-129</sup>.** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-126</sup>, EzrA<sup>24-128</sup> and EzrA<sup>24-129</sup> and co-purified *in vivo* proteolysis products. Only C-terminal proteolytic fragments are retained as EzrA<sup>24-126</sup>, EzrA<sup>24-128</sup> and EzrA<sup>24-129</sup> contain a C-terminal His<sub>6</sub>-tag. Species observed in mass spectra reveal residual fragments consistent with proteolysis between residues 58–59, liberating a 35 amino acid N-terminal fragment from each. Full length proteins are coloured blue, residual proteolysed EzrA<sup>24-126</sup> and EzrA<sup>24-128</sup> purple. The mass spectrum is not shown for EzrA<sup>24-129</sup>, which is coloured yellow. Calculated peptide masses are indicated in brackets.

#### 5.4.4 Examination of the foldedness of N- and C-terminally truncated *Staphylococcus aureus* EzrA

To establish the foldedness of the soluble truncated EzrA mutants, the proteins were prepared for proton NMR and sent to Dr Xun-Cheng Su at the State Key Laboratory of Elemento-organic Chemistry, Nankai University, Peoples Republic of China, to measure 1D and 2D TOCSY NMR spectra. 1D proton NMR gives various indications of protein foldedness (Section 4.4.5); in particular, chemical shift data for buried methyl protons ( $-0.5$  to  $1.5$  ppm),  $\beta$ -sheet  $C\alpha$  ( $5$ – $6$  ppm) and backbone amide ( $6$ – $10$  ppm) proton regions are useful (McDonald and Phillips, 1967; Christendat *et al.*, 2000; Rehm *et al.*, 2002). In the case of a wholly unfolded protein, every amino acid residue is solvent exposed and the chemical environments are essentially identical, resulting in each distinct amino acid producing highly similar proton NMR signals. In contrast, a folded protein imposes stricter conformational constraints, and consequently, unique chemical environments which give rise to distinct spectral peaks in NMR spectra.

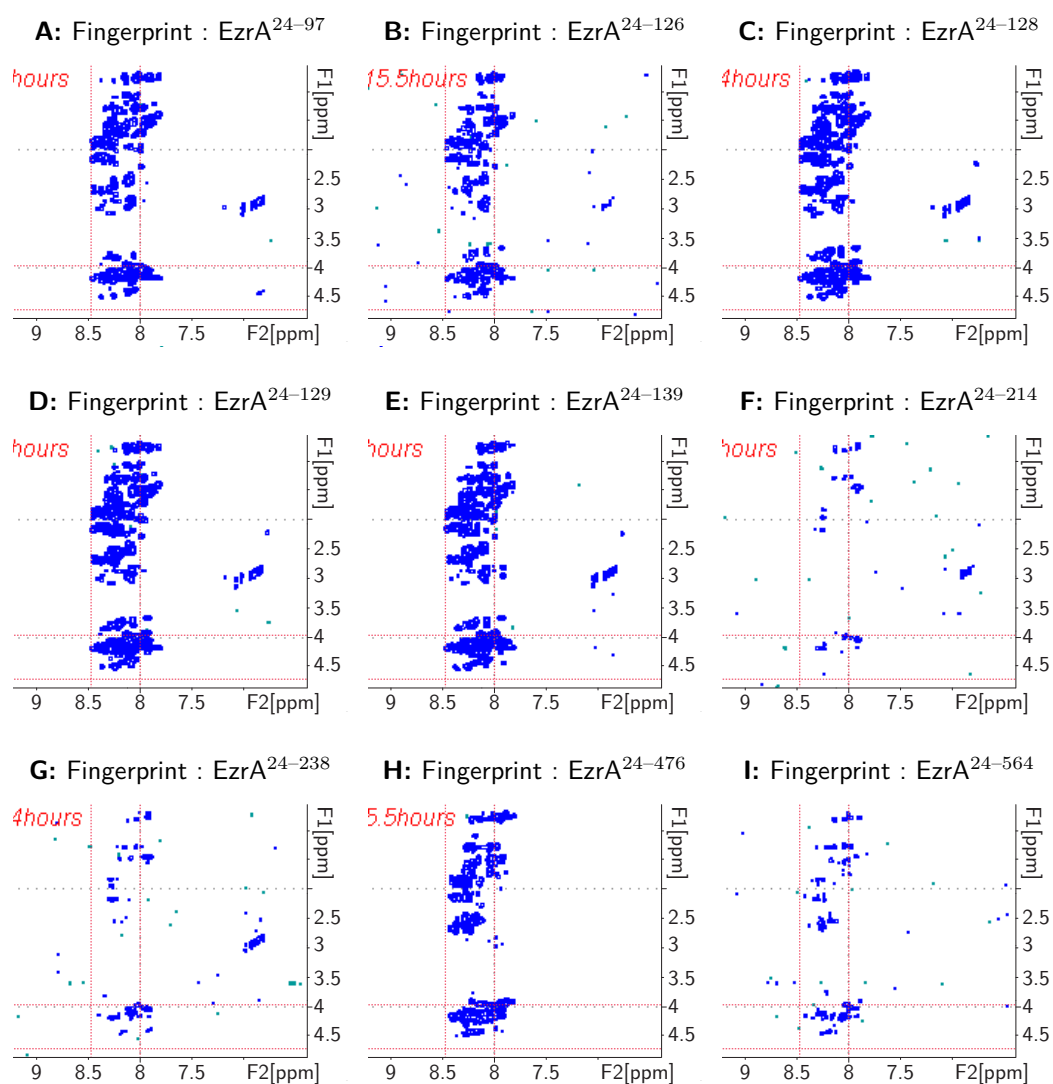
In the 2D NMR experiments performed here, spectra of large proteins show chemical shifts for flexible residues only, as large folded portions of proteins rotate too slowly in solution to give non-background signals (Section 4.4.5). However, as protein size decreases – and rotation speeds up — resonances from the folded core of a protein become visible in the “flexible” spectrum. In the case of these smaller proteins, the characteristics of observed chemical shifts can be useful in determining whether a protein is disordered, as evidenced by characteristic random-coil chemical shifts (Wishart *et al.*, 1995; Wishart, 2011).

One dimensional and 2D TOCSY NMR spectra was recorded for EzrA<sup>277–564</sup>, EzrA<sup>280–564</sup>, EzrA<sup>302–564</sup>, EzrA<sup>381–564</sup>, EzrA<sup>425–564</sup>, EzrA<sup>443–564</sup>, EzrA<sup>484–564</sup>, EzrA<sup>24–564</sup>, EzrA<sup>24–97</sup>, EzrA<sup>24–126</sup>, EzrA<sup>24–128</sup>, EzrA<sup>24–129</sup>, EzrA<sup>24–139</sup>, EzrA<sup>24–214</sup>, EzrA<sup>24–238</sup> and EzrA<sup>24–476</sup> (1D and complete TOCSY spectra, see Appendices E.1–E.6 and E.10–E.17). While some EzrA mutants appeared to be well folded by NMR, analysis of the TOCSY spectra makes unambiguous residue assignment difficult as the chemical shifts — which when flexible are very likely to be disordered — are at or near random-coil values and obscure each other due to poor spectral dispersion.

Truncated EzrA mutants, representing approximately the first domain (EzrA<sup>24–97</sup>, 10.4 kDa; EzrA<sup>24–126</sup>, 13.8 kDa; EzrA<sup>24–128</sup>, 14.0 kDa; EzrA<sup>24–129</sup>, 14.1 kDa), did not have methyl proton signals in their 1D NMR spectra (Appendices E.10B, E.11C, E.12C, E.13C), indicating that these are not folded. The 2D TOCSY spectra display significant signals at amino acid chemical shifts in the random-coil range (amide- $\alpha$ C proton correlations for C-terminally deleted EzrA mutants as shown in Figure 5.13). As these proteins are small, and no non-random-coil resonances are visible, they are likely completely unfolded. N-terminal proteolysis of these proteins (Section 5.4.3), would easily occur if they are poorly, or not folded.

Interestingly, EzrA<sup>24–139</sup> (15.3 kDa) appeared to be partially folded, as methyl signals are visible in the 1D NMR spectrum (Appendix E.14C). However, the protein also showed a significant random-coil signal in the TOCSY spectrum (Figure 5.13E). EzrA<sup>24–139</sup> does not undergo proteolysis between residues 58 and





**Figure 5.13: Two-dimensional TOCSY NMR fingerprint region in spectra of C-terminally deleted EzrA mutants.** C $\alpha$ -NH correlations in NMR spectra recorded for truncated EzrA mutants. 2D TOCSY NMR spectroscopy (mixing time 80 ms) was performed on EzrA mutants (**A–H**) and full-length cytoplasmic EzrA (**I**) in H<sub>2</sub>O with 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT by the research group of Dr Xun-Cheng Su, Nankai University, Peoples Republic of China. Red dashed lines indicate random-coil values for: vertical, amide proton and horizontal,  $\alpha$ -C proton chemical shifts (Wishart *et al.*, 1995; Wishart, 2011).

59 unlike EzrA<sup>24–126</sup>, EzrA<sup>24–128</sup>, and EzrA<sup>24–129</sup>, which is further evidence that the N-terminus of EzrA that incorporates this site is folded in EzrA<sup>24–139</sup>; no information can be inferred about the C-terminus of this protein.

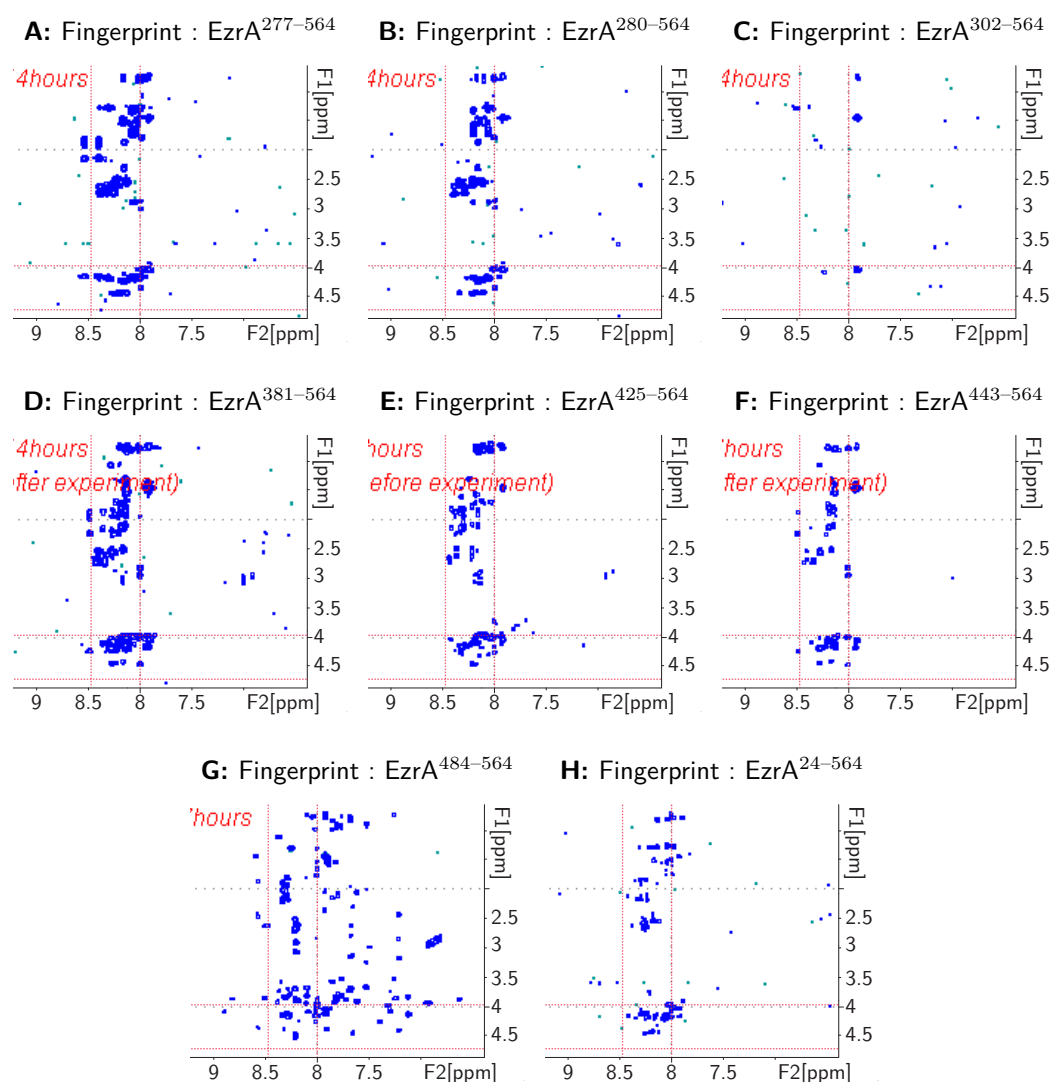
On the other hand, EzrA<sup>24–214</sup> (24.2 kDa) seemed to be well folded, with methyl signals in the 1D spectrum, and a minimal contribution of random-coil signals in the TOCSY spectrum (Figure 5.13F). Some of the resonances observed for EzrA<sup>24–214</sup> are likely from glutamic acid, and although unclear, several methyl containing residues may also have peaks in the TOCSY spectrum. It is not possible to distinguish if the flexible regions in EzrA<sup>24–214</sup> are at the N- or C-terminus from the NMR spectrum alone as such residues are present at both ends. The TOCSY spectrum of EzrA<sup>24–238</sup> (27.1 kDa) was highly similar to EzrA<sup>24–214</sup> (in addition to the 1D spectrum indicating that this protein is folded; Figure 5.13G; Appendix E.16C). However, EzrA<sup>24–238</sup> shows some additional random-coil chemical shifts at around 1.5 ppm/8.1–8.3 ppm (F1/F2). The additional unstructured residues cannot be positively identified, although, as resonances are not present in the spectrum of the shorter EzrA<sup>24–214</sup> mutant, the C-terminus of EzrA<sup>24–238</sup> is likely to contain a short, additional flexible region.

The large size of EzrA<sup>24–476</sup> (55.2 kDa) reduces the usefulness of NMR for investigating its structure, yet some random-coil resonances are still visible in the TOCSY NMR spectrum of this protein, indicating that not all of it is well folded (Figure 5.13H). Although not all, some of the resonances observed in the TOCSY spectrum of EzrA<sup>24–476</sup> are also observed in the spectrum of EzrA<sup>24–214</sup> and EzrA<sup>24–238</sup>, which may indicate that some of the flexible residues of EzrA<sup>24–476</sup>

are at the N-terminus, but these observations cannot rule out that the C-terminus of EzrA<sup>24–476</sup> is also poorly structured.

EzrA<sup>277–564</sup> (35.8 kDa), EzrA<sup>280–564</sup> (35.4 kDa) and EzrA<sup>302–564</sup> (32.8 kDa) are predicted to contain the complete three C-terminal domains and each mutant produced clear 1D NMR methyl peaks, indicative of them being folded proteins (Appendices E.1C, E.2B, E.3C). The TOCSY spectrum of EzrA<sup>302–564</sup> indicates that this protein is well folded, with only a single valine resonance in the random-coil region (amide- $\alpha$ C proton correlations for N-terminally deleted EzrA mutants are shown in Figure 5.14). Two Val residues are present near the N-terminus of the protein and one at the C-terminus, which makes it impossible to assign the location of the flexible region. However, all C-terminal EzrA fragments showed a Val resonance at an NH shift of  $\approx 7.95$  ppm (and indeed, full length EzrA has a similar signal), which may indicate that V561, at the extreme C-terminus of EzrA is mobile. EzrA<sup>280–564</sup> shows additional resonances in the TOCSY spectrum (Figure 5.14B), very likely due to the additional N-terminal residues compared to EzrA<sup>302–564</sup>, and EzrA<sup>277–564</sup> has a further two flexible Glu resonances (E277, E278: Figure 5.14C).

Mutants further N-terminally truncated (EzrA<sup>381–564</sup>, 23.5 kDa; EzrA<sup>425–564</sup>, 18.2 kDa; EzrA<sup>443–564</sup>, 16.2 kDa) contain folded protein cores as indicated by their 1D NMR methyl chemical shifts (Appendices E.4B, E.5B, E.6B), but also showed resonances in their flexible TOCSY spectra (Figures 5.14D, E, F). As the proteins get progressively shorter at the N-terminus, flexible resonances disappear from the TOCSY spectrum, where EzrA<sup>381–564</sup> has a larger contribution



**Figure 5.14: Two-dimensional TOCSY NMR fingerprints of N-terminally deleted EzrA mutants.** C $\alpha$ -NH correlations in NMR spectra recorded for truncated EzrA mutants. 2D TOCSY NMR spectroscopy (mixing time 80 ms) was performed on N-terminally truncated EzrA mutants (**A-G**) and full-length cytoplasmic EzrA (**H**) in H<sub>2</sub>O with 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT by the research group of Dr Xun-Cheng Su. Red dashed lines indicate random-coil values for: vertical, amide proton and horizontal,  $\alpha$ -C proton chemical shifts (Wishart *et al.*, 1995; Wishart, 2011).

of flexible residues while EzrA<sup>443-564</sup> seems to be well folded, indicating that the unstructured regions in these mutants is at the N-terminus. The sample used to record the spectrum of EzrA<sup>425-564</sup> contained a mixture of the full-length protein and the proteolytic fragment lacking 15 N-terminal residues (fragmented between residues 438 and 439), so its spectra are difficult to interpret. The TOCSY spectra of EzrA<sup>443-564</sup>, EzrA<sup>425-564</sup> and EzrA<sup>381-564</sup> have a readily identified Val resonance at an NH shift of  $\approx 7.95$  ppm, consistent with the other C-terminal fragments of EzrA. One-dimensional NMR indicates that EzrA<sup>484-564</sup> (11.2 kDa) has a folded protein core and the TOCSY spectrum — due to its small size — displays numerous resonances, the majority of which are not at random-coil values, indicating that EzrA<sup>484-564</sup> is well folded (Figure 5.14G).

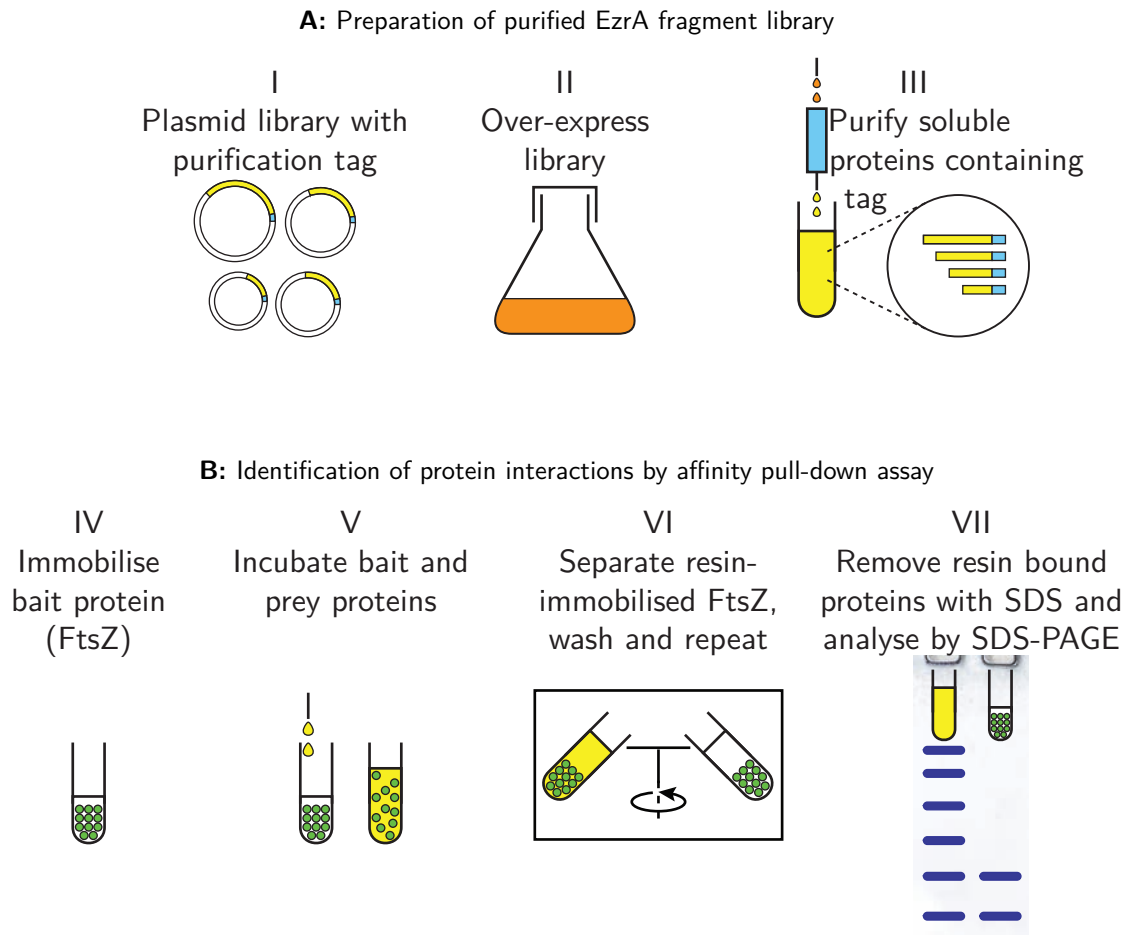
Due to the large size of EzrA<sup>24-564</sup> (65.4 kDa; Appendix E.18C and D) these NMR experiments can provide only limited detail about the full length cytoplasmic fragment. However, the structure appears predominantly  $\alpha$ -helical, based on the absence of C $\alpha$   $\beta$ -sheet proton signals (this is also the case for folded truncated EzrA mutants). Some Glu resonances are visible in the TOCSY spectrum of EzrA<sup>24-564</sup> that — assuming the flexibility is at an end of the protein — are possibly for residues E558, E559 and E560, as the N-terminus of EzrA is more sparsely populated with Glu. However, the EzrA<sup>24-564</sup> TOCSY spectrum also contains several methyl resonances that are hard to identify and these, along with the flexible Glu resonances, are absent from the TOCSY spectrum of EzrA<sup>302-564</sup>, which shares the same C-terminus. In comparison, spectra of the well folded N-terminal fragments EzrA<sup>24-214</sup> and EzrA<sup>24-238</sup> contain similar methyl resonances to EzrA<sup>24-564</sup> and a single Glu, suggesting that — these flexible residues at least

— are located in the N-terminal region of EzrA.

#### 5.4.5 Probing the EzrA–FtsZ interaction using affinity pull-down

Having produced a range of soluble, truncated EzrA proteins, we wished to utilise these to try to identify the site of interaction between EzrA and FtsZ. Aside from the terminal domains of EzrA, the proteins available from these EzrA libraries are not distinct domains, and instead many contain a series of presumably folded units, that are not by themselves useful for identifying which regions of the protein interacts with FtsZ. The advantage of using libraries of EzrA is that comparison of the relative binding of mutants from complementary N- and C-terminal deletion libraries should allow the site of interaction to be identified. In this case I used a synthetic library of the already-identified truncated genes, but in principle the same procedure could be used with naive libraries of truncations.

To identify the EzrA–FtsZ interactions a protein “pull-down” assay was used. Resin immobilised FtsZ was mixed with various pools of, or individual EzrA mutants and the resin then removed and washed; any EzrA proteins co-purified with the resin-bound FtsZ are therefore likely to have bound to FtsZ. The pools of soluble truncated EzrA were prepared in a straightforward manner (Figure 5.15A) by I) preparing truncated EzrA plasmid libraries with a C-terminal His<sub>6</sub>-tag. The protein libraries were then II) expressed, and the mixture of truncated EzrA mutants were III) purified from lysed cultures by IMAC.



