Searching for novel enzymes from microbial dark matter

Katy Alexandria Sophie Cornish

A thesis presented for the degree of Doctor of Philosophy

Department of Chemistry

Durham University

June 2022

Searching for novel enzymes from microbial dark matter

Katy Alexandria Sophie Cornish

The vast majority of inhabited environments on Earth are dominated by uncultivated microbes, described as 'microbial dark matter'. Metagenomic analyses of various ecological niches have unearthed an abundance of genomes from unculturable organisms that have evaded polymerase chain reaction (PCR)-based and cultivation-dependent isolation efforts. Many such organisms fall within a recently uncovered expanse of the bacterial domain, known as the Candidate Phyla Radiation (CPR), which is thought to harbour a third of all biodiversity within the tree of life. The genomes present within this mysterious microbial world present a major source of uncharted genetic diversity and are untapped sources of molecular tools for biological research, encoding swathes of novel proteins with innovation potential.

The Virus-X consortium endeavoured to probe the sequence diversity of extreme environments through metagenomics and identify commercially valuable enzymes. From the pool of over 50 million genes discovered during the Virus-X project, 18 from the CPR and five from a newly discovered giant phage dubbed the *Ubervirus* were selected as targets of particular interest. From 23 cloned targets, 18 proteins were successfully expressed in trials, leading to large-scale soluble expression of 12 and successful purification of 10 proteins. 9 of these targets were characterised through analytical size exclusion chromatography, mass spectrometry, and thermal shift analysis. Promising crystallisation conditions have been identified for four targets, leading to novel X-ray crystal structures for two CPR enzymes.

The structure of a DNA processing enzyme, CPR-DprA, is reported to a resolution of 2.10 Å (R/R_{free} 0.20/0.23), showing a core domain formed from an extended Rossmann fold. Also presented are three crystal structures of a hypothetical protein, CPR-C4, to a maximum resolution of 2.25 Å (R/R_{free} 0.18/0.23). Through remarkable structural homology to human vasohibin proteins, CPR-C4 was characterised as a cysteine protease utilising a noncanonical cysteine-histidine-leucine(carbonyl) catalytic triad, with protease activity confirmed using fluorescence-based assays. The structural and functional similarities between CPR-C4 and the human vasohibins point to an evolutionary relationship undetectable at the sequence level, which is addressed through phylogenetic analysis. The production and characterisation of a DnaK/ClpB bi-chaperone system from the CPR, with tangible biotechnology applications, is also reported, including preliminary cryo-electron microscopy analysis of the hexameric CPR-ClpB disaggregase.

Along with laying the groundwork for future commercialisation, the investigation of these CPR proteins takes strides towards addressing the substantial gaps in our knowledge of this little understood, yet pervasive, branch of the tree of life.

Table of contents

Declaration and Statement of Copyright....9
Acknowledgements....10
Dedication....11
Abbreviations....12

Chapter 1. The search for novel enzymes from uncultivated microbes in extreme environments ...15
1.1. A growing need for bioprospecting 15
1.2. The Virus-X Project....15
1.2.1. Exploring sequence space through metagenomics....15
1.2.2. The Virus-X workflow....16
1.3. Metagenomics....18
1.3.1. The introduction of cultivation-independent methods for genome sequencing....18
1.3.2. Metagenomics next generation sequencing 18
1.3.3. Remaining challenges for metagenomics....20
1.3.4. Metagenome retrieval during the Virus-X project....20
1.4. The Candidate Phyla Radiation....22
1.4.1. Microbial dark matter....22
1.4.2. Unusual features of CPR bacteria 24
1.4.3. Predicted symbiotic lifestyle of CPR bacteria 24
1.4.4. Cultivation of 'unculturable microbes' 25
1.4.5. The CPR as a source of biodiversity....25
1.5. Extremophiles 26
1.5.1. Organisms from extreme environments....26
1.5.2. Classification of extremophiles....26
1.5.3. Surviving extreme pH 27
1.5.4. Surviving extremes of temperature 27
1.5.5. Extremozymes....28
1.5.6. Uncultivated extremophiles as sources of novel enzymes....29
1.6. Structure determination of novel proteins....29
1.6.1. Functional insight from structure determination....29
1.6.2. Protein structure determination with X-ray crystallography....30
1.6.3. Structure determination within the Virus-X project....31
1.7. Project aims....32

Chapter 2. Selection of targets from the Candidate Phyla Radiation 33
2.1. Introduction 33
2.1.1 Virus-X target selection and categorisation 33
2.1.2. Targets from the CPR 34
2.2. Results and discussion 34
2.2.1. CPR target identification and contig analysis 34
2.2.2. Categorisation of CPR targets 37
2.2.3. Functional annotation of CPR targets 38
2.3. Conclusions 39

Chapter 3. Cloning and production trials of the Candidate Phyla Radiation targets 40
3.1. Introduction 40
3.2. Results and discussion 41
3.2.1. Cloning and transformation of CPR expression constructs 41
3.2.2. Production trials of CPR targets 42
3.2.3. Selection of targets for large-scale expression 47
3.3. Conclusions 48

Chapter 4. Production and characterisation of Candidate Phyla Radiation targets 49
4.1. Introduction 49
4.1.1. Tags and fusion proteins 49
4.1.2. Purification methods 49
4.1.3. Protein characterisation 50
4.1.4. Thermal shift analysis 51
4.1.5. General strategy for purification and characterisation of Virus-X targets 52
4.2. Results and discussion 53
4.2.1. Preliminary characterisation of CPR targets 53
4.2.2. CPR-hel-1 54
4.2.3. CPR-hel-2 54
4.2.4. CPR-exo-1 57
4.2.5. CPR-endo-1 and CPR-endo-2 63
4.2.6. CPR-B1 and CPR-B2 64
4.2.7. CPR-B3 66
4.2.8. Overall results of CPR target production and characterisation 69
4.3. Conclusions 70

Chapter 5. Structure determination and analysis of a DprA protein from the Candidate Phyla Radiation 72
5.1. Introduction 72
5.1.1. Nucleic acid processing enzymes 72
5.1.2. DNA processing protein, DprA 72
5.2. Results and discussion 73
5.2.1. Sequence analysis and annotation of CPR-DprA 73
5.2.2. Domain architecture of CPR-DprA 73
5.2.3. Production and characterisation of CPR-DprA 76
5.2.4. Structure determination of CPR-DprA 80
5.2.5. Crystal structure of CPR-DprA 83
5.2.6. Prediction of C-terminal domain structure using AlphaFold2 84
5.2.7. Dimer analysis 85
5.3. Conclusions 88

Chapter 6. Determination of the structure and function of hypothetical protein CPR-C4 from the Candidate Phyla Radiation 89
6.1. Introduction 89
6.1.1. Hypothetical proteins 89
6.1.2. 'Unknown' proteins as key Virus-X targets 89
6.2. Results and discussion 90
6.2.1. CPR-C4 sequence analysis with BLAST 90
6.2.2. CPR-C4 sequence analysis with 3DM 91
6.2.3. Production and characterisation of CPR-C4 92
6.2.4. Production of CPR-C4 using an alternative expression construct 99
6.2.5. Structure determination of CPR-C4 by X-ray crystallography 104
6.2.6. Crystal structures of CPR-C4 112
6.2.7. Structural analysis of the CPR-C4 Zn^{2+} binding sites 114
6.2.8. Structure analysis of CPR-C4 with 3DM 118
6.2.9. Activity analysis of CPR-C4 125
6.2.10. Phylogenetic analysis of CPR-C4 130
6.3. Conclusions 132

Chapter 7. Production and characterisation of a DnaK/ClpB bi-chaperone system from the Candidate Phyla Radiation 134
7.1. Introduction 134
7.1.1. Protein folding and molecular chaperones 134
7.1.2. DnaK/ClpB bi-chaperone system 135
7.1.3. Applications of chaperones in biotechnology 135

7.2. Results and discussion 136
7.2.1. Target identification and BLAST analysis 136
7.2.2. Production and characterisation of CPR-GrpE 137
7.2.3. Production and characterisation of CPR-DnaJ 145
7.2.4. Production and characterisation of CPR-DnaK 148
7.2.5. Production and characterisation of CPR-ClpB 155
7.3. Conclusions 160

Chapter 8. Single particle cryo-electron microscopy studies of the CPR-ClpB disaggregase 162
8.1. Introduction 162
8.1.1. Single particle cryo-electron microscopy 162
8.1.2. Negative staining 162
8.1.3. Cryo-EM studies of ClpB homologs 163
8.2. Results and discussion 164
8.2.1. Negative staining of CPR-ClpB samples 164
8.2.2. Single particle cryo-EM 165
8.2.3. Optimisation of CPR-ClpB samples 168
8.3. Conclusions 174

Chapter 9. Production and characterisation of targets from a thermophilic giant bacteriophage 176
9.1. Introduction 176
9.1.1. Exploring viral diversity through metagenomics 176
9.1.2. Discovery and implications of giant viruses 176
9.1.3. Jumbophages 177
9.1.4. The '*Ubervirus*' 177
9.2. Results and discussion 178
9.2.1 Sequence analysis of GX targets 178
9.2.2. Cloning and preliminary characterisation 179
9.2.3. Expression trials of GX targets 179
9.2.4. GX-hyp-4 181
9.2.5. GX-hyp-6 182
9.2.6. GX-hyp-16 187
9.3. Conclusions 193

Chapter 10. Conclusions 194
10.1. Contribution of CPR targets to the Virus-X project 194
10.2. Protein production 196
10.2.1. CPR targets 196

10.2.2. GX targets 197
10.3. CPR-DprA 197
10.4. CPR-C4 198
10.5. CPR-DnaK/ClpB bi-chaperone system 198
10.6. Thermophilic potential of CPR and GX targets 199
10.7. The impact of advanced structural prediction tools 199
10.8. Shining a light on microbial dark matter 200

Chapter 11. Experimental methods 201
11.1. General experimental information 201
11.2. Virus-X sampling, target selection, and annotation 201
11.3. GX and CPR cell stock production 201
11.3.1. General protocol 201
11.3.2. CPR-C4 202
11.4. Plasmid recovery 202
11.5. Plasmid transformations 202
11.5.1. General protocol 202
11.5.2. Transformation of CPR-B3 into SHuffle® cells 202
11.5.3. Transformation of pET28a(+)TEV-*CPRC4* 203
11.6. Agarose gel electrophoresis 203
11.7. DNA Sequencing 203
11.8. Small-scale protein expression trials 203
11.8.1. General protocol 203
11.8.2. Adjusted expression trials with CPR-B3 203
11.8.3. Expression trials of CPR-C4 with $ZnCl_2$ 204
11.8.4. Expression trials of CPR-C4 from pET28a(+)TEV-*CPRC4* 204
11.9. Full-scale protein expression 204
11.9.1. General protocol 204
11.9.2. Adjusted expression of CPR-hel-2 204
11.9.3. CPR-C4 expression with $ZnCl_2$ 204
11.9.4. CPR-C4 expression from pET28a(+)TEV-*CPRC4* 204
11.10. Immobilised metal ion affinity chromatography 205
11.10.1. General protocol 205
11.10.2. Optimised purification of CPR-C4 205
11.10.3. Purification of CPR-C4 from pET28a(+)TEV-*CPRC4* 205
11.10.4. Optimised purification of GX-hyp-16 206
11.11. Size exclusion chromatography 206
11.11.1. General protocol 206

11.11.2. Column Calibration 206
11.12. SDS-PAGE 207
11.12.1. General protocol 207
11.12.2. Sample preparation 208
11.13. Mass Spectrometry 208
11.13.1. ESI-TOF MS 208
11.13.2. Trypsin digest MS 208
11.14. Thermal shift analysis 208
11.15. Circular Dichroism 208
11.16. Protein Crystallisation 209
11.16.1. Crystallisation screening 209
11.16.2. Manual optimisation 209
11.16.3. Microseed matrix screening 209
11.16.4. CPR-DprA crystallisation 210
11.16.5. CPR-C4 crystallisation 210
11.17. Data collection, processing, and structural determination 210
11.17.1. Cryo-protection and harvesting of protein crystals 210
11.17.2. Structure determination of CPR-DprA 210
11.17.3. Structure determination of CPR-C4 211
11.18. CPR-C4 structural analysis with 3DM 211
11.19. CPR-C4 protease activity assays 212
11.20. CPR-C4 phylogenetics analysis 212
11.21. Electron microscopy studies of CPR-ClpB 212
11.21.1. Negative staining TEM 212
11.21.2. Glutaraldehyde crosslinking 212

Appendices 214
Appendix A: TSA Durham Screens® 214
Appendix B: CPR-exo-1 217
Appendix C: CPR-DprA 218
Appendix D: CPR-C4 221
Appendix E: CPR-GrpE 228
Appendix F: GX targets 230

Bibliography 232

Declaration and Statement of Copyright

The work in this thesis was carried out in the Department of Chemistry, Durham University, between October 2018 and June 2022. All work is the author's own unless otherwise stated. This work has not previously been submitted for a degree at this or any other institution.

Peer-reviewed publications from this thesis include the following, at the time of writing:

Cornish, K. A. S., Lange, J., Aevarsson, A. & Pohl, E. CPR-C4 is a highly conserved novel protease from the Candidate Phyla Radiation with remote structural homology to human vasohibins. *Journal of Biological Chemistry* **298**, 101919 (2022).

Aevarsson, A., Cornish. K. A. S. *et al.* Going to extremes – a metagenomic journey into the dark matter of life. *FEMS Microbiology Letters* **368**, (2021).

Acknowledgements

I have many people to thank, without whom this work would not have been possible.

Firstly, my supervisor, Ehmke, without whom I absolutely would not be where I am today. Thank you for your continued support and belief in me over the past eight years, particularly when I have not believed in myself. Thanks also to Stef for being a constant source of inspiration, and a wonderful travel companion and confidant.

Huge thanks must go to Ian for keeping everything up-and-running, and for always making me smile. Thanks to the occupants of Lab 229 and Office 231 for the laughs and entertaining tea-time discussions. Special thanks to past and present members of the Pohl Patrol (Becky, Charlie, Dan, Stef, Emma, Kate, Naomi, Dori, Abbey, Izzy, and Davide) for your encouragement and for making the lab such an enjoyable place to work even when times were tough. You will all be greatly missed!

Thanks to the Mass Spectrometry, Proteomics, and Sequencing teams for their assistance. Enormous thanks also to Arnaud for the use of his crystallisation robot whilst ours was out of action, and to Dan at the Astbury Centre for introducing me to the dauting world of cryo-EM.

Thanks to all members of the Virus-X consortium, with particular thanks to Arnþór and Joanna. I am very grateful to have been a part of this exciting project with such welcoming people.

Finally, thanks to my family for their love and constant encouragement, and for making it possible for me to pursue this PhD. A second round of thanks also to Charlie for supporting me in and out of work – I could not have done this without you.

Dedication

This thesis is dedicated to my grandfather, Derek, the original Dr Cornish.

Abbreviations

16S rRNA	Small subunit ribosomal RNA
3DM	Protein superfamily platform software
3D-MSA	Structure-based Multiple Sequence Alignment
ADP	Adenosine 5'-diphosphate
ÄKTA	FPLC system (Cytiva)
AMP-PNP	Adenylyl-imidodiphosphate
ASU	Asymmetric Unit
ATP	Adenosine 5'-triphosphate
ATPγS	Adenosine-5′-o-(3-thio-triphosphate)
BLAST	Basic Local Alignment Search Tool
BODIPY-FL	Green-fluorescent dye
bp	Base Pair (DNA)
BSA	Bovine Serum Albumin
Cas	CRISPR-associated protein
$CC_{1/2}$	Correlation coefficient 1/2 dataset
CCP4	Collaborative Computational Project 4
CD	Circular Dichroism
Coot	Crystallography Object-Oriented Toolkit
CPR	Candidate Phyla Radiation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cryo-EM	Cryo-Electron Microscopy
CTD	C-Terminal Domain
CTF	Contrast Transfer Function
CV	Column Volume
C_α	Alpha Carbon
DIALS	Diffraction Integration for Advanced Light Sources (software)
DLS	Diamond Light Source
DML1	Predicted Z-DNA binding domain of DprA proteins
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DS	Dropping Solution
DSC	Differential Scanning Calorimetry
DSF	Differential Scanning Fluorimetry
DUI	DIALS User Interface
Ec	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylenebis(oxyethylenenitrilo)tetraacetic acid
EM	Electron Microscopy
EMDB	Electron Microscopy Data Bank
EMGB	Virus-X Sequence Database
ESI	Electrospray Ionisation
EU	European Union
ex/em	Excitation and emission wavelengths
FP	Functional Priority
FPLC	Fast Protein Liquid Chromatography
GST	Glutathione-S-Transferase
GUI	Graphical User Interface
GX	Genome X

HEPES	(4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)
HMW	High Molecular Weight
Hp	*Helicobacter pylori*
HP	High Purity (HisTrap)
HPLC	High Performance Liquid Chromatography
HT	High Throughput
IEC	Ion Exchange Chromatography
IMAC	Immobilised Metal ion Affinity Chromatography
IPTG	Isopropyl ß-d-1-thiogalactopyranoside
JA/JLA	J and J-lightweight; fixed angle rotors for Beckmann centrifuges
JCSG	Joint Consortium for Structural Genomics (crystallisation condition screen)
JTT	Jones-Thornton-Taylor model (phylogenetics analysis)
K_{av}	Average distribution coefficient (SEC)
LB	Lysogeny Broth
LLG	Log-Likelihood Gain
LMW	Low Molecular Weight
MAD	Multiple wavelength Anomalous Dispersion
MALDI	Matrix-Assisted Laser Desorption/Ionisation
MBP	Maltose Binding Protein
MEGAX	Molecular Evolutionary Genetics Analysis software
MES	2-(N-morpholino)ethanesulfonic acid
MMS	Microseed Matrix Screening
MMT	DL-Malic acid, MES monohydrate, Tris
mNGS	Metagenomics Next Generation Sequencing
MOPS	3-(N-morpholino)propanesulfonic acid
MR	Molecular Replacement
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MSA	Multiple Sequence Alignment
MW	Molecular Weight
MWCO	Molecular Weight Cut-off
NAMI	TSA data analysis program
NBS	Non-Binding Surface, treatment for 96-well plates
NCBI	National Centre for Biotechnology Information
NCC	Normalised Cross-Correlation
NCS	Non-Crystallographic Symmetry
NEB	New England Biolabs (supplier)
NGS	Next Generation Sequencing
NME	N-terminal Methionine Excision
NMR	Nuclear Magnetic Resonance spectroscopy
NPS	Nitrate Phosphate Sulfate additive mixture (Morpheus Screen)
OP	Overall Priority (Virus-X targets)
OD_{600}	Optical Density at 600 nm
OP11	Obsidian Pool candidate phylum 11
PACT	pH, Anion and Cation (crystallisation condition screen)
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	Polyethylene Glycol
pET	Series of expression plasmids for use in *E. coli*
Pfam	Protein Family Database
pg	Preparation Grade (SEC columns)
pI	Isoelectric Point

PIPES	Piperazine-N,N′-bis(2-ethanesulfonic acid)
pJOE	Series of expression plasmids for use in *E. coli*
pLDDT	Predicted Local Distance Difference Test
PMTP	Sodium propionate, MES monohydrate, Bis-Tris propane
PTFE	Polytetrafluoroethylene
pUC	Series of plasmid cloning vectors for use in *E. coli*
RFU	Relative Fluorescence Units
RMSD	Root Mean Squared Deviation
RNA	Ribonucleic acid
rNCS	Rotational Non-Crystallographic Symmetry
rp	Ribosomal Protein
Rp	*Rhodopseudomonas palustris*
rpm	Revolutions per minute
rRNA	Ribosomal RNA
RT	Room Temperature
SAD	Single-wavelength Anomalous Dispersion
SAM	Sterile Alpha Motif (protein domain)
SAXS	Small-Angle X-ray Scattering
SDS	Sodium Dodecyl Sulphate
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SEC	Size Exclusion Chromatography
Seq. ID	Sequence Identity
SG	Structural Genomics
SGC	Structural Genomics Consortium
SOC	Super Optimal broth with Catabolite repression
SP	Structural Priority
Sp	*Streptococcus pneumoniae*
SSB	Single-Stranded DNA Binding Protein
SUMO	Small, Ubiquitin like Modifier (protein)
SVBP	Small Vasohibin-Binding Protein
TCEP	Tris(2-carboxyethyl)phosphine (reducing agent)
TEM	Transmission Electron Microscopy
TEV	Tobacco Etch Virus protease
T_m	Melt Temperature
TM7	Torf, Mittlere Schicht (candidate phylum)
TMAO	Trimethylamine N-oxide
TOF	Time-of-Flight (mass spectrometry)
tRNA	Transfer RNA
TSA	Thermal Shift Analysis
Tt	*Thermus thermophilus*
UV	Ultraviolet radiation
v/v	Volume per volume
VASH	Vasohibin/Tubulin-tyrosine carboxypeptidase
V_c	Geometric column volume
V_e	Elution volume
V_o	Void volume
w/v	Weight per volume
WH	Winged Helix (protein domain)
XDS	X-ray Detector Software (crystallographic data processing software)
ε	Extinction coefficient at 280 nm / M^{-1} cm^{-1}
θ	Ellipticity / mdeg (circular dichroism)
λ	Wavelength / nm

Chapter 1. The search for novel enzymes from uncultivated microbes in extreme environments

1.1. A growing need for bioprospecting

Bioprospecting is defined as 'the exploration of biodiversity for commercially valuable genetic resources and biochemicals'[1]. The natural world is an excellent resource of novel biotechnological and biomedical products, including enzymes with broad applications throughout the bioeconomy. In particular, enzymes sourced from extreme environments can serve as efficient biocatalysts under the harsh conditions often required for industrial processes, and with reduced environmental impact in comparison to conventional chemical methods[2–5]. As global focus continues to shift towards a more sustainable, bio-focussed economy with greener industrial processes, the requirement for such enzymes is continually increasing[2,6,7].

The vast majority of biological sequence diversity is currently unexplored[8]. However, with advances in genomics and sequencing technologies, natural biodiversity is increasingly accessible, including in the most extreme environments on the planet. Novel enzymes can be identified and thoroughly characterised without the need for cultivation of their source organisms[9,10]. In particular, the genomes of viruses and uncultivated microorganisms encode high proportions of uncharacterised proteins and are untapped sources of molecular tools for biological research[11–13]. Sequencing of these novel genomes provides an extensive and diverse collection of genes encoding proteins anticipated to have valuable and unique properties, with wide reaching applications across biosciences sectors[6,10,14,15]. There is therefore an increasing drive to explore natural environments using combined computational and structure-based approaches to identify novel extremozymes with innovation potential, and improve understanding of biological systems[2,8,14,16,17].

1.2. The Virus-X Project

1.2.1. Exploring sequence space through metagenomics

Virus-X was a European Union (EU)-funded consortium of 14 academic and industrial partners from eight European countries, collaborating to explore the vast and unexplored sequence space of organisms from extreme environments, such as geothermal hot springs and deep ocean ecosystems[18,19]. Launched in 2016, this bioprospecting project aimed to uncover commercially valuable enzymes through structural and biophysical characterisation of novel gene products sourced through a metagenomics-style approach to nucleic acid isolation from natural sources. The

consortium largely concentrated on viruses of bacteria and archaea, though also extended to the uncultivated cellular life present in the extreme habitats under investigation.

1.2.2. The Virus-X workflow

The Virus-X workflow followed a planned biodiscovery pipeline, outlined in Fig. 1.1, starting with the retrieval of genetic material and progressing towards an end goal of marketable products for the bioeconomy[18]. Bridging the gap between environmental sampling and commercial enzymes required high-throughput sequencing, genome construction, and gene annotation, followed by protein production, structure determination, and functional characterisation. The project was designed with this in mind, bringing together partners from various fields of expertise to collaborate. The activities were divided into four broad platforms, P1 - 4 (Table 1.1), with each partner contributing to and collaborating on at least two platforms.

Table 1.1: Platforms constituting the Virus-X workflow.