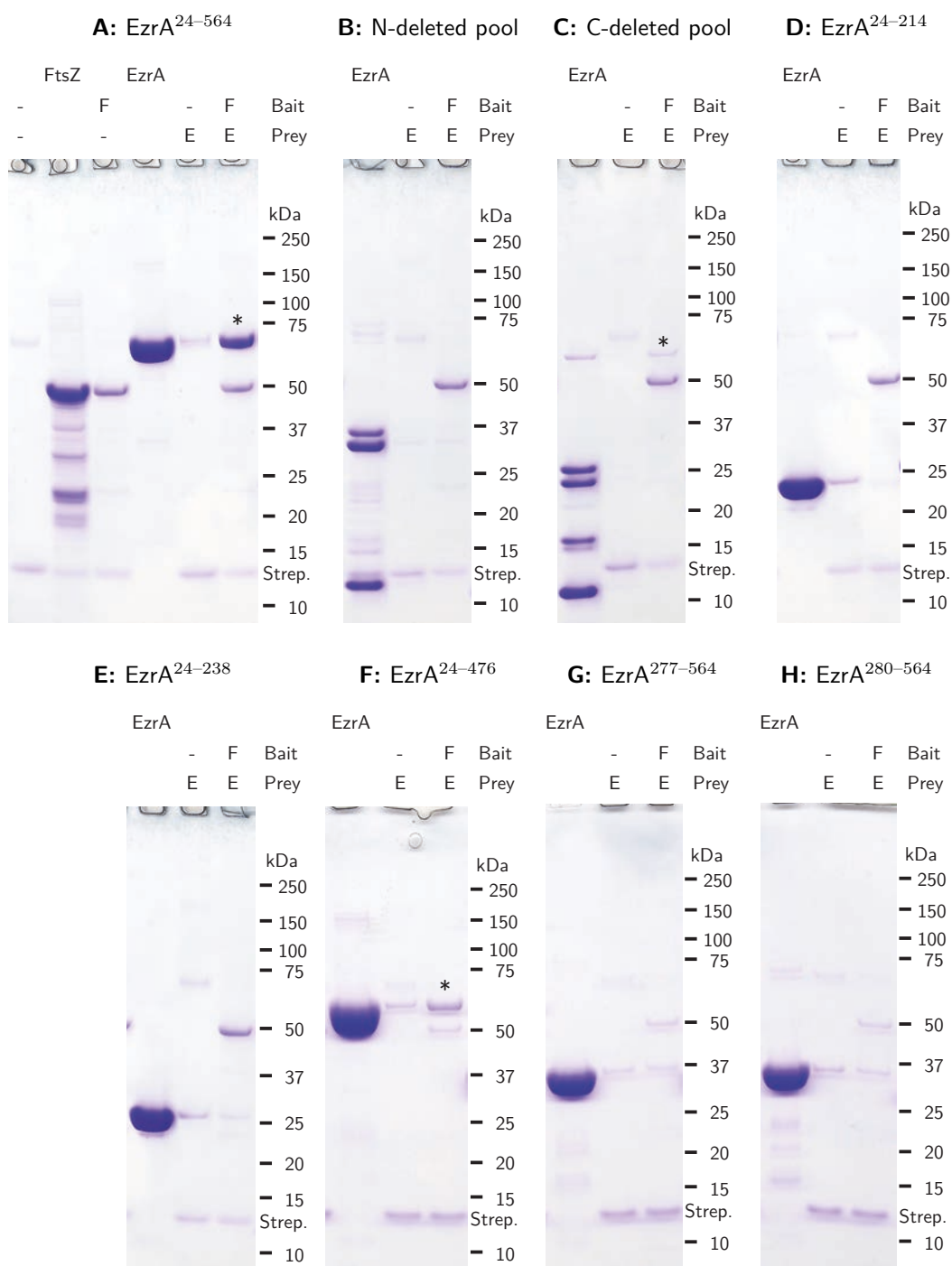
**Figure 5.15: Screening for FtsZ interaction using EzrA fragment libraries.** EzrA fragment libraries were made by **A: I–II**, over-expressing an EzrA-His<sub>6</sub> fragment library from plasmid pools in single cultures and **III**, purifying soluble over-expressed EzrA fragment proteins by IMAC. EzrA fragment libraries were then screened for mutants that bind to resin-immobilised FtsZ **B: IV**, N-terminally biotinylated FtsZ was incubated with streptavidin-coated resin and then unbound FtsZ was washed free. **V**, EzrA fragment libraries were incubated with resin-immobilised FtsZ on ice. Then **VI**, resin with attached FtsZ was separated from the EzrA library by centrifugation, and, fresh, protein free, ice cold buffer was added to wash unbound EzrA and the resin collected again. EzrA mutants bound to immobilised FtsZ were identified by **VII**, boiling streptavidin-coated resin in denaturing buffer and then analysis by SDS-PAGE.

With a pool of variable sized EzrA proteins, a simple pull-down experiment can then be used to identify fragments that interact with FtsZ. In these pull-down experiments, N-terminally biotinylated FtsZ “bait” was prepared by immobilisation on resin coated by streptavidin (IV; Figure 5.15B). V) The EzrA “prey” is then applied to the immobilised FtsZ, mixed and incubated on ice for 30 min to allow any interactions to form. As FtsZ is immobilised, VI) centrifugation of streptavidin resin can then separate and wash away the unbound EzrA.

To confirm that EzrA interacts with FtsZ, and that this pull-down technique is able to co-purify EzrA with immobilised FtsZ, a pull-down experiment was performed with EzrA<sup>24–564</sup>, and indeed, EzrA<sup>24–564</sup> was bound to FtsZ immobilised streptavidin resin, but not FtsZ naïve resin (Figure 5.16A). To then identify candidate EzrA fragments that interact strongly with FtsZ, purified N- and C-terminal libraries were investigated for interactions (Figure 5.16B–C). Of the proteins in the EzrA fragment libraries, only EzrA<sup>24–476</sup> appears to bind strongly to the FtsZ-immobilised resin. To validate the FtsZ-EzrA<sup>24–476</sup> interaction, the pull-down experiment was repeated using purified EzrA<sup>24–476</sup> (Figure 5.16F).

As previous reports placed the EzrA–FtsZ interaction site at the C-terminus of EzrA (Haeusser *et al.*, 2007) and yet Son and Lee (2013) suggested that the N-terminus of EzrA may be important for interaction with FtsZ, pull-down experiments were also performed with EzrA<sup>24–214</sup>, EzrA<sup>24–238</sup> (Figure 5.16D–E), EzrA<sup>277–564</sup> and EzrA<sup>280–564</sup> (Figure 5.16G–H). Only EzrA<sup>24–476</sup> bound FtsZ, suggesting that — compared to EzrA<sup>24–238</sup> and EzrA<sup>277–564</sup> — the region between residues 238 and 276 may be necessary for interaction with FtsZ.





**Figure 5.16: Identification of FtsZ–EzrA interaction by pull-down assay.** To identify candidate EzrA regions that interact with FtsZ, streptavidin- (13.2 kDa) coated resin was incubated with N-terminally biotinylated FtsZ (43.2 kDa), then free biotin and BSA (66.5 kDa). FtsZ-coated resin was then washed and incubated with EzrA and its truncation mutants on ice for 30 min. EzrA containing buffer was then separated from FtsZ-coated resin by centrifugation, and to remove free EzrA, twice replaced with TBS<sub>50</sub>, mixed and FtsZ-coated resin recovered again. FtsZ-coated resin, and any bound EzrA, was then incubated in SDS at 80°C for 10 min and analysed by SDS-PAGE. Polyacrylamide gels show purified protein samples marked FtsZ or EzrA and pull-down experiments. **A:** –/–, streptavidin resin incubated with BSA and free biotin only; F/– streptavidin resin incubated with biotinylated FtsZ, then BSA and free biotin. **A–H:** pull-down experiments performed with –/E, streptavidin resin was incubated with BSA and free biotin, washed, and then incubated with EzrA mutant, or F/E, streptavidin resin was incubated with biotinylated FtsZ, then BSA and free biotin, washed, and then incubated with an EzrA mutant. EzrA proteins that appear to bind FtsZ are marked \*.

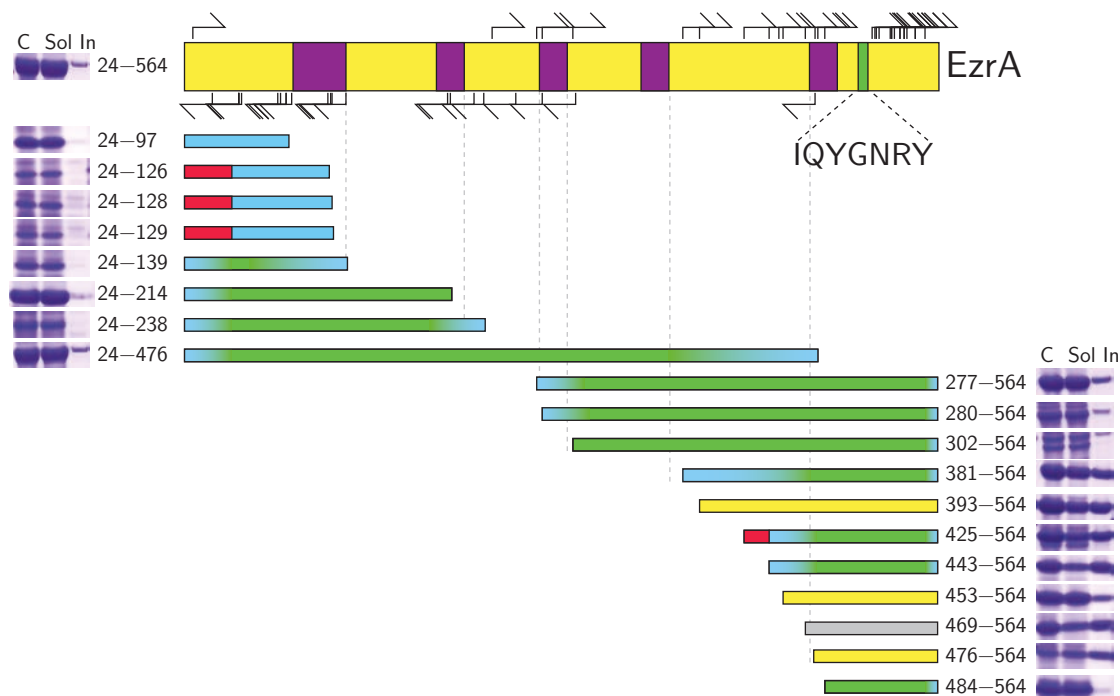
## 5.5 Discussion

### 5.5.1 Truncated EzrA and solubility

In this work both N- and C-terminally deleted libraries of *S. aureus* EzrA were made. The truncation strategy applied allowed identification of many soluble variants of EzrA and of these, several were selected for purification and determination of protein foldedness by NMR, showing that the selected mutant genes encode soluble proteins, some of which are very well folded (Figure 5.17). EzrA is predicted to contain five coiled-coil regions and some, but not all, soluble fragments of EzrA are truncated at the borders of these motifs. In general, the truncated proteins appear to have flexible residues at the non-truncated termini — as is common in proteins — and the truncated ends may be flexible even if not at a domain boundary. This is to be expected as the pragmatic solubility selection scheme used to identify these proteins is not biased for those without flexible tails.

Of the truncated mutants examined by NMR, EzrA<sup>24–214</sup>, EzrA<sup>302–564</sup>, and EzrA<sup>484–564</sup> were very well folded and likely truncated to the boundaries of domains. Recently, truncation of *S. aureus* EzrA by limited proteolysis was reported (EzrA<sup>164–564</sup>, EzrA<sup>250–564</sup> and EzrA<sup>278–564</sup>; Son and Lee, 2013). The authors, however, were not able to make soluble C-terminally truncated EzrA proteins by limited proteolysis, or produce the same degree of protein length

variation as the technique used here.



**Figure 5.17: Overview of soluble truncated *Staphylococcus aureus* EzrA mutant proteins.** At the top is a representation of the full length cytosolic part of EzrA (without the 23 residue N-terminal membrane anchor) with predicted coiled-coil domains coloured in purple and the QNR patch proposed to bind to FtsZ by Haeusser *et al.* (2007) in green. Arrows above and below the diagram of full length EzrA indicate the identity of putatively soluble truncated EzrA mutants. Also shown are truncated mutants of EzrA that have been over-expressed and purified; SDS-PAGE analysis of cellular, soluble and insoluble cell fractions from over-expressing cells are shown for each mutant. Protein foldedness as judged by NMR is indicated by colour: blue, unstructured and green, well structured; proteins not studied by NMR are coloured yellow. N-terminally proteolysed fragments are coloured red and EzrA<sup>469-564</sup> grey, as this mutant precipitated after purification.

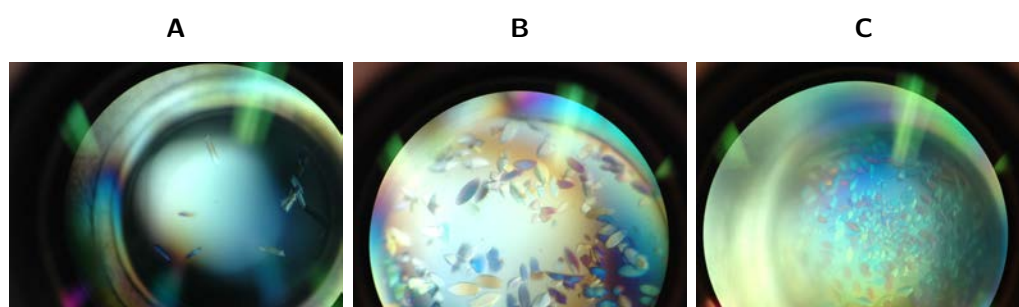
Overall, a short segment at the extreme N-terminus of EzrA is likely to be flexible since full-length cytoplasmic EzrA (EzrA<sup>24-564</sup>), EzrA<sup>24-214</sup>, and EzrA<sup>24-238</sup> appear to be well folded proteins, yet all appear to contain similar flexible residues in their TOCSY NMR spectra. However, this unstructured N-terminus does not extend past residue 58 as the partially folded EzrA<sup>24-139</sup> and longer mutants are protected from adventitious proteolysis at this site, unlike the poorly structured smaller

mutants. The extreme C-terminus of EzrA may also be slightly disordered as all C-terminal fragments of EzrA contain a flexible Val residue (V561; note that there is no Val residue in the presumably flexible purification tag). There appears to be little inter-domain flexibility in EzrA as EzrA<sup>24–214</sup> and EzrA<sup>302–564</sup> — representing all but the third predicted domain — contain no flexible residues other than those at the N- or C-termini of EzrA.

## 5.5.2 Structural insights into EzrA truncations

Members of our extended research group (Greg Harm, supervised by Dr Aaron Oakley) have successfully produced diffraction quality protein crystals of EzrA<sup>24–214</sup> (Figure 5.18), and the structure has now been solved (Dr Aaron Oakley, personal communication). These data support NMR observations that this mutant is well folded.

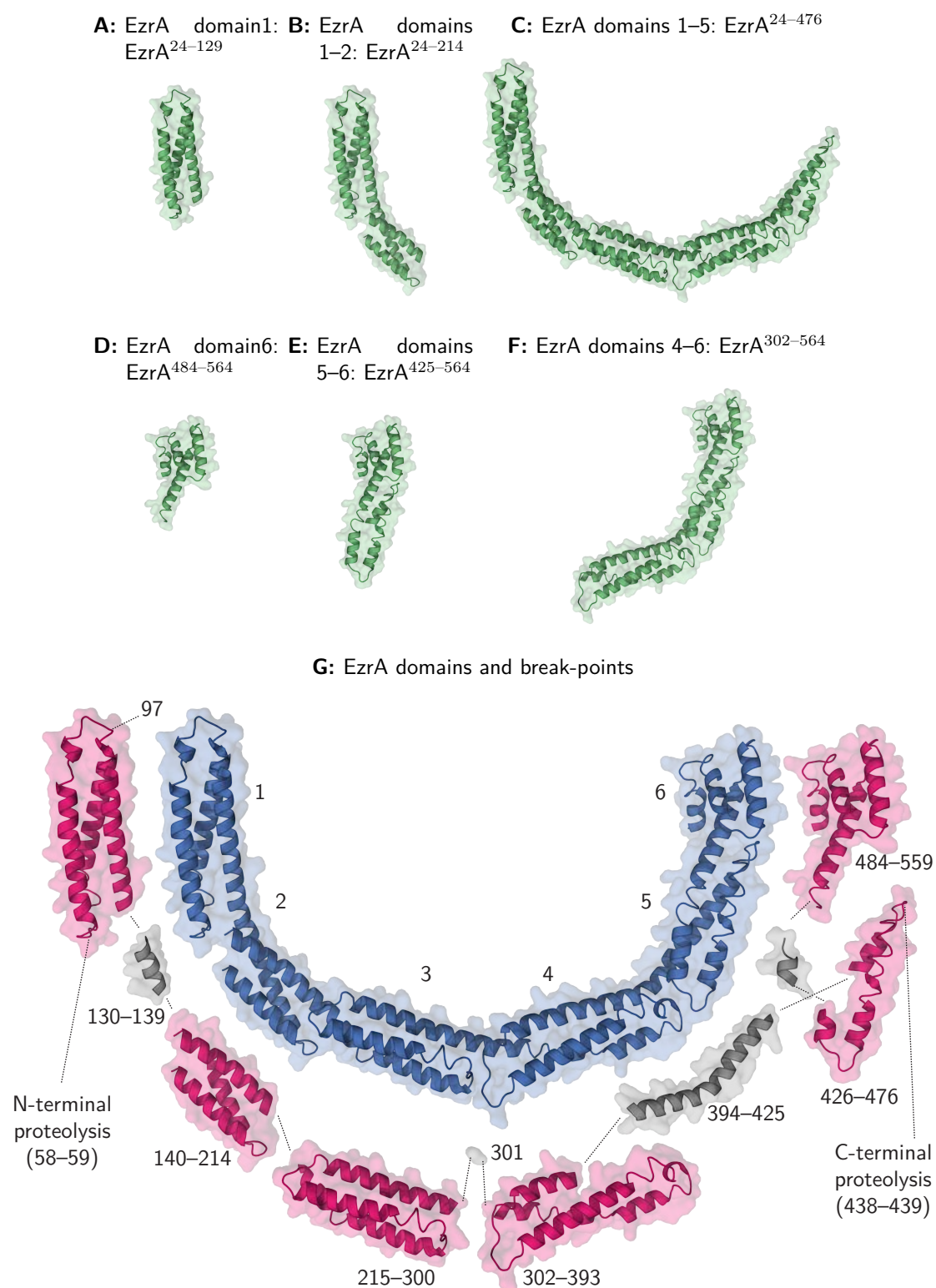
We were recently supplied with the coordinates of a crystal structure of *B. subtilis*



**Figure 5.18: Preliminary crystallography of EzrA<sup>24–214</sup>.** Purified EzrA<sup>24–214</sup> was subjected to protein crystallisation trials using the QIAGEN NeXtal JCSG+ suite in sitting drop configuration by Greg Harm. *Bona fide* protein crystals have been recovered in conditions: **A**, D3 (0.2 M NaCl, 0.1 M Na/K phosphate pH 6.2, 50% (v/v) PEG 200); **B**, D5 (0.1 M HEPES pH 7.5, 70% (v/v) MPD) and **C**, D8 (0.1 M Tris pH 8.0, 40% (v/v) MPD).

EzrA from the group of Professor Richard J. Lewis, University of Newcastle, United Kingdom (Figure 5.19; Cleverly *et al.*, 2014). EzrA has an extended coiled-coil structure, forming a half-ring composed exclusively of  $\alpha$ -helices that form six identifiable domains. Strikingly, many of the mutants identified using our truncation methodology map to positions between helical domains where the structure kinks around the ring (domain 1 ends at residue 129, domain 2 at residue 214, domain 3 at residue 300, domain 4 near residue 411 and domain 5 ends at 485).

Of the proteins examined by NMR, EzrA<sup>24-214</sup>, EzrA<sup>302-564</sup> and EzrA<sup>484-564</sup> are very well folded and are truncated nicely at inter-domain bends in the EzrA structure (Figure 5.19B, D, F). Although EzrA<sup>24-129</sup> appears to be disordered by NMR, this protein is truncated distinctly to the apparent boundary between the first and second domains of EzrA (as are EzrA<sup>24-126</sup> and EzrA<sup>24-128</sup>). This suggests that the long helix that extends across domains 1 and 2 is necessary to stabilise domain 1. The solubility selection methodology using EGFP fusion to report on protein solubility is proposed to be useful in identifying protein domains (in combination with gene truncation), as insoluble and/or aggregation-prone proteins should drag EGFP into aggregates *in vivo* where it can neither fold nor develop green fluorescence. If this were the case with EzrA<sup>24-129</sup>, mutants slightly shorter than this (also likely to be unfolded), would not be expected to perturb development of EGFP fluorescence. Yet, it appears this was not the case for C-terminally deleted proteins truncating within the first EzrA domain. Perhaps, *in vivo*, where molecular crowding is higher, the domain represented by EzrA<sup>24-129</sup> may be able to form, whereas proteins representing only part of this domain may perturb development of mature EGFP.



**Figure 5.19: Cytoplasmic EzrA domain architecture.** The unpublished crystal structure of the cytoplasmic regions of *B. subtilis* EzrA was kindly provided by Professor Richard J. Lewis (University of Newcastle, United Kingdom; Cleverly *et al.*, 2014). To examine the identity of truncations of *S. aureus* EzrA produced in this work, *B. subtilis* and *S. aureus* EzrA protein sequences were aligned with ClustalW (Thompson *et al.*, 2002) to approximate the *S. aureus* residue positions. Truncated mutants of EzrA (**A–F**) express well and appear to be truncated at or near domain boundaries. **G**: complete EzrA structure (blue) and folds delineated by breakpoints identified in this Chapter (pink). The N-terminus of EzrA is on the left side of the image, and the six domains are indicated.

Interestingly, no soluble EzrA mutant was identified at the boundary of domains 4 and 5. Like the other domains (1–4), these are joined by a long  $\alpha$ -helix that extends through both domains. Nevertheless, mutants including (EzrA<sup>393–564</sup>) or excluding most of (EzrA<sup>425–564</sup>; nine residues of the helix are included) this helix were identified as soluble (Figure 5.19G). Some molecules of EzrA<sup>425–564</sup> were proteolysed between residues 438–439 but EzrA<sup>381–564</sup> was not. Although EzrA<sup>381–564</sup> contains a flexible N-terminus, it seems that these residues are protected from proteolysis. EzrA<sup>24–476</sup> completely contains the first four domains of EzrA, and a C-terminal extension containing, what appears to be, the majority of the fifth domain (Figure 5.19C). This protein contains significant flexibility (by NMR) which is very likely attributed to the fifth domain. That domain 5 appears to be short, folded in EzrA<sup>443–564</sup> but not in EzrA<sup>24–476</sup> may point to a requirement for its stabilisation by domain 6 to become structured, or the least residues 477 to  $\sim$  483.

At the N-terminus of EzrA, EzrA<sup>24–126</sup>, EzrA<sup>24–128</sup> and EzrA<sup>24–129</sup> appear not to be folded, yet addition of just a few residues, as is the case with EzrA<sup>24–139</sup>, produces a protein that is partially structured based on observations by NMR. This is interesting as residues 130–139 represent an extension into domain 2 of the  $\alpha$ -helix at the base of the  $\alpha$ -helical bundle making up domain 1 (Figure 5.19G). These additional  $\alpha$ -helical residues may help stabilise the fold of domain 1, although the domain is still apparently somewhat unstable. Further support for this proposition is that EzrA<sup>24–126</sup>, EzrA<sup>24–128</sup> and EzrA<sup>24–129</sup>, but not EzrA<sup>24–139</sup>, proteolysed between residue 58 and 59, which appear to be in a loop adjacent to

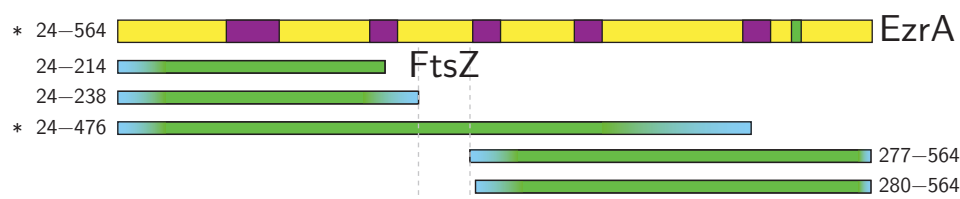
the helical residues at 130–139.

Of course, we should recognise that this analysis is predicated on the *B. subtilis* structure being a reliable model for *S. aureus* EzrA. It is possible that subtle differences between the two protein structures may be revealed in future work.

### 5.5.3 The interaction between EzrA and FtsZ

This work successfully showed evidence for the interaction of EzrA and FtsZ using a streptavidin pull-down assay, and used this method to further narrow down the region of EzrA involved. Due to the methodology of the assay used — which involves washing unbound EzrA by dilution — only interactions with off-rates slower than the washing time can be reliably recovered. Performing the EzrA pull-down experiment using a library of truncated EzrA proteins produced only one candidate interacting fragment, EzrA<sup>24–476</sup>; this interaction with FtsZ was confirmed by performing a pull-down using purified EzrA<sup>24–476</sup>. The pull down experiments used here also showed that EzrA<sup>24–214</sup>, EzrA<sup>24–238</sup>, EzrA<sup>277–564</sup> and EzrA<sup>280–564</sup> do not interact strongly with FtsZ, suggesting that a necessary component for a long lived EzrA–FtsZ interaction is present between residues 238 and 277 (Figure 5.20). According to examination of the protein’s structural homologue and protein foldedness by NMR, the four non-FtsZ-interacting EzrA fragments do not contain a folded domain 3, so the specific site of the interaction may lie beyond residues 238 and 277, but very likely involves domain 3.





**Figure 5.20: FtsZ pull-down of purified truncated EzrA proteins.** Full length EzrA (without the N-terminal membrane anchor) and EzrA<sup>24-238</sup> interact with FtsZ-coated resin, but EzrA<sup>24-214</sup>, EzrA<sup>24-238</sup>, EzrA<sup>277-564</sup> and EzrA<sup>280-564</sup> do not, suggesting that the region between residues 238 and 276 in EzrA is important for the interaction with FtsZ.

The interaction of EzrA<sup>24-476</sup> and FtsZ is interesting as previous work by Haeusser *et al.* (2007) showed that deletion of the near-C-terminal EzrA QNR patch (*B. subtilis* residues 506-510) — not present in EzrA<sup>24-476</sup> — stops mid-cell localisation of EzrA in *B. subtilis*. However, no direct interaction has been observed between EzrA and FtsZ at this site in any published study; maybe the QNR patch mediates another interaction at the divisome.

Further, Son and Lee (2013) attempted to narrow down the *S. aureus* FtsZ-EzrA interaction. These authors used FtsZ-EzrA pull-down assays with immobilised N-terminal GST tagged FtsZ<sup>351-390</sup> and FtsZ<sup>370-390</sup> to show that EzrA<sup>67-564</sup> does interact with FtsZ but that EzrA<sup>164-564</sup> does not. Moving on from these results Son and colleagues attempted to co-purify His<sub>6</sub>-EzrA<sup>25-163</sup> and GST-FtsZ but were unsuccessful in doing so using GST affinity chromatography. They did observe its association with GST-FtsZ using IMAC. These authors also performed analytical gel filtration using EzrA<sup>25-163</sup> and FtsZ and failed to observe co-elution of these proteins. These results purportedly show a weak interaction between the C-terminus of FtsZ and domain 1 of EzrA, but further work is necessary to understand this interaction fully.

Possibly relevant to the apparently contradictory reports of the EzrA–FtsZ interaction are the interactions of other divisome proteins, that show evidence of multiple, weak and redundant interactions. For example, *E. coli* FtsN appears to have multiple, synergistic, interactions with PBP1b. FtsN<sup>58–232</sup> and FtsN<sup>58–319Δ129–242</sup> have both been shown by pull-down assay to interact with PBP1b (Müller *et al.*, 2007). These individual FtsN–PBP1b interactions are of course much weaker than the interaction of full length FtsN. The divisome — a very large protein complex populated by many protein–protein interactions — may be regulated by multivalent interactions like those of FtsZ and PBP1b, as these would provide a more dynamic multi-protein machine than is possible using the alternative individual, very strong interactions.

Perhaps then, the various reports of EzrA interactions with FtsZ suggest that this interaction too is mediated not by a single strong interaction, but by several weaker interactions that cooperatively bind the proteins together.

#### 5.5.4 General discussion

Generating protein truncations by our methodology appears, in the case of EzrA, to be superior to limited proteolysis in identifying protein truncation variants and soluble protein fragments. This work gives access to truncated EzrA proteins that have previously been unobtainable, some of which are now predicted to be truncated directly at or near domain boundaries. In addition to identifying soluble

fragments of EzrA, we now have a very useful EzrA protein library pull-down tool for narrowing down the folding units of EzrA involved in numerous interactions within the divisome.

If a single, long lived EzrA–FtsZ interaction exists, the work presented here suggests that the third domain in EzrA is responsible. Confirmation of this interaction will require synthesis of domain 3 alone, using mutations identified here.

An intriguing target for identification of EzrA interactions using the protein fragments generated here is the important *S. aureus* PBP2 interaction. Depletion of EzrA in *S. aureus* strains leads to distribution of PBP2 — normally localised to the division septum — throughout the cell wall and appears to halt peptidoglycan synthesis (Steele *et al.*, 2011). These data suggest that EzrA helps localise and activate peptidoglycan synthesis by PBP2, with both of these properties being enticing targets for antibiotic effects. Delocalisation of peptidoglycan production in *S. aureus* should severely inhibit daughter cell production and may even lead to cell death. Interfering with the action of PBP2 by mimicking EzrA might also allow inhibition of, or even activation of peptidoglycan synthesis, which would have deleterious effects for cell reproduction and viability.

## Chapter 6

### Concluding remarks

This Thesis presents a new technique for protein domain identification that works by uni-directionally truncating genes using *ExoIII* and then expressing truncated mutants with a solubility reporting fusion protein. These experiments result in soluble, truncated proteins whose deletions did not hinder protein expression, folding or solubility and therefore are unlikely to contain substantial poorly folded regions. Chapter 3 presents the rationale behind gene deletion and solubility selection, as well as the various plasmids that were constructed to produce the truncation libraries. This new technique was then used to identify new soluble fragments of both DnaG primase (Chapter 4) and the divisome regulator EzrA (Chapter 5).

### 6.0.5 Caveats

Some proteins are not directly amenable to application of the technique presented in this Thesis. The solubility selection methods applied in this Thesis are biased to select well expressed proteins that fold efficiently. Potentially useful proteins that express poorly, or that suffer significant aggregation after correct folding will be under-represented. This technique is also poorly suited for proteins that contain multiple domains that fold in tandem such as *Bacillus amyloliquefaciens* RNase (Neira and Fersht, 1999) and barley chymotrypsin inhibitor 2 (Neira *et al.*, 1997) that fold only when the near C-terminal residues are present; or proteins which require a complementary protein for correct domain folding (San Martin *et al.*, 1995; Kamada *et al.*, 2003; Zhang *et al.*, 2003; Klammt *et al.*, 2006; Katzen *et al.*, 2008; Wu and Swartz, 2008). It is also unlikely to be useful for proteins that have a discontinuous domain structure, *i.e.*, where all or part of domain(s) are inserted within other folded domains.

The technique developed here is capable of providing an abundance of truncated proteins that are soluble and do not aggregate. However, identification of mutants that do not contain flexible, unfolded ends is not certain. Structural modelling of the ZBD and RNAP domain of DnaG (Chapter 4) suggests that soluble truncations occur at the start and end of protein folds, yet at these domain boundaries, inclusion of flexible extensions and deletion of short secondary structural elements is tolerated. As the basis of selection in this technique is a high level of over-expression and high solubility of protein fragments, flexible ends and deletion of short secondary structural elements do not necessarily negatively

affect these properties. In fact, it is plausible that in some cases, protein sequences that are not present in the mature protein domain will play a role in the folding pathway of domains, which may result in positive selection. Depending on the intended experimental use of a truncated protein, the fact that some proteins contain flexible ends and can be produced in soluble form may be more important than identifying proteins without flexible regions. NMR techniques can be used subsequently to identify these regions, providing soluble samples can be produced in the first place.

#### 6.0.6 Usefulness of this method

This new technique has advantages over previously reported physical domain identification methods and does not require prior knowledge of the protein. Gene deletion from a single terminus provides gene libraries of manageable size that can be fully sampled easily, in contrast to other genetic truncation methodologies used for this purpose (Miyazaki, 2002; Dyson *et al.*, 2008). Although uni-directional gene deletion does not sample all of the available truncated sequences, the straightforward methodology can potentially identify all protein domain boundaries in a protein. To then isolate internal domains of a protein, this information can be used to design new domain mutants.

Soluble mutants of *A. baylyi* DnaG RPD–HBD and numerous domain mutants of EzrA are truncated directly at or near domain boundaries, which highlights the success of this new domain identification methodology. In addition, as this new

technique allows straightforward production of soluble truncated protein pools — as performed in Chapter 5 — these can be used for initial studies to identify interesting regions of proteins, such as sites of interaction with other proteins.

Unlike limited proteolysis, this new technique does not require soluble expression of the target protein, which greatly increases the applicability of the method. Further, limited proteolysis necessitates that flexible, inter-domain segments are accessible to proteolytic enzyme active sites, which is not always the case. In limited proteolysis experiments performed for *S. aureus* EzrA (Son and Lee, 2013), the core of the protein is resistant to proteolysis, and these authors were unable to produce soluble C-terminally deleted mutants. Yet using our new technique, we have identified numerous domain truncations and soluble C-terminally deleted EzrA mutants with greater variability. Also, protein breakpoints identified by limited proteolysis do not account for regions required for nascent protein folding, which this technique most certainly does, since it identifies protein solubility following over-expression.

Work described in this Thesis has shown the utility of this new method of soluble protein domain identification and provided, in a straightforward manner, expressible soluble protein domains that were previously unavailable. This technique should be applicable to many proteins which currently are recalcitrant to soluble expression.

# Bibliography

- Adams, D. and Errington, J. **(2009)** Bacterial cell division: assembly, maintenance and disassembly of the Z ring. *Nature Reviews in Microbiology* **7**: 642–653.
- Allen, S., Polazzi, J., Gierse, J., and Easton, A. **(1992)** Two novel heat shock genes encoding proteins produced in response to heterologous protein expression in *Escherichia coli*. *Journal of Bacteriology* **174**: 6938–6947.
- Apic, G., Gough, J., and Teichmann, S. **(2001)** Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* **310**: 311–325.
- Ausubel, F., Brent, R., Kingston, R., Moore, D., Seidman, J., Smith, J., and Struhl, K. **(1987)** Current Protocols in Molecular Biology. John Wiley and Sons, New York.
- Bach, H., Mazor, Y., Shaky, S., Shoham-Lev, A., Berdichevsky, Y., Gutnick, D., and Benhar, I. **(2001)** *Escherichia coli* maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *Journal of Molecular Biology* **312**: 79–93.
- Bae, T., Banger, A., Wallace, A., Glass, E., Åslund, F., Schneewind, O., and Missiakas, D. **(2004)** *Staphylococcus aureus* virulence genes identified by *Bursa aurealis* mutagenesis and nematode killing. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 12312–12317.
- Barrow, E. W., Bourne, P. C., and Barrow, W. W. **(2004)** Functional cloning of *Bacillus anthracis* dihydrofolate reductase and confirmation of natural resistance to trimethoprim. *Antimicrobial Agents and Chemotherapy* **48**: 4643–4649.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S., Griffiths-Jones, S., Howe, K., Marshall, M., and Sonnhammer, E. **(2002)** The



- Pfam protein families database. *Nucleic Acids Research* **30**: 276–280.
- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000) The Pfam protein families database. *Nucleic Acids Research* **28**: 263–266.
- Beckett, D., Kovaleva, E., and Schatz, P. J. (1999) A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Science* **8**: 921–929.
- Begg, K. J. and Donachie, W. D. (1985) Cell shape and division in *Escherichia coli*: experiments with shape and division mutants. *Journal of Bacteriology* **163**: 615–622.
- Benkert, P., Biasini, M., and Schwede, T. (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**: 343–350.
- Benkert, P., Tosatto, S. C., and Schomburg, D. (2008) QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics* **71**: 261–277.
- Benkovic, S. J., Valentine, A. M., and Salinas, F. (2001) Replisome-mediated DNA replication. *Annual Review of Biochemistry* **70**: 181–208.
- Bergogne-Berezin, E. and Towner, K. (1996) *Acinetobacter spp.* as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clinical Microbiology Reviews* **9**: 148.
- Berman, A. L., Kolker, E., and Trifonov, E. N. (1994) Underlying order in protein sequence organization. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 4044–4047.
- Bhavsar, A. P., Truant, R., and Brown, E. D. (2005) The TagB protein in *Bacillus subtilis* 168 is an intracellular peripheral membrane protein that can incorporate glycerol phosphate onto a membrane-bound acceptor *in vitro*. *Journal of Biological Chemistry* **280**: 36691–36700.
- Bi, E. and Lutkenhaus, J. (1991) FtsZ ring structure associated with division in *Escherichia coli*. *Nature* **354**: 161.
- Bird, L. E., Pan, H., Soultanas, P., and Wigley, D. B. (2000) Mapping protein-protein interactions within a stable complex of DNA primase and DnaB helicase from *Bacillus stearothermophilus*. *Biochemistry* **39**: 171–182.

- Biswas, N. and Weller, S. K. (1999) A mutation in the C-terminal putative Zn<sup>2+</sup> finger motif of UL52 severely affects the biochemical activities of the HSV-1 helicase-primase subcomplex. *Journal of Biological Chemistry* **274**: 8068–8076.
- Björklund, A., Ekman, D., Light, S., Frey-Skött, J., and Elofsson, A. (2005) Domain rearrangements in protein evolution. *Journal of Molecular Biology* **353**: 911.
- Bollen, Y. J., Sánchez, I. E., and van Mierlo, C. P. (2004) Formation of on- and off-pathway intermediates in the folding kinetics of *Azotobacter vinelandii* apoflavodoxin. *Biochemistry* **43**: 10475–10489.
- Bouché, J.-P., Rowen, L., and Kornberg, A. (1978) The RNA primer synthesized by primase to initiate phage G4 DNA replication. *Journal of Biological Chemistry* **253**: 765–769.
- Bremer, H. and Dennis, P. P. (1996) Modulation of chemical composition and other parameters of the cell by growth rate, in: *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, edited by F. C. Neidhardt, R. I. Curtiss, J. L. Ingraham, C. C. Lin, Edmund, K. B. Low, B. Magananik, W. S. Reznikoff, M. Riley, M. Scharchter, *et al.*, vol. 2, American Society for Microbiology Press, Washington, DC, pp. 1553–1569.
- Brockwell, D. J. and Radford, S. E. (2007) Intermediates: ubiquitous species on folding energy landscapes? *Current Opinion in Structural Biology* **17**: 30–37.
- Canaves, J. M., Page, R., Wilson, I. A., and Stevens, R. C. (2004) Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: Maximum clustering strategy for structural genomics. *Journal of Molecular Biology* **344**: 977–991.
- Carballido-López, R. (2006) Orchestrating bacterial cell morphogenesis. *Molecular Microbiology* **60**: 815–819.
- Carrio, M., Corchero, J., and Villaverde, A. (1998) Dynamics of *in vivo* protein aggregation: building inclusion bodies in recombinant bacteria. *FEMS Microbiology Letters* **169**: 9–15.
- Cha, H. J., Srivastava, R., Vakharia, V. N., Rao, G., and Bentley, W. E. (1999) Green fluorescent protein as a noninvasive stress probe in resting *Escherichia coli* cells. *Applied and Environmental Microbiology* **65**: 409–414.
- Chandonia, J.-M., Kim, S.-H., and Brenner, S. E. (2006) Target selection and

- deselection at the Berkeley Structural Genomics Center. *Proteins: Structure, Function, and Bioinformatics* **62**: 356–370.
- Chaudhuri, R., Allen, A., Owen, P., Shalom, G., Stone, K., Harrison, M., Burgis, T., Lockyer, M., Garcia-Lara, J., Foster, S., *et al.* (2009) Comprehensive identification of essential *Staphylococcus aureus* genes using transposon-mediated differential hybridisation (TMDH). *BMC Genomics* **10**: 291.
- Chaudhuri, T. K., Farr, G. W., Fenton, W. A., Rospert, S., and Horwich, A. L. (2001) GroEL/GroES-mediated folding of a protein too large to be encapsulated. *Cell* **107**: 235–246.
- Chen, L., Oughtred, R., Berman, H. M., and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**: 2860–2862.
- Cheng, C.-H. and Lee, W.-C. (2010) Protein solubility and differential proteomic profiling of recombinant *Escherichia coli* overexpressing double-tagged fusion proteins. *Microbial Cell Factories* **9**: 63.
- Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustar, V., Klvana, M., Medek, P., *et al.* (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Computational Biology* **8**: e1002708.
- Christ, D. and Winter, G. (2006) Identification of protein domains by shotgun proteolysis. *Journal of Molecular Biology* **358**: 364–371.
- Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., *et al.* (2000) Structural proteomics of an archaeon. *Nature Structural & Molecular Biology* **7**: 903–909.
- Claessen, D., Emmins, R., Hamoen, L. W., Daniel, R. A., Errington, J., and Edwards, D. H. (2008) Control of the cell elongation–division cycle by shuttling of PBP1 protein in *Bacillus subtilis*. *Molecular Microbiology* **68**: 1029–1046.
- Clarke, A. R. (1996) Molecular chaperones in protein folding and translocation. *Current Opinion in Structural Biology* **6**: 43–50.
- Cleverly, R. M., Bui, N. K., Solovyova, A., Vollmer, W., and Lewis, R. J. (2014) A bacterial homologue of spectrin regulates the Z-ring in cell division, submitted.
- Cole, P. A. (1996) Chaperone-assisted protein expression. *Structure* **4**: 239–242.

- Corn, J. E., Pease, P. J., Hura, G. L., and Berger, J. M. (2005) Crosstalk between primase subunits can act to regulate primer synthesis in *trans*. *Molecular Cell* **20**: 391–401.
- Cortazzo, P., Cerveñansky, C., Marín, M., Reiss, C., Ehrlich, R., and Deana, A. (2002) Silent mutations affect *in vivo* protein folding in *Escherichia coli*. *Biochemical and Biophysical Research Communications* **293**: 537–541.
- Daniel, R. and Errington, J. (2003) Control of cell morphogenesis in bacteria: two distinct ways to make a rod-shaped cell. *Cell* **113**: 767.
- Davis, G. D., Elisee, C., Newham, D. M., and Harrison, R. G. (1999) New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnology and Bioengineering* **65**: 382–388.
- Deuerling, E., Patzelt, H., Vorderwulbecke, S., Rauch, T., Kramer, G., Schaffitzel, E., Mogk, A., Schulze-Specking, A., Langen, H., and Bukau, B. (2003) Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Molecular Microbiology* **47**: 1317–1328.
- Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A., and Bukau, B. (1999) Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature* **400**: 693–696.
- di Guan, C., Li, P., Riggs, P. D., and Inouye, H. (1988) Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* **67**: 21–30.
- Dijkshoorn, L., Nemec, A., and Seifert, H. (2007) An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nature Reviews in Microbiology* **5**: 939–951.
- Dill, K. and Chan, H. (1997) From Levinthal to pathways to funnels. *Nature Structural Biology* **4**: 10–19.
- Dinner, A., Sali, A., Smith, L., Dobson, C., and Karplus, M. (2000) Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences* **25**: 331–339.
- Dobson, C. (2003) Protein folding and misfolding. *Nature* **426**: 884–890.
- Dobson, C. M., Šali, A., and Karplus, M. (1998) Protein folding: A perspective from theory and experiment. *Angewandte Chemie International Edition* **37**: 868–893.

- Dyson, M., Shadbolt, S., Vincent, K., Perera, R., and McCafferty, J. (2004) Production of soluble mammalian proteins in *Escherichia coli*: Identification of protein features that correlate with successful expression. *BMC Biotechnology* **4**.
- Dyson, M. R., Perera, R. L., Shadbolt, S. P., Biderman, L., Bromek, K., Murzina, N. V., and McCafferty, J. (2008) Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucleic Acids Research* **36**: e51–e51.
- Ekman, D., Björklund, Å., Frey-Skött, J., and Elofsson, A. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of Molecular Biology* **348**: 231–243.
- El-Samad, H., Kurata, H., Doyle, J. C., Gross, C. A., and Khammash, M. (2005) Surviving heat shock: control strategies for robustness and performance. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 2736–2741.
- Ellis, R. J. and Hartl, F. U. (1996) Protein folding in the cell: competing models of chaperonin function. *FASEB Journal* **10**: 20–26.
- Elvin, C., Dixon, N., and Rosenberg, H. (1986) Molecular cloning of the phosphate (inorganic) transport (*pit*) gene of *Escherichia coli* K12. *Molecular and General Genetics* **204**: 477–484.
- Eramian, D., Eswar, N., Shen, M.-Y., and Sali, A. (2008) How well can the accuracy of comparative protein structure models be predicted? *Protein Science* **17**: 1881–1893.
- Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., Madhusudhan, M. S., Yerkovich, B., and Sali, A. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Research* **31**: 3375–3380.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2006) Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics* **Chapter 5**: Unit 5.6.
- Felitsyn, N., Peschke, M., and Kebarle, P. (2002) Origin and number of charges observed on multiply-protonated native proteins produced by ESI. *International Journal of Mass Spectrometry* **219**: 39–62.

- Fenton, W., Beechem, J., and Horwich, A. (1996) Characterization of the active intermediate of a GroEL–GroES-mediated protein folding reaction. *Cell* **84**: 481–490.
- Fenton, W. A. and Horwich, A. L. (2003) Chaperonin-mediated protein folding: fate of substrate polypeptide. *Quarterly Reviews of Biophysics* **36**: 229–256.
- Fenton, W. A., Kashi, Y., Furtak, K., and Horwich, A. L. (1994) Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* **371**: 614–619.
- Fersht, A. and Daggett, V. (2002) Protein folding and unfolding at atomic resolution. *Cell* **108**: 573.
- Fierke, C. A., Johnson, K. A., and Benkovic, S. J. (1987) Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from *Escherichia coli*. *Biochemistry* **26**: 4085–4092.
- Finn, R., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2010) The Pfam protein families database. *Nucleic Acids Research* **38**: D211–D222.
- Fiser, A., Do, R. K., and Sali, A. (2000) Modeling of loops in protein structures. *Protein Science* **9**: 1753–1773.
- Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. (2004) Probing protein structure by limited proteolysis. *Acta Biochimica Polonica (English Edition)* **51**: 299–322.
- Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J. M., *et al.* (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Molecular Microbiology* **43**: 1387–1400.
- Fox, J. D., Kapust, R. B., and Waugh, D. S. (2001) Single amino acid substitutions on the surface of *Escherichia coli* maltose-binding protein can have a profound impact on the solubility of fusion proteins. *Protein Science* **10**: 622–630.
- Frick, D. N. and Richardson, C. C. (2001) DNA primases. *Annual Review of Biochemistry* **70**: 39–80.
- Fujiwara, K., Ishihama, Y., Nakahigashi, K., Soga, T., and Taguchi, H. (2010) A systematic survey of *in vivo* obligate chaperonin-dependent substrates. *EMBO Journal* **29**: 1552–1564.