Platform	Objective	Activities
P1	Metagenome retrieval	Environmental sampling Metagenome sequencing
P2	Bioinformatics	Assembly and sorting Annotation Ecosystem dynamics
P3	Protein characterisation	Cloning and expression optimisation Protein production Activity screening and characterisation Structural determination
P4	Invention to innovation	Demonstration Dissemination

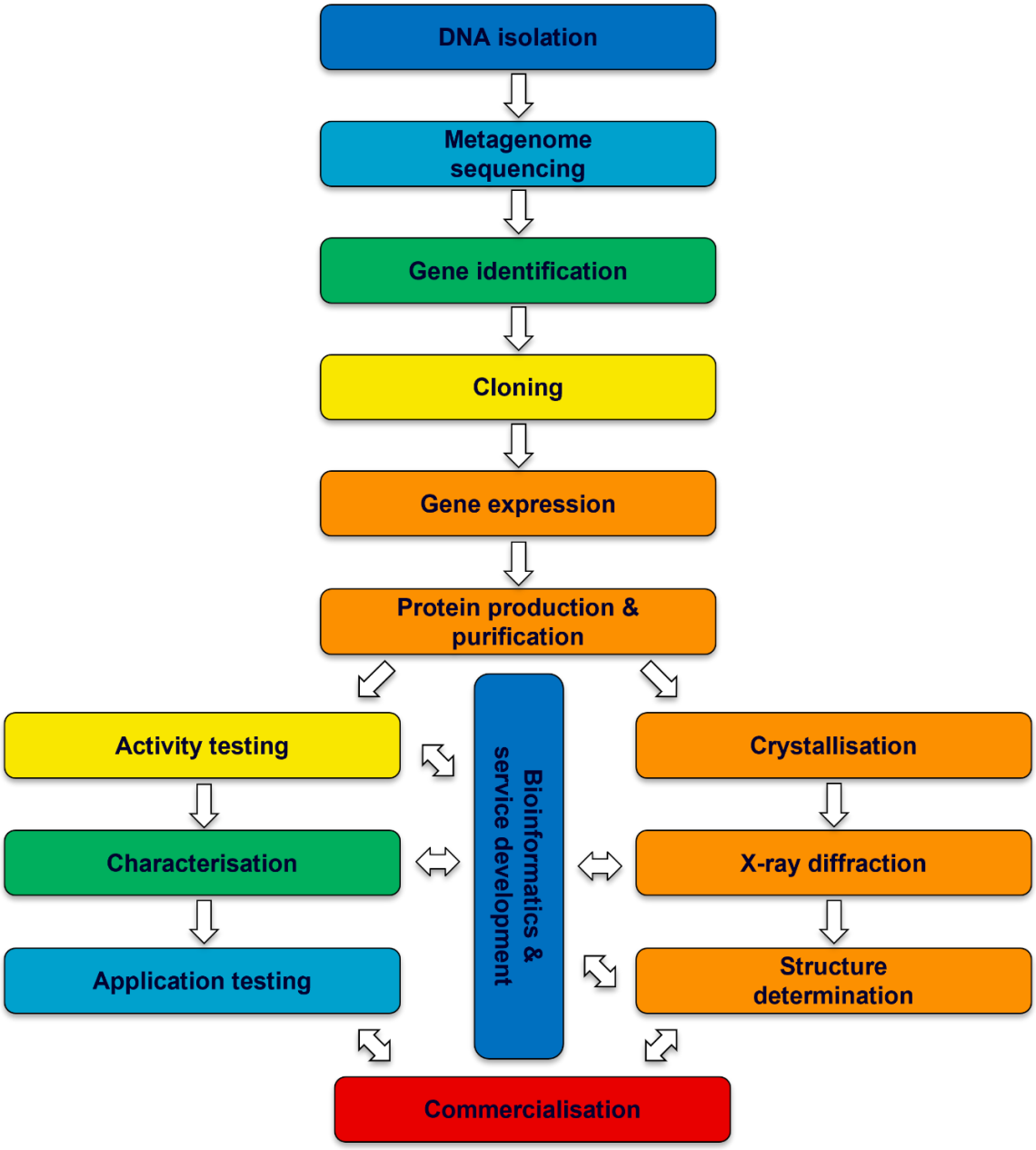

Fig. 1.1: The streamlined Virus-X biodiscovery pipeline[18]. The main activities performed by team members at Durham University are shown in orange, falling within platform P3 (Table 1.1).

This pipeline (Fig. 1.1) combining high-throughput bioinformatics technologies with experimental structural and functional analysis enabled efficient analysis of large numbers of novel genomes and their encoded enzymes on an accelerated schedule. Gene products with desirable characteristics were carried forward from one stage in the pipeline to the next, creating a funnel-effect aiming towards a handful of commercial enzymes[18]. The project results would also serve to better our understanding of the uncultivated organisms and viruses within these extreme environments[20–22], and stimulate the development of new technologies for metagenomics analyses[18,23].

1.3. Metagenomics

1.3.1. The introduction of cultivation-independent methods for genome sequencing

Prior to developments in sequencing methods, investigation of microbial communities depended on culture-based methods[24]. However, this approach resulted in a biased depiction of microbial diversity, since the vast majority of microorganisms (> 99%) resist cultivation with standard laboratory techniques[24–27]. For example, approximately two thirds of microbial research published between 1991 and 1997 focussed on only eight genera each with human health implications, including *Escherichia*, *Pseudomonas,* and *Streptococcus*, all of which can readily be cultured[25].

Advances in molecular sequencing methods resolved some of the issues presented by the cultivation bottleneck[27–29]. Environmental DNA sequencing by PCR methods was pivotal for expanding the diversity of known microorganisms[24,28,30,31]. This technique is based on detection and amplification of highly conserved genetic markers, the most widely targeted of which is the small subunit ribosomal RNA (16S rRNA) gene[31,32]. The 16S rRNA gene is considered 'universal' in the bacterial and archaeal kingdoms, and variation in sequences is considered a reliable reflection of genome divergence[33,34]. 16S rRNA sequencing was long considered the 'gold standard' for surveying microbial communities and has been used to study numerous environments[31,35,36], revealing an expanse of microbial diversity far beyond that of the cultured minority[37–40]. However, this sequencing method comes with limitations. For example, the results only convey the composition and taxonomic diversity within communities, without providing insight into the specific biological roles of the microorganisms present. The use of PCR also introduces bias towards species containing the expected marker sequences, giving an incomplete picture of true biodiversity. Predictions of community composition can also be skewed towards organisms containing more than one copy of the 16S rRNA gene[31,32].

Two alternative approaches that bypass the need for both cultivation and PCR are single-cell genomics and metagenomics. Both methods involve 'shotgun sequencing', whereby whole genomes are sequenced by first shearing them into smaller fragments, rather than targeting a specific marker for amplification. In single-cell genomics, DNA fragments from one cell or group of cells are sequenced, with all fragments being part of a single genome[41]. In metagenomics, the entire nucleic acid content of a microbial community is extracted from an environmental sample and sequenced[42]. This can be used to identify the types and distribution of different species present in a sample, as well as variations within a particular population. It often exposes previously unknown genes, having the capability to give near-complete genomes suitable for metabolic and phylogenetic analyses[43,44].

1.3.2. Metagenomics next generation sequencing

The first examples of shotgun sequencing of bacterial communities from environmental samples were conducted in 2004[44,45], sparking a boom in the use of metagenomics or 'community genomics' that has greatly expanded the breadth of genomic data available for microorganisms[41]. Dramatic

reductions in sequencing costs from over $5000 to less than $0.01 per Mbases over the past 20 years[46], coupled with computational advances and improvements in throughput, have also helped expedite the sequencing of increasing numbers of microbial genomes[47,48].

In typical metagenomics studies (Fig. 1.2), all of the DNA within a sample of interest is extracted and fragmented. The fragments are then sequenced in parallel using metagenomic next generation sequencing (mNGS) technologies[41]. The resulting sequences, known as reads, can range in length from tens to thousands of base pairs. Through identifying overlapping regions, reads are assembled into longer sequences known as contigs, which are then grouped to construct draft genomes in a process called binning[18,44]. This approach can be used to create complete or near-complete draft genomes for organisms throughout the tree of life[13,42,43,49], as well as for viruses and plasmids[50].

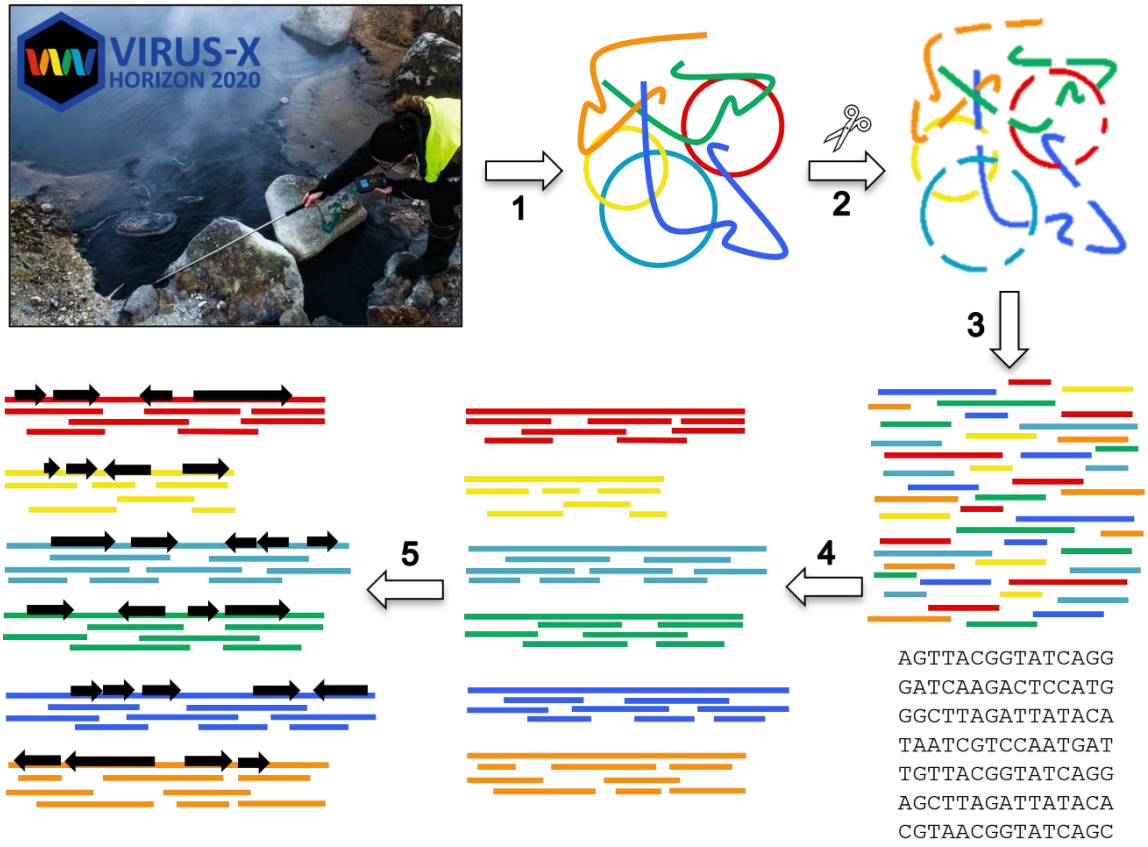

Fig. 1.2: An overview of the mNGS workflow. 1: Total DNA is extracted from an environmental sample of interest. 2: DNA is randomly fragmented. 3: Fragments are individually sequenced using NGS technologies, generating reads. 4: Reads are assembled into contigs and binned to reconstruct genomes. 5: Taxonomic classification and functional analysis of assembled genomes to detect open reading frames (ORFs) and annotate with predicted functions.

The resulting draft genomes can be aligned to existing sequences within a database for taxonomic classification of the organism, and used to study microbial interactions[51,52], metabolic processes[42,53], and roles in global biogeochemical cycles[53]. Further interrogation of diverse environments, from the

human body to the most extreme locations on the planet, will increase the number of available genomes and expand our knowledge of diversity throughout the tree of life.

1.3.3. Remaining challenges for metagenomics

The use of mNGS to recover genomes from environmental samples has rapidly increased our awareness of microbial diversity, though some challenges remain. A prominent issue is contamination with DNA sequences that arise from sampling instruments and reagents or human contamination. This issue is being addressed with development of improved software to identify and remove contaminant DNA sequences, and better quality control to ensure reagents and equipment are pure[54].

Difficulties can arise when trying to identify rare genomes that are under-represented in a sample, as current methods are less suited for studying species present in only low numbers[55]. Most microbial communities comprise many low abundance species, meaning a significant amount of diversity can be overlooked. Highly similar strains within a sample are also not always well distinguished, resulting in composite draft genomes that incorporate several sequence variants[41,56]. Despite the challenges, mNGS brings significant advantages over classical PCR-based sequencing methods, and ongoing improvements to software used for metagenome data analysis means the field will continue to progress[18,57–59].

1.3.4. Metagenome retrieval during the Virus-X project

The overarching theme of the Virus-X project was the exploration of uncharted sequence diversity contained within extreme environments. Such environments host thriving microbial communities and present a source of enzymes adapted to function under extreme conditions. A desire for enzymes with unique properties, including thermostability and extreme pH tolerance, fuels the interrogation of organisms from such environments formerly considered uninhabitable, and the biotechnological appeal of such enzymes was one of the primary motivating factors behind the selection of sampling sites by the Virus-X consortium[18]. Examples of the natural systems explored during the project include Icelandic geothermal hot springs, as shown in Fig. 1.3, and the Loki's Castle hydrothermal vent system on the Mid-Atlantic ridge between Greenland and Norway[18]. The use of metagenomics enabled the total genetic content of such environments to be isolated and analysed, including genomes of viruses and uncultivated microorganisms.

Fig. 1.3: Bioprospecting by Virus-X team members at the geothermal hot springs in Hveragerði, located ca. 45 km east of Reykjavik, Iceland.

50 sampling sites across 12 hydrothermal regions in Iceland were explored, including terrestrial hot springs with temperatures reaching 96 °C and spanning the pH spectrum (pH 2.0 to 9.3)[18]. Hot springs were also explored in western Georgia, Japan, and Italy, with generally acidic conditions (pH 1.0 to 8.0) and temperatures ranging from 50 °C to 96 °C. Many other high temperature locations were surveyed, including deep-sea hydrothermal vents in the Arctic Mid-Ocean Ridge, along with colder environments in the Norwegian Sea, with temperatures as low as −3.9 °C[18]. Over the course of the Virus-X project, 54 million genes were identified, sequenced, and annotated using a mNGS approach, from a mixture of viral and cellular genomes, including bacterial lineages from the Candidate Phyla Radiation (CPR) that have previously evaded detection by typical cultivation methods and PCR analyses[18,60].

1.4. The Candidate Phyla Radiation

1.4.1. Microbial dark matter

The use of cultivation-independent methods, including 16S rRNA gene sequencing and metagenomics, has resulted in a more than five-fold expansion of the bacterial domain of life[41]. The ensuing uncultivated phyla, constituting the vast majority of microbial diversity, are often referred to as 'microbial dark matter' as so little is known about them[40,61,62]. Many have been found to diverge phylogenetically from cultivated phyla, forming new bacterial divisions with low sequence identity to existing 16S rRNA gene sequences (< 80%), and variation in sequences typically used as signatures for bacteria (30 - 40% mismatch)[12,36]. These 'candidate phyla' that lack isolated representatives were initially named after their origins, such as TM7 ('Torf, Mittlere Schicht', the middle layer of a German peat bog) and OP11 ('Obsidian Pool', a hot spring at Yellowstone National Park)[30,36,63].

A large collection of candidate phyla was found to form a monophyletic group, appearing to subdivide the bacterial domain[41,43,64]. Draft genomes from these uncultivated phyla were first reported in 2012[42], though in the decade since then the number of sequences has soared into the thousands[41].This portion of the bacterial domain is described as the Candidate Phyla Radiation (CPR)[13,40,42,49,62], the scale and diversity of which only becomes evident when taking a metagenomics-based view of the tree of life, as shown in Fig. 1.4[64,65].

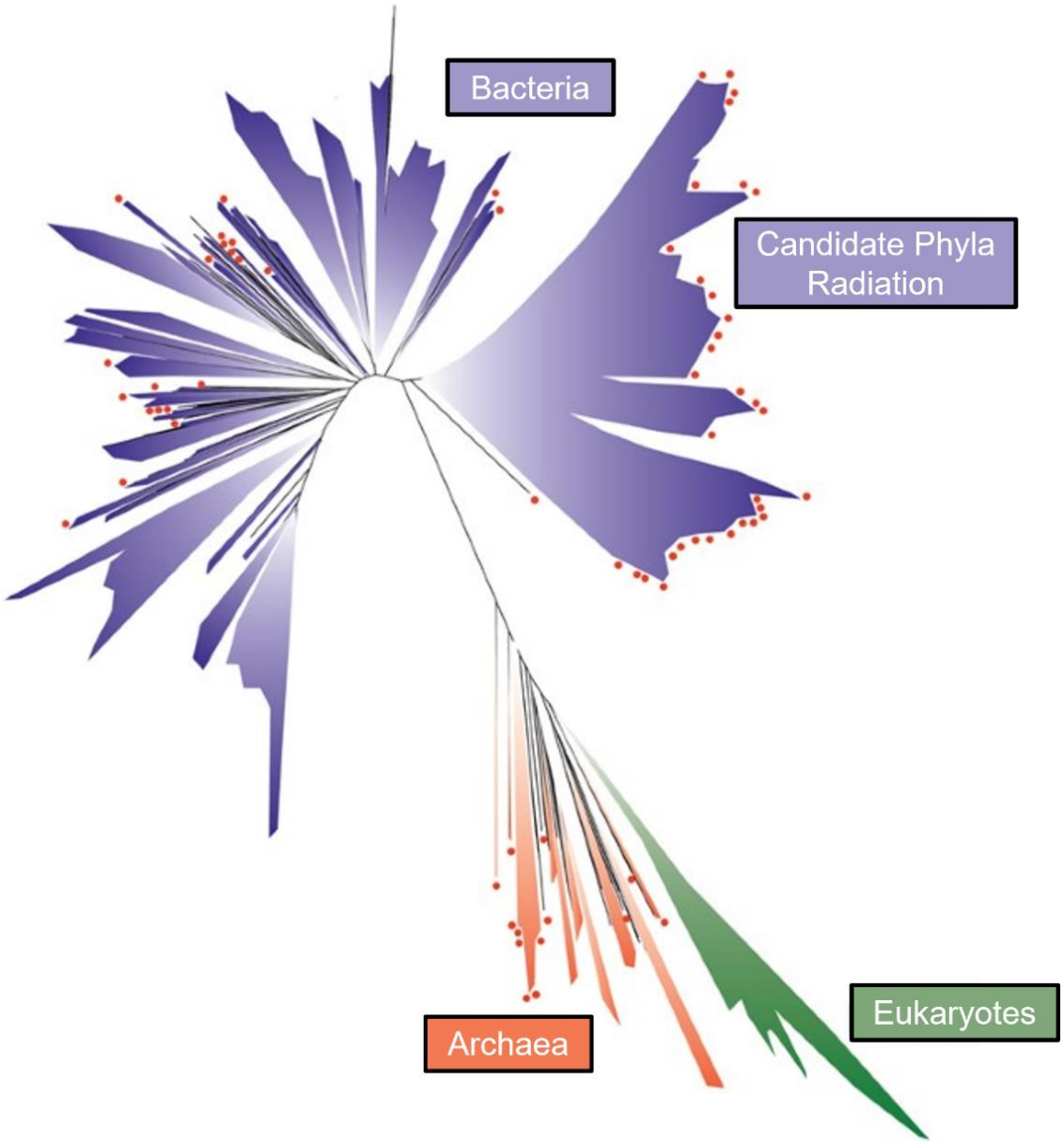

Fig. 1.4: Simplified representation of a recent view of the tree of life, including sequenced genomes from culture-independent surveys that demonstrate the magnitude of the CPR within the bacterial domain[64,65]. Red dots indicate lineages without an isolated representative. Adapted with permission (Springer Nature).

There is still uncertainty over the scale of the CPR[47,64,66]. A tree deduced from sequences of 16 ribosomal proteins (Fig. 1.4) suggested it makes up ~ 50% of all bacterial phyla[64], whereas analyses using a set of 120 conserved proteins suggested a smaller ~26%[47], and earlier 16S rRNA studies predicted a maximum of 25%[66]. Despite the ambiguity over its magnitude, there is no doubt that the CPR contains significant diversity worthy of exploration and highlights the importance of cultivation-independent methods[47,64,66]. The continued progression of phylogenetic methods and an increase in available genome sequences will help to delineate the CPR and bring us closer to a complete representation of the microbial diversity present on the planet[47,66].

1.4.2. Unusual features of CPR bacteria

Whilst genome sizes and metabolic capabilities within a phylum are usually highly varied, organisms within the CPR have consistently small genomes (from 0.7 – 1.2 Mbp)[13] and share a limited set of biosynthetic pathways[42]. Most CPR bacteria have little capability for *de novo* synthesis of numerous essential metabolites including amino acids, cofactors, fatty acids, and even nucleotides, and none have been found to contain machinery required to synthesise lipids for the cell membrane[41]. Despite generally harbouring a range of enzymes for glycolysis and fermentation[51,67,68], CPR bacteria have largely been found to lack complete electron transport chain and tricarboxylic acid cycles[41] necessary for aerobic growth, suggesting predominantly anaerobic, fermentation-based lifestyles[68]. At present, it is unknown whether the reduced genomes of CPR organisms are due to rapid gene loss occurring across the radiation, or if the CPR arose from very early ancestors with small genomes from times of anaerobic conditions on Earth[41].

The use of an unusual genetic code has been observed in several candidate phyla[13] whereby the UGA stop codon instead codes for glycine[69]. Additionally, CPR bacteria have been found to have unusual ribosome properties. Divergence in 16S rRNA sequences mean that half to all organisms sampled from certain CPR phyla would remain undetected in typical PCR surveys[43]. rRNA sequences have been reported that contain self-splicing introns and proteins encoded within them, which are highly unusual traits in bacteria[43]. Some groups of candidate phyla are also lacking ribosomal proteins (rp) that were previously thought to be universal, such as rpL9 and rpL1[43,70]. Notably, the ribosomal protein rpL30 is also missing, which is a common feature among symbiotic bacteria[70,71]. These unusual ribosome properties highlight the poorly understood and atypical biology of this part of the bacterial domain[43].

1.4.3. Predicted symbiotic lifestyle of CPR bacteria

It is widely agreed that species within the CPR are obligate symbionts due to their small size and genomes, reduced biosynthetic abilities, and the conserved absence of rpL30. A few cases of symbiosis with eukaryotes are known[72], though most are predicted to have bacterial or archaeal hosts based on their abundance and diversity in environments dominated by prokaryotes[41,68]. Some CPR species have been shown to be intracellular symbionts[72]; however, the majority are thought to live on the surface of host cells as epibionts. This is supported by evidence of physical attachment to host cell membranes[63,73–75], and enrichment of CPR organisms in 0.2 μm filtrates[41].

Host organisms provide CPR bacteria with access to nucleic acids and other metabolites they are incapable of synthesising, whilst likely benefiting from fermentation products in return[13,43]. Pili-like structures have been observed on the surfaces of CPR cells, which could provide the link between symbiotic partners for metabolite exchange[76], along with a potential role in host recognition[73]. A collection of protein families related to pili structures and DNA uptake are also markedly more

widespread within the CPR than in non-CPR bacteria[77], highlighting the importance of intercellular interactions for the lifestyle of these metabolically deficient organisms.

Due to the difficulties with cultivation, the relationships between CPR bacteria and their hosts are poorly understood. There is evidence that members of the TM7 phylum (*Candidatus Saccharibacteria*) can cause the death of their *Actinobacteria* hosts[74,78,79], suggesting a parasitic rather than mutualistic relationship[74]. However, this is a rare example of an established host for a specific CPR lineage, and is unlikely to be representative of all CPR-host relationships. Considerable research effort and capital are needed to assess the diverse range of intercellular partnerships adopted by CPR members, and the roles these associations play in their respective environments. The question also remains as to whether the symbiotic lifestyle of CPR bacteria arose as a result of their notably limited metabolisms, or whether symbiosis stimulated a loss of metabolic-related genes[65].

1.4.4. Cultivation of 'unculturable microbes'

The limited metabolisms of bacteria within the CPR, and their dependency on a host organism, are largely responsible for the difficulties associated with their cultivation. Nevertheless, much work has been conducted to isolate and culture bacterial species from candidate phyla, to enable in-depth phenotypic analysis[40]. Thus far, the only CPR bacteria to be isolated effectively have been from the human oral microbiome. The first success in 2015 was a representative phylotype from the mysterious TM7/*Candidatus Saccharibacteria* division, one of the first candidate phyla proposed from 16S rRNA gene sequences in the 1990s[63,80] and identified in many diverse habitats globally[49]. The representative, designated TM7x (*Candidatus Nanosynbacter lyticus)*, was cultured attached to the surface of an *Actinomyces odontolyticus* bacterium[74]. Subsequent successful cocultures have been limited to *Candidatus Saccharibacteria* with *Actinobacteria* hosts[73,79].

More recently, TM7 strains have also been cultivated using an approach dubbed 'reverse-genomics', along with another uncultured lineage, SR1[75]. Reverse-genomics involves capturing CPR bacteria from communities using antibodies engineered to specifically bind membrane proteins on the surface of the desired cells[75]. Whilst robust cultivation conditions still need to be identified for isolated cells[75], this approach could increase the number of cultured species from the CPR and enable deeper characterisation of host interactions and phenotypic traits.

1.4.5. The CPR as a source of biodiversity

The CPR is thought to harbour a third of all biodiversity on Earth[43,62,64,68], with members identified in a wide range of ecological niches, including soil[81,82], groundwater[13,43], hot springs[36,60,83], and even the human body[74,79,84]. CPR genomes have also been shown to contain high proportions of hypothetical proteins with no predicted function[13]. Many of these may be found to complete metabolic pathways that currently appear to be broken or missing completely from CPR genomes[13]. Others could have unique and valuable functions with novel applications. The vast amount of

unexplored sequence space therefore makes the CPR an attractive bioprospecting opportunity. Additionally, continued exploration of the CPR and deeper radiation-wide analysis will further our understanding of the tree of life, providing insights into evolution, metabolism, and the roles of microbial interactions in global ecosystems[41].

1.5. Extremophiles

1.5.1. Organisms from extreme environments

Life inhabits practically all conceivable ecological niches on the planet[85]. Thriving and taxonomically diverse microbial communities have been discovered in locations with exceptionally harsh conditions, including extremes of temperature, pH, pressure, and salinity. The study of these extremophiles and their proteins sheds light on the very edges of the envelope for life on Earth. Metagenomics analyses of such environments has provided a wealth of microbial genomes encoding valuable enzymes adapted to function under extreme conditions[86].