- Gaitanaris, G. A., Vysokanov, A., Hung, S. C., Gottesman, M. E., and Gragerov, A. (1994) Successive action of *Escherichia coli* chaperones *in vivo*. *Molecular Microbiology* **14**: 861–869.
- Gales, A. C., Jones, R. N., and Sader, H. S. (2006) Global assessment of the antimicrobial activity of polymyxin B against 54 731 clinical isolates of Gram-negative bacilli: report from the SENTRY antimicrobial surveillance programme (2001–2004). *Clinical Microbiology and Infection* **12**: 315–321.
- Gardner, K. H., Zhang, X., Gehring, K., and Kay, L. E. (1998) Solution NMR studies of a 42 KDa *Escherichia coli* maltose binding protein/ $\beta$ -cyclodextrin complex: Chemical shift assignments and analysis. *Journal of the American Chemical Society* **120**: 11738–11748.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., and Bairoch, A. (2005) Protein identification and analysis tools on the ExPASy server, in: *The Proteomics Protocols Handbook*, edited by J. M. Walker, Springer, pp. 571–607.
- Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M., Somera, A., Kyrpides, N., Anderson, I., Gelfand, M., *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology* **185**: 5673–5684.
- Gerstein, M. and Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Science* **7**: 445–456.
- Ghuysen, J. M. (1991) Serine  $\beta$ -lactamases and penicillin-binding proteins. *Annual Review of Microbiology* **45**: 37–67.
- Gil, R., Silva, F. J., Peretó, J., and Moya, A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews* **68**: 518–537.
- Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., Smith, H. O., and Venter, J. C. (2006) Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 425–430.
- Goehring, N. and Beckwith, J. (2005) Diverse paths to midcell: assembly of the bacterial cell division machinery. *Current Biology* **15**: R514–R526.

- Goh, C.-S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *Journal of Molecular Biology* **336**: 115–130.
- Goloubinoff, P., Gatenby, A. A., and Lorimer, G. H. (1989) GroE heat-shock proteins promote assembly of foreign prokaryotic ribulose biphosphate carboxylase oligomers in *Escherichia coli*. *Nature* **337**: 44–47.
- Gough, J. (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics* **21**: 1464–1471.
- Griep, M. A., Adkins, B. J., Hromas, D., Johnson, S., and Miller, J. (1997) The tyrosine photophysics of a primase-derived peptide are sensitive to the peptide's zinc-bound state: proof that the bacterial primase hypothetical zinc finger sequence binds zinc. *Biochemistry* **36**: 544–553.
- Griep, M. A. and Lokey, E. R. (1996) The role of zinc and the reactivity of cysteines in *Escherichia coli* primase. *Biochemistry* **35**: 8260–8267.
- Güntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics* **31**: 145–237.
- Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database* **2013**: article ID bap001.
- Haeusser, D., Garza, A., Buscher, A., and Levin, P. (2007) The division inhibitor EzrA contains a seven-residue patch required for maintaining the dynamic nature of the medial FtsZ ring. *Journal of Bacteriology* **189**: 9001–9010.
- Haeusser, D., Schwartz, R., Smith, A., Oates, M., and Levin, P. (2004) EzrA prevents aberrant cell division by modulating assembly of the cytoskeletal protein FtsZ. *Molecular Microbiology* **52**: 801–814.
- Hammarström, M., Hellgren, N., van den Berg, S., Berglund, H., and Hård, T. (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Science* **11**: 313–321.
- Hamoen, L., Meile, J., De Jong, W., Noirot, P., and Errington, J. (2005) SepF, a novel FtsZ-interacting protein required for a late step in cell division. *Molecular Microbiology* **59**: 989–999.



- Hartl, F. U. (1996) Molecular chaperones in cellular protein folding. *Nature* **381**: 571–580.
- Hartl, F. U., Bracher, A., and Hayer-Hartl, M. (2011) Molecular chaperones in protein folding and proteostasis. *Nature* **475**: 324–332.
- Heim, R., Prasher, D. C., and Tsien, R. Y. (1994) Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 12501–12504.
- Henikoff, S. (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28**: 351–359.
- Henriques, A. O., Glaser, P., Piggot, P. J., and Moran Jr, C. P. (1998) Control of cell shape and elongation by the *rodA* gene in *Bacillus subtilis*. *Molecular Microbiology* **28**: 235–247.
- Hernández, H. and Robinson, C. V. (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nature Protocols* **2**: 715–726.
- Herschlag, D. (1988) The role of induced fit and conformational changes of enzymes in specificity and catalysis. *Bioorganic Chemistry* **16**: 62–96.
- Hesterkamp, T., Hauser, S., Lütcke, H., and Bukau, B. (1996) *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 4437–4441.
- Hlodan, R., Tempst, P., and Hartl, F. U. (1995) Binding of defined regions of a polypeptide to GroEL and its implications for chaperonin-mediated protein folding. *Nature Structural Biology* **2**: 587–595.
- Hoheisel, J. D. (1993) On the activities of *Escherichia coli* exonuclease. *Analytical Biochemistry* **209**: 238–246.
- Horwich, A. L., Apetri, A. C., and Fenton, W. A. (2009) The GroEL/GroES *cis* cavity as a passive anti-aggregation device. *FEBS Letters* **583**: 2654–2662.
- Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F., and Hartl, F. U. (1999) Identification of *in vivo* substrates of the chaperonin GroEL. *Nature* **402**: 147–154.
- Hsin, K., Sheng, Y., Harding, M., Taylor, P., and Walkinshaw, M. (2008) MESPEUS: a database of the geometry of metal sites in proteins. *Journal of*

- Applied Crystallography* **41**: 963–968.
- Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**: 2165–2169.
- Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) Nature and consequences of GroEL-protein interactions. *Biochemistry* **34**: 14581–14587.
- Ivankov, D. and Finkelstein, A. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 8942–8944.
- Jahn, T. R. and Radford, S. E. (2005) The Yin and Yang of protein folding. *FEBS Journal* **272**: 5962–5970.
- Jahn, T. R. and Radford, S. E. (2008) Folding versus aggregation: Polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics* **469**: 100–117.
- Jardetzky, O. and Roberts, G. C. K. (1981) NMR in Molecular Biology. Academic Press, New York.
- Jergic, S., Ozawa, K., Williams, N., Su, X.-C., Scott, D., Hamdan, S., Crowther, J., Otting, G., and Dixon, N. (2007) The unstructured C-terminus of the  $\tau$  subunit of *Escherichia coli* DNA polymerase III holoenzyme is the site of interaction with the  $\alpha$  subunit. *Nucleic Acids Research* **35**: 2813–2824.
- Ji, Y., Zhang, B., Van Horn, S. F., Warren, P., Woodnutt, G., Burnham, M. K. R., and Rosenberg, M. (2001) Identification of critical *staphylococcal* genes using conditional phenotypes generated by antisense RNA. *Science* **293**: 2266–2269.
- Jia, Y., Gregory Dewey, T., Shindyalov, I. N., and Bourne, P. E. (2004) A new scoring function and associated statistical significance for structure alignment by CE. *Journal of Computational Biology* **11**: 787–799.
- Jones, L., Carballido-López, R., and Errington, J. (2001) Control of cell shape in bacteria: helical, actin-like filaments in *Bacillus subtilis*. *Cell* **104**: 913–922.
- Jorge, A., Hoiczky, E., Gomes, J., and Pinho, M. (2011) EzrA contributes to the regulation of cell size in *Staphylococcus aureus*. *PLoS One* **6**: e27542.
- Jung, S., Honegger, A., and Plückthun, A. (1999) Selection for improved protein stability by phage display. *Journal of Molecular Biology* **294**: 163–180.
- Kamada, K., Hanaoka, F., and Burley, S. K. (2003) Crystal structure of the

- MazE/MazF complex: molecular bases of antidote-toxin recognition. *Molecular Cell* **11**: 875–884.
- Kapust, R. B. and Waugh, D. S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science* **8**: 1668–1674.
- Katayama, T., Murakami, Y., Wada, C., Ohmori, H., Yura, T., and Nagata, T. (1989) Genetic suppression of a *dnaG* mutation in *Escherichia coli*. *Journal of Bacteriology* **171**: 1485–1491.
- Katzen, F., Fletcher, J. E., Yang, J. P., Kang, D., Peterson, T. C., Cappuccio, J. A., Blanchette, C. D., Sulchek, T., Chromy, B. A., Hoepflich, P. D., *et al.* (2008) Insertion of membrane proteins into discoidal membranes using a cell-free protein expression approach. *Journal of Proteome Research* **7**: 3535–3542.
- Kawasaki, M. and Inagaki, F. (2001) Random PCR-based screening for soluble domains using green fluorescent protein. *Biochemical and Biophysical Research Communications* **280**: 842–844.
- Kay, B. K., Thai, S., and Volgina, V. V. (2009) High-throughput biotinylation of proteins. *Methods in Molecular Biology* **498**: 185–196.
- Keck, J. L., Roche, D. D., Lynch, A. S., and Berger, J. M. (2000) Structure of the RNA polymerase domain of *E. coli* primase. *Science* **287**: 2482–2486.
- Keeler, J. (2011) Understanding NMR Spectroscopy. John Wiley & Sons, Inc.
- Kelly, S., Jess, T., and Price, N. (2005) How to study proteins by circular dichroism. *Biochimica et Biophysica Acta—Proteins and Proteomics* **1751**: 119–139.
- Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H.-C., Stines, A. P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., and Mann, M. (2005) Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in *Escherichia coli*. *Cell* **122**: 209–220.
- Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczyk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M. R., and Cebur, S. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8**: 163–163.
- Klammt, C., Schwarz, D., Löhr, F., Schneider, B., Dötsch, V., and Bernhard, F. (2006) Cell-free expression as an emerging technique for the large scale production of integral membrane protein. *FEBS Journal* **273**: 4141–4153.

- Ko, K., Lee, J., Song, J., Baek, J., Oh, W., Chun, J., and Yoon, H. (2006) Screening of essential genes in *Staphylococcus aureus* N315 using comparative genomics and allelic replacement mutagenesis. *Journal of Microbiology and Biotechnology* **16**: 623–632.
- Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., *et al.* (2003) Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 4678–4683.
- Koch, H., Gräfe, N., Schiess, R., and Plückthun, A. (2006) Direct selection of antibodies from complex libraries with the protein fragment complementation assay. *Journal of Molecular Biology* **357**: 427–441.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* **235**: 1501–1531.
- Kuchta, R. D. and Stengel, G. (2010) Mechanism and evolution of DNA primases. *Biochimica et Biophysica Acta—Proteins and Proteomics* **1804**: 1180–1189.
- Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F., and Mackay, J. P. (2011) Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal* **278**: 687–703.
- LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L., Schendel, P. F., and McCoy, J. M. (1993) A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Nature Biotechnology* **11**: 187–193.
- Lecroisey, A., Martineau, P., Hofnung, M., and Delepierre, M. (1997) NMR studies on the flexibility of the poliovirus C3 linear epitope inserted into different sites of the maltose-binding protein. *Journal of Biological Chemistry* **272**: 362–368.
- Lesley, S. A., Graziano, J., Cho, C. Y., Knuth, M. W., and Klock, H. E. (2002) Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Engineering* **15**: 153–160.
- Levin, P. A., Kurtser, I. G., and Grossman, A. D. (1999) Identification and characterization of a negative regulator of FtsZ ring formation in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 9642–9647.

- Levinthal, C. (1969) How to fold gracefully, in: Mossbauer Spectroscopy in Biological Systems, edited by P. Debrunner, J. Tsibris, and E. Muck, University of Illinois Press, Urbana, IL, pp. 22–24.
- Liberek, K., Wall, D., and Georgopoulos, C. (1995) The DnaJ chaperone catalytically activates the DnaK chaperone to preferentially bind the  $\sigma^{32}$  heat shock transcriptional regulator. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 6224–6228.
- Liu, J.-W., Boucher, Y., Stokes, H., and Ollis, D. L. (2006) Improving protein solubility: The use of the *Escherichia coli* dihydrofolate reductase gene as a fusion reporter. *Protein Expression and Purification* **47**: 258–263.
- Lorimer, G. H. (1996) A quantitative assessment of the role of the chaperonin proteins in protein folding *in vivo*. *FASEB Journal* **10**: 5–9.
- Louis, J. M., McDonald, R. A., Nashed, N. T., Wondrak, E. M., Jerina, D. M., Oroszlan, S., and Mora, P. T. (1991) Autoprocessing of the HIV-1 protease using purified wild-type and mutated fusion proteins expressed at high levels in *Escherichia coli*. *European Journal of Biochemistry* **199**: 361–369.
- Lu, C., Stricker, J., and Erickson, H. P. (1998) FtsZ from *Escherichia coli*, *Azotobacter vinelandii*, and *Thermotoga maritima*—quantitation, GTP hydrolysis, and assembly. *Cell Motility and the Cytoskeleton* **40**: 71–86.
- Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. (2006) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnology* **25**: 117–124.
- Lutkenhaus, J. and Addinall, S. G. (1997) Bacterial cell division and the Z ring. *Annual Review of Biochemistry* **66**: 93–116.
- Maass, S., Sievers, S., Zühlke, D., Kuzinski, J., Sappa, P. K., Muntel, J., Hessling, B., Bernhardt, J., Sietmann, R., Völker, U., *et al.* (2011) Efficient, global-scale quantification of absolute protein amounts by integration of targeted mass spectrometry and two-dimensional gel-based proteomics. *Analytical Chemistry* **83**: 2677–2684.
- Machida, S., Yu, Y., Singh, S. P., Kim, J. D., Hayashi, K., and Kawata, Y. (1998) Overproduction of  $\beta$ -glucosidase in active form by an *Escherichia coli* system coexpressing the chaperonin GroEL/ES. *FEMS Microbiology Letters* **159**: 41–46.
- Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Sali, A.

- (2000) Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure* **29**: 291–325.
- McDonald, C. C. and Phillips, W. D. (1967) Manifestations of the tertiary structures of proteins in high-frequency nuclear magnetic resonance. *Journal of the American Chemical Society* **89**: 6332–6341.
- Melo, F. and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *Journal of Molecular Biology* **277**: 1141–1152.
- Melo, F., Sanchez, R., and Sali, A. (2002) Statistical potentials for fold assessment. *Protein Science* **11**: 430–448.
- Mendelman, L. V., Beauchamp, B. B., and Richardson, C. C. (1994) Requirement for a zinc motif for template recognition by the bacteriophage T7 primase. *EMBO Journal* **13**: 3909.
- Mercer, K. and Weiss, D. (2002) The *Escherichia coli* cell division protein FtsW is required to recruit its cognate transpeptidase, FtsI (PBP3), to the division site. *Journal of Bacteriology* **184**: 904–912.
- Mertens, N., Remaut, E., and Fiers, W. (1995) Tight transcriptional control mechanism ensures stable high-level expression from T7 promoter-based expression plasmids. *Nature Biotechnology* **13**: 175–179.
- Miller, E. and Nickoloff, J. (1995) *Escherichia coli* electrotransformation. *Methods in Molecular Biology* **47**: 105–113.
- Miller, J. H. (1972) Experiments in Molecular Genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Mitkova, A. V., Khopde, S. M., and Biswas, S. B. (2003) Mechanism and stoichiometry of interaction of DnaG primase with DnaB helicase of *Escherichia coli* in RNA primer synthesis. *Journal of Biological Chemistry* **278**: 52253–52261.
- Miyazaki, K. (2002) Random DNA fragmentation with endonuclease V: application to DNA shuffling. *Nucleic Acids Research* **30**: e139–e139.
- Mössner, E., Koch, H., and Plückthun, A. (2001) Fast selection of antibodies without antigen purification: adaptation of the protein fragment complementation assay to select antigen-antibody pairs. *Journal of Molecular Biology* **308**: 115–122.
- Müller, P., Ewers, C., Bertsche, U., Anstett, M., Kallis, T., Breukink, E., Fraipont, C., Terrak, M., Nguyen-Distèche, M., and Vollmer, W. (2007) The essential

- cell division protein FtsN interacts with the murein (peptidoglycan) synthase PBP1B in *Escherichia coli*. *Journal of Biological Chemistry* **282**: 36394–36402.
- Mushegian, A. R. and Koonin, E. V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 10268–10273.
- Nakayama, M. and Ohara, O. (2003) A system using convertible vectors for screening soluble recombinant proteins produced in *Escherichia coli* from randomly fragmented cDNAs. *Biochemical and Biophysical Research Communications* **312**: 825–830.
- Neira, J. L. and Fersht, A. R. (1999) Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments. *Journal of Molecular Biology* **285**: 1309.
- Neira, J. L., Itzhaki, L. S., Ladurner, A. G., Davis, B., de Prat Gay, G., and Fersht, A. R. (1997) Following co-operative formation of secondary and tertiary structure in a single protein module. *Journal of Molecular Biology* **268**: 185–197.
- Neylon, C., Brown, S. E., Kralicek, A. V., Miles, C. S., Love, C. A., and Dixon, N. E. (2000) Interaction of the *Escherichia coli* replication terminator protein (Tus) with DNA: A model derived from DNA-binding studies of mutant proteins by surface plasmon resonance. *Biochemistry* **39**: 11989–11999.
- Nilsson, B., Moks, T., Jansson, B., Abrahmsen, L., Elmblad, A., Holmgren, E., Henrichson, C., Jones, T. A., and Uhlen, M. (1987) A synthetic IgG-binding domain based on *staphylococcal* protein A. *Protein Engineering* **1**: 107–113.
- Nygren, P.-Å., Stefan, S., and Uhlén, M. (1994) Engineering proteins to facilitate bioprocessing. *Trends in Biotechnology* **12**: 184–188.
- Oakley, A. J., Loscha, K. V., Schaeffer, P. M., Liepinsh, E., Pintacuda, G., Wilce, M. C. J., Otting, G., and Dixon, N. E. (2005) Crystal and solution structures of the helicase-binding domain of *Escherichia coli* primase. *Journal of Biological Chemistry* **280**: 11495–11504.
- Ogawa, T. and Okazaki, T. (1980) Discontinuous DNA replication. *Annual Review of Biochemistry* **49**: 421–457.
- Oh, M.-K. and Liao, J. C. (2000) DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli*. *Metabolic Engineering* **2**: 201–209.
- Oldfield, C. J., Ulrich, E. L., Cheng, Y., Dunker, A. K., and Markley, J. L. (2005)

- Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins: Structure, Function, and Bioinformatics* **59**: 444–453.
- Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. (1996) Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **273**: 1392–1395.
- Ostermeier, M. (2003) Theoretical distribution of truncation lengths in incremental truncation libraries. *Biotechnology and Bioengineering* **82**: 564–577.
- Ostermeier, M. and Benkovic, S. J. (2000) Evolution of protein function by domain swapping. *Advances in Protein Chemistry* **55**: 29–77.
- Ostermeier, M., Nixon, A. E., Shim, J. H., and Benkovic, S. J. (1999) Combinatorial protein engineering by incremental truncation. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 3562–3567.
- Ouyang, Z. and Liang, J. (2009) Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Science* **17**: 1256–1263.
- Pan, H., Bird, L. E., and Wigley, D. B. (1999) Cloning, expression, and purification of *Bacillus stearothermophilus* DNA primase and crystallization of the zinc-binding domain. *Biochimica et Biophysica Acta—Gene Structure and Expression* **1444**: 429–433.
- Pan, H. and Wigley, D. B. (2000) Structure of the zinc-binding domain of *Bacillus stearothermophilus* DNA primase. *Structure* **8**: 231–239.
- Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology* **224**: 461–471.
- Pédélecq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2005) Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology* **24**: 79–88.
- Pédélecq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S., Terwilliger, T. C., and Waldo, G. S. (2002) Engineering soluble proteins for structural genomics. *Nature Biotechnology* **20**: 927–932.
- Perez, F., Hujer, A. M., Hujer, K. M., Decker, B. K., Rather, P. N., and Bonomo, R. A. (2007) Global challenge of multidrug-resistant *Acinetobacter baumannii*. *Antimicrobial Agents and Chemotherapy* **51**: 3471–3484.
- Pettitt, C. S., McGuffin, L. J., and Jones, D. T. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics* **21**:



3509–3515.

- Pinho, M. and Errington, J. **(2003)** Dispersed mode of *Staphylococcus aureus* cell wall synthesis in the absence of the division machinery. *Molecular Microbiology* **50**: 871–881.
- Pinho, M., Filipe, S., de Lencastre, H., and Tomasz, A. **(2001)** Complementation of the essential peptidoglycan transpeptidase function of penicillin-binding protein 2 (PBP2) by the drug resistance protein PBP2A in *Staphylococcus aureus*. *Journal of Bacteriology* **183**: 6525–6531.
- Platt, A., Woodhall, R., and George, A. **(2007)** Improved DNA sequencing quality and efficiency using an optimized fast cycle sequencing protocol. *BioTechniques* **43**: 58.
- Pouwels, L. J., Zhang, L., Chan, N. H., Dorrestein, P. C., and Wachter, R. M. **(2008)** Kinetic isotope effect studies on the *de novo* rate of chromophore formation in fast-and slow-maturing GFP variants. *Biochemistry* **47**: 10111–10122.
- Power, R. F., Conneely, O. M., McDonnell, D. P., Clark, J. H., Butt, T., Schrader, W. T., and O'Malley, B. **(1990)** High level expression of a truncated chicken progesterone receptor in *Escherichia coli*. *Journal of Biological Chemistry* **265**: 1419–1424.
- Punta, M., Coghill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al.* **(2012)** The Pfam protein families database. *Nucleic Acids Research* **40**: D290–D301.
- Putney, S. D., Benkovic, S. J., and Schimmel, P. R. **(1981)** A DNA fragment with an  $\alpha$ -phosphorothioate nucleotide at one end is asymmetrically blocked from digestion by exonuclease III and can be replicated *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America* **78**: 7350–7354.
- Qian, J., Luscombe, N., and Gerstein, M. **(2001)** Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology* **313**: 673–681.
- Qimron, U., Lee, S.-J., Hamdan, S. M., and Richardson, C. C. **(2006)** Primer initiation and extension by T7 DNA primase. *EMBO Journal* **25**: 2199–2208.
- Randall, A. and Baldi, P. **(2008)** SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC*

*Structural Biology* **8**: 52–52.

- Raran-Kurussi, S. and Waugh, D. S. (2012) The ability to enhance the solubility of its fusion partners is an intrinsic property of maltose-binding protein but their folding is either spontaneous or chaperone-mediated. *PLoS One* **7**: e49589.
- Rehm, T., Huber, R., and Holak, T. A. (2002) Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. *Structure* **10**: 1613–1618.
- Reid, B. G. and Flynn, G. C. (1997) Chromophore formation in green fluorescent protein. *Biochemistry* **36**: 6786–6791.
- Richarme, G. and Caldas, T. D. (1997) Chaperone properties of the bacterial periplasmic substrate-binding proteins. *Journal of Biological Chemistry* **272**: 15607–15612.
- Robinson, A., Brzoska, A., Turner, K., Withers, R., Harry, E., Lewis, P., and Dixon, N. (2010) Essential biological processes of an emerging pathogen: DNA replication, transcription, and cell division in *Acinetobacter spp.* *Microbiology and Molecular Biology Reviews* **74**: 273–297.
- Robinson, A., Ruiz, S., and Dixon, N. (2007–2011) Unpublished.
- Romberg, L., Simon, M., and Erickson, H. P. (2001) Polymerization of FtsZ, a bacterial homolog of tubulin: Is assembly cooperative? *Journal of Biological Chemistry* **276**: 11743–11753.
- Rosenow, M., Patel, H., and Wachter, R. (2005) Oxidative chemistry in the GFP active site leads to covalent cross-linking of a modified leucine side chain with a histidine imidazole: implications for the mechanism of chromophore formation. *Biochemistry* **44**: 8303–8311.
- Rüdiger, S., Germeroth, L., Schneider-Mergener, J., and Bukau, B. (1997) Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO Journal* **16**: 1501–1507.
- Rye, H. S., Burston, S. G., Fenton, W. A., Beechem, J. M., Xu, Z., Sigler, P. B., and Horwich, A. L. (1997) Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL. *Nature* **388**: 792–798.
- Saavedra-Alanis, V., Rysavy, P., Rosenberg, L., and Kalousek, F. (1994) Rat liver mitochondrial processing peptidase. Both  $\alpha$ - and  $\beta$ -subunits are required for activity. *Journal of Biological Chemistry* **269**: 9284–9288.