1.5.2. Classification of extremophiles

Organisms that are optimised for growth under extreme conditions are dubbed extremophiles[87]. Not only do these organisms tolerate such conditions, they thrive in them, adapted to live and metabolise optimally in their respective environments. Extremophiles are categorised by the physical and biogeochemical parameters for which they are optimised. The parameter thresholds for these classifications are poorly defined[85,86,88], though approximate ranges are outlined in Table 1.2.

Table 1.2: Names and optimal growth conditions of extremophile categories[85].

Extremophile Class	Conditions of Habitat
Psychrophile	< 20 °C
Thermophile	> 45 °C
Hyperthermophile	> 80 °C
Acidophile	pH ≤ 5
Alkaliphile	pH ≥ 9
Piezophile (Barophile)	≥ 50 MPa pressure
Halophile	≥ 0.2 M NaCl

Many environments, such as acidic hot springs and deep-sea hydrothermal vents, present extremes of more than one physiochemical parameter[89]. Organisms that flourish under multiple types of extremes are referred to as polyextremophiles. For such organisms, tolerance of a temperature extreme is often accompanied or even reinforced by adaptation to other extremes[85]. For example, tolerance of high salt concentrations supports survival of halopsychrophiles in cold environments by

lowering the freezing point of water[90], whilst hyperthermopiezophiles in deep-sea hydrothermal vents use the contrasting effects of high temperature and high pressure to maintain structural integrity[91,92].

Another common type of polyextremophile are thermoacidophiles (and hyperthermoacidophiles), which grow optimally under acidic conditions at high temperatures[93,94]. The acidic, high temperature conditions found at the majority of Virus-X sampling sites were therefore expected to be dominated by such organisms, including several uncultivated CPR bacteria[18,60]. It is estimated that > 80% of microbial cells in such environments are from uncultivated lineages[85,95].

1.5.3. Surviving extreme pH

Acidophiles and alkaliphiles are present in all three domains of life[2,89]. However, environments with conditions nearer the ends of the pH scale are dominated by prokaryotes, which are categorised based their optimal pH for growth (Fig. 1.5).

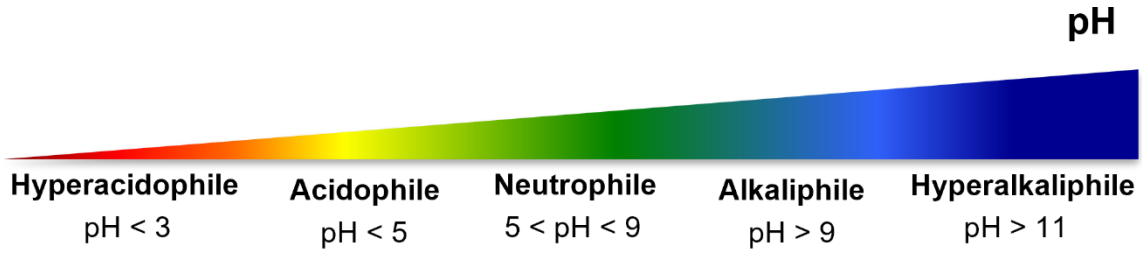

Fig. 1.5: Classification of extremophiles based on optimum pH for growth.

Currently, the most extreme pH values at which life has been identified are pH 0 and pH 12.5[85,94,96]. However, in order to sustain essential cellular functions, the cytoplasmic pH of all microbes must be kept close to neutral[85,97,98], regardless of external pH. The cytoplasmic pH is consequently preserved at ~ 5 to 6 for acidophiles and ~7.2 to 8.7 for alkaliphiles[97]. These moderate pH values are regulated by strict proton homeostasis, largely through proton transport[85,99]. Acidophilic and alkalophilic microbes can also somewhat alter the pH of their immediate surroundings through the excretion of metabolites, such as lactic acid[85,100]. Intracellular and, to an extent, local extracellular pH is therefore controlled through metabolism, meaning adaptation of individual intracellular proteins to function at extreme pH values is not required[86]. However, proton regulation and translocation systems, and extracellular enzymes are adapted to withstand the severe pH encountered in the growth environment[86,101].

1.5.4. Surviving extremes of temperature

Enzyme function and metabolic processes are limited by a susceptibility to temperature change. As a result, temperature is crucial for determining the range of biogeographical limits for many organisms, and therefore the species composition and diversity of communities inhabiting particular

environments. Microbial life is known to extend to absolute extremes of −25 °C[102] and 130 °C[103], though metabolic activity has only been detected in the range −20 °C[104] to 122 °C[91]. The classification of organisms based on their optimum growth temperature is shown in Fig. 1.6.

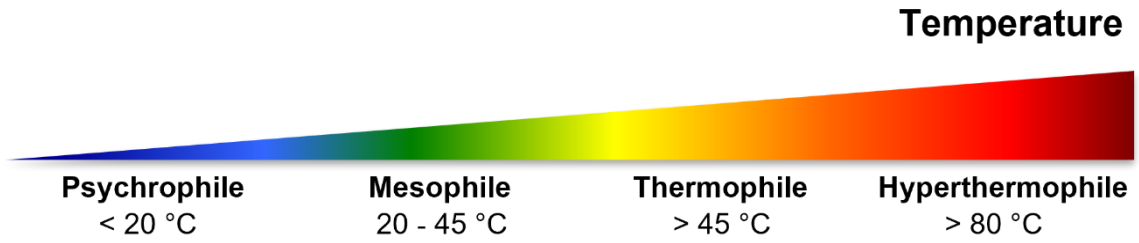

Fig. 1.6: Classification of extremophiles based on optimum growth temperature.

At temperatures above 60 °C, only a selection of prokaryotes survives[105]. The hyperthermophiles are predominantly archaea, although certain species of bacteria are known to grow at temperatures up to 100 °C[105,106], including uncultured strains identified from Obsidian Pool in Yellowstone National Park[107]. At the other end of the scale, various organisms have successfully colonised cold environments, particularly those with no temperature regulation such as bacteria, yeasts, unicellular algae, and fungi[108].

Thermophiles and psychrophiles have internal temperatures close to, if not matching, their surrounding environment. For species to survive at such extreme temperatures, their entire suite of cellular components must be able to withstand and even function optimally under these conditions. For proteins, this is in part achieved through various structural adaptations that compensate for the negative effects of high or low temperature. Thermophilic enzymes maintain rigid, densely packed structures with increased hydrophobicity at their core and increased surface charge[86,109,110]. In contrast, psychrophilic proteins have decreased numbers of ionic interactions, hydrogen bonds, and hydrophobic interactions, along with longer surface loops for greater flexibility and increased catalytic function at low temperatures[111,112].

1.5.5. Extremozymes

There is constant demand for new biocatalysts that can withstand the conditions required in industrial processes[113]. Currently, most marketed enzymes have mesophilic sources; however, the conditions required for typical industrial processes are often far from optimal for such enzymes. Biocatalysts therefore represent only a small fraction of the current industrial market for organic reaction catalysts, though their number is increasing with the continuous discovery of new enzymes from extremophilic microorganisms[106].

Enzymes from organisms that can withstand and even thrive in extreme environments are called extremozymes. Like their source organisms, extremozymes can typically withstand harsher temperatures, pH, and pressure than their mesophilic counterparts, as well as non-aqueous

environments and water-solvent mixtures in cases, making them more suited to typical industrial environments[4,106]. Extremozymes therefore have high economic potential and widen the scope of biocatalysis, with a broad range of applications in agriculture, pharmaceuticals, bioremediation, and beyond[7,86,87,106]. An improved understanding of their natural adaptations to harsh conditions will also guide the tuning of biocatalysts through rational protein engineering[3], working towards the ability to tailor enzymes with the required traits for any chosen application.

1.5.6. Uncultivated extremophiles as sources of novel enzymes

It is thought that extremophiles are the most abundant lifeforms on planet[85], though only a fraction of their sequence diversity has been explored. Metagenomics analyses continue to provide a growing inventory of diverse extremophilic genomes from which to search for novel enzymes[15,101]. However, many unique and exciting functions are likely missed in the swathes of hypothetical proteins these genomes contain, going undetected during typical functional screening processes. It can also be assumed that some of these hypothetical proteins support the life of the organism in its particular extreme environment[86]. Production of proteins from uncultivated extremophilic microbes in a heterologous expression host is required to evaluate their structures, functions, and biocatalytic properties[114,115]. Developments in metagenomics analyses, functional screening of genes, and heterologous protein production with improved systems will all expedite the identification of novel extremozymes with wide-reaching applications[15].

1.6. Structure determination of novel proteins

1.6.1. Functional insight from structure determination

The structures of proteins at atomic or near-atomic resolution reveal insights into their function and mechanisms of action, including the roles of individual side chains and interactions with substrates, cofactors, and other molecules[116]. The number of protein structures deposited in the Protein Data Bank (PDB) now exceeds 160,000[117]; nevertheless, this is only a small fraction of all known protein sequences, the number of which is continually increasing with the discovery of novel genomes, such as those of uncultivated microorganisms from metagenomics studies[10,23,86,118].

The functions of novel proteins are routinely predicted through comparison of their amino acid sequences against databases of sequences for previously characterised proteins[116,119,120]. Significant sequence or motif identity between a protein of interest and a functionally annotated protein is an indicator of shared function. However, every newly sequenced genome contains gene products found to have no annotated proteins with significant sequence similarity, prohibiting functional prediction[119,121]. A considerable number of such protein sequences are identified in the metagenomes of uncultivated microbes[13]. Determining the function of these proteins then becomes more challenging, though their three-dimensional structures can provide clues. A prediction of function can be made by comparing the structure of a protein of interest with the structures of previously

characterised proteins in databases such as the Protein Data Bank (PDB). Structure is better conserved during evolution than amino acid sequence, and proteins sharing folds and structural domains can often have relatively little sequence identity[116,119,122]. If structural similarities are identified between proteins, they can be anticipated to share similar function, which can then be validated experimentally.

A structure-based approach to functional assignment of novel proteins is common in structural genomics (SG) initiatives[122–124]. SG was instigated in the late 1990s after the advent of sequencing techniques meant larger numbers of available genomes were available to researchers[125]. It aimed to solve representative structures for all protein families, whilst also advancing technologies from genome bioinformatic analysis, recombinant protein production, crystallisation, and structure solution, to improve the speed and reduce the costs at which protein structures can be determined[123–127]. The Virus-X project builds on this international effort to 'fill sequence-structure space', providing insight into the functions of novel proteins through structure determination[18,125]. This in turn assists characterisation of other hypothetical proteins in future by expanding the variety of homologous structures available for comparison.

1.6.2. Protein structure determination with X-ray crystallography

X-ray crystallography has been described as the 'workhorse' of the structural biology field, and is responsible for the majority of structures deposited in the PDB (87% as of June 2022)[117,128]. The availability of synchrotron facilities, along with progress in automation of data collection and processing, continue to make X-ray crystallography the most popular and accessible choice for a wide-ranging community of structural biologists[129]. Whilst cryo-electron microscopy (cryo-EM) advances at a remarkable rate[130–132], and other complementary techniques such as small-angle X-ray scattering (SAXS) and Nuclear Magnetic Resonance (NMR) spectroscopy are available[128], crystallography remains the technique of choice for experimentally determining the structures of individual proteins, particularly in the context of structural genomics[126].

Crystal structure determination of novel protein requires an involved process of cloning, recombinant expression, purification, crystallisation, followed by diffraction data collection, processing, phasing, model building, and refinement[133]. The bottleneck in this process is typically the production of high quality, diffracting crystals, as protein crystallisation is often a time-consuming process of trial-and-error[133,134]. However, a variety of rationally designed, commercially available screens that apply both random and systematic variation of additives based on known crystallisation conditions allow a large amount of crystallisation space to be screened in a high-throughput fashion[128,134–137]. Obtaining high quality crystals often requires successive rounds of screening to identify chemical and physical conditions that produce crystalline material, followed by rational improvement through systematic changes of these promising conditions[134]. Advancements in liquid handling robotics have further expedited the process of crystallisation by dramatically reducing the time and sample volume

required to screen a protein of interest against hundreds of conditions[134,138]. This has enabled crystallography experiments to be conducted in high throughput, with greater accuracy, and with fewer errors in comparison to manual experiments[126,129].

On achieving diffracting crystals, the structure of a protein of interest can be solved by determining the phases of the diffracted waves, which are experimentally inaccesible[139]. This is, where possible, done by molecular replacement (MR), which uses the atomic coordinates of a structural homolog to estimate the phases of the new structure[140,141]. If suitable structural homologs are not available, alternative techniques can be applied that focus on identifying the positions of heavy atoms either intrinsic to the protein structure[60,142], or that have been incorporated artificially[143], such as by substituting methionine residues with selenomethionine[144]. These atoms produce an anomalous signal, captured in either single- or multi-wavelength anomalous dispersion (SAD/MAD) experiments, that allow their positions to be calculated, and subsequently phases to be estimated for the protein as a whole[129,139].

1.6.3. Structure determination within the Virus-X project

The primary focus of the Virus-X project at Durham University was the structure determination of target gene products by X-ray crystallography. It was anticipated that comprehensive understanding of novel target structures at a near-atomic level would illuminate unique characteristics or functionalities undetectable from sequence alone[122,127,145]. A greater bank of structural knowledge will help to develop the process of inferring biochemical function from structure, particularly in the case of hypothetical proteins. Structure determination can also be used concurrently to verify functional annotation and improve the bioinformatics techniques used to analyse protein sequences. Even for proteins where a confident prediction of function can be made based on sequence similarity, structure determination can enable comparisons between proteins from novel genomes and other organisms to reveal specific adaptations conferring additional functionalities, such as enhanced or alternate substrate specificity, or resistance to extremes of temperature and pH. This makes X-ray crystallography a vital method for studying novel enzymes and extremozymes with potential innovation value.

1.7. Project aims

The genomes of uncultivated bacteria from the CPR offer a treasure trove of unexplored sequence diversity. The discovery of CPR metagenomes from an Icelandic hot spring (75 °C, pH 5) during the Virus-X project presents an excellent opportunity to search for novel enzymes with potentially unique functionalities and innovation potential. These proteins are of added interest due to the lack of knowledge regarding their source organism, along with their potential thermophilic properties and tolerance of mildly acidic conditions.

Since the functions and desirable traits of enzymes for commercialisation are not always detectable from sequence alone, structure determination of these proteins will help to characterise their biochemical functions. The major focus of this work will therefore be the structure determination of target CPR proteins selected by Virus-X partners, principally using X-ray crystallography.

In the case of hypothetical proteins, three-dimensional structures will be used for initial assessment of function. For targets that can be functionally annotated through sequence alone, structures will help to verify this predicted function and identify unique characteristics through comparisons with homologs. The production and structure determination of these predicted extremozymes will deepen our understanding of how thermotolerance is achieved in nature through comparisons with mesophilic homologs[146]. Characterisation of these proteins will also give insights into the lifestyle of their source organism from the little understood CPR.

To fulfil these aims and maximise the number of protein targets that can be explored within the scope of the project, the biodiscovery pipeline shown in Fig. 1.1 will be followed, in collaboration with other Virus-X consortium members. The input of a large number of targets to the pipeline will provide flexibility in selecting the most promising for full characterisation and structural analysis.

Chapter 2. Selection of targets from the Candidate Phyla Radiation

2.1. Introduction

2.1.1 Virus-X target selection and categorisation

The full biodiscovery pipeline of the Virus-X project has previously been described in detail[18]. In total, nearly 40 million contigs (38,417,735) and 54 million predicted genes (54,106,508) were identified from a total assembly size of 23 Gbases. With over 50 million genes to choose from, careful selection of potential candidates for structural and functional characterisation was required. This was facilitated by the EMGB data browser developed by Virus-X partners at CeBiTec at Bielefeld University, which collated all the annotated open reading frames for screening. Part of the functional annotation of predicted genes in the Virus-X workflow included a prediction of functional domains using the Pfam database[147]. Pfam is a large collection of protein families and domains used to analyse novel genomes and guide annotation of proteins, which proved highly useful for target selection and classification of gene products from Virus-X metagenomes.

Gene products were divided into three categories (A, B, and C) based on how confidently functions could be assigned through sequence alignments.

A: known function **B**: putative function **C**: unknown function

Category A targets had 'known function', assigned through significant amino acid sequence identity, and category B targets were assigned 'putative function' based on weak sequence similarity. Targets placed in category C were those with 'unknown function', where sequence alone was insufficient to predict even a putative function[18].

Proteins in categories A and B were evaluated and ranked according to the interests of Virus-X consortium members. Each was allocated structural priority (SP) and functional priority (FP) scores from 1 (low priority) to 3 (high priority) based on their expected commercialisation potential, giving them a combined overall priority (OP) score of up to 6. Gene products with high biotechnological potential were of particular interest, with desirable criteria including heat or salt tolerance, high substrate specificity and affinity, DNA and RNA processing capabilities, high fidelity, and the ability to withstand harsh reaction condition [18] s. However, nucleotide sequences alone were insufficient to determine if targets will display such characteristics; experimental characterisation and structural

analysis was therefore vital to uncover these features and, in turn, allow for improvement of the initial annotation methods used to assign predicted functions.

Many of the targets within the Virus-X database are hypothetical proteins allocated to category C. Special focus was given to those targets that showed extended conservation across genomes, suggesting important and possibly novel functions.

2.1.2. Targets from the CPR

The primary focus of the Virus-X project was viral gene products; however, the metagenomics strategy employed also inexorably identified cellular genetic material. Some of the cellular genes discovered were found to be from the recently uncovered expansion of the bacterial domain known as the Candidate Phyla Radiation[64]. These particularly small bacteria[148,149] evaded the 0.45 μm filtration steps intended to remove cells from metagenomic samples whilst still allowing large virions to be captured[18]. Very little is known of the CPR, in large part due to the difficulties associated with cultivation, and so it presents unexplored genetic territory and abundance of novel proteins with potential innovation value comparable to that of the virosphere. CPR targets were therefore allocated high priority for cloning and expression due to the lack of knowledge surrounding their source organisms, and the potential for useful and novel functionalities they present. Targets were selected from across all three categories (A, B, and C) to offer a mixture of proteins with known, valuable functions and some with unknown but potentially unique functions.

2.2. Results and discussion

2.2.1. CPR target identification and contig analysis

A total of 18 CPR targets were selected from the Virus-X database for expression to enable structural and functional characterisation. These were all identified through metagenomic sequencing of a water sample from a terrestrial hot spring in Iceland, with conditions of 75 °C and pH 5.0. In the case of the category A and B targets, selection was conducted by Virus-X consortium members based largely on the requirements of commercial partners, with a particular focus on nucleic acid processing enzymes due to their high marketability[18]. For category C targets, extensive conservation throughout CPR bacterial genomes was used as an indicator of important function during selection.

The total metagenome from this sample contains 442,373 sequenced genes (82,285 complete genes) across 312,710 contigs. All but one of the selected targets were from a single contig of over 400 kbp and comprised of 432 genes, classified as belonging to a CPR bacterium through taxonomic annotation. The full-length contig and positions of target genes are shown in Fig. 2.1.

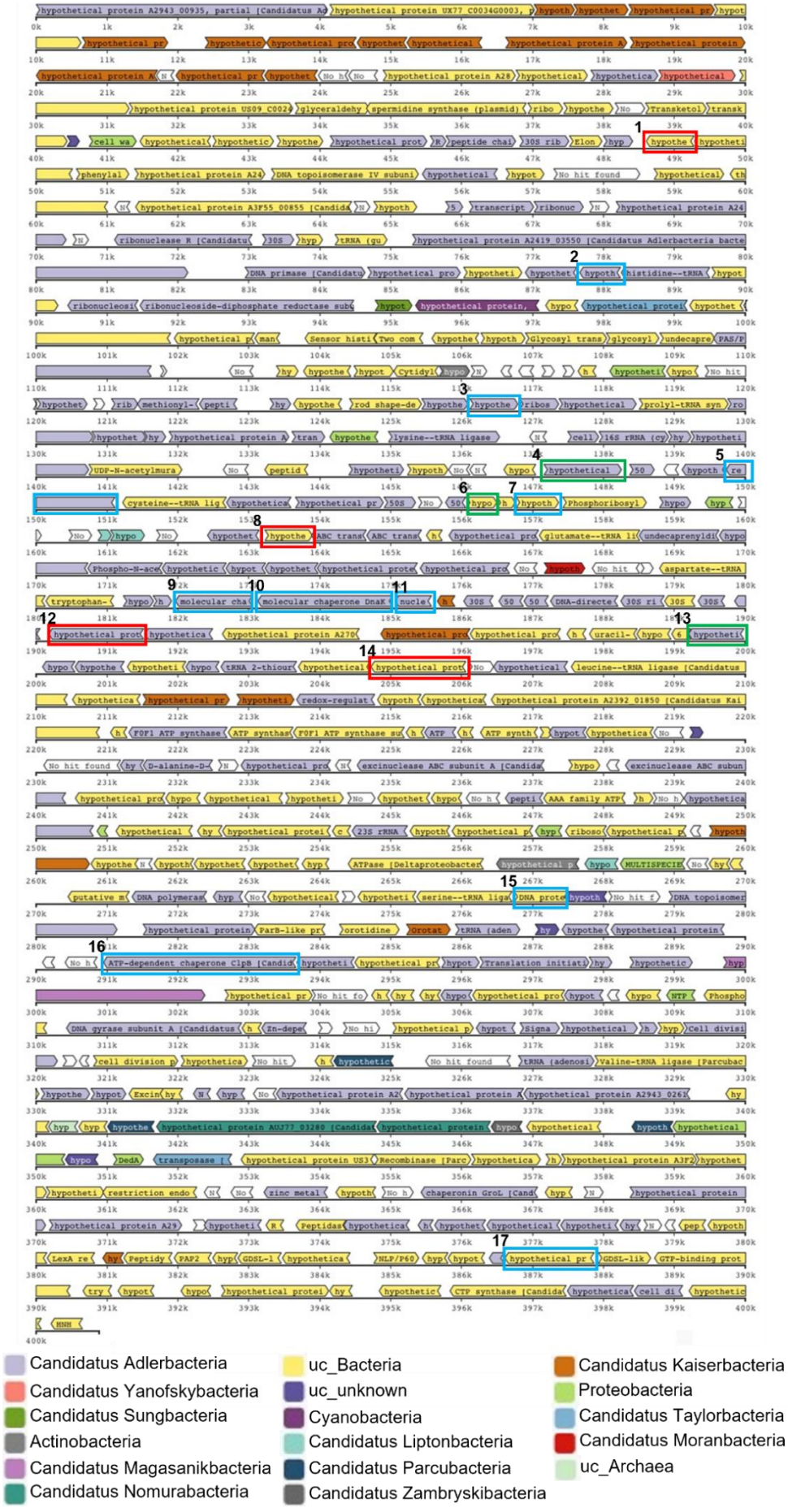

Fig. 2.1: Full-length contig and selected CPR targets, from the contig viewer of the Virus-X project data viewer EMGB[18]. Predicted open reading frames are depicted as arrows following the strand direction for

transcription and colour coded by the phylum classification level of their most significant BLAST hit, with the colour key shown underneath; uc_Bacteria = uncharacterised bacterial phylum, white = unknown phylum. Category A targets are outlined in blue, category B targets in green, and category C targets in red. Targets are numbered as follows: 1 = CPR-C1, 2 = CPR-hel-3, 3 = CPR-endo-2, 4 = CPR-B2, 5 = CPR-hel-2, 6 = CPR-B3, 7 = CPR-exo-1, 8 = CPR-C4, 9 = CPR-DnaJ, 10 = CPR-DnaK, 11 = CPR-GrpE, 12 = CPR-C3, 13 = CPR-B1, 14 = CPR-C2, 15 = CPR-DprA, 16 = CPR-ClpB, 17 = CPR-hel-1.

An additional target was selected from a second, shorter contig identified in the same metagenome, with overall length 165 kbp and containing 179 genes (Fig. 2.2).

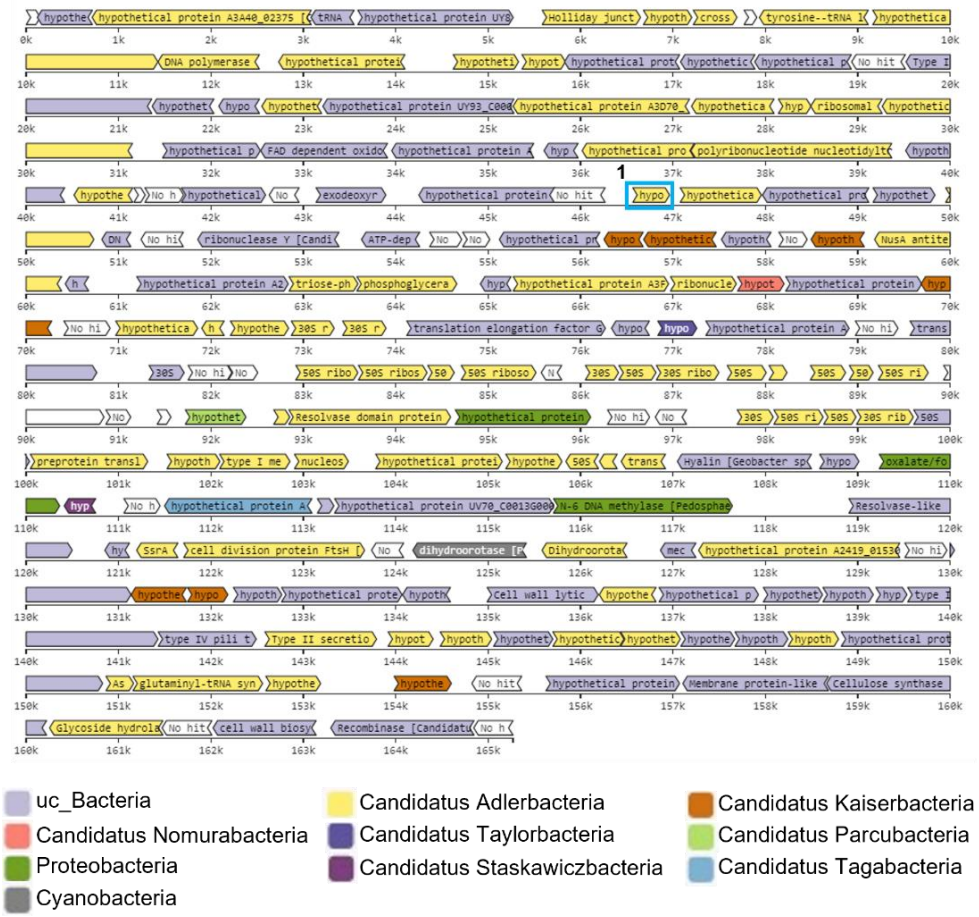

Fig. 2.2: Full-length CPR contig, from the contig viewer of the web-based Virus-X project data viewer EMGB[18]. Predicted genes are depicted as arrows following the strand direction for transcription and colour coded by the phylum classification level of their most significant BLAST hit, with the colour key shown underneath; uc_Bacteria = uncharacterised bacterial phylum, white = unknown phylum. Target numbered 1 and outlined in blue is CPR-endo-1.

Whilst CPR genomes are consistently short, these contigs (0.4 Mbp and 0.17 Mbp) are significantly smaller than previously identified closed CPR genomes (0.7 - 1.2 Mbp)[150], suggesting they do not cover complete genomes. Basic Local Alignment Search Tool (BLAST)[151] searches enabled assignment of known or predicted functions to only 30% of the gene products identified across the contigs. It was not possible to identify functions for the remaining 70% through sequence alone,

resulting in their annotation as hypothetical proteins. When classifying the genes based on the phylum classification level of their most significant BLAST hit, almost three quarters were categorised as from either *Candidatus Adlerbacteria* or an uncharacterised bacterial phylum, with the remainder mostly from other candidate phyla. This added confidence to the taxonomic classification of these contigs, which are most likely from the *Candidatus Adlerbacteria* phylum within the CPR.

2.2.2. Categorisation of CPR targets

The 18 selected proteins include a mixture of category A, B, and C targets, with varying levels of functional annotation. The categorisation and priority scores allocated by Virus-X consortium members for each target are outlined in Table 2.1. Percentage sequence identities (% seq. ID) shared with functionally annotated proteins from BLAST searches and structures deposited in the PDB are also shown where available.

Table 2.1: Details of CPR targets selected for cloning and expression.

Target	Category	Structural priority	Functional Priority	Overall Priority	% seq. ID by BLAST search	% seq. ID by PDB search
CPR-hel-1	A	1	2	3	79.7	46.2
CPR-hel-2	A	1	2	3	83.8	47.8
CPR-hel-3	A	1	2	3	63.7	-
CPR-endo-1	A	1	2	3	57.7	-
CPR-endo-2	A	1	2	3	75.0	40.7
CPR-exo-1	A	1	2	3	73.7	25.3
CPR-DprA	A	1	2	3	69.3	41.9
CPR-GrpE	A	0	2	2	57.8	31.2
CPR-DnaJ	A	0	3	3	78.7	41.2
CPR-DnaK	A	0	3	3	84.8	56.4
CPR-ClpB	A	0	2	2	85.7	47.7
CPR-B1	B	3	2	5	86.0	-
CPR-B2	B	2	2	4	89.8	-
CPR-B3	B	2	2	4	54.3	-
CPR-C1	C	3	2	5	-	-
CPR-C2	C	3	2	5	-	-
CPR-C3	C	3	2	5	-	-
CPR-C4	C	3	2	5	-	-

The category A targets have low structural priority due to the general availability of homologous structures in the PDB, which give some indication of the target structures without experimental determination. This also makes most of the A targets suitable candidates for X-ray crystal structure determination by molecular replacement (MR) techniques. Higher structural priority is given to the B targets, which have putative functions but lack significant homology with proteins of known structure. Consequently, structural solutions from X-ray diffraction data would require the use of alternative phasing methods to conventional MR. This is also the case for the C targets, which are hypothetical proteins with completely unknown structures, hence the highest structural priority allocated to each (SP 3, Table 2.1). These targets are of particular interest as they could have potentially distinct structures, with accordingly novel functions.

2.2.3. Functional annotation of CPR targets

The category A targets selected for characterisation largely have DNA/RNA processing functions, including helicases and nucleases. These functions were particularly sought after by Virus-X consortium members, with advancements in genomic research techniques creating a continual demand for novel enzymes that can manipulate and modify nucleic acids under non-standard conditions[152,153]. Additionally, proteins forming a DnaK/ClpB bi-chaperone system (CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB) were identified and chosen for analysis, since homologs have been widely exploited during recombinant protein expression to assist folding (Chapter 7)[154,155].

Three category B targets (CPR-B1, CPR-B2, and CPR-B3) were also selected, again with putative functions related to DNA processing but based on sequence similarity to proteins with only predicted or poorly defined functions. CPR-B1 was allocated to the Pfam PDDEXK_1, a functionally diverse superfamily of nucleases with various biological activities sharing a conserved PD-(D/E)XK motif[156]. CPR-B2 was annotated as part of the RmuC family, believed to be involved in DNA recombination[157]. No Pfam could be assigned to target CPR-B3, which falls on the border between categories B and C. The majority of homologs for CPR-B3 identified through BLAST searches were hypothetical proteins with unknown function, though significant similarity to some nucleases was also identified.

As well as the A and B targets selected for their desirable predicted functions, four category C targets with unknown function were chosen for analysis. These proteins could not be assigned to protein families; however, BLAST searches revealed hundreds of hypothetical proteins with high sequence similarity for each C target. These ‘hits’ were largely also from candidate phyla, indicating that these proteins are broadly distributed and likely have important roles within the CPR.

2.3. Conclusions

The selection of a manageable number of targets for cloning and expression from a database containing over 50 million genes presented a challenge. There will undoubtedly be proteins with useful functions that are overlooked. However, the careful selection and prioritisation of targets from across categories A, B, and C maximised the chance of finding desirable enzymes with innovation potential, as well as some with novel functions.

Due to their small size, CPR genomes could be expected to contain particularly high percentages of genes with known functions, dominated by universal enzymes thought to be crucial for survival. However, less than one third of the genes comprising the CPR contigs shown in Fig. 2.1 and Fig. 2.2 could be assigned predicted functions through sequence alignments, compared to up to 70% of a typical newly sequenced bacterial genome[158]. The exploration of novel gene products from the CPR is key to improving our understanding of this newly uncovered yet expansive part of the tree of life[64]. Newly sequenced CPR genomes are likely to be untapped sources of molecular tools for biological research. The category B and C targets are of particular interest in this regard, with possible unique or unusual functions, and the A targets were selected with biotechnology applications in mind. There is also the potential for additional useful features, such as tolerance to extremes of temperature, pH, and salt concentration, that could make these proteins especially valuable[159,160]. Enzymes that can withstand high temperatures are particularly desirable and have been applied to a number of biotechnological applications due to their enhanced stability. Enzymes from thermophilic organisms are generally intrinsically thermostable[161], and the source of the CPR targets is a thermophilic bacterium surviving at > 70 °C. It is therefore plausible that the selected targets will show intrinsic thermostability that will add to their innovation potential.

Chapter 3. Cloning and expression trials of the Candidate Phyla Radiation targets

3.1. Introduction

In order to characterise and study proteins of interest, they must be produced in sufficient quantities and with suitable purity for analysis. The heterologous expression of proteins in a cultured medium has enabled major advancements in biological research and biotechnology through the harnessing of host cell machinery to manufacture recombinant proteins from specially designed DNA vectors[162]. The desired protein can then be extracted and purified for investigation.

Common heterologous expression systems make use of both prokaryotic and eukaryotic organisms, including *Escherichia coli*, *Saccharomyces* yeast strains, insect cells systems, and mammalian cell cultures, as well as cell-free extracts[162–166]. These well-studied systems allow for proteins of interest to be produced outside of their normal environment, and in greater quantities than could be extracted from their natural sources. In the case of the CPR protein targets under investigation, the host organisms cannot be cultured using currently available techniques, and so recombinant expression in a host system is vital to enable their characterisation.

There are many factors that influence the selection of a suitable expression system for a specific target, such as the type of protein, the quantity required, and the folding environment necessary to produce a functionally active protein. A particularly well-established and popular expression host is *E. coli*, which grows to high cell densities with short culturing time, in inexpensive and readily available media[162,167]. Historic issues such as differences in codon usage, an inability to perform many eukaryotic post-translational modifications, misfolding, and target protein toxicity have been addressed with host strains incorporating rare codon tRNAs, additional chaperones to aid folding, and tightly regulated promoter systems, enabling even complex eukaryotic targets to be successfully overexpressed[168–171].

Another benefit of using *E. coli* expression hosts is the availability of well-engineered plasmid vectors with externally controllable promoters. The timing and levels of target protein expression can be controlled by these promoters through regulation with an associated inducer. The most popular expression vector systems feature a viral T7 promoter upstream of the gene of interest, and require specific *E. coli* host strains, such as BL21(DE3), that carry a chromosomal copy of the phage T7 RNA polymerase gene under control of the *lacUV5* promoter[172–174]. Addition of the inducer isopropyl ß-D-1-thiogalactopyranoside (IPTG) causes expression of the T7 RNA polymerase, which in turn

results in expression of the target gene through the T7 promoter. Whilst the T7 expression system is robust, undesirable basal transcription in the absence of IPTG inducer can occur. This is particularly detrimental when expressing toxic proteins, where 'leaky' expression can cause death of the host cells and mutations of the target protein. The use of *E. coli* strains containing T7 lysozyme encoded by the pLysS plasmid, which reduces the basal expression of the target protein by degradation of the T7 polymerase, can be used to add an additional layer of regulation. An alternative strategy is the use of plasmid vectors with a slower response and therefore minimal basal transcription, such as the pJOE vectors that utilise a positively regulated L-rhamnose operon system[169,175–178]. In this system, heterologous expression is under the control of the versatile $rhaP_{BAD}$ promoter. Transcription is positively controlled by two transcriptional activators, RhaR and RhaS, which are upregulated upon induction with L-rhamnose and lead to expression of the desired protein. The $rhaP_{BAD}$ promoter is also subject to catabolite repression[179], resulting in tight regulation in the absence of L-rhamnose and with the addition of D-glucose[169,179]. High levels of heterologous expression can be achieved using this system with undetectable basal transcription. It has also been noted to give better results than other vectors, particularly in cases of very high expression levels that can otherwise result in large quantities of insoluble protein[169,176,177].

E. coli was chosen as a suitable host for recombinant expression of the bacterial CPR proteins, facilitating quick and inexpensive expression. An L-rhamnose-inducible pJOE vector system was selected to limit undesired basal transcription, and for internal consistency and scalability with other Virus-X team members expressing potentially toxic viral proteins. Small-scale expression trials were first conducted to determine if the target proteins could be recombinantly expressed with these systems, which would then allow scaled-up overexpression in large quantities for downstream characterisation.

3.2. Results and discussion

3.2.1. Cloning and transformation of CPR expression constructs

Due to the substantial number of targets selected for investigation, cloning was outsourced to save considerable time and expense. The primary vector used for cloning by Virus-X members was the pBR322-based pJOE5751.1 plasmid, the features of which are outlined in Table 3.1.

Table 3.1: Details of the pJOE5751.1 plasmid used for recombinant protein production.

Plasmid	Genotype	Length / bp
pJOE5751.1	Ori_{pBR322}, *rop*, *bla*, $rhaP_{BAD}$-His_6-*eGFP*-ter_{rrnB}	4260

This vector was chosen for its tightly controlled L-rhamnose-inducible promoter, $rhaP_{BAD}$, and contains an ampicillin antibiotic resistance marker (*bla*) for recombinant cell selection. Repressor of primer (*rop*), in combination with the pBR322 origin of replication, also maintains a low plasmid

copy number to limit basal transcription of the target genes. Cloning was conducted so a His-tag (MTMITHHHHHHGS) from the pJOE5751.1 plasmid would be appended to the N-terminus of the proteins, enabling purification by immobilised metal ion affinity chromatography (IMAC).

3.2.2. Production trials of CPR targets

Cloning and transformations of the CPR constructs into competent *E. coli* cells were shown to be successful by agarose gel electrophoresis and confirmatory Sanger sequencing of purified plasmids from transformed cells. Small-scale (10 ml) expression tests were then performed to determine if the target proteins could be overexpressed. Tests were conducted both with and without the addition of 0.2% L-rhamnose to evaluate basal expression levels and make the identification of overexpression easier by whole cell sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) through expected molecular weight (MW).

Category A targets

For more manageable analysis of the category A targets, they were grouped by function: helicases, nucleases and DNA binding, and chaperones. The SDS-PAGE results for expression tests of the three helicase targets (CPR-hel-1, CPR-hel-2, and CPR-hel-3) are shown in Fig. 3.1.

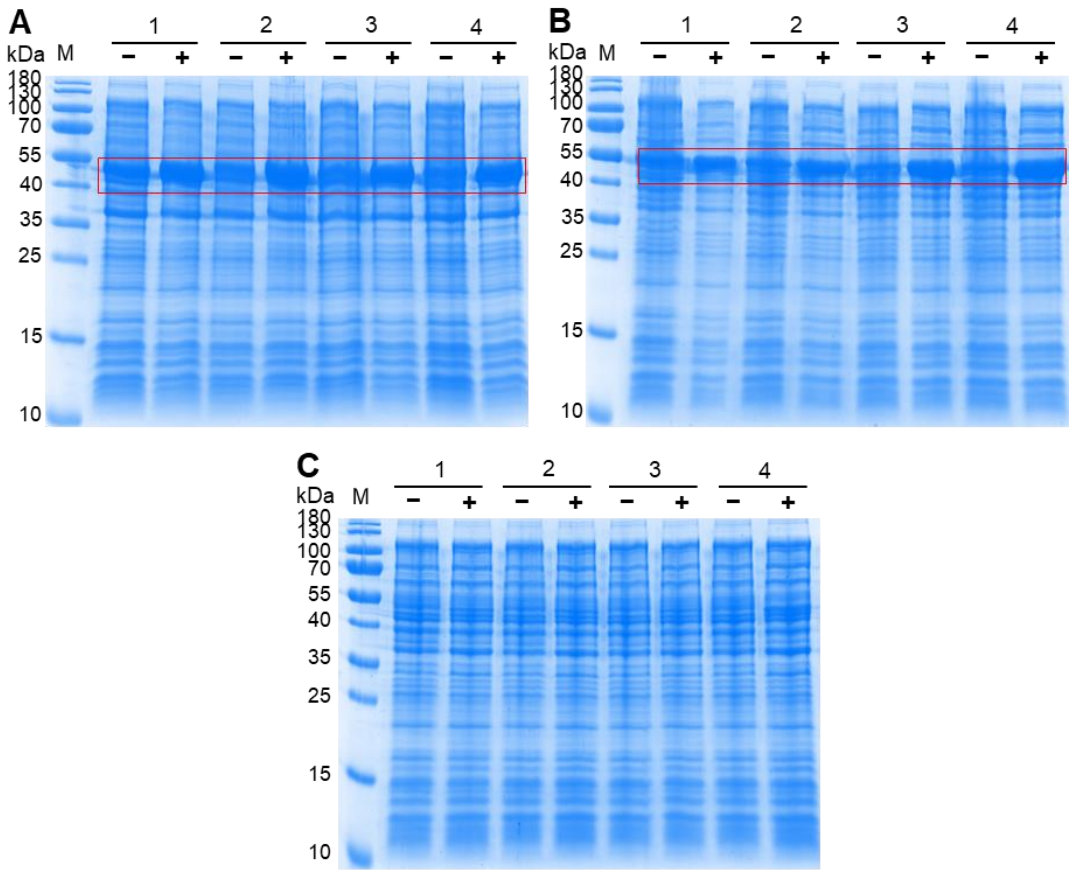

Fig. 3.1: Whole-cell SDS-PAGE results from small-scale expression tests of category A helicase targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to target protein expression are indicated with a red box. **(A)** CPR-hel-1, MW = 47.6 kDa. **(B)** CPR-hel-2, MW = 53.2 kDa. **(C)** CPR-hel-3, MW = 23.4 kDa; no bands corresponding to CPR-hel-3 overexpression were identified.

There is strong overexpression of both CPR-hel-1 (Fig. 3.1A) and CPR-hel-2 (Fig. 3.1B), even in the uninduced cultures, indicating leaky expression from the pJOE5751.1 vector. No expression of CPR-hel-3 (MW 23.4 kDa) could be detected in any test cultures, and close observation of the 15 kDa to 35 kDa portion of the CPR-hel-3 expression test gel (Fig. 3.1C) did not reveal any distinct bands that were absent in Fig. 3.1A/B. It is possible that CPR-hel-3 is being expressed in quantities too low to be detected by SDS-PAGE, requiring optimisation for expression in higher quantities.

The second group of category A targets contains those with nucleic acid processing functions, such as nuclease activity (CPR-exo-1, CPR-endo-1, and CPR-endo-2) or DNA binding and stabilisation capabilities (CPR-DprA). These four targets were successfully overexpressed to varying degrees in the small-scale tests, as can be seen in Fig. 3.2.

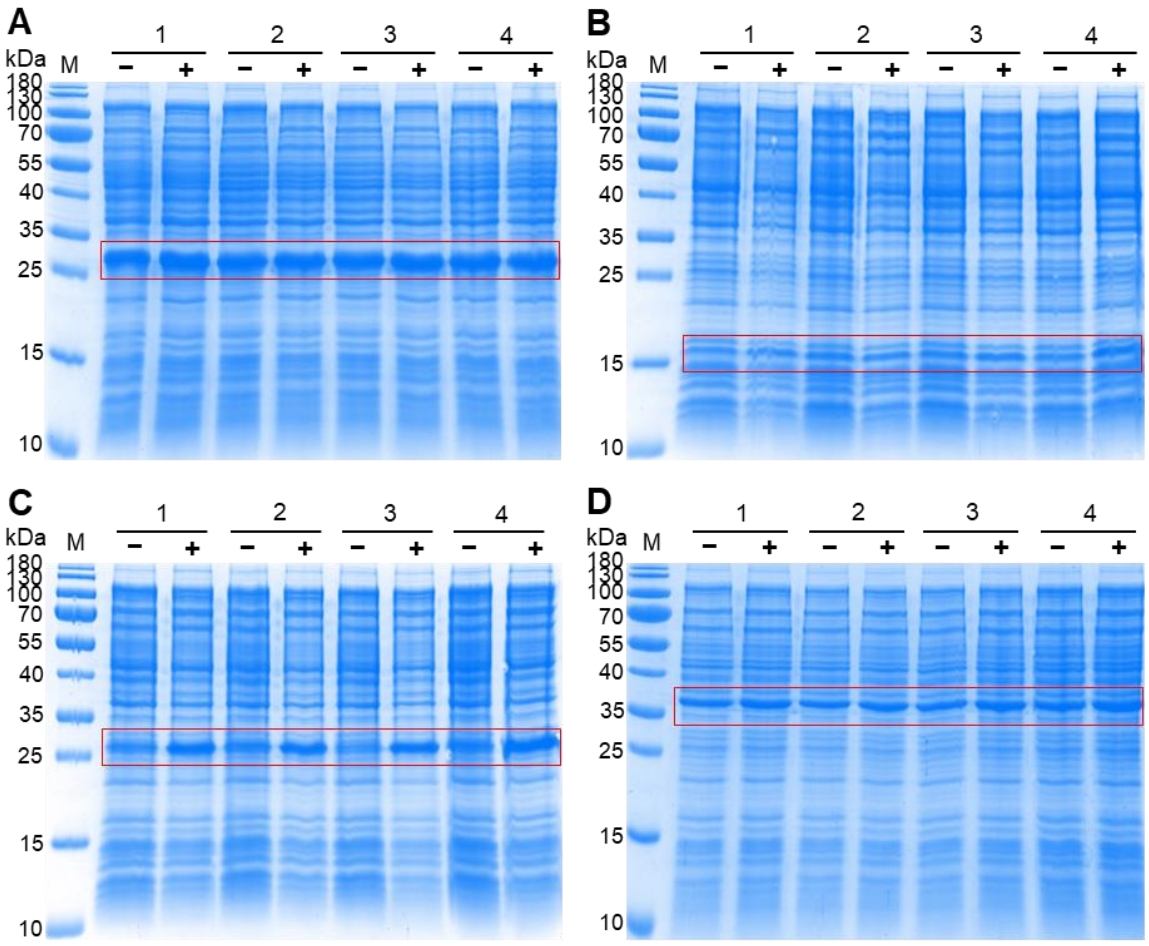

Fig. 3.2: Whole-cell SDS-PAGE results from small-scale expression tests of category A DNA processing targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to target protein expression are indicated with a red box. **(A)** CPR-exo-1, MW = 23.4 kDa. **(B)** CPR-endo-1, MW = 16.9 kDa. **(C)** CPR-endo-2, MW = 27.1 kDa. **(D)** CPR-DprA, MW = 32.2 kDa.

CPR-exo-1 expressed extremely well (Fig. 3.2A), though expression levels without L-rhamnose were almost as high as in the induced cultures. Basal and induced expression levels were also similar for CPR-endo-1 (Fig. 3.2B) and CPR-DprA (Fig. 3.2D), though the overall expression levels of these targets were lower than for CPR-exo-1. In the case of CPR-endo-1, bands around the 15 kDa marker needed to be examined across all expression test gels to determine which likely corresponded to CPR-endo-1 as opposed to native *E. coli* proteins. The addition of L-rhamnose did considerably improve expression levels of CPR-endo-2 (Fig. 3.2C), though significant basal expression is still evident.

The remaining category A targets are those comprising the DnaK/ClpB bi-chaperone system: CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB. Whole-cell SDS-PAGE results from expression tests of these four targets are shown in Fig. 3.3.

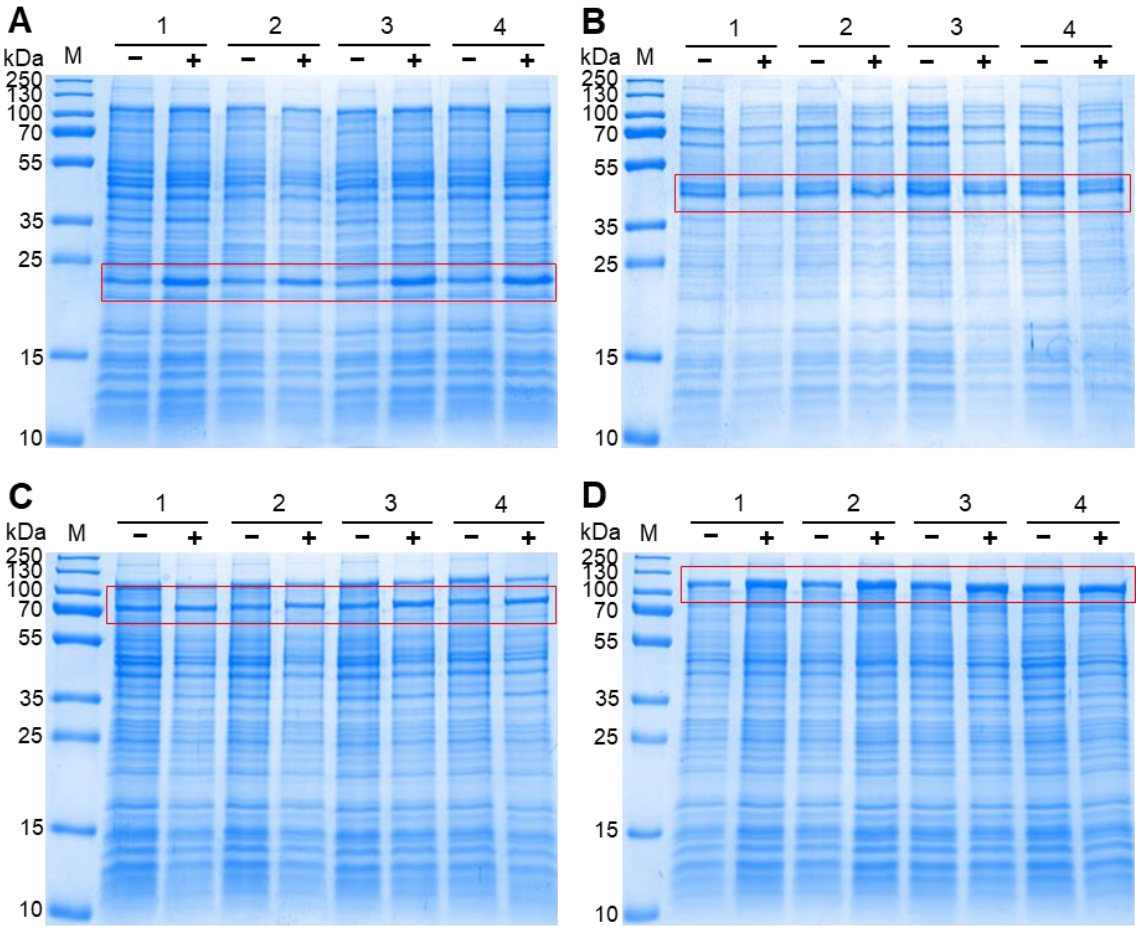

Fig. 3.3: Whole-cell SDS-PAGE results from small-scale expression tests of category A chaperone targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to target protein expression are indicated with a red box. **(A)** CPR-GrpE, MW = 20.8 kDa. **(B)** CPR-DnaJ, MW = 40.6 kDa; the corresponding band is the lower MW of the two visible within the red box. **(C)** CPR-DnaK, MW = 69.7 kDa. **(D)** CPR-ClpB, MW = 102.1 kDa.