- Sachdev, D. and Chirgwin, J. M. (1998) Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. *Protein Expression and Purification* **12**: 122–132.
- Saibil, H. R., Zheng, D., Roseman, A. M., Hunter, A. S., Watson, G. M. F., Chen, S., auf der Mauer, A., O'Hara, B. P., Wood, S. P., Mann, N. H., *et al.* (1993) ATP induces large quaternary rearrangements in a cage-like chaperonin structure. *Current Biology* **3**: 265–273.
- Salema, V. and Fernández, L. Á. (2013) High yield purification of nanobodies from the periplasm of *E. coli* as fusions with the maltose binding protein. *Protein Expression and Purification* **91**: 42–48.
- Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**: 779–815.
- Sambrook, J., Fritsch, E., and Maniatis, T. (1989) Molecular Cloning: a Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2<sup>nd</sup> edn.
- Samudrala, R. and Moul, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology* **275**: 895–916.
- Samuelsson, E., Moks, T., Uhlen, M., and Nilsson, B. (1994) Enhanced *in vitro* refolding of insulin-like growth factor I using a solubilizing fusion partner. *Biochemistry* **33**: 4207–4211.
- San Martin, M. C., Stamford, N. P. J., Dammerova, N., Dixon, N. E., and Carazo, J. M. (1995) A structural model for the Escherichia coli DnaB helicase based on electron microscopy data. *Journal of Structural Biology* **114**: 167–176.
- Savageau, M. A. (1986) Proteins of *Escherichia coli* come in sizes that are multiples of 14 kDa: domain concepts and evolutionary implications. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 1198–1202.
- Sawaya, M. R. and Kraut, J. (1997) Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry* **36**: 586–603.
- Schechter, I. and Berger, A. (1967) On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications* **27**: 157–162.
- Scheffers, D. and Pinho, M. (2005) Bacterial cell wall synthesis: new insights from

- localization studies. *Microbiology and Molecular Biology Reviews* **69**: 585–607.
- Schrödinger, LLC (2012) *The PyMOL Molecular Graphics System*, Version 1.5.7.
- Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H. M., Jones, D., Brenner, S. E., Burley, S. K., Das, R., Dokholyan, N. V., *et al.* (2009) Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17**: 151–159.
- Shindyalov, I. N. and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11**: 739–747.
- Sieber, V., Plückthun, A., and Schmid, F. X. (1998) Selecting proteins with improved stability by a phage-based method. *Nature Biotechnology* **16**: 955–960.
- Sippl, M. J. (2009) Fold space unlimited. *Current Opinion in Structural Biology* **19**: 312–320.
- Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., and Godzik, A. (2007) The challenge of protein structure determination—lessons from structural genomics. *Protein Science* **16**: 2472–2482.
- Smith, D. B. and Johnson, K. S. (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* **67**: 31–40.
- Sniegowski, J., Lappe, J., Patel, H., Huffman, H., and Wachter, R. (2005a) Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. *Journal of Biological Chemistry* **280**: 26248–26255.
- Sniegowski, J., Phail, M., and Wachter, R. (2005b) Maturation efficiency, trypsin sensitivity, and optical properties of Arg96, Glu222, and Gly67 variants of green fluorescent protein. *Biochemical and Biophysical Research Communications* **332**: 657–663.
- Son, S. H. and Lee, H. H. (2013) The N-terminal domain of EzrA binds to the C terminus of FtsZ to inhibit *Staphylococcus aureus* FtsZ polymerization. *Biochemical and Biophysical Research Communications* **433**: 108–114.
- Steele, V. R., Bottomley, A. L., Garcia-Lara, J., Kasturiarachchi, J., and Foster, S. J. (2011) Multiple essential roles for EzrA in cell division of *Staphylococcus aureus*. *Molecular Microbiology* **80**: 542–555.

- Stemmer, W. P. C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**: 389–391.
- Stoller, G., Rücknagel, K., Nierhaus, K., Schmid, F., Fischer, G., and Rahfeld, J. (1995) A ribosome-associated peptidyl-prolyl *cis/trans* isomerase identified as the trigger factor. *EMBO Journal* **14**: 4939.
- Studier, F. (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *Journal of Molecular Biology* **219**: 37–44.
- Studier, F. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expression and Purification* **41**: 207–234.
- Studier, F. and Moffatt, B. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology* **189**: 113–130.
- Taguchi, H., Ueno, T., Tadakuma, H., Yoshida, M., and Funatsu, T. (2001) Single-molecule observation of protein-protein interactions in the chaperonin system. *Nature Biotechnology* **19**: 861–865.
- Teter, S. A., Houry, W. A., Ang, D., Tradler, T., Rockabrand, D., Fischer, G., Blum, P., Georgopoulos, C., and Hartl, F. U. (1999) Polypeptide flux through bacterial Hsp70: DnaK cooperates with trigger factor in chaperoning nascent chains. *Cell* **97**: 755–765.
- Thomas, J. G., Ayling, A., and Baneyx, F. (1997) Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli*. To fold or to refold. *Applied Biochemistry and Biotechnology* **66**: 197–238.
- Thompson, J. D., Gibson, T., Higgins, D. G., *et al.* (2002) Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics* : 2–3.
- Tomoyasu, T., Ogura, T., Tatsuta, T., and Bukau, B. (2002) Levels of DnaK and DnaJ provide tight control of heat shock gene expression and protein repair in *Escherichia coli*. *Molecular Microbiology* **30**: 567–581.
- Tosatto, S. C. (2005) The victor/FRST function for model quality estimation. *Journal of Computational Biology* **12**: 1316–1327.
- Tougu, K. and Marians, K. J. (1996) The interaction between helicase and primase sets the replication fork clock. *Journal of Biological Chemistry* **271**: 21398–21405.
- Tougu, K., Peng, H., and Marians, K. J. (1994) Identification of a domain of

- Escherichia coli* primase required for functional interaction with the DnaB helicase at the replication fork. *Journal of Biological Chemistry* **269**: 4675.
- van den Berg, B., Ellis, R., and Dobson, C. (1999) Effects of macromolecular crowding on protein folding and aggregation. *EMBO Journal* **18**: 6927–6933.
- Vanbogelen, R., Sankar, P., Clark, R., Bogan, J., and Neidhardt, F. (2005) The gene-protein database of *Escherichia coli*: Edition 5. *Electrophoresis* **13**: 1014–1054.
- Villegas, M. V. and Hartstein, A. I. (2003) *Acinetobacter* outbreaks, 1977–2000. *Infection Control and Hospital Epidemiology* **24**: 284–295.
- Vogel, C., Teichmann, S. A., and Pereira-Leal, J. (2005) The relationship between domain duplication and recombination. *Journal of Molecular Biology* **346**: 355–366.
- Waldo, G. (2003) Genetic screens and directed evolution for protein solubility. *Current Opinion in Chemical Biology* **7**: 33–38.
- Waldo, G., Standish, B., Berendzen, J., and Terwilliger, T. (1999) Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology* **17**: 691–695.
- Wallner, B. and Elofsson, A. (2003) Can correct protein models be identified? *Protein Science* **12**: 1073–1086.
- Weart, R. B. and Levin, P. A. (2003) Growth rate-dependent regulation of medial FtsZ ring formation. *Journal of Bacteriology* **185**: 2826–2834.
- Weiner, J., Beaussart, F., and Bornberg-Bauer, E. (2006) Domain deletions and substitutions in the modular protein evolution. *FEBS Journal* **273**: 2037–2047.
- Williams, N. K., Prosselkov, P., Liepinsh, E., Line, I., Sharipo, A., Littler, D. R., Curmi, P. M. G., Otting, G., and Dixon, N. E. (2002) *In vivo* protein cyclization promoted by a circularly permuted *Synechocystis* sp. PCC6803 DnaB mini-intein. *Journal of Biological Chemistry* **277**: 7790–7798.
- Wishart, D. S. (2011) Interpreting protein chemical shift data. *Progress in Nuclear Magnetic Resonance Spectroscopy* **58**: 62–87.
- Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995)  $^1\text{H}$ ,  $^{13}\text{C}$  and  $^{15}\text{N}$  random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of Biomolecular NMR* **5**: 67–81.

- Wu, C. A., Zechner, E. L., and Marians, K. J. (1992) Coordinated leading- and lagging-strand synthesis at the *Escherichia coli* DNA replication fork. I. Multiple effectors act to modulate Okazaki fragment size. *Journal of Biological Chemistry* **267**: 4030–4044.
- Wu, Y., Vadrevu, R., Kathuria, S., Yang, X., and Matthews, C. R. (2007) A tightly packed hydrophobic cluster directs the formation of an off-pathway sub-millisecond folding intermediate in the alpha subunit of tryptophan synthase, a TIM barrel protein. *Journal of Molecular Biology* **366**: 1624–1638.
- Wuu, J. J. and Swartz, J. R. (2008) High yield cell-free production of integral membrane proteins without refolding or detergents. *Biochim Biophys Acta* **1778**: 1237–1250.
- Xu, D. and Nussinov, R. (1998) Favorable domain size in proteins. *Folding and Design* **3**: 11–17.
- Xu, Z., Horwich, A. L., and Sigler, P. B. (1997) The crystal structure of the asymmetric GroEL-GroES-(ADP)<sup>7</sup> chaperonin complex. *Nature* **388**: 741–750.
- Yumerefendi, H., Tarendeau, F., Mas, P. J., and Hart, D. J. (2010) ESPRIT: An automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *Journal of Structural Biology* **172**: 66–74.
- Zhang, J., Zhang, Y., and Inouye, M. (2003) Characterization of the interactions within the *mazEF* addiction module of *Escherichia coli*. *Journal of Biological Chemistry* **278**: 32300–32306.
- Zhang, Y., Olsen, D. R., Nguyen, K. B., Olson, P. S., Rhodes, E. T., and Mascarenhas, D. (1998) Expression of eukaryotic proteins in soluble form in *Escherichia coli*. *Protein Expression and Purification* **12**: 159–165.
- Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* **11**: 2714–2726.

# Appendices



## Appendix A

### Apparatus for generating exonuclease III libraries

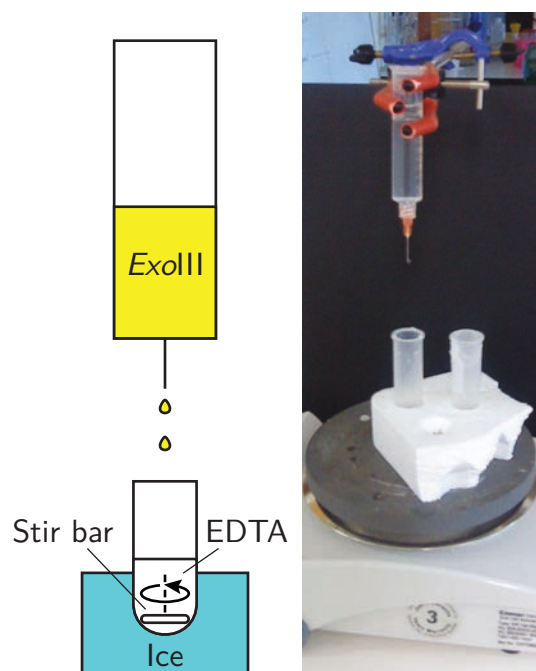
A library truncation method was constructed by setting up an apparatus to constantly produce samples from an *ExoIII* reaction into a chilled EDTA reaction stop solution (Figure A.1). The apparatus was constructed from a 10 mL syringe and a 25 gauge needle supported above a 10 mL centrifuge tube containing a magnetic stir bar. The 10 mL centrifuge tube and stir bar were placed in a chilled container on a magnetic stirrer where the *ExoIII* droplet apparatus produced 10 droplets per 7 s, with a combined volume of 70  $\mu\text{L}$ .

The *ExoIII* reactions were initiated by mixing pre-warmed DNA and enzyme solutions. DNA solutions were prepared with an appropriate amount of DNA in *ExoIII* reaction buffer (60 mM Tris-HCl pH 7.6, 5 mM  $\text{MgCl}_2$ , 1 mM DTT and 100  $\mu\text{g}\cdot\text{mL}^{-1}$  BSA). A separate *ExoIII* solution was prepared by mixing 100 units of *ExoIII* per  $\mu\text{g}$  DNA in 1 mL of *ExoIII* buffer.

*ExoIII* reactions were initiated by mixing DNA and *ExoIII* solutions together and

placing the resulting solution in the drip apparatus. The *ExoIII* treated droplets were collected in a 10 mL centrifuge tube containing 1/20 *ExoIII* reaction volume of 500 mM EDTA and a magnetic stir bar so that the final EDTA concentration would be 25 mM at completion of the *ExoIII* experiment.

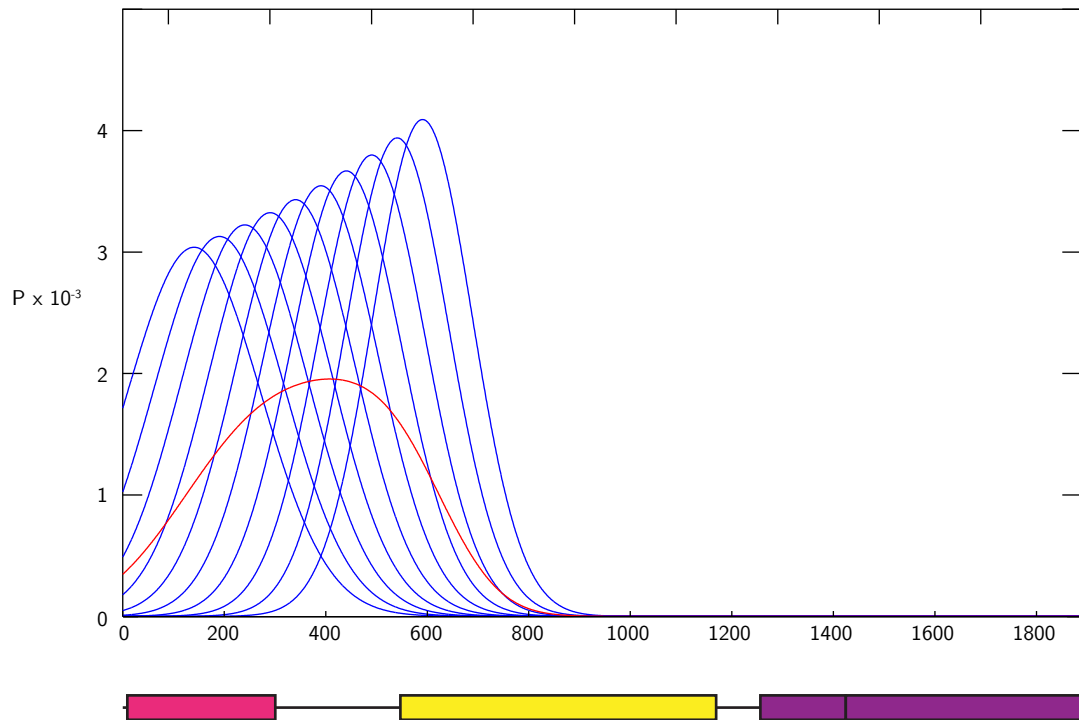
Following generation of an *ExoIII* truncated DNA library, the DNA solution was concentrated to a volume of 250  $\mu\text{L}$  in a pre-washed Centricon 10 k MWCO centrifugal filter unit. The concentrated DNA was then precipitated by addition of 750  $\mu\text{L}$  100 % ethanol for 30 min. Precipitated DNA was collected by centrifugation at  $21,000 \times g$  for 15 min. The DNA pellet was washed with chilled 70 % ethanol and re-centrifuged. The purified DNA pellet was air dried prior to re-suspension in TE.



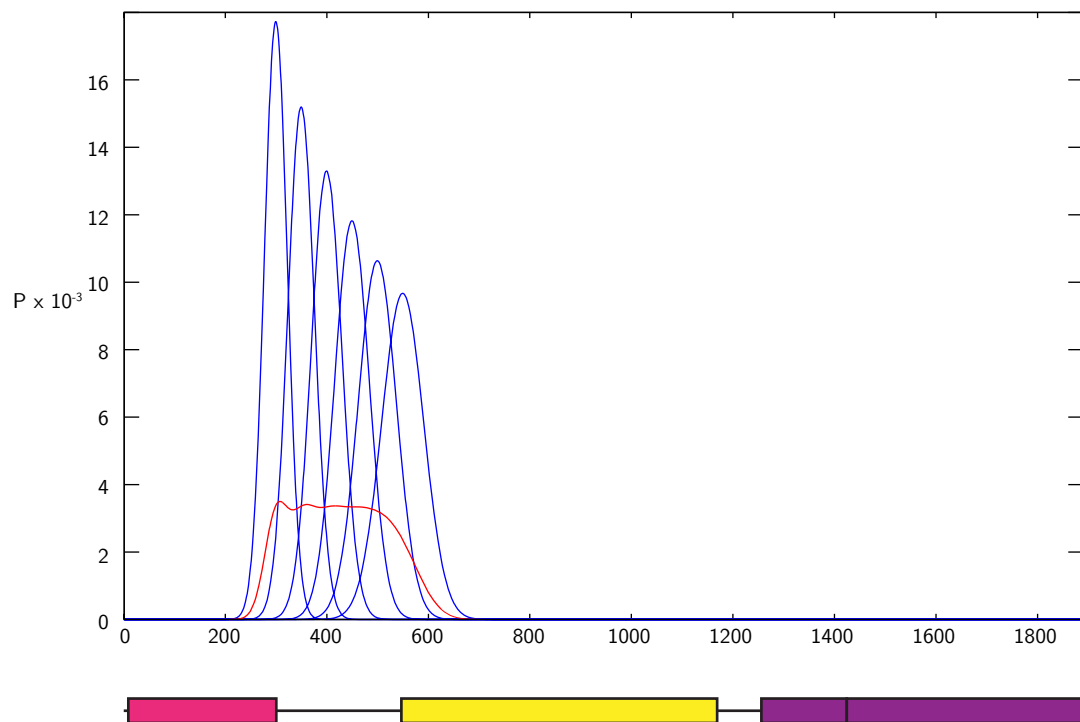
**Figure A.1:** Apparatus for generating exonuclease III libraries.

## Appendix B

### Library population prediction



**Figure B.1: Distribution of truncation lengths for an ideal uni-directional exonuclease III truncation library.** Truncation was modelled using Equation 3.1 (Hoheisel, 1993; Ostermeier, 2003) where  $L$  = time points samples for C-terminal deletion of *A. baylyi dnaG* (blue; sampled every 15 s between 6 min 30 s and 8 min 45 s of the *ExoIII* reaction) and  $c = 0.075$ . The theoretical population from pooling each time point sample is represented in red.



**Figure B.2: Distribution of truncation lengths for an ideal uni-directional exonuclease III truncation library.** Truncation was modelled using Equation 3.1 (Hoheisel, 1993; Ostermeier, 2003) where  $L$  = time points samples for N-terminal deletion of *A. baylyi dnaG* (blue; sampled every 15 s between 1 min 25 s and 2 min 40 s of the *ExoIII* reaction) and  $c = 0.075$ . The theoretical population from pooling each time point sample is represented in red.

# Appendix C

## Oligonucleotides

**Table C.1: Oligonucleotides used in this work.**

Oligonucleotide	Sequence (5'–3')
PET3	CGACTCACTATAGGGAGACCACAAC
PET4	CCTTTCGGGCTTTGTTAGCAG
137	TATGGGATCTAGCGGATCCTCAGGTGGG
138	AATTCCCACCTGAGGATCCGCTAGATCCCA
139	AATTCTCATCACCATCACCATCACCAATTGAGTAC
140	TCAATTGGTGATGGTGATGGTGATGAG
141	AATTGAGGTAGTTCTAAGGTACCA
142	CATGTGGTACCTTAGAACTACCTC
151	GGATGACTGGGAATCGGTATTTCAGCGAATTTACGATGCTGA– TGCGCAGAACTCTC
152	GAGAGTTCTGCGCATCAGCATCGTGAAATTCGCTGAATACCG– ATTCCCAGTCATCC
153	GGAAATCTCACAATTGATCAGTCTGATTGCGGCGTTAGC
154	CTATACAAAAGGTACCTTACCGCCGCTCCAGAATCTCAAAGC
184	TTTAAGAAGGAGAGAATTCTATGGCTGGTCTGAACGAC
205	ATGATCTAGAGTCGCGGGTACCTTACTTGTACAGCTCGTC
206	ATCCACCGGTCGCCCAATTGGTGAGCAAGGGCGAGGAG
207	CTCCTCGCCCTTGCTCACCAATTGATCTTTGCACAGCTGCGC– GCTCAT
208	ACGACGACAAGAGGAATTCTGAAGATGTGGATGAA
209	CTCCTCGCCCTTGCTCACCAATTGATGTTTCGTGCCATTC
235	CTCGCCCTTGCTCACC

Continued on next page...

**Table C.1** – continued from previous page

Oligonucleotide	Sequence (5'-3')
301	TACCATGGCGACCTCGTGAAGGGTGTGCCTGAGTACATATGGT- AGAGGCTTTGCTATTTCAGCGTTTGATGAATGAGGATCCTCTGGG
302	AATTCCCAGAGGATCCTCATTTCATCAAACGCTGAATAGCAAAG- CCTCTACCATATGTACTCAGGCACACCCTTCACGAGGTCGCCA- TGG
305	GATCTGTGTGCCTGAGGGATGGATCCTAAGTAACTAACCATGG- CTCCTCTGGG
306	AATTCCCAGAGGAGCCATGGTTAGTTACTTAGGATCCATCCCT- CAGGCACACA
307	GATCTGTGTGCCTGAGGGATGGATCCTAAGTAACTAACCTCAG- CGGGCTCCTCTGGG
308	AATTCCCAGAGGAGCCCGCTGAGGTTAGTTACTTAGGATCCAT- CCCTCAGGCACACA
337	TTTTTTTTTTTTTTTCATATGGCTATTCCGCAGCATACC
338	TTTTTTTTTTTTTTTGGATCCAGATAATAATCTTAAG
420	AATTGTAAGCTTAC
549	CAGGATGGGCACCACC

# Appendix D

## Enzyme buffers

**Table D.1: Enzyme buffers.**

Buffer	working concentration
NEB 1	10 mM Bis-Tris-Propane-HCl, 10 mM MgCl <sub>2</sub> , 1 mM DTT: pH 7.0
NEB 2	10 mM Tris-HCl, 50 mM NaCl, 10 mM MgCl <sub>2</sub> , 1 mM DTT: pH 7.9
NEB 3	50 mM Tris-HCl, 100 mM NaCl, 10 mM MgCl <sub>2</sub> , 1 mM DTT: pH 7.9
NEB 4	20 mM Tris-acetate, 50 mM potassium acetate, 10 mM magnesium acetate, 1 mM DTT: pH 7.9
Fermentas T4 DNA Ligase	40 mM Tris-HCl, 10 mM MgCl <sub>2</sub> , 10 mM DTT, 0.5 mM ATP: pH 7.8 at 25°C).
NEB T4 DNA Ligase	50 mM Tris-HCl, 10 mM MgCl <sub>2</sub> , 1 mM ATP, 10 mM DTT: pH 7.5 at 25°C
Promega T4 DNA Ligase	30 mM Tris-HCl, 10 mM MgCl <sub>2</sub> , 10 mM DTT and 1 mM ATP: pH 7.8 at 25°C
BIOTAQ Red	67 mM Tris-HCl, 16 mM (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 10 mM KCl, 0.1% stabiliser: pH 8.8 at 25°C
ACUSURE	60 mM Tris-HCl, 6 mM (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 10 mM KCl, 2 mM MgSO <sub>4</sub> : pH 8.3 at 25°C.
ACCUZYME	60 mM Tris-HCl, 6 mM (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 10 mM KCl, 2 mM MgSO <sub>4</sub> : pH 8.3 at 25°C.
VELOCITY	Proprietary Hi-Fi Reaction Buffer

## Appendix E

### Mass spectra and NMR analysis of truncated EzrA proteins



**Table E.1:** Mass spectrometry of His<sub>6</sub>-tagged EzrA truncated proteins.

EzrA	Predicted mass (– methionine; Da)	Observed mass (Da)	N-met	Fragment mass (Da)	Remainder mass (Da)	Notes
EzrA <sup>24–564</sup>	65353.9 (65222.7)	65435.0	✓			
EzrA <sup>277–564</sup>	35752.8 (35621.6)	35760.2	✓			
EzrA <sup>280–564</sup>	35381.4 (35250.2)	35260.1	X			
EzrA <sup>302–564</sup>	32809.6 (32678.4)	32687.6	X			
EzrA <sup>381–564</sup>	23476.2 (23345.0)	23482.0	✓			
EzrA <sup>425–564</sup>	18236.5 (18105.3)	18109.9	X	16434.5	1675.4 <sup>1</sup>	–15 aa
EzrA <sup>443–564</sup>	16161.1 (16029.9)	16033.8	X			
EzrA <sup>453–564</sup>	14902.6 (14771.4)	14905.5	✓			
EzrA <sup>476–564</sup>	12103.5 (11972.3)	12105.0	✓			
EzrA <sup>484–564</sup>	11220.4 (11089.2)	11222.4	✓			
EzrA <sup>24–97</sup>	10406.7 (10275.5)	10407.5	✓			
EzrA <sup>24–126</sup>	13806.2 (13675.0)	13807.6	✓	9655.4 <sup>2</sup>	4152.1 <sup>3</sup>	–35 aa
EzrA <sup>24–128</sup>	14020.4 (13889.2)	14022.0	✓	9870.4 <sup>2</sup>	4151.8 <sup>3</sup>	–35 aa
EzrA <sup>24–129</sup>	14134.5 (14003.3)	14137.5	✓	9984.8 <sup>2</sup>	4152.8 <sup>3</sup>	–35 aa
EzrA <sup>24–139</sup>	15309.8 (15178.6)	15311.7	✓			
EzrA <sup>24–214</sup>	24238.9 (24107.7)	24242.9	✓			
EzrA <sup>24–238</sup>	27082.1 (26950.9)	27088.6	✓	25703.6 <sup>2</sup>	1385.2	
EzrA <sup>24–238</sup>				21440.1 <sup>2</sup>	5648.7 <sup>4</sup>	–48 aa
EzrA <sup>24–238</sup>				21311.6 <sup>2</sup>	5777.2 <sup>5</sup>	–49 aa
EzrA <sup>24–476</sup>	55205.4 (55074.2)	55223.5	✓			

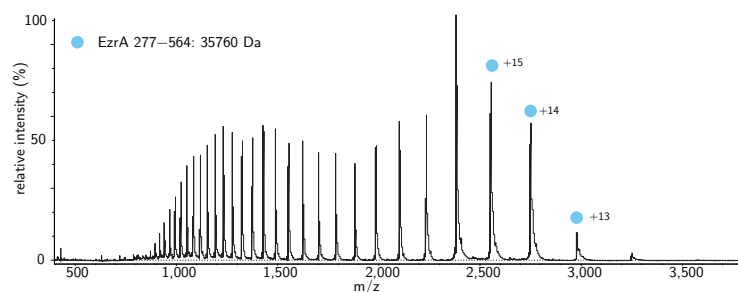
<sup>1</sup> mass is consistent with proteolysis liberating the 15 N-terminal amino acids: SKKEEVYRRLLASN (predicated mass 1692.9 Da).

<sup>2</sup> fragments from secondary chromatography peak.

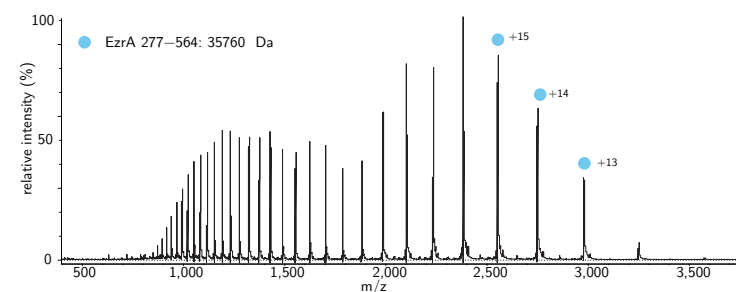
<sup>3</sup> mass is consistent with proteolysis liberating the 35 N-terminal amino acids: (M)RSNKRQIEKAIERKNEIETLPFDQNLAQLSKLN (predicated mass 4169.8 Da).

<sup>4</sup> mass is consistent with proteolysis liberating the 48 N-terminal amino acids: (M)RSNKRQIEKAIERKNEIETLPFDQNLAQLSKLNLKGETKTKYDAMK (predicated mass 5664.5 Da).

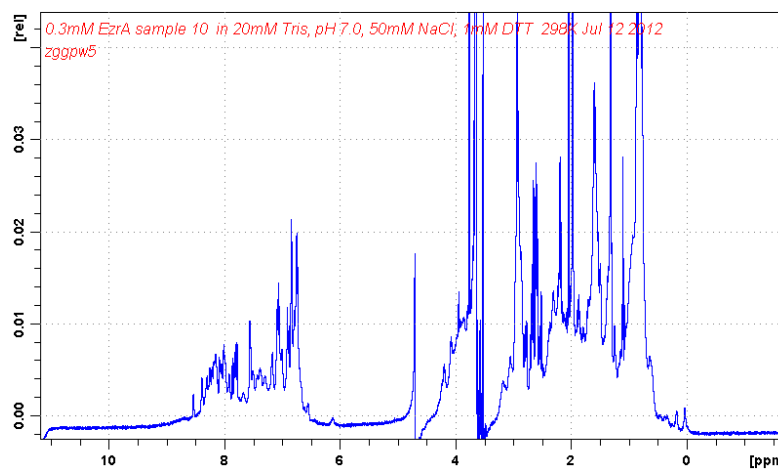
<sup>5</sup> mass is consistent with proteolysis liberating the 49 N-terminal amino acids: (M)RSNKRQIEKAIERKNEIETLPFDQNLAQLSKLNLKGETKTKYDAMKK (predicated mass 5792.7 Da).



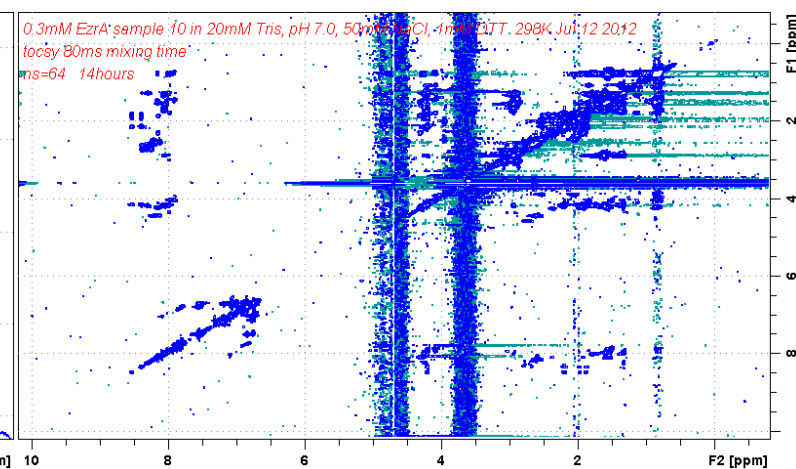
**A:** Mass spectrum of EzrA<sup>277-564</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>277-564</sup>: late anion-exchange chromatography peak

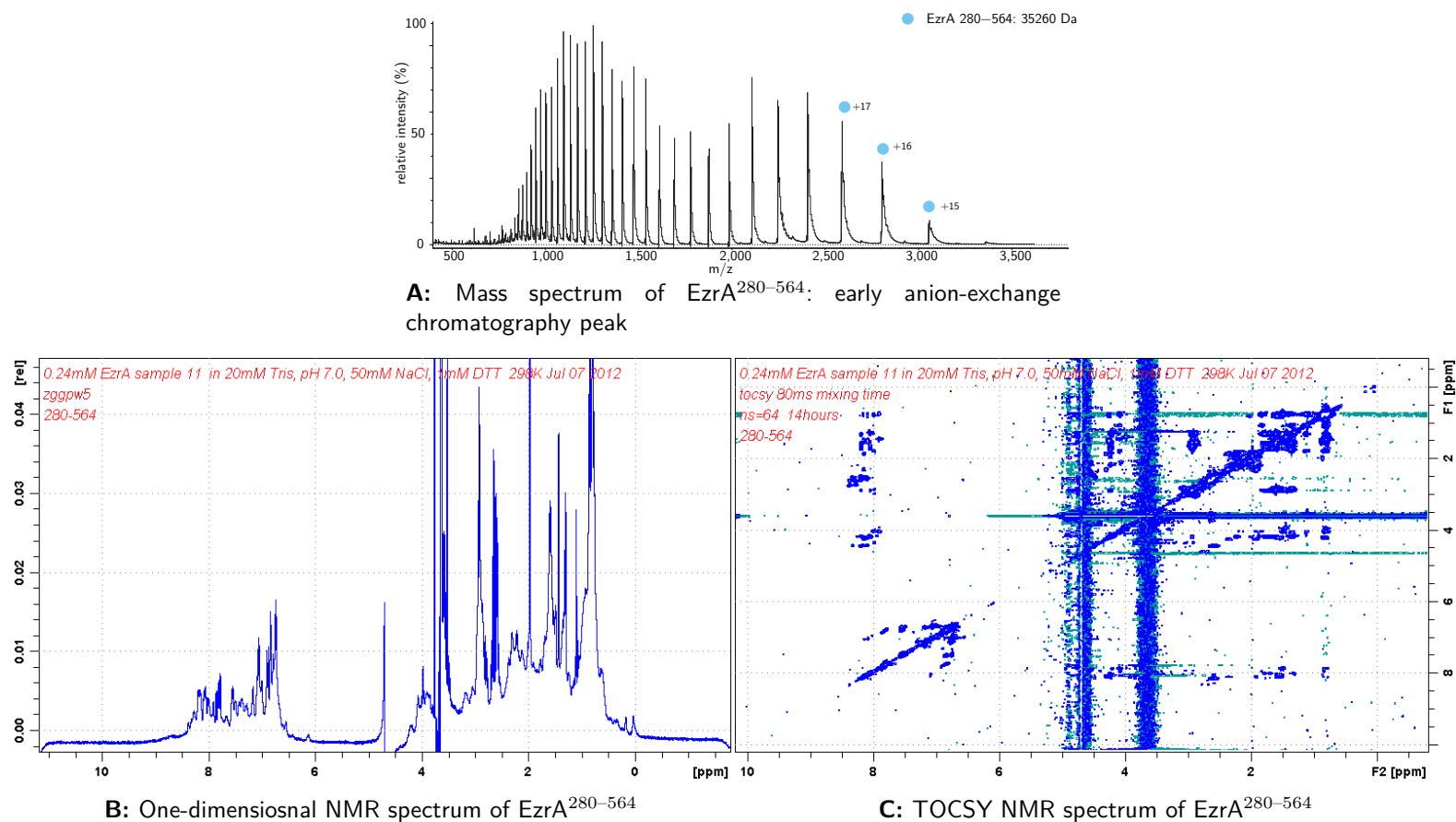


**C:** One-dimensionals NMR spectrum of EzrA<sup>277-564</sup>

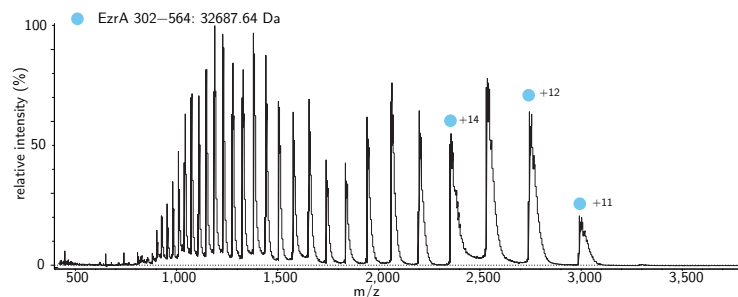


**D:** TOCSY NMR spectrum of EzrA<sup>277-564</sup>

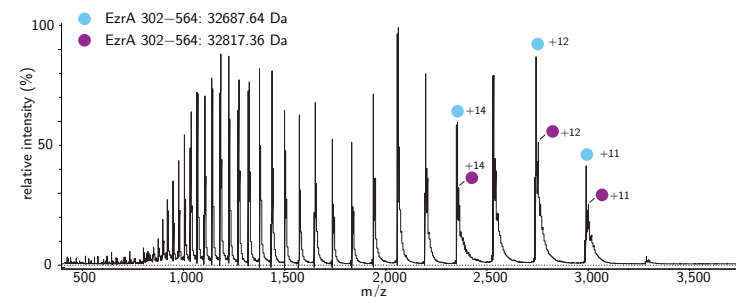
**Figure E.1: Mass and NMR spectral analysis of EzrA<sup>277-564</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>277-564</sup>. To see if purified EzrA<sup>277-564</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



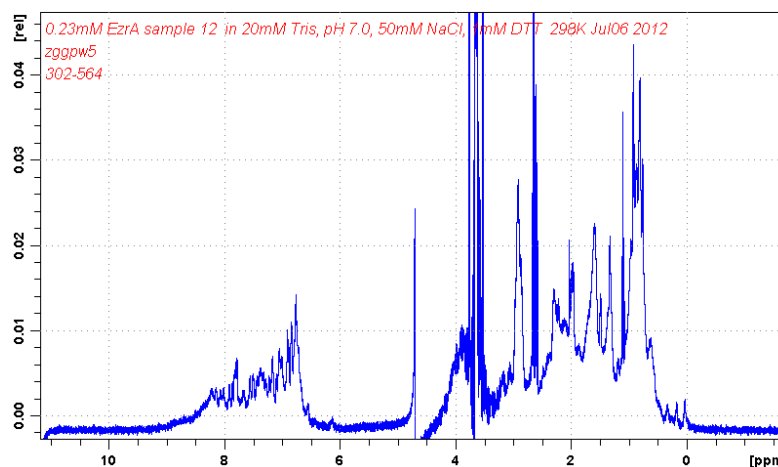
**Figure E.2: Mass and NMR spectral analysis of EzrA<sup>280-564</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>280-564</sup>. To see if purified EzrA<sup>280-564</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



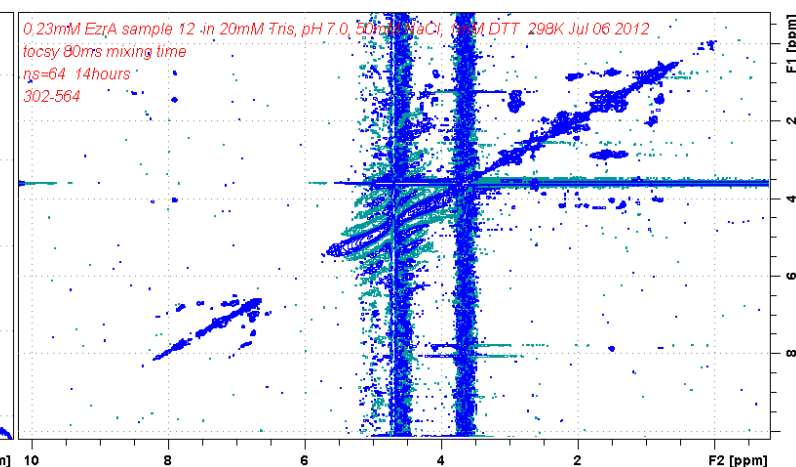
**A:** Mass spectrum of EzrA<sup>302-564</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>302-564</sup>: late anion-exchange chromatography peak

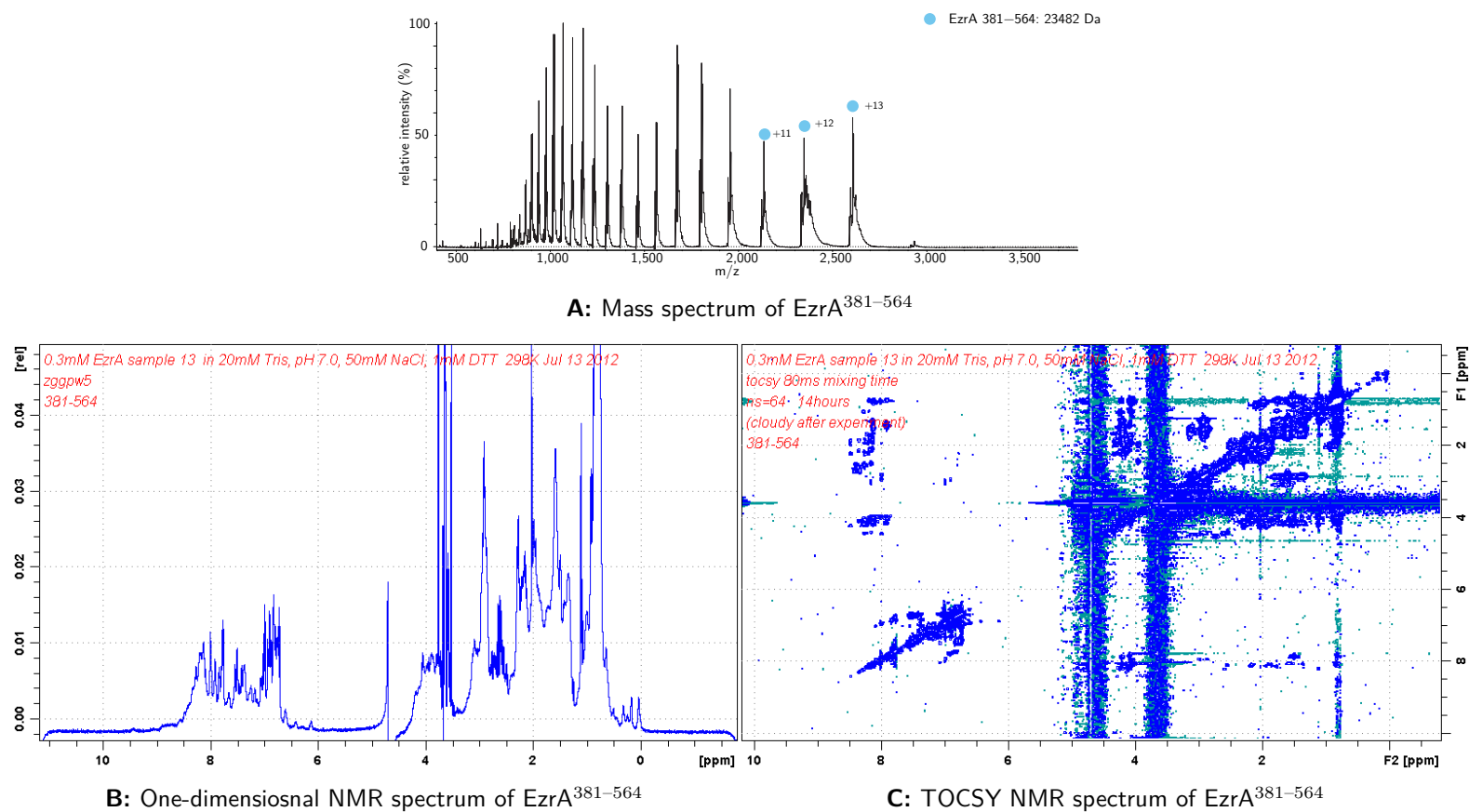


**C:** One-dimensional NMR spectrum of EzrA<sup>302-564</sup>

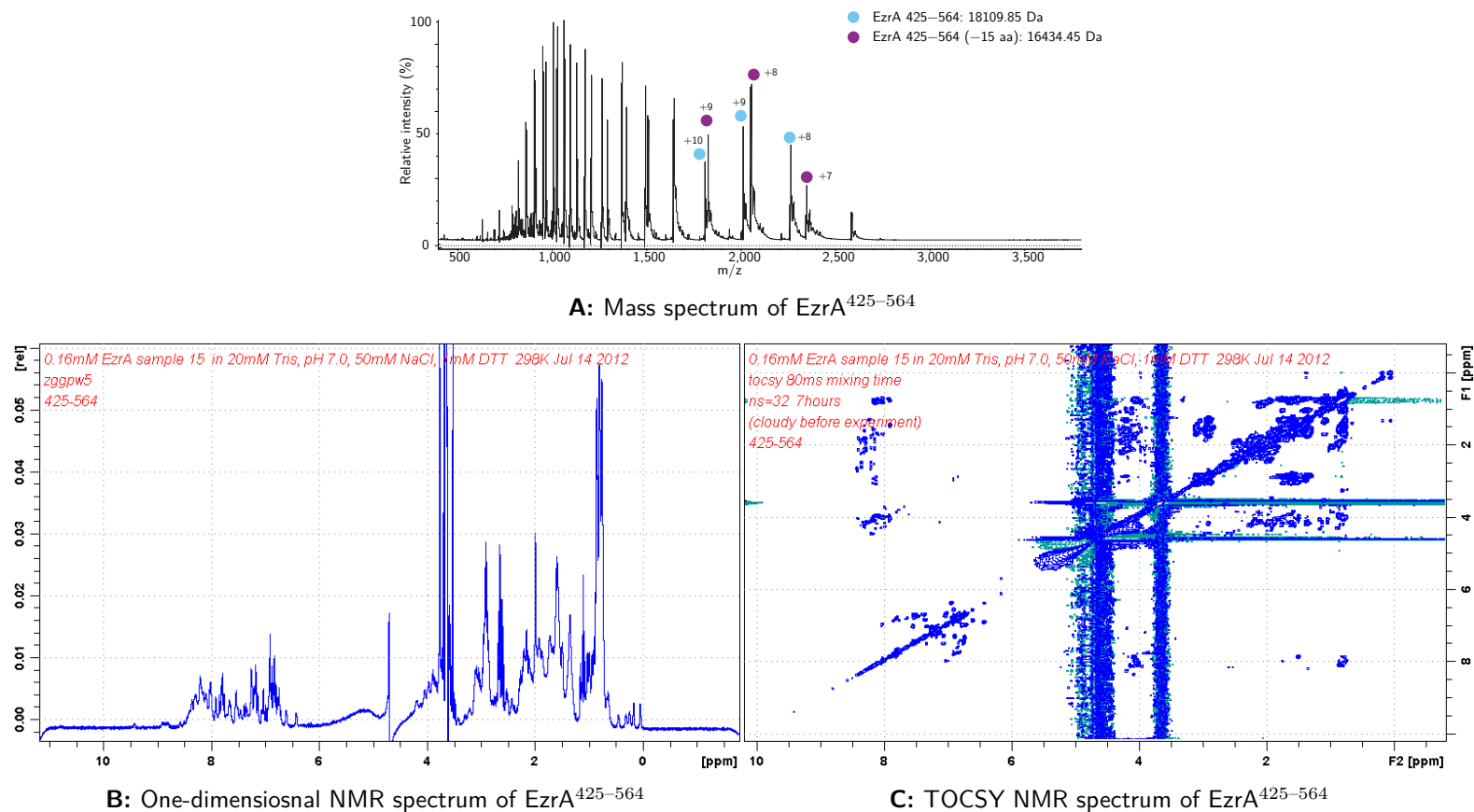


**D:** TOCSY NMR spectrum of EzrA<sup>302-564</sup>

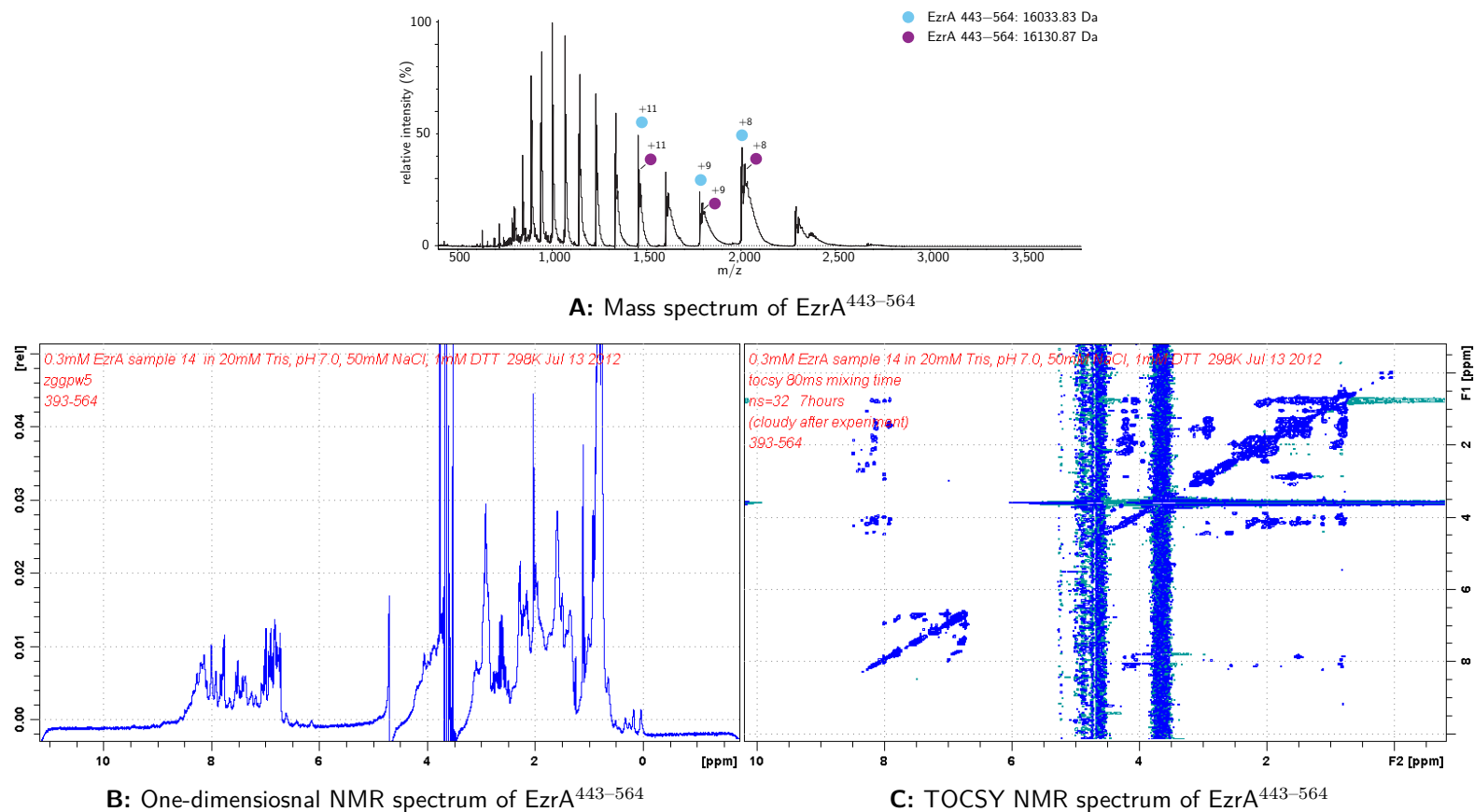
**Figure E.3: Mass and NMR spectral analysis of EzrA<sup>302-564</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>302-564</sup>. To see if purified EzrA<sup>302-564</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