Overall expression levels for the chaperone system proteins were moderate in comparison to other A targets. Careful analysis of the 35 kDa to 55 kDa region of Fig. 3.3B was required to identify the band probably corresponding to CPR-DnaJ expression. Basal expression of all four proteins is also evident, though addition of L-rhamnose did increase the expression levels of CPR-GrpE (Fig. 3.3A) and CPR-ClpB (Fig. 3.3D).

Category B targets

SDS-PAGE analyses of small-scale expression tests with the three category B targets (CPR-B1, CPR-B2, and CPR-B3) are shown in Fig. 3.4.

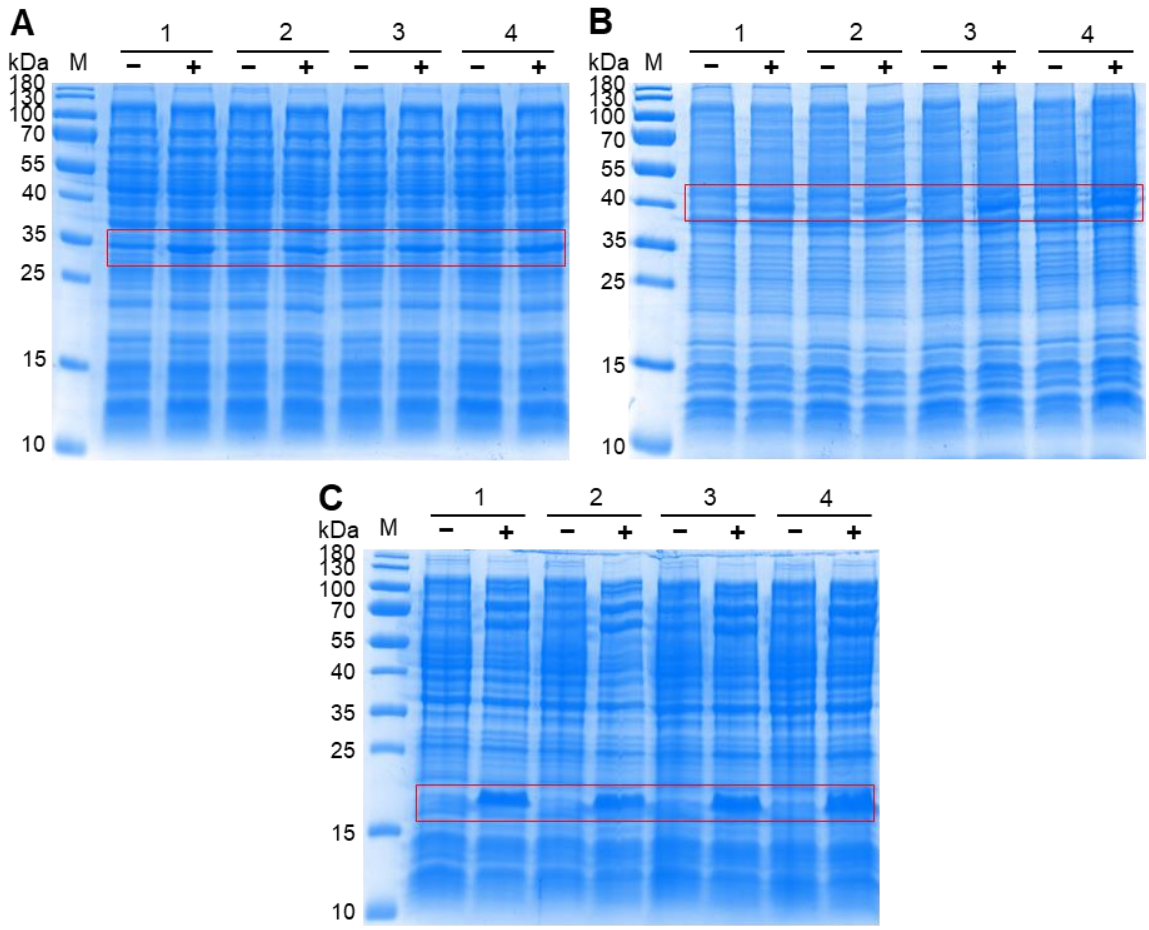

Fig. 3.4: Whole-cell SDS-PAGE results from small-scale expression tests of category B targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to target protein expression are indicated with a red box. **(A)** CPR-B1, MW = 30.6 kDa. **(B)** CPR-B2, MW = 42.3 kDa. **(C)** CPR-B3, MW = 16.3 kDa.

Bands indicating overexpression were observed for all three B targets, and in each case the addition of L-rhamnose induced a clear increase in expression levels. This is particularly evident for CPR-B3, with negligible basal expression and strong bands visible in induced culture samples on the gel

(Fig. 3.4C). Expression levels appear lower for CPR-B1 (Fig. 3.4A) and CPR-B2 (Fig. 3.4B), though bands matching the expected MWs of the target proteins in induced cultures are clear.

Category C targets

Expression tests of the four category C targets (CPR-C1, CPR-C2, CPR-C3, and CPR-C4) were less successful than the A and B targets, as can be seen from the whole cell SDS-PAGE results in Fig. 3.5.

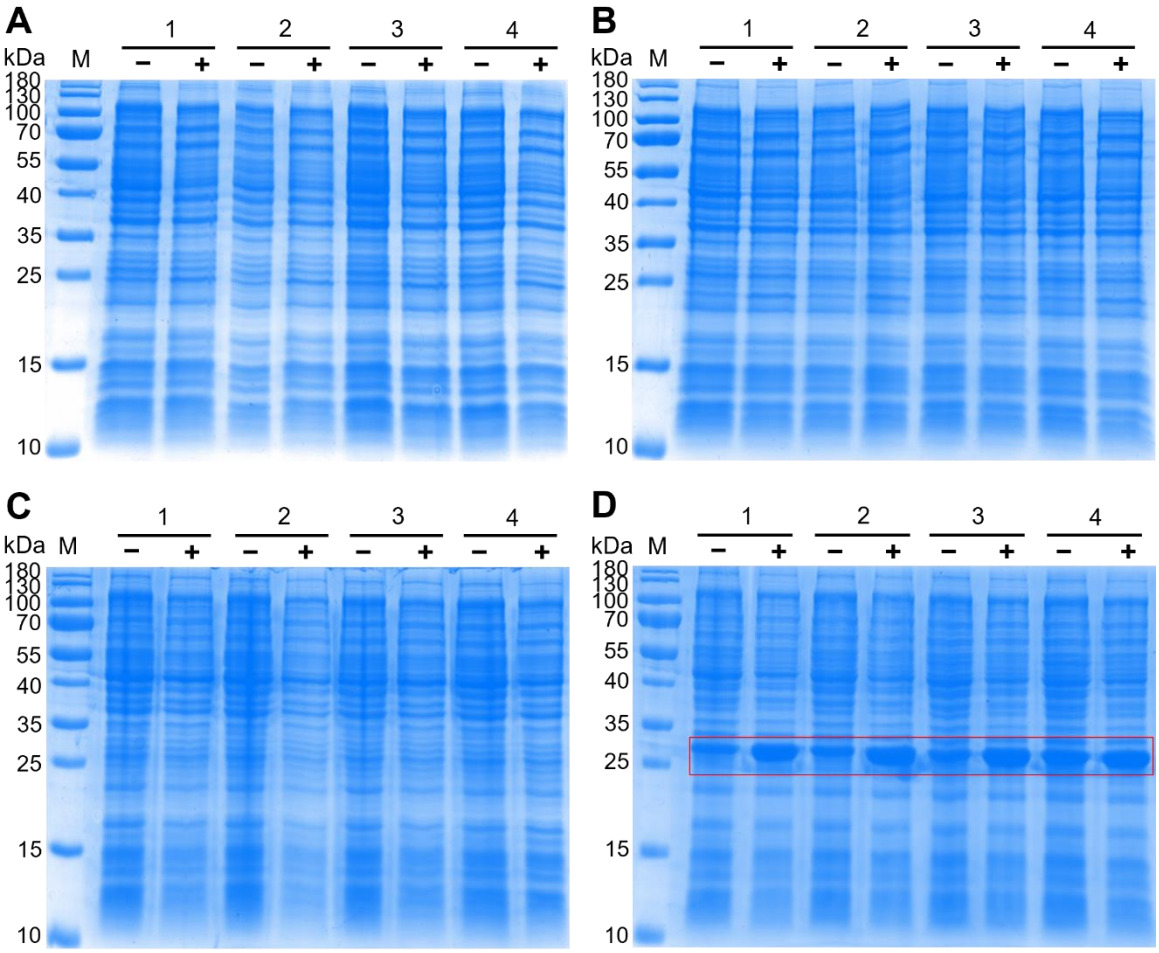

Fig. 3.5: Whole-cell SDS-PAGE results from small-scale expression tests of category C targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. **(A)** CPR-C1, MW = 26.2 kDa. **(B)** CPR-C2, MW = 53.2 kDa. **(C)** CPR-C3, MW = 48.0 kDa. **(D)** CPR-C4, MW = 27.1 kDa; positions of bands corresponding to CPR-C4 expression are indicated with a red box.

On studying the C target gels, distinct bands corresponding to CPR-C1, CPR-C3, or CPR-C3 could not be identified. Due to the hypothetical nature of these targets, it is possible that they are not naturally expressing proteins. Expression could also be occurring in levels too low to be detected by SDS-PAGE. For CPR-C3 (Fig. 3.5C), gel bands of induced cultures are discernibly fainter than for their uninduced counterparts. This could suggest the target protein is toxic, causing cell death when

induced to express. Only CPR-C4 is clearly overexpressed (Fig. 3.5D), with low levels of basal expression relative to induced expression.

3.2.3. Selection of targets for large-scale expression

The whole cell SDS-PAGE analysis of the small-scale expression tests served as a qualitative assessment of expression levels for each target. Table 3.2 outlines the relative levels of basal and induced expression, based on comparisons of band strengths across the gels in Figs. 3.1 to 3.5: − signifies no detectable expression, + represents weak expression, + + for moderate expression, and + + + for high expression.

Table 3.2: Results of small-scale test expressions of CPR targets.

Target	MW / kDa	Basal expression	Induced expression
CPR-hel-1	47.6	+ +	+ + +
CPR-hel-2	53.2	+ +	+ + +
CPR-hel-3	23.4	−	−
CPR-exo-1	23.4	+ + +	+ + +
CPR-endo-1	16.9	+	+
CPR-endo-2	27.1	+	+ +
CPR-DprA	32.3	+	+
CPR-GrpE	21.0	+ +	+ + +
CPR-DnaJ	40.6	+	+
CPR-DnaK	69.7	+ +	+ +
CPR-ClpB	102.1	+ +	+ + +
CPR-B1	30.6	+	+ +
CPR-B2	42.3	+	+ + +
CPR-B3	16.3	−	+ + +
CPR-C1	26.2	−	−
CPR-C2	53.2	−	−
CPR-C3	48.0	−	−
CPR-C4	27.1	+	+ + +

14 of the 18 CPR targets were seen to overexpress on addition of 0.2 % L-rhamnose, with basal expression also evident in almost all cases. Fortunately, basal expression did not have any noticeable effect on the host cell viability and could be disregarded at this stage. This point presented the first opportunity to select which targets would continue along the Virus-X biodiscovery pipeline. The 14 targets found to readily overexpress (Table 3.2) were forwarded to large-scale (2 – 6 l) expression experiments to produce protein for characterisation and structure determination.

All pJOE5751.1-*CPR* expression constructs were shown to contain the desired insert sequences in frame for transcription. The construct sequences were also codon optimised for expression in *E. coli* to remove tRNA availability as a limiting factor. This suggests that the apparent lack of expression for CPR-hel-3, CPR-C1, CPR-C2, and CPR-C3 was not due to issues with their respective constructs. Whilst attempts could be made to express these targets, it was decided that a more effective use of time and resources overall was to focus on the successfully expressing proteins.

3.3. Conclusions

The small-scale expression trials were useful to establish which CPR targets could be overexpressed in the *E. coli* expression host; these targets were then advanced to larger scale production to enable characterisation and structure determination.

The amount of basal transcription detected in these expression tests was surprising, since the L-rhamnose system should minimise expression in the absence of inducer. In several cases, the expression levels from non-induced cultures matched those from induced cultures, indicating heavily leaky expression from the pJOE5751.1 vector (Fig. 3.2 A and D). Whilst the $rhaP_{BAD}$ promoter system did not work as anticipated, its versatility still affords additional control if necessary. In cases with low levels of basal expression, target protein expression can be fine-tuned by varying the concentration of L-rhamnose used during induction. An extra layer of regulation can also be introduced by adding a small amount of D-glucose to cultures, which tightly represses expression of target proteins. These adjustments can be implemented where necessary if undesirable basal expression proves to be an issue during scaled-up experiments.

Over the course of the Virus-X project, 659 genes were cloned into expression vectors and production tests were conducted for 478, of which ~ 65% were successfully recombinantly expressed[18]. The success rate seen with the CPR targets is notably higher, with 78% (14 / 18) of the cloned proteins effectively overexpressing. Of the four targets showing no detectable overexpression, three were category C targets (CPR-C1/C2/C3). These are high priority targets due to their unknown functions, which could prove to be novel or have high innovation potential. However, the hypothetical nature of these targets also means they may not be naturally expressed *in vivo*. Rather than expending effort on troubleshooting the expression of these targets, the decision was made to instead focus on characterising the one successfully overexpressed C target, CPR-C4.

Chapter 4. Production and characterisation of Candidate Phyla Radiation targets

4.1. Introduction

4.1.1. Tags and fusion proteins

After expression in a heterologous host, recombinant proteins need to be recovered in a pure, soluble, and active form. This is typically achieved with the addition of a 'tag' to the protein of interest, attached to the N- or C-terminus, or both, via a short linker sequence, creating what is known as a fusion protein[180]. Popular tags include glutathione-S-transferase (GST-tag) and polyhistidine (His-tag), which facilitate purification of a target protein from crude cell extracts through affinity to an immobilised substrate[181–185]. The use of these tags and others, including maltose binding protein (MBP)[171], thioredoxin[186], and small ubiquitin-like modifier (SUMO) tags[187,188], can also increase the solubility of target proteins by promoting correct folding. The benefits of such tags are clear; however, their presence can affect the activity and crystallisation of proteins. For example, the dimeric GST-tag promotes dimerisation of the attached target protein. The large GST and MBP tags are also common contaminants found to nucleate in place of recombinantly expressed proteins during crystallisation experiments[189]. Consequently, expression constructs are often designed with a protease recognition sequence in the linker to allow removal of affinity tags using site specific proteases such as thrombin or Tobacco Etch Virus protease (TEV)[190,191]. This is not always necessary, however, particularly with short and typically inconspicuous His-tags[180,192], or in cases where crystallisation is dependent on the presence of a tag[193,194].

4.1.2. Purification methods

The incorporation of affinity tags has become the standard procedure for efficient purification of recombinant proteins. In particular, His-tags deliver excellent yields of target protein through binding to an inexpensive, high-capacity resin in a process called immobilised metal ion affinity chromatography (IMAC)[184,185]. IMAC is based on the high affinity of the histidine-rich tag for metal ions immobilised on a chromatography medium. IMAC therefore captures the tagged fusion protein, whilst allowing the removal of impurities and contaminants through washing with common buffer components. As a result, it is possible to obtain high purity samples directly from crude cell lysate.

Additional purification methods exist that take advantage of unique properties of target proteins, such as charge, size, and hydrophobicity, to separate them from others in the sample. Ion-exchange chromatography (IEC) exploits electrostatic interactions between the target protein and fixed charges

on the surface of the chromatography medium, separating proteins based on their charge[195,196]. Size exclusion chromatography (SEC), also known as gel filtration, separates proteins based on differences in size as they pass through a porous, unreactive matrix formed from spherical beads of a dextran and agarose composite[197,198]. The extent to which molecules access the pores between beads varies depending on their size, with smaller proteins passing more slowly than the larger constituents of a sample. SEC is particularly useful for removing fragments and aggregates of the target protein, which are difficult to separate by IMAC and IEC.

The purity requirements for a protein of interest are dependent on its final application. Suitable purity for activity assays and characterisation may only require a single purification step, usually IMAC. For structural studies and crystallisation, the purity demands are higher. Samples must be homogenous to form high quality crystals that diffract well for structure determination[134]. An additional polishing step, typically SEC, is therefore often required to achieve the desired purity, with more steps needed in some cases[199]. Increasing the number of steps will improve the overall purity but at the cost of yield, and potentially reduced activity as the time required for purification increases. Careful selection and optimisation of techniques is therefore required for efficient purification, with success reliant on numerous factors including sample composition, concentration, volume, pH, temperature, and ionic strength. Such optimisation has been simplified by advances in automated chromatography systems such as the ÄKTA range by Cytiva, enabling reliable, repeatable, and straightforward purification[199].

4.1.3. Protein characterisation

Following purification, it is necessary to characterise the protein of interest to confirm its identity, and ensure it is of suitable quality for downstream functional and structural studies[199]. The extent of characterisation required for a target protein depends on its final usage. For structural determination by crystallisation, rigorous characterisation is required, since chemical and physical heterogeneities can prevent crystallisation[134]. Preliminary biophysical analysis can therefore guide optimisation of protein quality to improve crystallisation potential.

Sodium dodecyl-sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is ubiquitously used to quickly confirm that a protein is the expected MW, and give an indication of purity[200,201]. The intensity of bands upon staining is typically proportional to the amount of protein present in the sample, providing an indication of target protein to contaminant ratios[202]. A more quantitative method is the Bradford assay, which determines the protein concentration of a sample relative to bovine serum albumin (BSA) standards through binding to Coomassie Blue dye[202]. Alternatively, the UV absorption of the protein sample at 280 nm (A_{280}) can be measured to estimate concentration quickly and conveniently, using molar extinction coefficients (ε) predicted using the ExPASy ProtParam tool[203].

As well as being a useful separation method, size exclusion chromatography (SEC) can be used to evaluate the oligomeric state of target proteins and the homogeneity of samples, which is important for crystallisation. Analytical SEC using a column calibrated against known standards can reveal protein aggregation and multiple oligomeric states, which can be nonspecific or biologically relevant[199,204].

Mass spectrometry (MS) is another crucial analytical technique for authenticating proteins of interest. Intact-protein MS requires species to be vaporised and ionised intact using 'soft ionisation' methods, typically electrospray ionisation (ESI) or matrix-assisted laser desorption/ionisation (MALDI), with time of flight (TOF) analysis used to determine species mass[205–208]. The resulting mass spectra give MWs of species to within a few Da accuracy. Contaminants, degradation, and the occurrence of chemical or post-translational modifications can also be detected, and the sensitivity of MS techniques allows proteins to be detected at very low concentrations, requiring only small amounts of sample[205]. ESI-TOF MS is particularly suited for analysis of large numbers of protein targets due to straightforward sample preparation protocols and sensitivity within the kDa range[199]. Sample digestion with proteolytic enzymes, such as trypsin, followed by tandem MS analysis of the resulting peptides can also confirm the identity of a protein species through comparison of distinct fragmentation patterns to databases of amino acid sequences[209].

4.1.4. Thermal shift analysis

Buffer composition can strongly influence the stability and homogeneity of a protein sample. Purified proteins can undergo precipitation that worsens with increasing concentration[199], which is problematic when high concentrations are required for crystallisation. In these cases, it is prudent to find conditions that stabilise the protein, through optimisation of pH, salt concentration, and the addition of ligands[199]. The buffer environment also contributes to the protein adopting an active conformation, which is essential for functional assays.

One such method for determining stabilising conditions is thermal shift analysis (TSA), also known as the Thermofluor assay or differential scanning fluorimetry (DSF)[210–212]. This technique involves measuring changes in the thermal denaturation of a protein under different conditions using a fluorescent dye. A wide range of conditions can be screened in parallel using a high-throughput 96-well format and a standard real-time PCR machine, enabling it to be performed routinely in most laboratories. Commercially available screens, such as the Durham Screens® by Molecular Dimensions, systematically vary the pH and buffer compositions, salt environment and ionic strength, and test a variety of 'chemical chaperone' additives that are known to stabilise and refold disordered proteins in stress environments[210,211,213].

The fluorescent dyes employed are sensitive to their environment, with a high fluorescence output in hydrophobic environments that is rapidly quenched in polar environments. As a result, the fluorescence intensity tracks the state of protein folding. When folded, the hydrophobic core of the

protein is buried and inaccessible to the dye, resulting in low fluorescence. As the temperature increases and the protein unfolds, hydrophobic regions are exposed, enabling binding of the dye, and causing the fluorescence output to increase. The temperature at the inflection point of the resulting melt curve, also known as the melt temperature, T_m, is then used as a quantitative measure of protein thermostability under various conditions.

TSA is useful for assessing the stability of large numbers of targets in a high-throughput manner. Comparisons of T_m values from screening can determine favourable, stabilising buffer components and additives. These results can then be used throughout the protein production process, from improving soluble expression, to guiding notoriously time-consuming and sample intensive crystallisation experiments[214–216].

4.1.5. General strategy for purification and characterisation of Virus-X targets

The production of pure, soluble CPR proteins in milligram quantities was required to facilitate crystallisation experiments. A general protocol was established as a starting point for expression and purification, which was then optimised for challenging but high value targets[18]. The strategy chosen for the CPR targets follows the consensus approach described by the Structural Genomics Consortium (SGC) and other protein production research groups, based on comparisons of their optimised approaches[199]. This strategy consists of overexpression of targets as fusion proteins with an N-terminal His-tag, with induction at low temperatures (15 – 25 °C) to minimise formation of insoluble inclusion bodies. Overexpression is followed by initial capture and purification by IMAC with well-buffered, high ionic strength salt solutions (20 mM HEPES pH 7.5, 300 mM NaCl) that have been successfully applied by structural genomics platforms. If required, polishing purification by SEC is conducted. Proteins are then characterised by SDS-PAGE, analytical SEC, ESI-TOF MS, and UV absorbance at 280 nm to validate their identities, and estimate purity and concentration for downstream structural studies. TSA will also be conducted, serving the dual purpose of assessing intrinsic thermotolerance as a highly desirable feature for biotechnology applications[106,217,218], and identifying favourable conditions that improve thermostability[210,211].

It was expected that some CPR targets would accumulate as insoluble aggregates owing to misfolding during overexpression, as is a common issue in heterologous protein production in *E. coli*. Whilst there are methods available to solubilise and refold these aggregates, they are time-consuming, cumbersome, and often give only poor recovery[219,220]. Therefore, only those CPR targets found to readily express in soluble form would be carried forward in the first instance. This ruthless approach maximised the time and resources available to characterise the largest number of targets. However, in cases where the targets were particularly high value, more involved attempts were made to optimise protein production.

4.2. Results and discussion

4.2.1. Preliminary characterisation of CPR targets

The CPR targets were produced as fusion proteins with an N-terminal His-tag and short linker sequence without a protease recognition sequence (MTMITHHHHHHGS). Whilst the tag could not be removed, it was suitably small and deemed unlikely to influence the structure or activity of the proteins[192,221].

Physical and chemical parameters, including molecular weight (MW), molar extinction coefficient (ε) at 280 nm, and isoelectric point (pI), were calculated using the ExPASy ProtParam tool[203,222] to assist purification and characterisation of the CPR proteins. These are outlined in Table 4.1.

Table 4.1: Parameters for the CPR targets. The molar extinction coefficient, ε, is at 280 nm.

Target	Category	MW / kDa	ε / M^{-1} cm^{-1}	pI
CPR-hel-1	A	47.6	7450	10.11
CPR-hel-2	A	53.2	34380	5.69
CPR-endo-1	A	16.9	13980	9.22
CPR-endo-2	A	27.1	24410	9.99
CPR-exo-1	A	23.4	27960	5.49
CPR-DprA	A	32.3	8940	6.88
CPR-GrpE	A	20.8	11460	5.45
CPR-DnaJ	A	40.4	13410	8.28
CPR-DnaK	A	69.6	15930	5.32
CPR-ClpB	A	102.0	40800	6.66
CPR-B1	B	30.6	47900	6.30
CPR-B2	B	42.3	38850	5.85
CPR-B3	B	16.3	16960	6.55
CPR-C4	C	27.1	50420	9.39

Based on the tag sequence, it was predicted that N-terminal methionine excision (NME) would occur[223,224], and this was considered when calculating the expected MWs. These MWs facilitated identification of target proteins during SDS-PAGE and MS analysis, and the extinction coefficients enabled spectrophometric estimations of concentration through absorbance at 280 nm.

4.2.2. CPR-hel-1

Expression and IMAC purification

CPR-hel-1 was expressed on a 2 litre scale, followed by IMAC purification. SDS-PAGE analysis of fractions from IMAC purification are shown in Fig. 4.1.

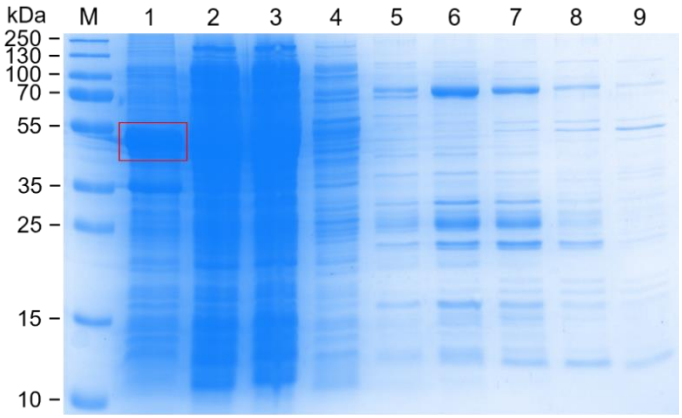

Fig. 4.1: SDS-PAGE analysis of CPR-hel-1 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 9 = protein fractions. The band corresponding to insoluble CPR-hel-1 (MW = 47.6 kDa) in the cell pellet is indicated with a red box.

CPR-hel-1 was found to express insolubly, with large quantities identified in the cell pellet (Fig. 4.1 lane 1). The cell lysate (lane 2) is too overloaded on the gel to identify soluble CPR-hel-1, though no appropriate MW bands were detected in eluted fractions (lanes 5 - 9) from IMAC purification. Due to a lack of soluble CPR-hel-1 expression, the target was assigned low priority and not pursued at this stage.

4.2.3. CPR-hel-2

Expression and IMAC purification

CPR-hel-2 initially expressed in such enormous quantities that precipitation was visible in the lysate, which subsequently could not be filtered. This made standard IMAC purification impractical due to column pressure issues. The expression protocol was therefore adapted to reduce expression levels of CPR-hel-2 to more manageable levels by inducing for only 4 h at 18 °C, rather than the standard 16 h (overnight) at 25 °C. Even with the reduced temperature and expression time, CPR-hel-2 overexpressed in very large quantities, yielding roughly 10 mg of soluble protein per litre of culture. The reduced quantities of CPR-hel-2 produced with the altered protocol enabled smoother IMAC purification, with the SDS-PAGE results shown in Fig. 4.2.

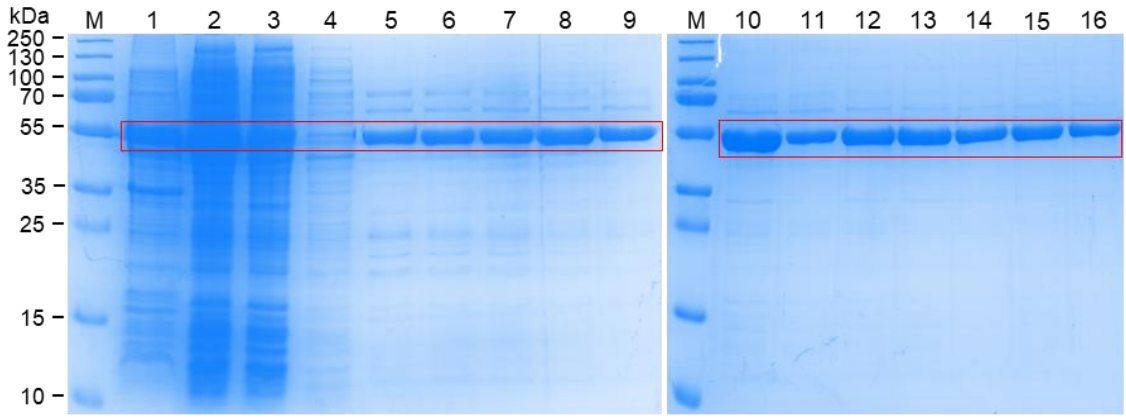

Fig. 4.2: SDS-PAGE analysis of CPR-hel-2 IMAC purification after induction at 18 °C for 4 h. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 16 = protein fractions. Bands corresponding to CPR-hel-2 are indicated with a red box; MW = 53.2 kDa.

A strong band at the appropriate MW for CPR-hel-2 was identified in the cell pellet, indicating significant insoluble expression (Fig. 4.2 lane 1). Despite this, ample soluble CPR-hel-2 was also present in the cell lysate and retrieved through IMAC purification with reasonable purity. Most of the protein precipitated within 1 h of purification, even with storage at 4 °C, revealing major instability and sensitivity to the purification conditions.

SEC analysis and characterisation

After centrifugation and removal of the precipitated protein, the remaining CPR-hel-2 was further purified by SEC to improve the homogeneity of the sample. The resulting chromatogram and SDS-PAGE analysis are shown in Fig. 4.3.

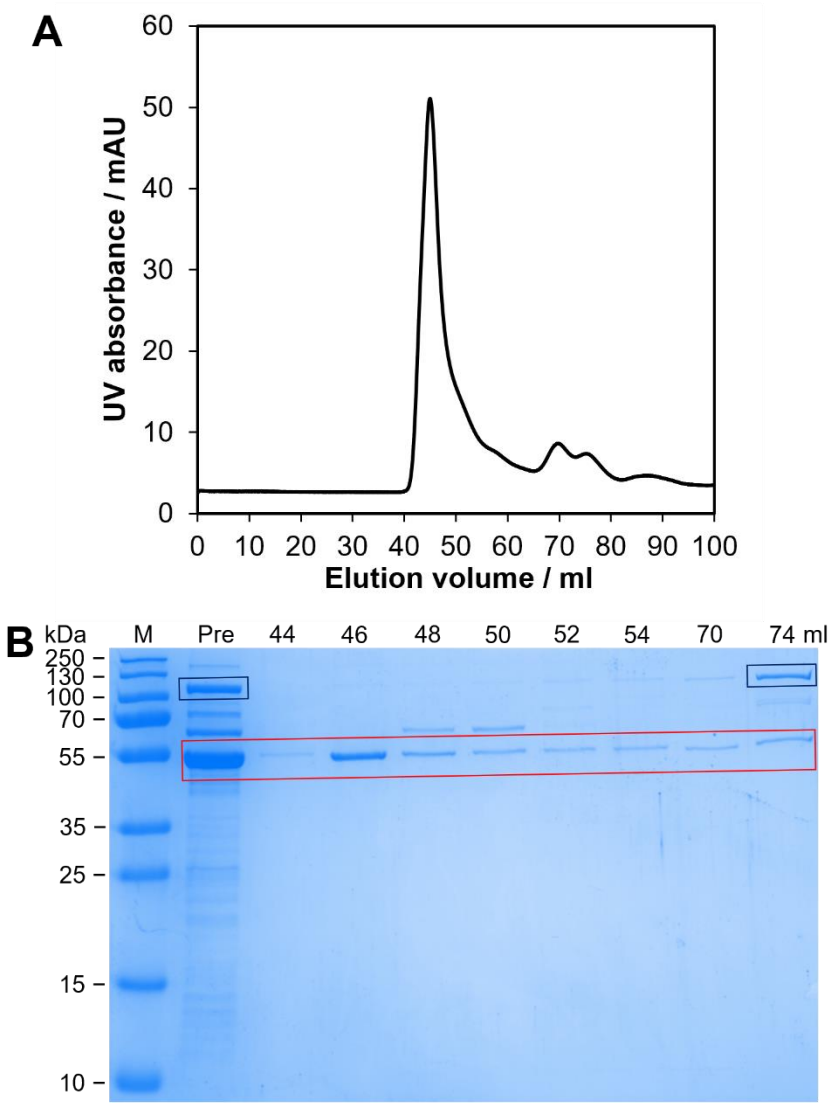

Fig. 4.3: (A) Chromatogram from analytical SEC of CPR-hel-2. **(B)** SDS-PAGE analysis of elution fractions from SEC. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above. CPR-hel-2 bands are indicated with a red box; MW = 53.2 kDa. The bands outlined in black are predicted to be the CPR-hel-2 dimer; MW = 106.4 kDa.

The minor peaks at 69.7 ml and 75.2 ml correspond to ~ 149.6 kDa and 94.3 kDa species, which could be the CPR-hel-2 trimer (159.6 kDa) and dimer (106.4 kDa), respectively. The major species present in the sample is aggregated protein, eluting at the void volume (V_o = 45.0 ml) of the column. This aggregated protein was confirmed by SDS-PAGE to be CPR-hel-2 (Fig. 4.3B), correlating with the heavy precipitation observed throughout purification. A high MW band was also identified during SDS-PAGE analysis (Fig. 4.3B) that was not observed during IMAC purification (Fig. 4.2). Based on the MW from SDS-PAGE, this likely corresponds to a CPR-hel-2 dimer.

Attempts to obtain a mass spectrum for CPR-hel-2 were unsuccessful due to its instability and tendency to precipitate. The protein continued to precipitate during and after SEC, leaving insufficient quantities for TSA. It is possible that CPR-hel-2 is incorrectly folded due to its notably

rapid overexpression, and the aggregated protein will likely be inactive. This challenging target was therefore put to one side to prioritise other CPR proteins.

4.2.4. CPR-exo-1

Expression and IMAC purification

CPR-exo-1 overexpressed and purified very well, yielding around 8 mg of protein per litre of culture despite significant losses in the flow through and column wash, as well as in the insoluble cell pellet (Fig. 4.4).

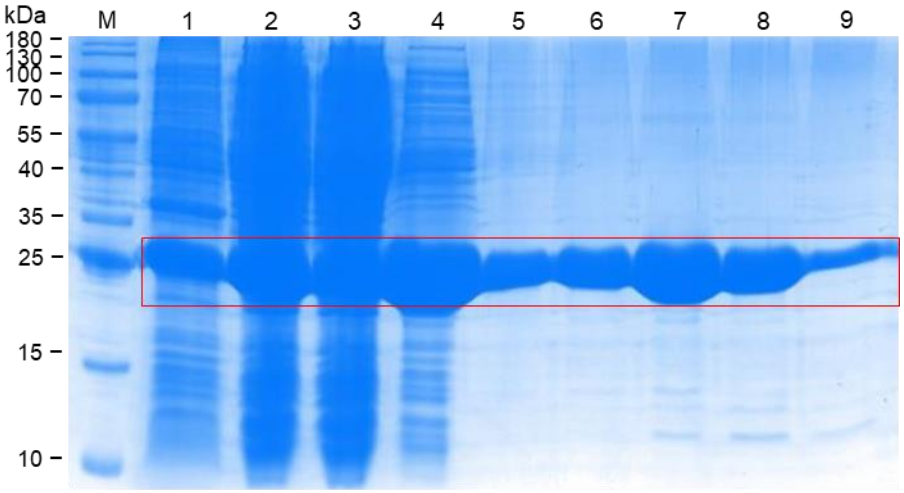

Fig. 4.4: SDS-PAGE analysis of CPR-exo-1 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 9 = protein fractions. Bands corresponding to CPR-exo-1 are indicated with a red box; MW = 23.4 kDa.

Despite these losses, the amount of soluble protein obtained from IMAC was more than required for initial characterisation. The purity of CPR-exo-1 after this single purification step was also very high and suitable for preliminary crystallisation trials; however, further analysis with SEC was required to assess the homogeneity of the sample and reveal any aggregation.

SEC analysis and characterisation

During SEC, CPR-exo-1 eluted as a single species (Fig. 4.5) with an elution volume (V_e) of 57.6 ml corresponding to a 49.2 kDa species, concluded to be dimeric CPR-exo-1 (expected MW = 46.8 kDa).

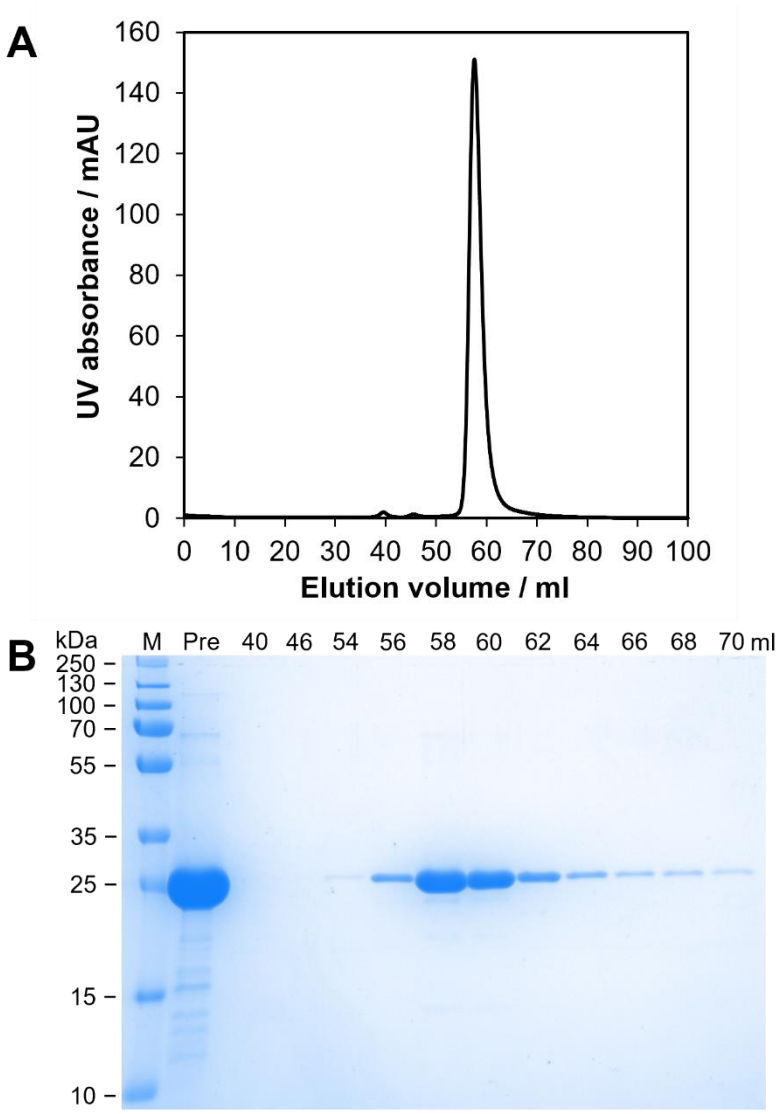

Fig. 4.5: (A) Chromatogram from SEC of CPR-exo-1. **(B)** SDS-PAGE analysis of elution fractions from SEC of CPR-exo-1; MW = 23.4 kDa. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The single symmetrical peak in the chromatogram (Fig. 4.5A) indicates a highly homogenous sample suitable for crystallisation trials. The identity of the species was confirmed to be CPR-exo-1 based on expected MW by SDS-PAGE, which also demonstrated the high purity of the target protein (Fig. 4.5B). ESI-TOF MS further verified that the correct protein had been purified (Appendix B). The accurate mass from MS also correlated with the loss of a methionine from the full-length sequence, confirming the occurrence of NME.

Thermal shift analysis

CPR-exo-1 purified in sufficient quantities to enable TSA with the Durham screens at 1.1 mg/ml. The reference melt temperature, T_m, in water was found to be ~ 45 °C. Whilst this indicates the protein will not be denatured at the temperatures encountered during production, it is low when considering that the source genome was isolated at 75 °C. When removed from the stabilising

mechanisms present *in vivo*, such as molecular chaperones and compatible solutes, proteins from thermophiles are not necessarily intrinsically thermostable[225]. TSA was therefore used to identify favourable conditions to increase the thermostability of CPR-exo-1.

pH screening

TSA with the Durham pH Screen®[211] indicates that the protein is most stable at middling pH values, with minor stabilisation relative to the reference T_m in water seen at weakly acidic pH. As can be seen in Fig. 4.6, CPR-exo-1 was very destabilised at extremes of pH, with a large drop in T_m. This was particularly pronounced in strongly alkaline conditions, which produced uninterpretable melt curves due to denaturation, and so were not included in comparisons.

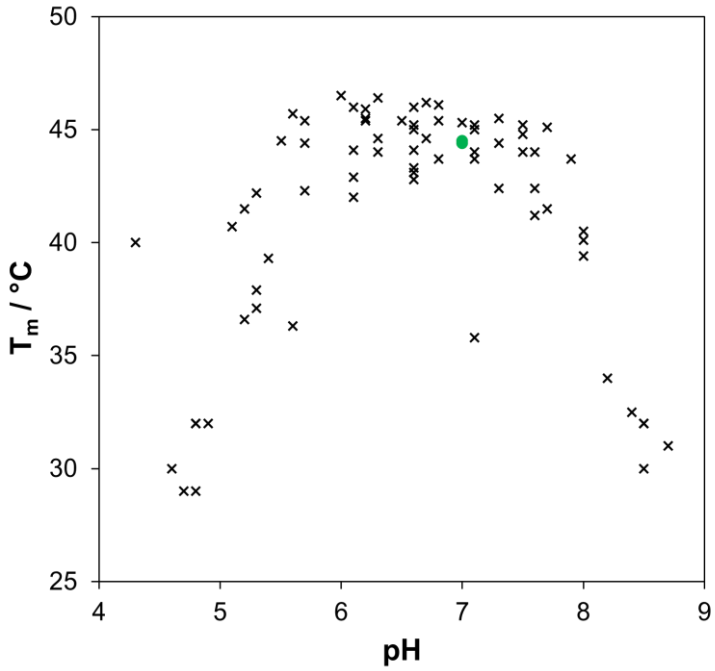

Fig. 4.6: Results from TSA using the Durham pH Screen® showing the effect of pH on the T_m of CPR-exo-1. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the same pH value are from different buffer systems at identical pH.

The predicted isoelectric point of the protein is 5.5, which could explain the rapid fall in T_m around this pH. CPR-exo-1 was slightly destabilised in the mildly alkaline buffers recommended during IMAC, with the exception of PIPES and HEPES buffers. The standard HEPES-based SGC buffer used during production was therefore retained during subsequent purification experiments.

The buffer system itself was also found to have an effect, with a range of T_m values observed at identical pH values. This is very pronounced at pH 7.1 (Fig. 4.6): four of the five buffers tested at this pH have T_m values close to the reference T_m, whilst 100 mM imidazole results in an almost 10 °C decrease in T_m to 35.8 °C. This result highlighted the need for quick removal of imidazole from the protein buffer following IMAC purification of CPR-exo-1.

Salt screening

TSA with the Durham Salt Screen® found that CPR-exo-1 was affected by both salt concentration and composition, with the results summarised in Fig. 4.7.

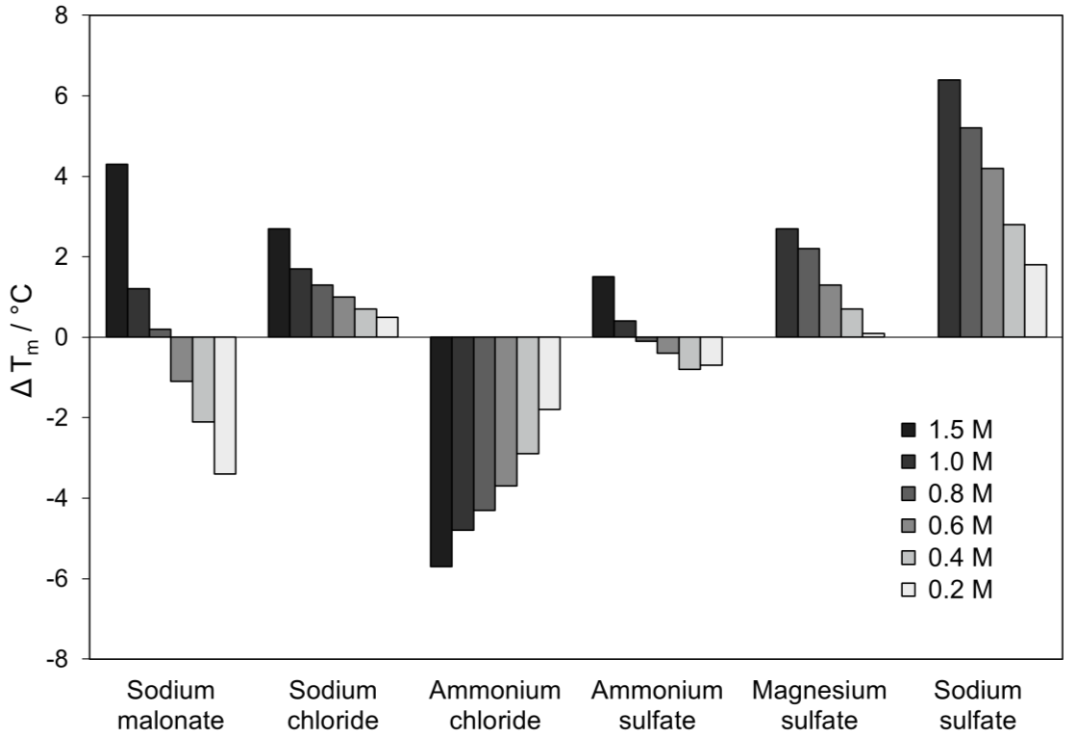

Fig. 4.7: Results from TSA of CPR-exo-1 with the Durham Salt Screen® showing changes in melt temperature, T_m, upon addition of increasing concentrations of various salts relative to the control in water.

The salt composition had a clear impact on CPR-exo-1 stability. Some salts caused an increase in T_m (sodium chloride, magnesium sulfate, and sodium sulfate) whilst others caused a decrease (ammonium chloride), with ΔT_m increasing in tandem with concentration. Intriguingly, sodium malonate and ammonium sulfate were found to be stabilising at high concentrations but destabilising at lower concentrations. Evaluating the results as a whole, sulfate and sodium ions appear to be generally stabilising, whilst ammonium and malonate ions are destabilising. This could explain the concentration-dependent switch in effect for sodium malonate and ammonium sulfate. Destabilising malonate or ammonium ions prevail at low concentrations, but are overcome by the stabilising effects of sodium or sulfate ions at high concentrations.

The Durham Salt Screen® also includes a variety of metal salts, which were of particular interest for CPR-exo-1 due to its homology to ribonuclease T (RNaseT) identified through BLAST[151] searches (Chapter 2). RNaseT is known to require a divalent metal cation to catalyse the removal of nucleotides from the 3' end of nucleic acid strands[226,227]. The preferred metal cation for RNaseT is Mg^{2+}, but activity is also seen with Mn^{2+} or Co^{2+}. It was proposed that if CPR-exo-1 similarly required a metal cation for activity, a stabilising effect may be seen in TSA and could give an indication of

preference. Most cations, including Co^{2+}, in fact led to a decrease in T_m, though Mg^{2+}, Mn^{2+}, Ca^{2+} and Sr^{2+} caused a slight increase, as shown in Fig. 4.8.

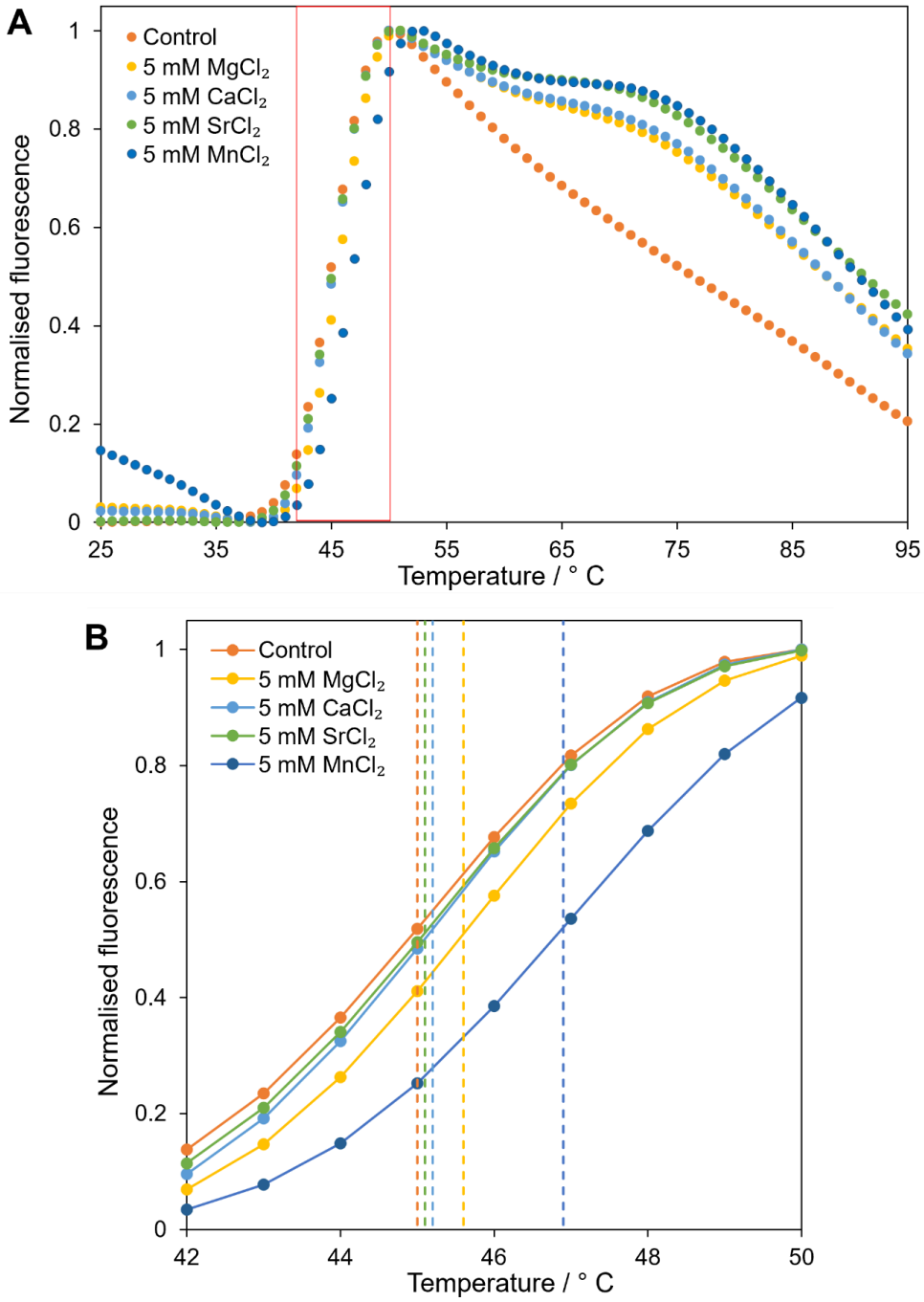

Fig. 4.8: (A) Melt curves showing the normalised fluorescence intensities at 590 nm of CPR-exo-1 in water (control) and with the addition of 5 mM divalent metal chloride salts from TSA with the Durham Salt Screen®. Raw experimental data are shown as individual points. **(B)** Closer view of the region outlined in red in (A) showing the inflection points used to calculate T_m values; dashed vertical lines indicate the calculated T_m values in water (control) and with the addition of 5 mM divalent metal chloride salts, following the same colour scheme as the melt curves.