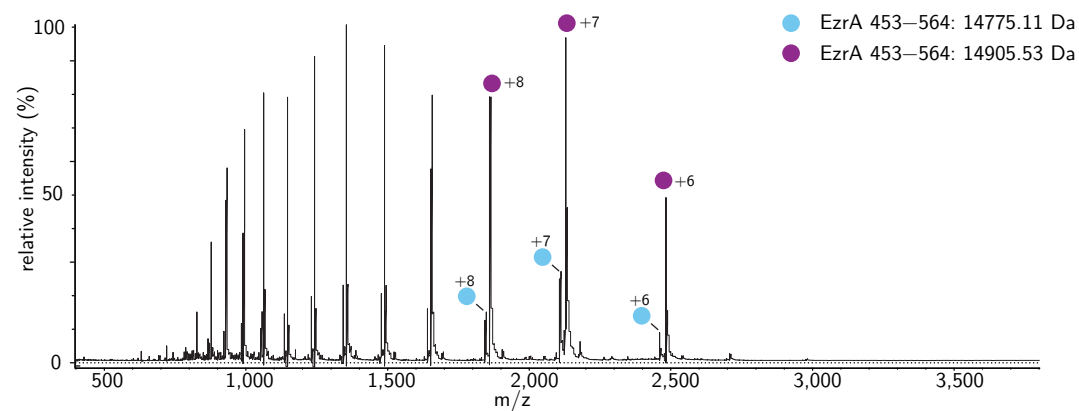
**Figure E.4: Mass and NMR spectral analysis of EzrA<sup>381-564</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>381-564</sup>. To see if purified EzrA<sup>381-564</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



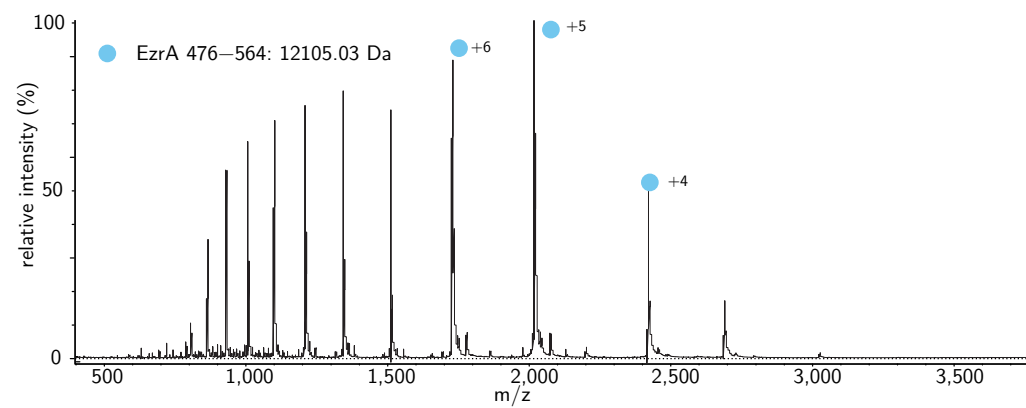
**Figure E.5: Mass and NMR spectral analysis of EzrA<sup>24-139</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>24-139</sup>. To see if purified EzrA<sup>24-139</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



**Figure E.6: Mass and NMR spectral analysis of EzrA<sup>443-564</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>443-564</sup>. To see if purified EzrA<sup>443-564</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.

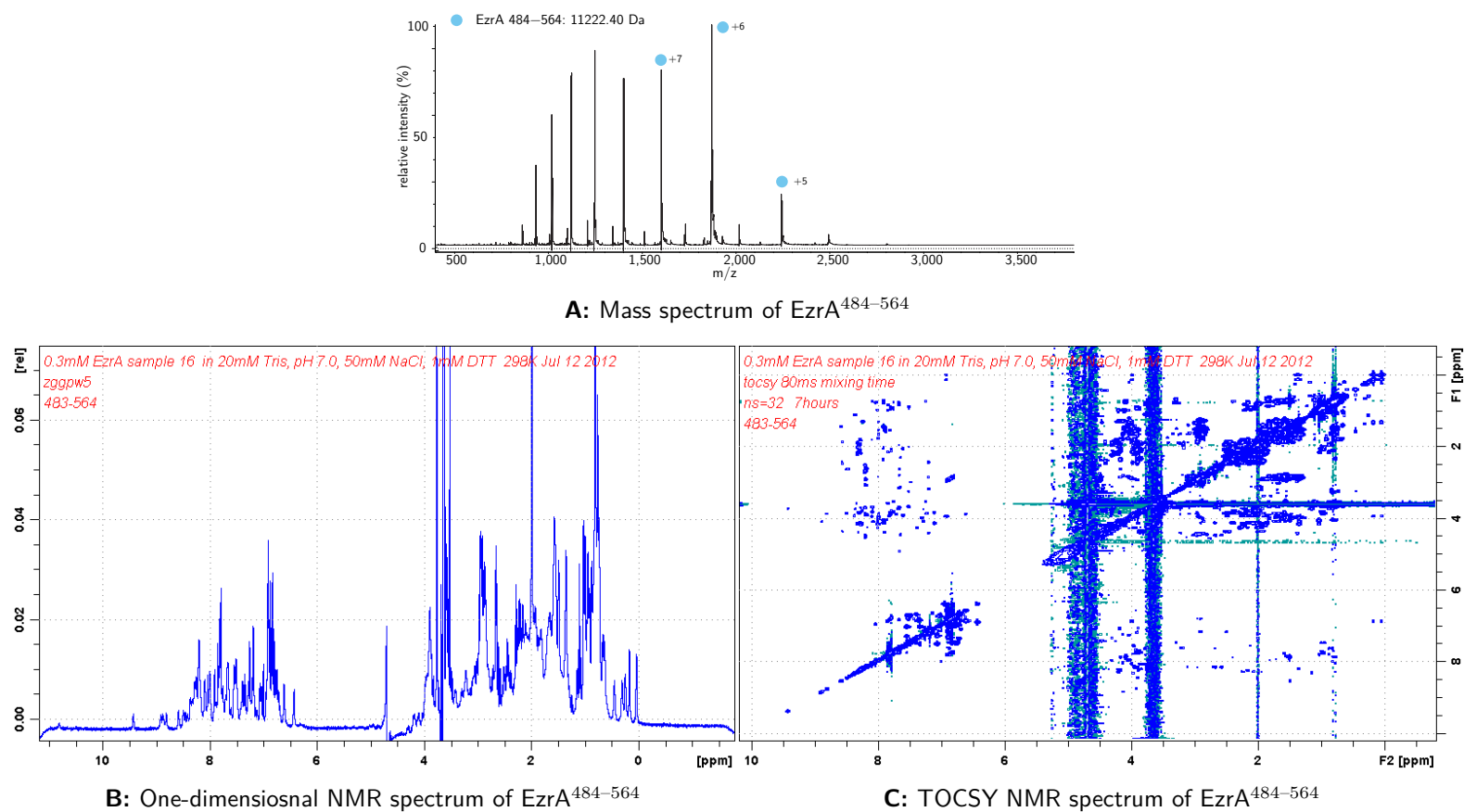


**Figure E.7: Mass analysis of EzrA<sup>453–564</sup>.** Positive ion electrospray mass spectrum of denatured EzrA<sup>453–564</sup>.

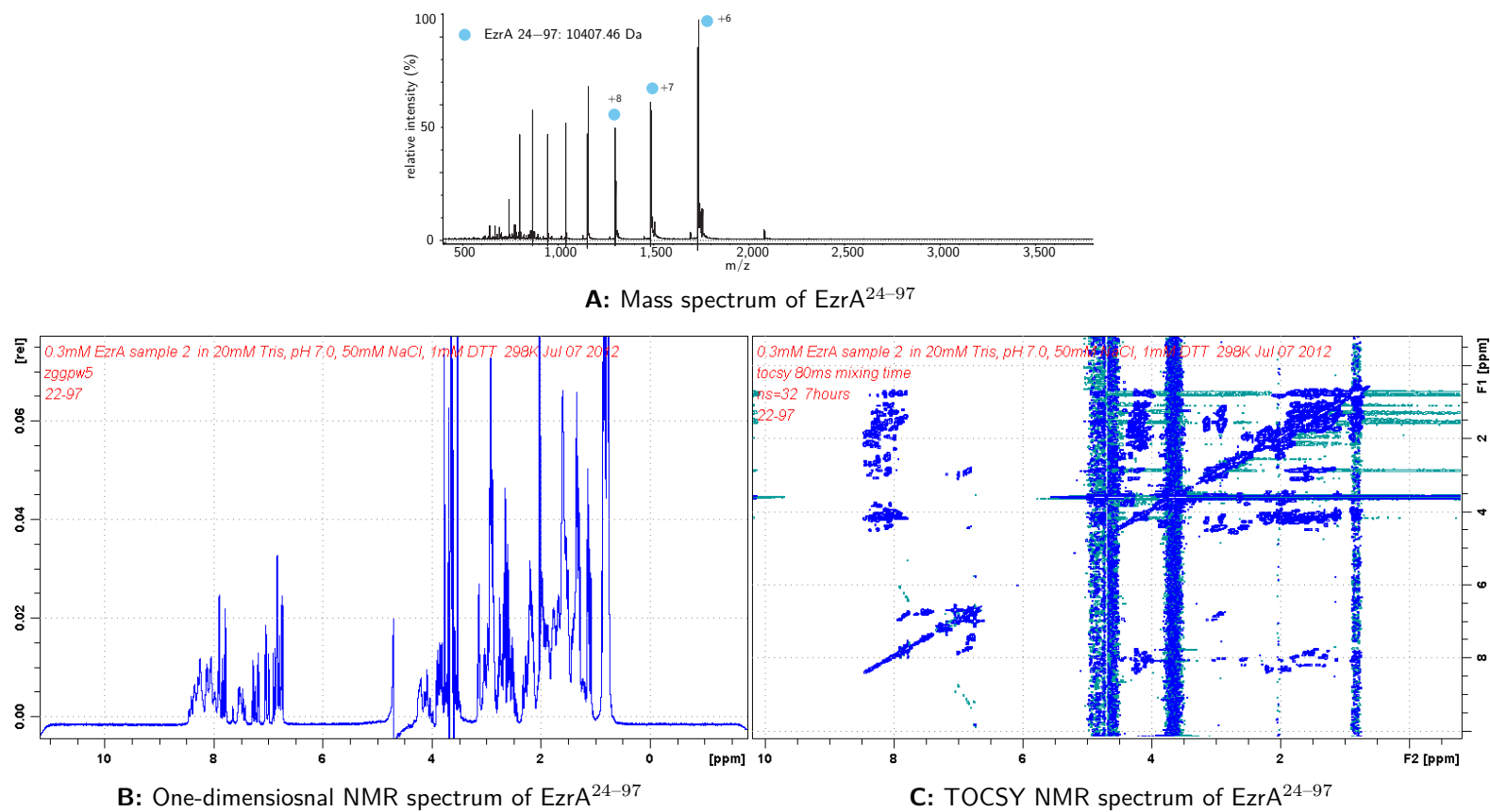


**Figure E.8: Mass analysis of EzrA<sup>476–564</sup>.** Positive ion electrospray mass spectrum of denatured EzrA<sup>476–564</sup>.

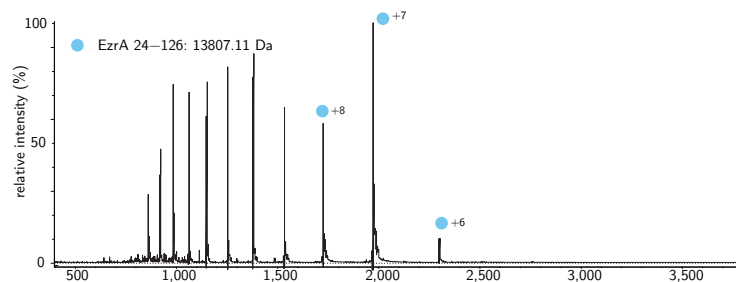




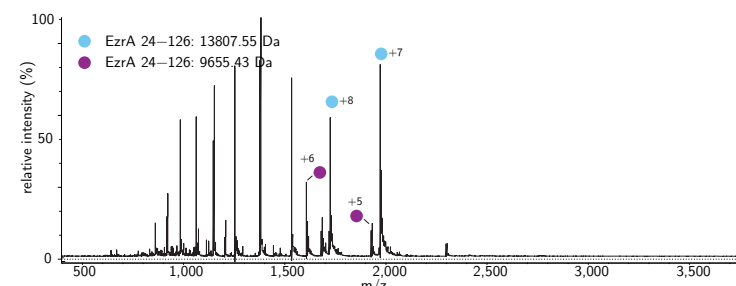
**Figure E.9: Mass and NMR spectral analysis of EzrA<sup>484-564</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>484-564</sup>. To see if purified EzrA<sup>484-564</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



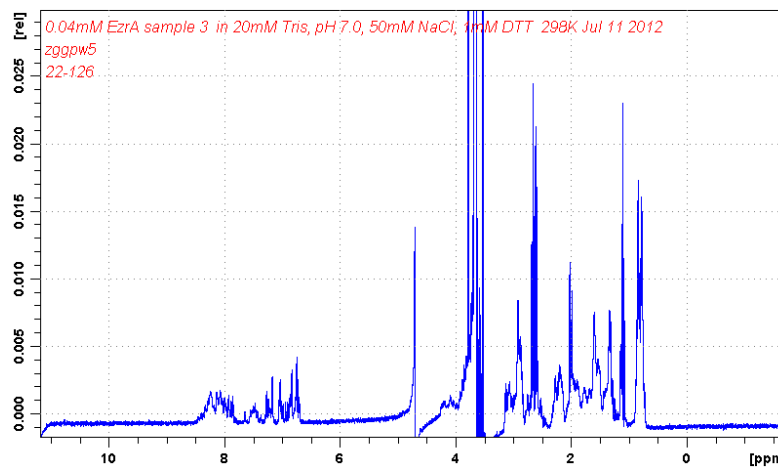
**Figure E.10: Mass and NMR spectral analysis of EzrA<sup>24-97</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>24-97</sup>. To see if purified EzrA<sup>24-97</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



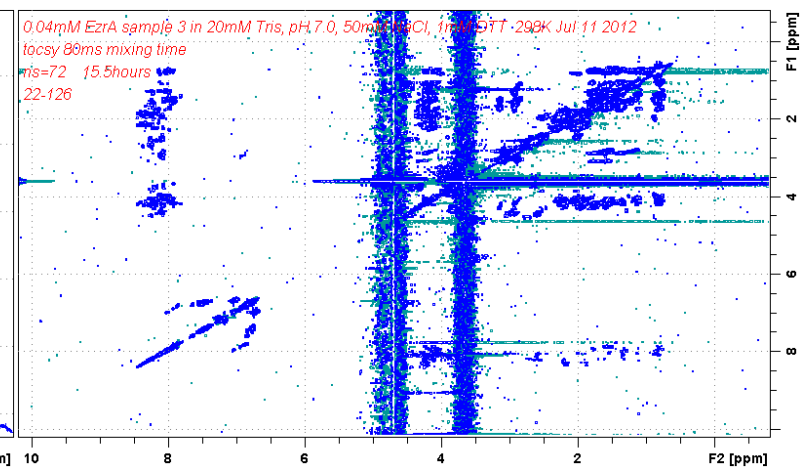
**A:** Mass spectrum of EzrA<sup>24-126</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-126</sup>: late anion-exchange chromatography peak

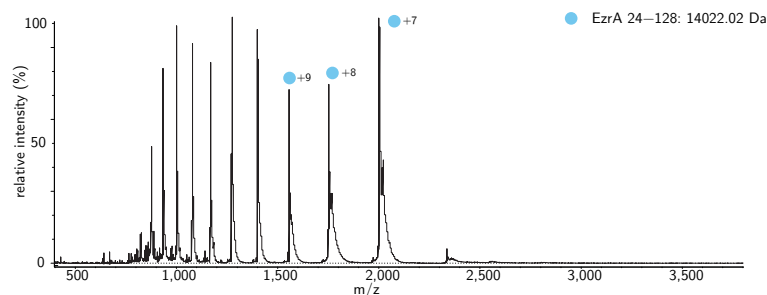


**C:** One-dimensionosnal NMR spectrum of EzrA<sup>24-126</sup>

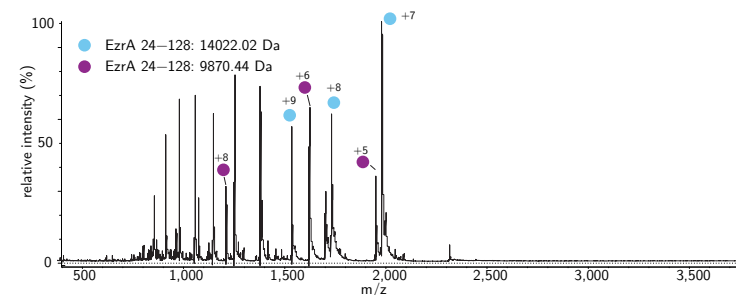


**D:** TOCSY NMR spectrum of EzrA<sup>24-126</sup>

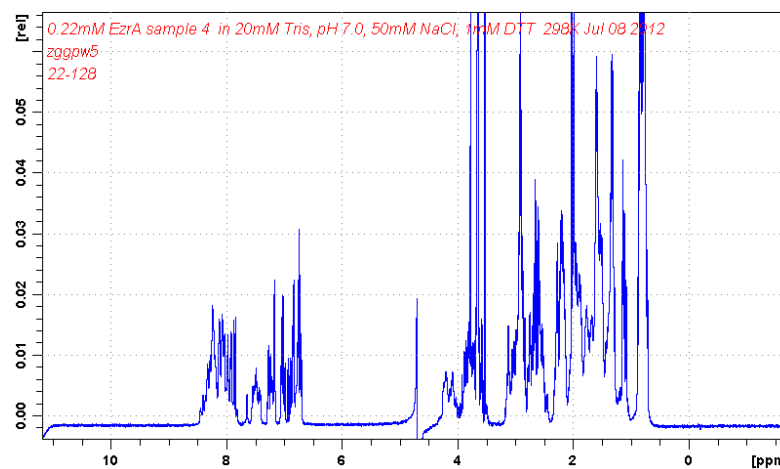
**Figure E.11: Mass and NMR spectral analysis of EzrA<sup>24-126</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-126</sup>. To see if purified EzrA<sup>24-126</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



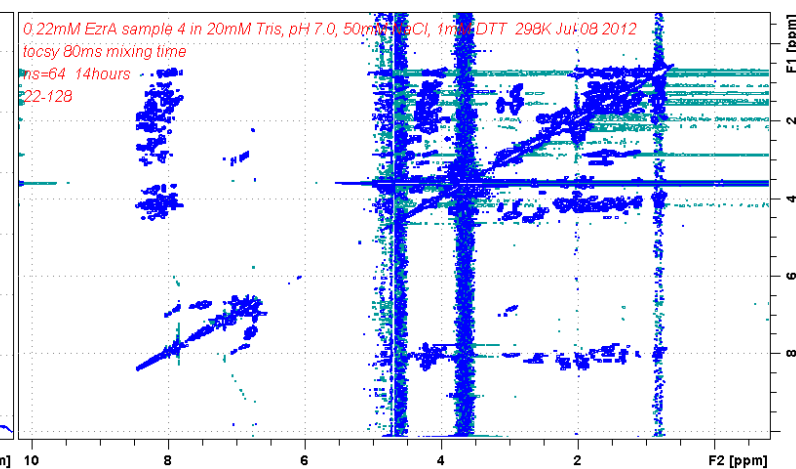
**A:** Mass spectrum of EzrA<sup>24-128</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-128</sup>: late anion-exchange chromatography peak

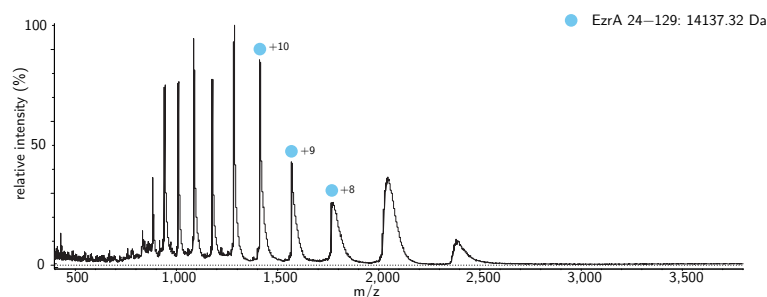


**C:** One-dimensional NMR spectrum of EzrA<sup>24-128</sup>

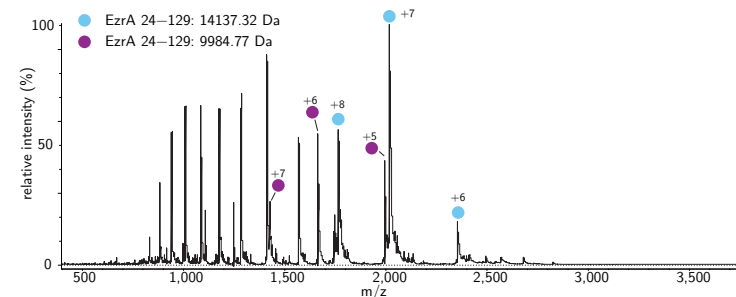


**D:** TOCSY NMR spectrum of EzrA<sup>24-128</sup>

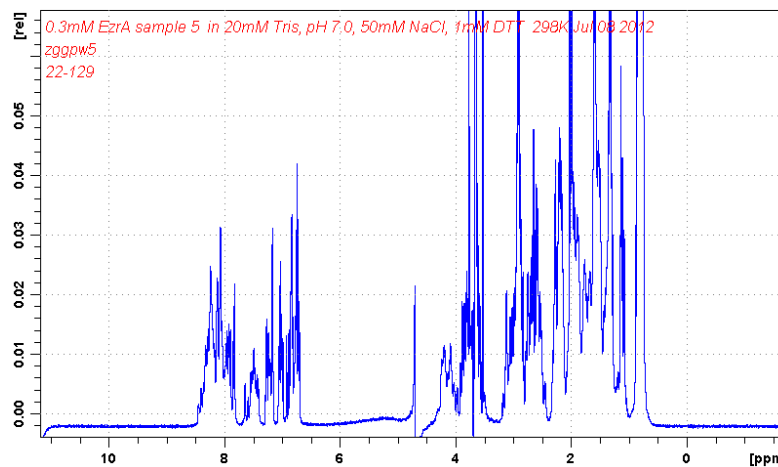
**Figure E.12: Mass and NMR spectral analysis of EzrA<sup>24-128</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-128</sup>. To see if purified EzrA<sup>24-128</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



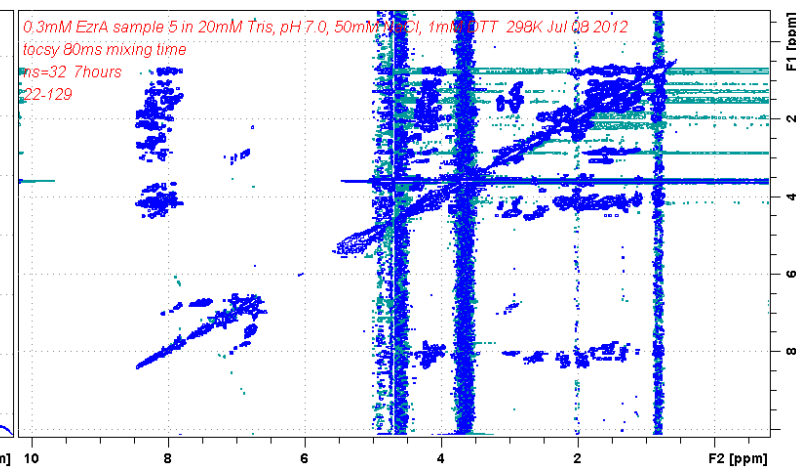
**A:** Mass spectrum of EzrA<sup>24-129</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-129</sup>: late anion-exchange chromatography peak

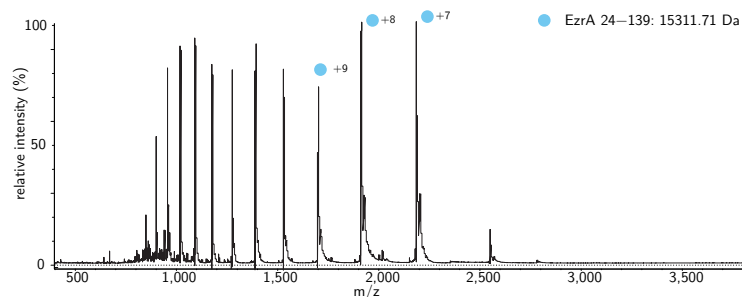


**C:** One-dimensionalsal NMR spectrum of EzrA<sup>24-129</sup>

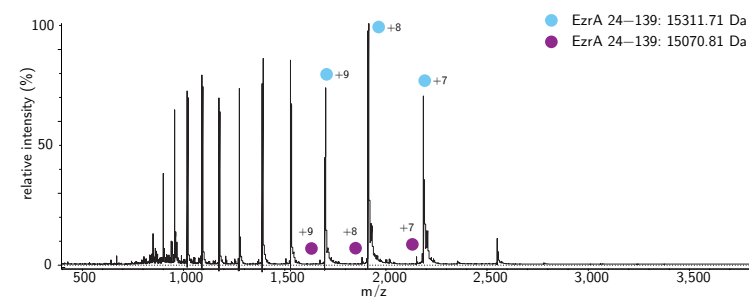


**D:** TOCSY NMR spectrum of EzrA<sup>24-129</sup>

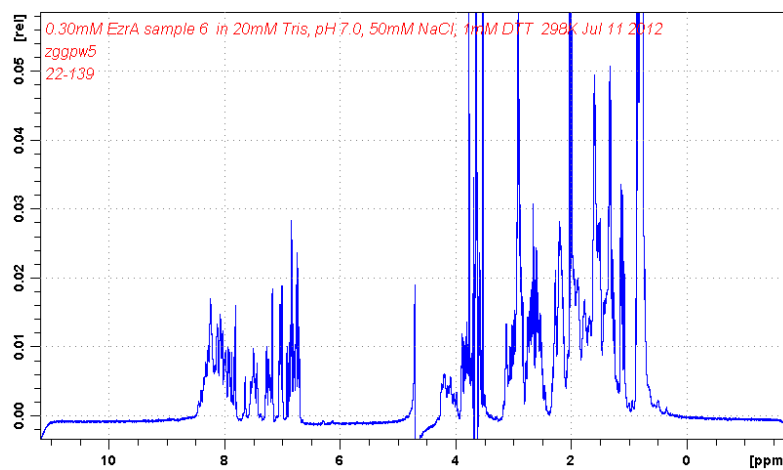
**Figure E.13: Mass and NMR spectral analysis of EzrA<sup>24-129</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-129</sup>. To see if purified EzrA<sup>24-129</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



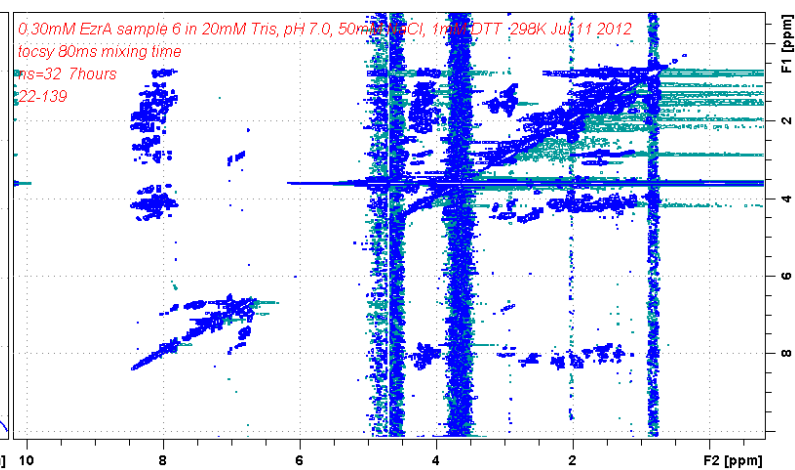
**A:** Mass spectrum of EzrA<sup>24-139</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-139</sup>: late anion-exchange chromatography peak

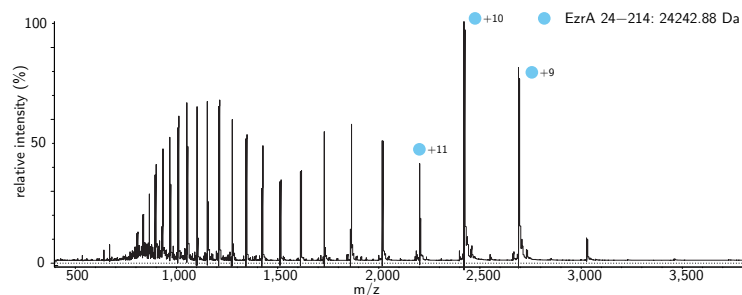


**C:** One-dimensionosnal NMR spectrum of EzrA<sup>24-139</sup>

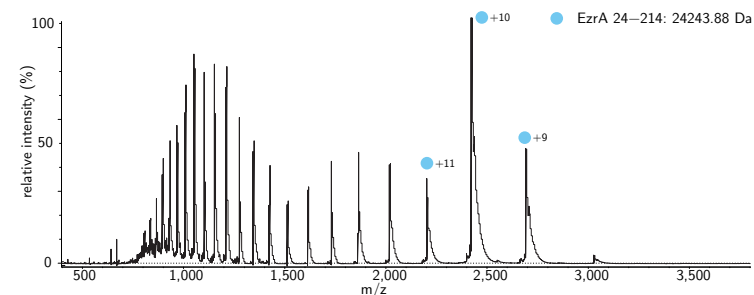


**D:** TOCSY NMR spectrum of EzrA<sup>24-139</sup>

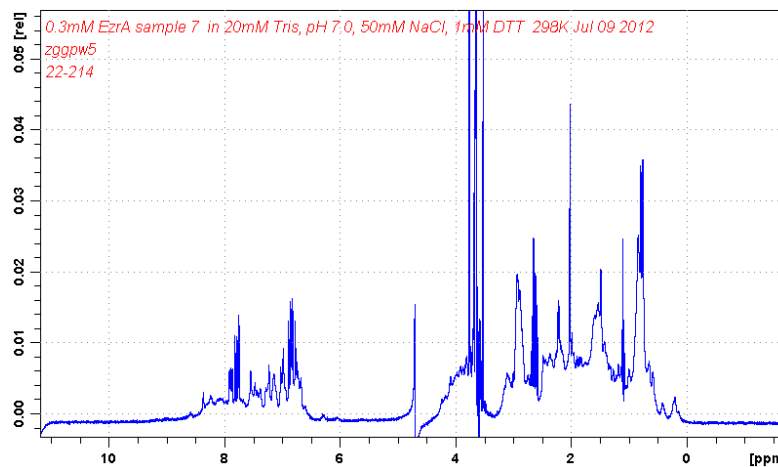
**Figure E.14: Mass and NMR spectral analysis of EzrA<sup>24-139</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-139</sup>. To see if purified EzrA<sup>24-139</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



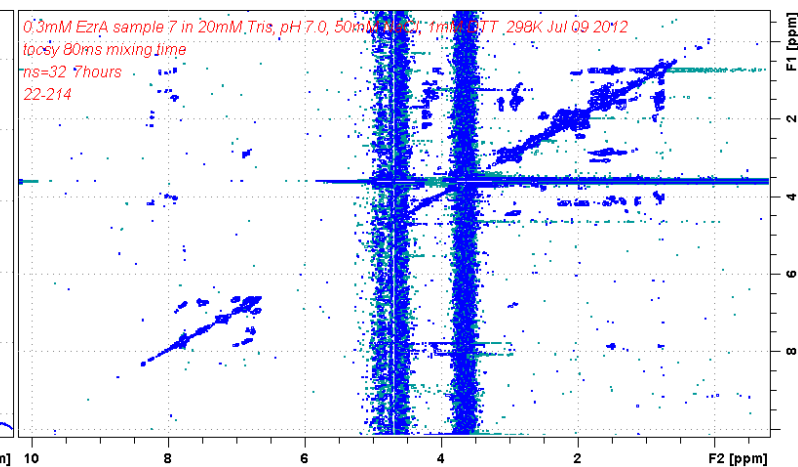
**A:** Mass spectrum of EzrA<sup>24-214</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-214</sup>: late anion-exchange chromatography peak

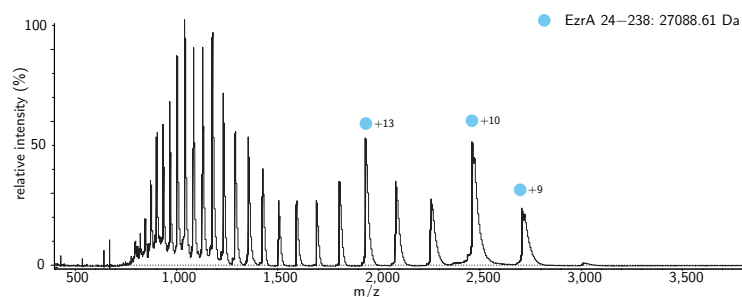


**C:** One-dimensionals NMR spectrum of EzrA<sup>24-214</sup>

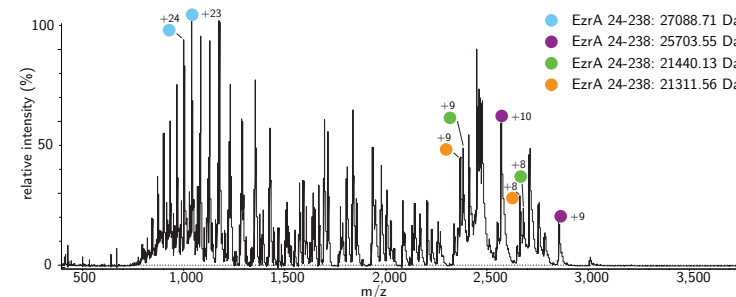


**D:** TOCSY NMR spectrum of EzrA<sup>24-214</sup>

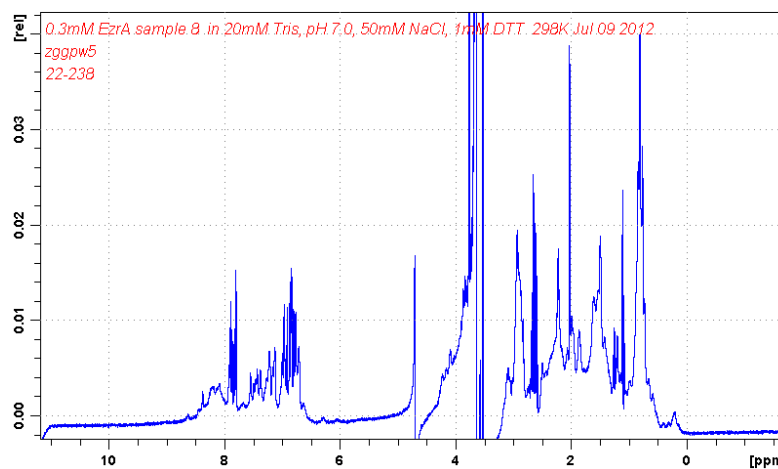
**Figure E.15: Mass and NMR spectral analysis of EzrA<sup>24-214</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-214</sup>. To see if purified EzrA<sup>24-214</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.



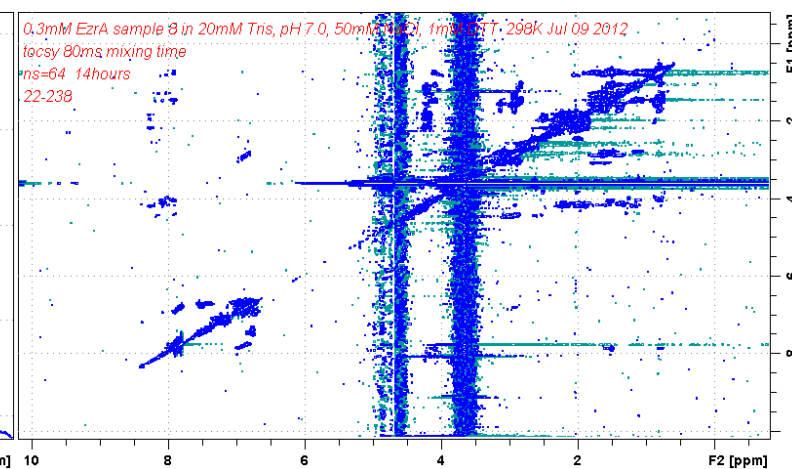
**A:** Mass spectrum of EzrA<sup>24-238</sup>: early anion-exchange chromatography peak



**B:** Mass spectrum of EzrA<sup>24-238</sup>: late anion-exchange chromatography peak



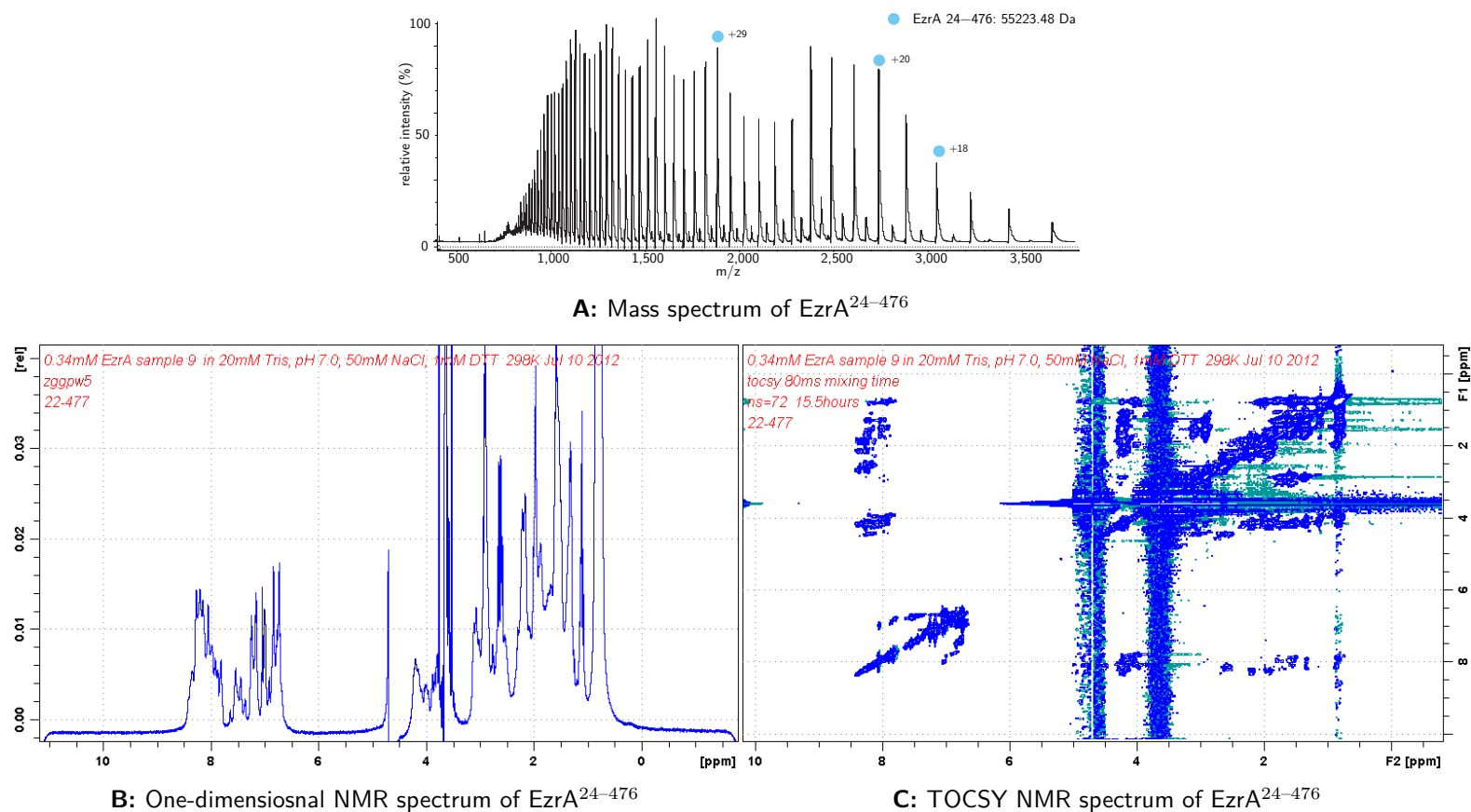
**C:** One-dimensionalsal NMR spectrum of EzrA<sup>24-238</sup>



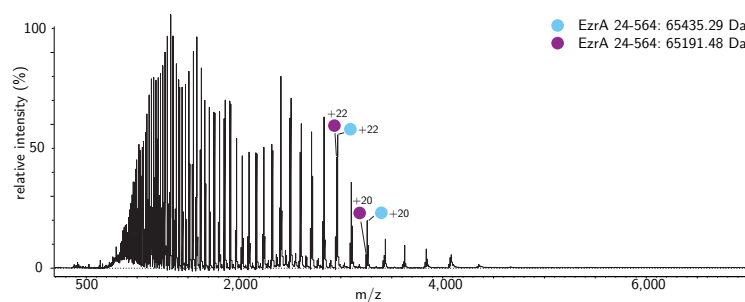
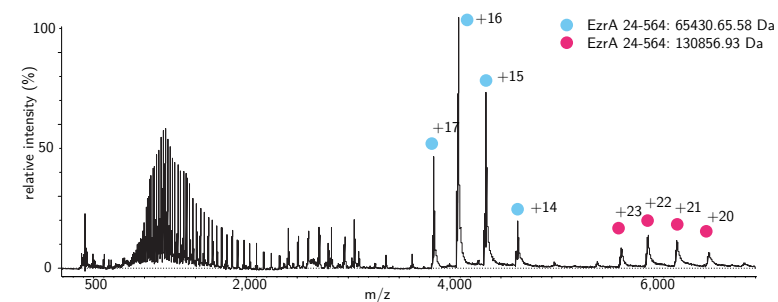
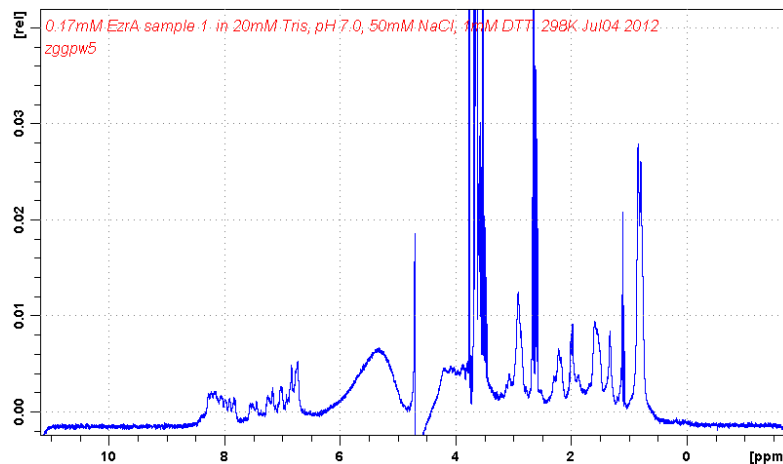
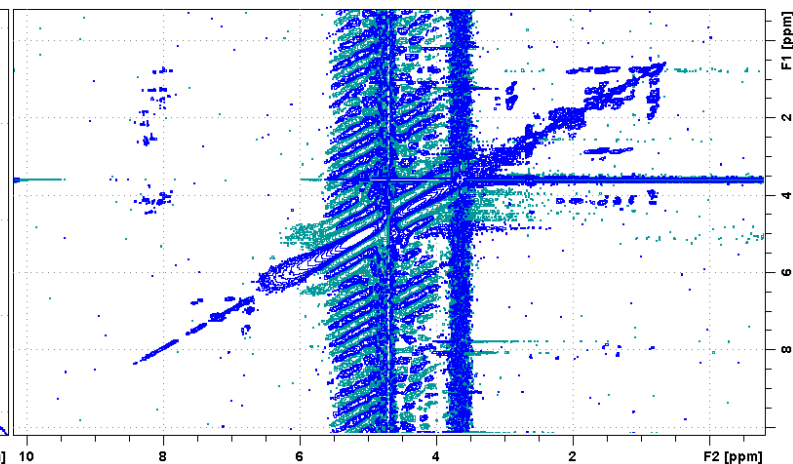
**D:** TOCSY NMR spectrum of EzrA<sup>24-238</sup>

**Figure E.16: Mass and NMR spectral analysis of EzrA<sup>24-238</sup>.** **A,B,** Positive ion electrospray mass spectrum of denatured EzrA<sup>24-238</sup>. To see if purified EzrA<sup>24-238</sup> was folded: **C,** 1D NMR and **D,** TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.





**Figure E.17: Mass and NMR spectral analysis of EzrA<sup>24-476</sup>.** **A**, Positive ion electrospray mass spectrum of denatured EzrA<sup>24-476</sup>. To see if purified EzrA<sup>24-476</sup> was folded: **B**, 1D NMR and **C**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.

A: Mass spectrum of EzrA<sup>24-564</sup>B: Mass spectrum of folded EzrA<sup>24-564</sup>C: One-dimensional NMR spectrum of EzrA<sup>24-564</sup>D: TOCSY NMR spectrum of EzrA<sup>24-564</sup>

**Figure E.18: Mass and NMR spectral analysis of EzrA<sup>24-564</sup>.** Positive ion electrospray mass spectrum of **A**, denatured and **B**, folded EzrA<sup>24-564</sup>. To see if purified EzrA<sup>24-564</sup> was folded: **C**, 1D NMR and **D**, TOCSY NMR spectra were recorded in 20 mM Tris-HCl pH 7.0, 50 mM NaCl and 1 mM DTT.