The stabilising metal cations had unusual melt curve profiles, with a secondary peak in fluorescence output at around 70 °C (Fig. 4.8A). This could be due to the metal ions inducing additional stabilisation in a region of the protein, which then unfolds at a higher temperature. Considering the greatest stabilisation was seen with Mn^{2+}, it could be that this is the preferred metal for CPR-exo-1. However, it is worth noting that the increase in T_m with these metal ions is small, and the addition of chelators such as EDTA and EGTA did not alter the T_m. It is therefore unlikely that CPR-exo-1 contains structurally integral metal ions. Due to the amount of protein required to conduct TSA, these assays were only performed once. The results can therefore only be used as a qualitive guide rather than to draw specific conclusions about the activity of the protein, but do provide useful insights that will be beneficial when activity analysis is conducted. Detailed functional characterisation is required to determine the role, if any, of metal ions in CPR-exo-1.

Crystallisation Trials

Production of CPR-exo-1 had proved to be successful, yielding large quantities of pure, soluble protein, verified to be CPR-exo-1 by ESI-TOF MS and characterised by SDS-PAGE, SEC, and TSA. The protein was found to be stable and could be concentrated to > 5 mg/ml without precipitation. CPR-exo-1 could therefore be carried forward to crystallisation screening for structure determination with a wide range of commercially available 96-condition screens. Primary crystallisation trials were conducted with protein at 5.0 mg/ml of purity shown in Fig. 4.9.

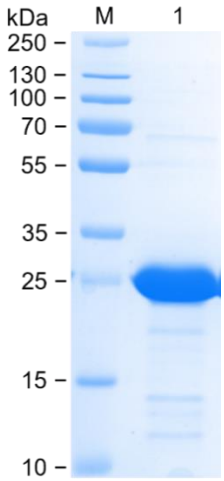

Fig. 4.9: SDS-PAGE analysis showing the purity of CPR-exo-1 used for initial crystallisation screening; MW = 23.4 kDa. M = marker, with MWs of bands on the left; 1 = CPR-exo-1 in SGC buffer for crystallisation.

After one month, the majority of crystallisation experiments remained clear. However, a promising condition yielding possible crystalline material was found in the Index HT™ Screen (Hampton), as shown in Fig. 4.10.

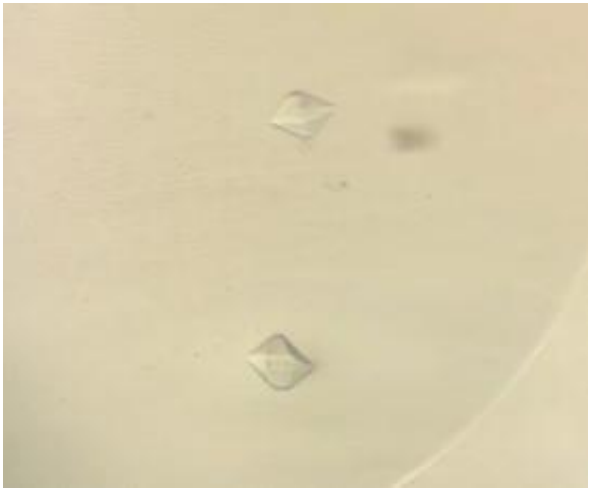

Fig. 4.10: CPR-exo-1 crystals grown in 0.2 M ammonium sulfate, 0.1 M HEPES pH 7.5, 25% w/v PEG 3350 viewed after 4 weeks. Crystals are approximately 20 μm at their widest dimension.

These potential crystals were too small to harvest for diffraction experiments, and were found to have redissolved a few weeks later, possibly due to temperature fluctuations in the environment. Attempts to reproduce and optimise the crystals by varying the pH, salt, and precipitant concentrations of the original promising condition were unsuccessful. Since most drops from original crystallisation trials remained clear, additional screening experiments with increased protein concentration (9.7 mg/ml) were conducted but did not yield crystals for structure determination. Thus far, no diffracting crystals to enable structure determination for CPR-exo-1 have been obtained.

4.2.5. CPR-endo-1 and CPR-endo-2

Expression and IMAC purification

Soluble heterologous expression was not detected during scaled-up expression of CPR-endo-1. In contrast, CPR-endo-2 was found to overexpress well, though most was insoluble and trapped in the cell pellet. The SDS-PAGE results from IMAC purification to retrieve soluble CPR-endo-2 are shown in Fig. 4.11.

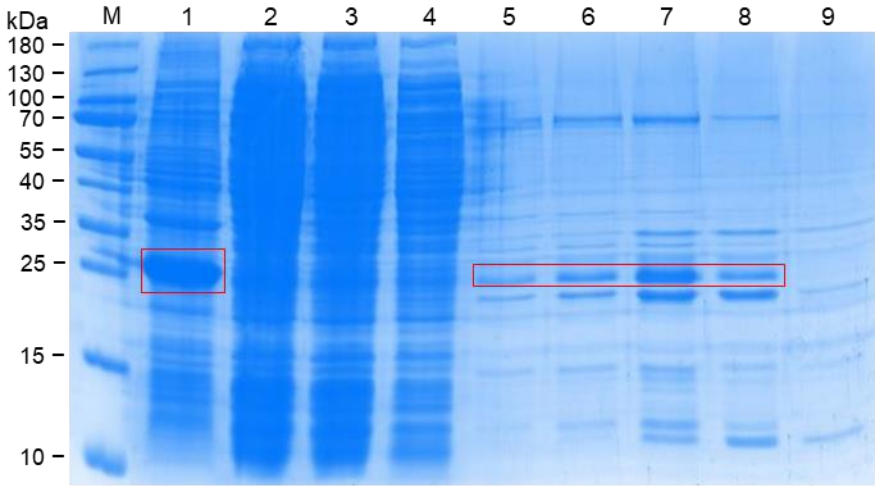

Fig. 4.11: SDS-PAGE analysis of CPR-endo-2 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 9 = protein fractions. Bands corresponding to CPR-endo-2 (MW = 27.1 kDa) are indicated with a red box.

The strong band in lane 1 of Fig. 4.11 indicates that the majority of CPR-endo-2 was in the cell pellet, likely as insoluble inclusion bodies. Whilst SDS-PAGE also indicated that some soluble protein of an appropriate MW for CPR-endo-2 was recovered through IMAC, it was of low purity and precipitated rapidly, preventing characterisation or further purification. Since neither endonuclease readily expressed in a stable, soluble form, these targets were deprioritised.

4.2.6. CPR-B1 and CPR-B2

Expression and IMAC purification

All three category B targets were found to overexpress in small-scale tests and so were carried forward for full-scale expression and purification. Unfortunately, large quantities of CPR-B1 and CPR-B2 were identified in the insoluble cell pellet, as shown in Fig. 4.12.

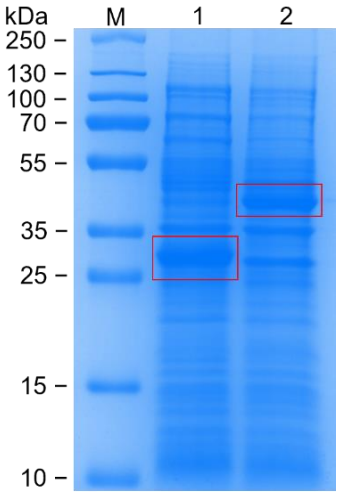

Fig. 4.12: SDS-PAGE analysis of the CPR-B1 and CPR-B2 insoluble cell pellet fractions from full-scale expression. M = marker, with MWs of bands on the left; 1 = CPR-B1 pellet; 2 = CPR-B2 pellet. Bands corresponding to the target proteins are indicated with red boxes; CPR-B1 = 30.6 kDa; CPR-B2 = 42.3 kDa.

The cell lysate was purified by IMAC to recover any soluble CPR-B1 and CPR-B2. Little, if any, of the desired proteins were identified by SDS-PAGE analysis of the resulting fractions, the results of which are shown in Fig. 4.13.

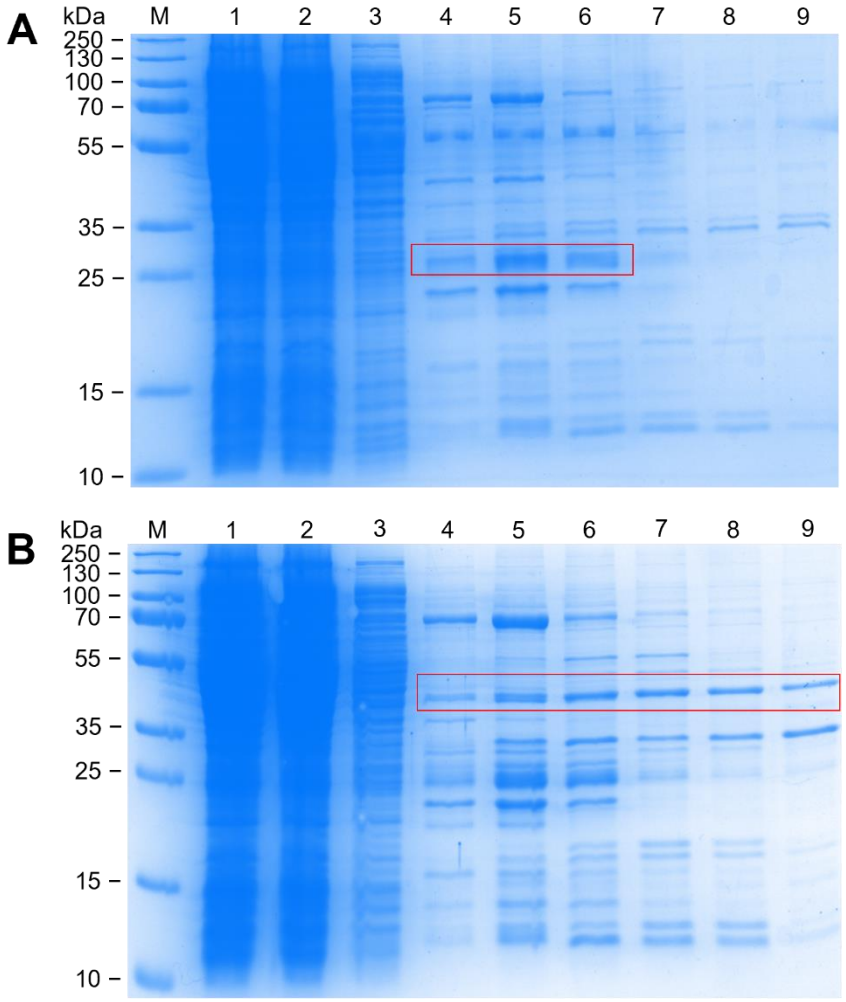

Fig. 4.13: (A) SDS-PAGE analysis of CPR-B1 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell lysate; 2 = flow through; 3 = wash; 4 - 9 = protein fractions. Bands corresponding to CPR-B1 (MW = 30.6 kDa) are indicated with a red box. **(B)** SDS-PAGE analysis of CPR-B2 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell lysate; 2 = flow through; 3 = wash; 4 - 9 = protein fractions. Bands corresponding to CPR-B2 (MW = 42.3 kDa) are indicated with a red box.

Bands at the expected MWs of the target proteins were identified and are highlighted in Fig. 4.13. However, the quantities and purity of these proteins is very low, which made further purification or characterisation impractical. Since the vast majority of overexpressed CPR-B1 and CPR-B2 was found to be insoluble, and insufficient soluble protein could be retrieved, focus was shifted to other targets at this point.

4.2.7. CPR-B3

Expression and IMAC purification

Despite high levels of expression and minimal basal expression in small-scale tests, no expression of CPR-B3 was identified through SDS-PAGE analysis after scaled-up expression (Fig. 4.14).

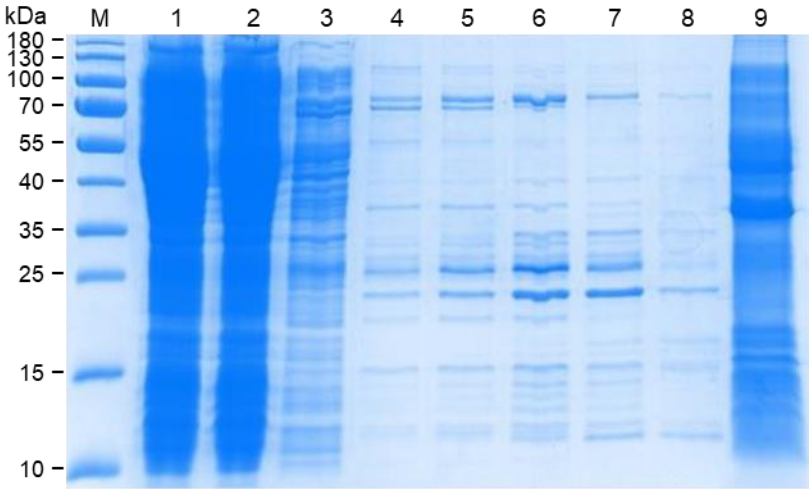

Fig. 4.14: SDS-PAGE analysis of CPR-B3 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell lysate; 2 = flow through; 3 = wash; 4 - 8 = protein fractions; 9 = cell pellet. Expected MW of CPR-B3 = 16.3 kDa.

In contrast to the strong bands observed in the initial expression tests (Fig. 3.4C), distinct bands at an appropriate MW for CPR-B3 were not identified in either the cell pellet or lysate fractions, and no soluble target protein was detected in fractions from IMAC. Other CPR proteins exhibiting similar issues had been discarded at this stage; however, CPR-B3 was a higher priority target, on the border between categories B and C with potentially novel or unique functionalities and structure (2.2.3.). Additional measures were therefore taken to improve soluble protein production for this high value target.

Optimised expression trials for CPR-B3

As well as the change in culture volume, the scale-up from 10 ml tests to full 1 litre expression cultures introduced additional variations in growth conditions and experimental set-up. During the initial tests, cultures were left for only 4 h post-induction to gauge if desired overexpression was inducible and assess basal expression. In contrast, the 1 litre cultures were left for ~ 16 h overnight to maximise protein production for downstream use. Growth conditions, such as the amount of aeration, are also affected by culture volume and can influence expression levels and solubility during recombinant expression[199]. The decision was therefore made to return to the smaller culture volumes with which expression had been detected. Samples were then taken at various time points at or after induction to determine the effects of growth duration on CPR-B3 overexpression. Fig. 4.15 shows the results of whole-cell SDS-PAGE on samples taken at the point of induction (0 h), as well as after

4 h and 16 h at 30 °C or 37 °C, both with and without addition of L-rhamnose. SDS-PAGE samples were prepared with equivalent optical densities (OD_{600}) to allow better comparison of band strengths as a qualitative indication of expression levels.

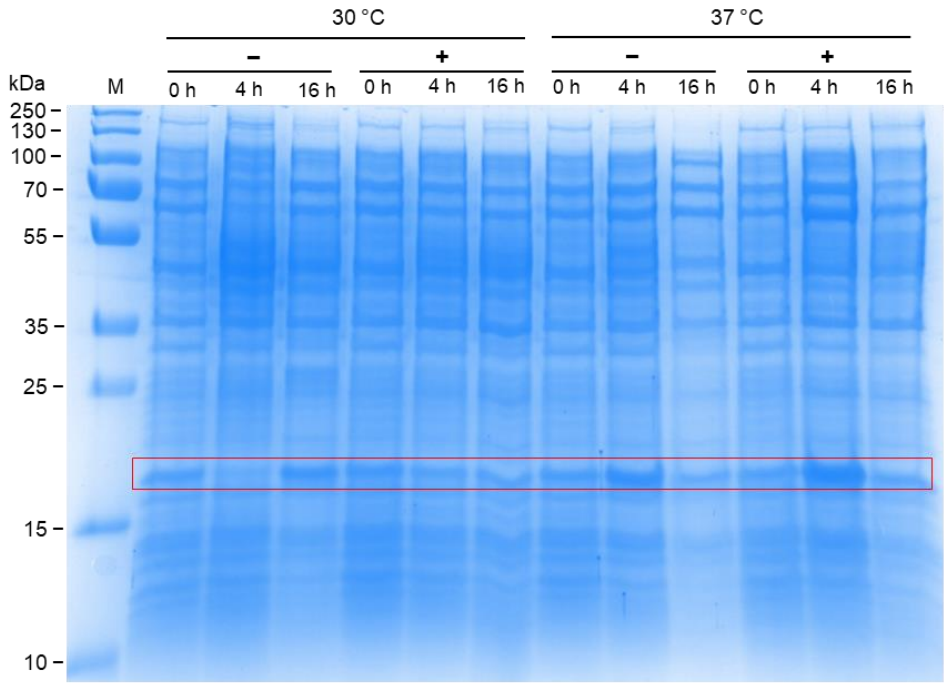

Fig. 4.15: Whole-cell SDS-PAGE results from small-scale expression tests of CPR-B3, showing samples taken at induction (0 h), 4 h post induction, and 16 h post induction, with (+) and without (−) addition of L-rhamnose. M = marker, with MWs in kDa to the left. Positions of bands corresponding to CPR-B3 expression (MW = 16.3 kDa) are indicated with a red box.

The amounts of CPR-B3 present in the cultures appears to fluctuate over time. The most intense bands, corresponding to the strongest expression levels, are seen after 4 h incubation at 37 °C, with the presence of L-rhamnose having a relatively minor effect on expression levels. Interestingly, the amount of protein present after 16 h is vastly reduced, which suggests that CPR-B3 could be unstable and is degrading in the host cells.

To assess the solubility of CPR-B3, the expression test samples were lysed, and SDS-PAGE was conducted on the resulting cell pellet and lysate (Fig. 4.16).

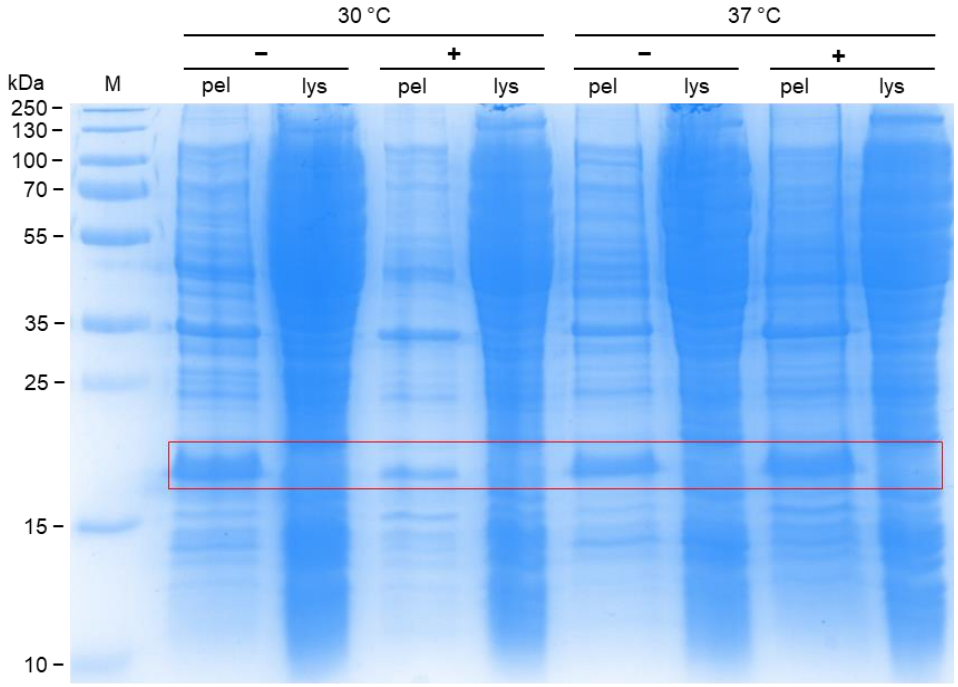

Fig. 4.16: Whole-cell SDS-PAGE results from solubility tests of CPR-B3 after small-scale expression. Noninduced (−) and induced samples with 0.2% L-rhamnose (+). M = marker, with MWs in kDa to the left; pel = cell pellet; lys = cell lysate. Positions of bands corresponding to CPR-B3 (MW = 16.3 kDa) are indicated with an arrow.

Regardless of the level of overexpression, CPR-B3 was found to only be present in the cell pellet and not the desired lysate. The protein was therefore inaccessible without demanding refolding efforts, trapped as insoluble inclusion bodies in the cell pellet.

Optimising soluble expression of CPR-B3 using SHuffle® cells

In a bid to resolve the issue of insolubility, the pJOE5751.1-*CPRB3* expression construct was transformed into an alternative *E. coli* host strain, SHuffle® (New England Biolabs), designed to aid solubilisation of recombinant proteins. The main features of this strain are enhancements to promote correct folding of recombinant proteins with disulfide bonds[170]. This is less relevant for bacterial targets, in which disulfide bonds as far less common than in eukaryotic proteins[228]. Nevertheless, SHuffle® cells also constitutively express a chaperone protein, DsbC, that can assist folding of proteins without disulfide bonds[229], and so could potentially stimulate expression of correctly folded, soluble CPR-B3. Standard expression tests (11.8.1.) were repeated in this new expression strain, and the results of whole-cell SDS-PAGE are shown in Fig. 4.17.

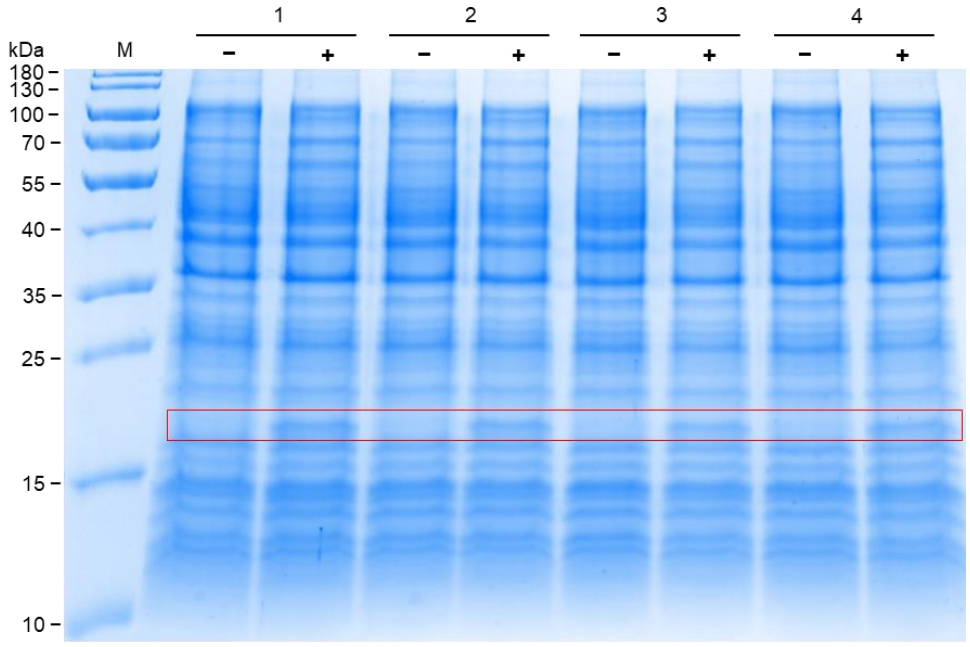

Fig. 4.17: Whole-cell SDS-PAGE results from small-scale expression tests of CPR-B3 in SHuffle® cells. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to CPR-B3 (MW = 16.3 kDa) expression are indicated with a red box.

As in the original expression tests (Fig. 3.4C), basal expression appears to be absent, with CPR-B3 only detected in the samples induced with L-rhamnose. However, much fainter bands, corresponding to relatively weak expression, were detected in the SHuffle® cell expression tests than in previous tests. Despite the enhanced folding capacity of the SHuffle® cells, the CPR-B3 was again found to be insoluble. At this point, the decision was made to pause work on CPR-B3 and prioritise better expressing CPR targets.

4.2.8. Overall results of CPR target production and characterisation

Full-scale expression and purification experiments were also conducted for the remaining CPR targets, the details of which are discussed in detail in upcoming chapters: CPR-DprA (Chapter 5), CPR-C4 (Chapter 6), and the four chaperone system proteins (CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB; Chapter 7). These targets were successfully overexpressed and purified, yielding large quantities of pure, soluble protein that enabled more in-depth characterisation, TSA, and crystallisation experiments.

The overall results of upscaled expression of the CPR targets, as well as purification and biophysical characterisation where possible, are outlined in Table 4.2.

Table 4.2: Summary of large-scale expression, purification, and characterisation of the CPR target proteins. No detectable expression is represented as − and high levels of expression as + + +. MS denotes ESI-TOF mass spectrometry; MS* denotes trypsin digest mass spectrometry.

Target	Insoluble expression	Soluble expression	Successful purification	Characterisation
CPR-hel-1	+ + +	−	−	−
CPR-hel-2	+ + +	+ +	✓	SEC
CPR-exo-1	+ + +	+ + +	✓	SEC, MS, TSA
CPR-endo-1	+	−	−	−
CPR-endo-2	+ +	+	−	−
CPR-DprA	+	+ + +	✓	SEC, MS, TSA
CPR-GrpE	+	+ +	✓	SEC, MS, MS*, TSA
CPR-DnaJ	++	+ + +	✓	SEC, MS*, TSA
CPR-DnaK	+ +	+ + +	✓	SEC, MS*, TSA
CPR-ClpB	+ +	+ + + +	✓	SEC, MS*, TSA
CPR-B1	+ +	−	−	−
CPR-B2	+ + +	+	−	−
CPR-B3	+	−	−	−
CPR-C4	+ + +	+	✓	SEC, MS, MS*, TSA

All of the CPR targets that successfully expressed in a small-scale format were also found to express on the larger, litre culture scale. However, undesirable insoluble expression was observed in each case, with greater levels of insoluble than soluble expressed observed for eight of the 14 targets (Table 4.2). Soluble expression was detected for ten of the targets through SDS-PAGE, with eight effectively purified in sufficient quantities for downstream analysis. With the exception of highly unstable CPR-hel-2, the identities of these proteins were successfully verified by SDS-PAGE and MS techniques. Sample purity and homogeneity was also assessed by SDS-PAGE and SEC, confirming their suitability for crystallisation experiments.

4.3. Conclusions

In total, 44% (8 / 18) of the CPR targets that underwent small-scale production trials were produced in suitable quantities in a soluble form for downstream characterisation, which is approximately in line with the 42% (201 / 478) success rate seen across the Virus-X project as a whole[18]. The use of a standard expression and purification protocol for all targets worked well, facilitating streamlined production and rapid assessment of their suitability for further analysis. The combination of biophysical and analytical techniques also gave confidence that the correct proteins had been isolated prior to investing time and resources on crystallisation.

Though the small-scale test expressions (3.2.2.) were a useful predictive tool for full-scale growth, the expression levels observed in the initial tests were not always reflected in larger cultures (Table 4.2). The expression levels and solubility of recombinant proteins can be influenced by parameters that do not scale with culture volume, such as the amount of aeration[199]. Consequently, proteins expressed insolubly during tests can be expressed solubly on a larger scale, and *vice versa*. In the case of the CPR targets, the scale-up revealed a case where expression observed in small-scale tests was not directly reproduced in large-scale tests (CPR-B3). Similarly, false negatives can occur, where a protein is found to express in large culture volumes despite a lack of expression in small-scale tests. Had more time been available, this supposition could have been tested through full-scale expression of the targets discarded following trials.

The rejection of challenging targets with little effort to improve production may appear uneconomical. In effect, however, it enabled a larger number of targets to be assessed and greater focus to be placed on those that expressed and purified most successfully, with little or no solubility and precipitation issues. These were the targets most likely to be correctly folded and active, so providing the best opportunities for crystallisation and structure determination.

Chapter 5. Structure determination and analysis of a DprA protein from the Candidate Phyla Radiation

5.1. Introduction

5.1.1. Nucleic acid processing enzymes

Amongst the most highly sought-after enzymes for biotechnological exploitation are those performing DNA and RNA processing functions, such as polymerases, recombinases, nucleases, and nucleic acid binding proteins[153,230–232]. The range of applications necessitating these enzymes is continually expanding; however, the diversity within existing libraries is insufficient for engineering any desired characteristic into a particular enzyme. Expanding the catalogue of such proteins through the discovery of novel enzymes will provide a larger pool from which to select or rationally engineer enzymes with desired traits, such as thermostability, fidelity, specificity, or the ability to act on modified DNA/RNA substrates[230,233–235]. There is a clear correlation between the identification of these novel DNA handling enzymes, most notably polymerases, and the major progressions that underpin much of modern molecular biology[230]. Accordingly, it is expected that the discovery and characterisation of new nucleic acid processing enzymes will further stimulate advancements in molecular biology and biotechnology [10,18,236–238], making these proteins a key focus of the Virus-X project.

5.1.2. DNA processing protein, DprA

DNA processing protein A (DprA) is required for natural transformation[239,240,241]: the process by which bacteria can exchange genetic information between species through homologous recombination, and a crucial driver of genetic diversity[242–244]. DprA is capable of non-specific binding to DNA, with a preference for single stranded DNA (ssDNA)[241,245]. It is proposed to protect incoming DNA during transformation, and has been demonstrated to interact with the RecA recombinase, facilitating loading of both naked and SSB-bound ssDNA[241,246,247]. DprA proteins share a highly conserved domain (Pfam02481) that is widespread throughout diverse bacterial genera, including both transformable and non-transformable species, suggesting important functions beyond facilitating transformation[239,241,248].

One of the CPR targets selected for production during the Virus-X project is CPR-DprA, a predicted DprA homolog. Its suspected ability to bind and stabilise DNA, as well as interact with recombinases and SSBs, made it an appealing target for characterisation, in the hope that novel and valuable functionalities can be identified.

5.2. Results and discussion

5.2.1. Sequence analysis and annotation of CPR-DprA

The CPR-DprA gene was identified on the bacterial contig discussed in 2.2.1., from a CPR bacterial genome identified in an Icelandic hot spring at 75 °C and pH 5.

The CPR-DprA sequence was analysed using the Basic Local Alignment Search Tool (BLAST)[151], returning over 100 'hit' sequences with > 45% sequence identity. These were predominantly also from candidate bacterial phyla or unclassified bacteria, indicating that the CPR-DprA sequence is more conserved within the CPR than in bacteria from cultivated phyla. The most significant BLAST hit (69.3% sequence identity with 95% query cover; E value 8e-127) was annotated as a DNA protecting protein, DprA, from a bacterium of the uncultured *Candidatus Adlerbacteria* phylum (NCBI accession: txid1797243). The overwhelming majority of the 'hit' sequences were similarly described as 'DNA protecting/processing protein, DprA', leading to the classification of CPR-DprA as a category A target, prioritised for structural determination due to potentially valuable function.

5.2.2. Domain architecture of CPR-DprA

BLAST[151] and InterPro[249] analysis of the CPR-DprA sequence determined that the target belongs to the DprA protein family (IPR003488) and is made up of two distinct domains. The first is an N-terminal domain matching that of Pfam02481, known as DNA_processg_A, which typifies the DprA protein family. The second is a winged helix-like DNA binding domain at the C-terminus. Fig. 5.1 compares the domain assembly of CPR-DprA with homologous DprA proteins from *Streptococcus pneumonia*, *Rhodopseudomonas palustris*, and *Helicobacter pylori*.

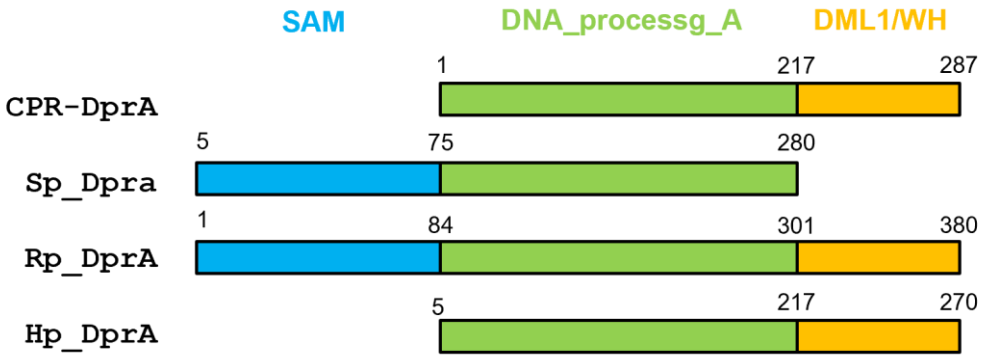

Fig. 5.1: Domain architecture of CPR-DprA and homologous DprA proteins from *S. pneumonia* (*Sp*DprA), *R. palustris* (*Rp*DprA), and *H. pylori* (*Hp*DprA)[250]. N-terminal SAM domain present in *Sp*DprA and *Rp*DprA is shown in blue; conserved DNA_processg_A domain shown in green; C-terminal DML1-like domain (DML1) in *Rp*DprA and *Hp*DprA and winged helix-liked DNA binding domain (WH) in CPR-DprA are shown in yellow.

Proteins within the DprA family have been observed with various domain configurations[250]. The four homologs shown in Fig. 5.1 all contain the core DprA domain (DNA_processg_A), though have varying combinations of additional N- or C-terminal extradomains. CPR-DprA, along with *Hp*DprA, is noted to lack the N-terminal five helix SAM-like domain found in *Sp*DprA and *Rp*DprA, which is thought to participate in protein-protein interactions[251]. In addition, *Rp*DprA and *Hp*DprA have a C-terminal domain with similarities to the DML1 domain, proposed to be involved in Z-DNA binding[252]. CPR-DprA is predicted to have a C-terminal winged helix-like domain, which is also typically responsible for DNA binding[245]. Unlike the other three DprA proteins, *Sp*DprA has no C-terminal extradomain. Previous studies of DprA proteins from across the bacterial kingdom showed that a three-domain protein, as observed for *Rp*DprA, is most typical, but that two domain proteins are also widely occurring[250]. The fact that these extradomains are not present in all DprA proteins suggests that they are not crucial for general protein function, but they could produce highly diversified and species-specific additional functions.

Based on domain architecture, CPR-DprA most closely resembles *Hp*DprA, which similarly lacks the N-terminal SAM-like domain and has a C-terminal extradomain (CTD) with predicted DNA binding capabilities. However, sequence alignments with ClustalW[253] (Fig. 5.2) found that CPR-DprA shares only 31.1% sequence similarity with *Hp*DprA, compared to 40.7% and 33.3% for *Sp*DprA and *Rp*DprA, respectively. The higher similarity between CPR-DprA and *Sp*DprA is surprising since these have the most dissimilar domain architectures, only sharing the well conserved core DprA domain.

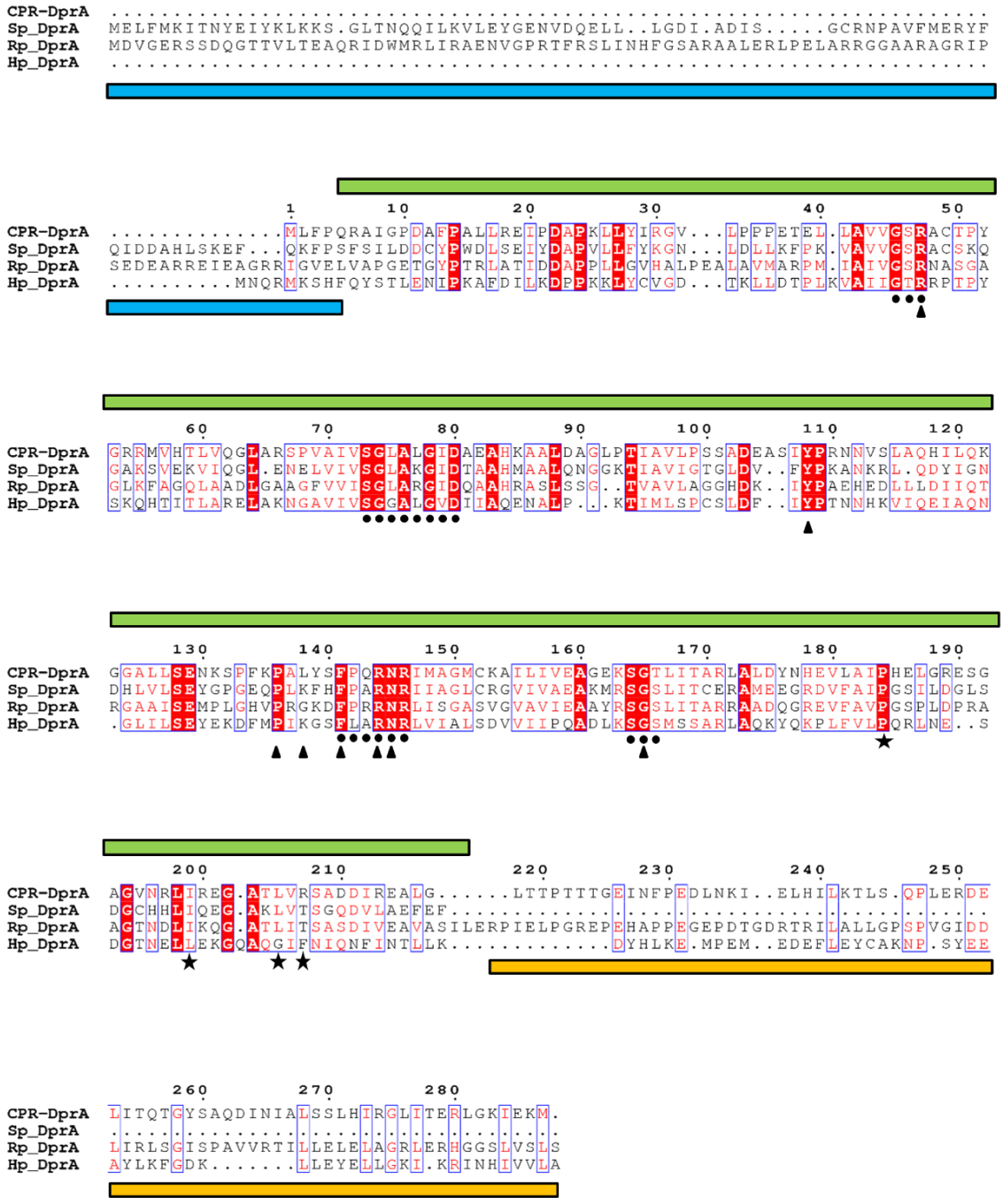

Fig. 5.2: Sequence alignment of CPR-DprA with homologous DprA sequences from *S. pneumonia* (*Sp*DprA), *R. palustris* (*Rp*DprA), and *H. pylori* (*Hp*DprA)[250]. Domains are coloured as in Fig. 5.1. Key residues known to form interactions in the DprA homologs are indicated with black symbols: residues involved in dimer formation are shown as stars, conserved residues near the binding pocket are shown as dots, and residues directly contacted by ssDNA as triangles[250].

The core DprA domain is clearly well conserved between the four proteins. Key residues identified in the DprA homologs are also present in CPR-DprA, including several stretches of residues involved in DNA binding and three residues known to form a hydrophobic dimerisation interface (Pro184, Ile199, and Leu205), implying that these residues are similarly important in CPR-DprA.

The additional N- and C-terminal domains are considerably more diverse among the DprA homologs. This again implies that these extradomains are responsible for differences in the specific mechanisms of the various DprA proteins. Structural analysis of the CPR-DprA CTD is required to draw comparisons to the DML1-like domains of *Rp*DprA and *Hp*DprA that have been previously determined by X-ray crystallography and NMR studies, respectively[245,250].

5.2.3. Production and characterisation of CPR-DprA

Expression and IMAC purification

Following on from effective small-scale tests, CPR-DprA was expressed on a larger scale for characterisation and structural analysis. Large-scale overexpression experiments with CPR-DprA were very successful, producing 15 mg of soluble protein per litre of culture. SDS-PAGE analysis following IMAC purification is shown in Fig. 5.3 below.

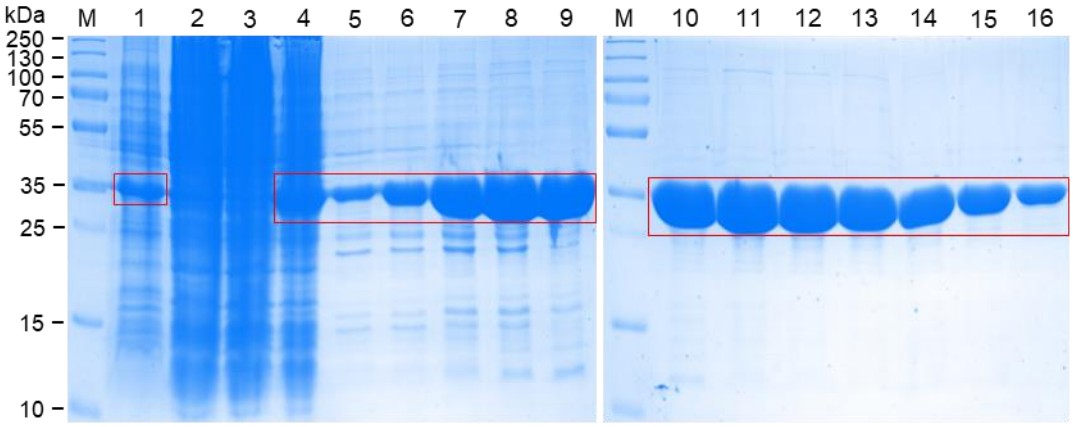

Fig. 5.3: SDS-PAGE analysis of CPR-DprA IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 16 = protein fractions. Bands corresponding to CPR-DprA (MW = 32.3 kDa) are indicated with a red box.

A small band corresponding to insoluble protein in the cell pellet can be seen in lane 1 of Fig. 5.3, though the overwhelming majority of CPR-DprA was found to be soluble. Whilst some minor contaminants remain after IMAC, the purity of CPR-DprA was sufficient for biophysical characterisation.

Whilst a significant amount of target protein was lost during column washing (Fig. 5.3, lane 4), ample protein was collected during IMAC. It was therefore not considered necessary to lower the imidazole concentration in the washing buffer, as the enhanced sample purity offset any loss in quantity.

SEC analysis

SEC was conducted with CPR-DprA to assess its oligomeric state. It also served as an additional polishing purification step prior to crystallisation. The resulting chromatogram and SDS-PAGE analysis of eluted fractions are shown in Fig. 5.4.

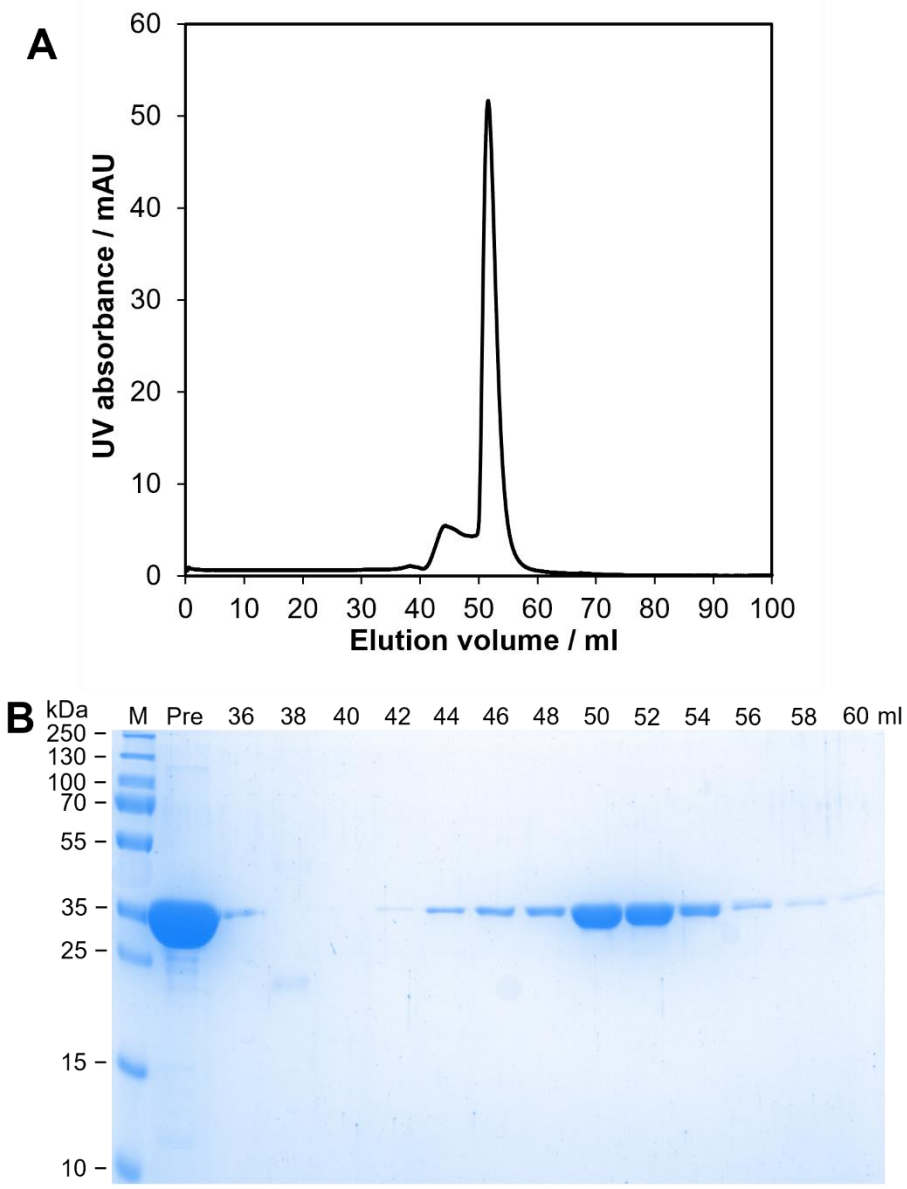

Fig. 5.4: (A) Chromatogram from SEC of CPR-DprA. **(B)** SDS-PAGE analysis of elution fractions from SEC of CPR-DprA. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

CPR-DprA eluted as a single species with a sharp, symmetrical peak in the chromatogram (Fig. 5.4A), indicative of a homogenous and pure sample. This was substantiated through SDS-PAGE (Fig. 5.4B). The elution volume (V_e) for CPR-DprA was 51.6 ml, corresponding to a species with a MW of ~ 71.9 kDa. This is roughly twice the monomeric MW of CPR-DprA (32.3 kDa), indicating that the protein is dimeric. This correlates with DprA homologs, which are known to form functional dimers[250,254]. The minor peak at 44.1 ml is approximately at the void volume (V_o) of the column (44.6 ml) and likely corresponds to aggregated protein.

Thermal shift analysis

TSA was conducted on CPR-DprA to assess its intrinsic thermotolerance and identify stabilising conditions for production. TSA using the Durham Screens®[210,255] returned a melt temperature (T_m) for CPR-DprA of 52.6 °C.

pH screening

The results of TSA with the Durham pH Screen® revealed that CPR-DprA tended to favour weakly acidic conditions, as can be seen in Fig. 5.5.

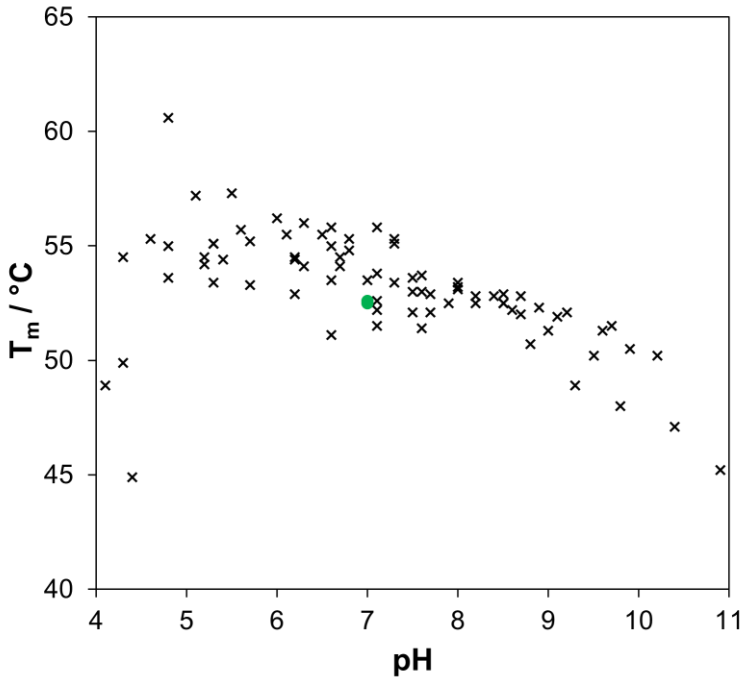

Fig. 5.5: Results from TSA using the Durham pH Screen®[210,255] showing the effect of pH on the T_m of CPR-DprA. The two reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the same pH value are from different buffer systems at identical pH.

As the pH of screened conditions moved from acidic to alkaline, there was a general decrease in T_m, corresponding to destabilisation of CPR-DprA. The most promising buffer identified was citric acid, which resulted in a consistent stabilisation of ≥ 3 °C across the three pH values tested relative to the control in water. Citric acid would therefore be a suitable choice for storage or assay buffers in future.

Salt screening

TSA with the Durham Salt Screen® revealed a general stabilising effect of a range of common salts on CPR-DprA. For each of the salts shown in Fig. 5.6, the extent of CPR-DprA stabilisation tended to increase with salt concentration. This is particularly pronounced with ammonium sulfate, which caused a 3 °C increase in T_m at 0.2 M, rising to ~ 14 °C (T_m = 66.9 °C) at 1.5 M.

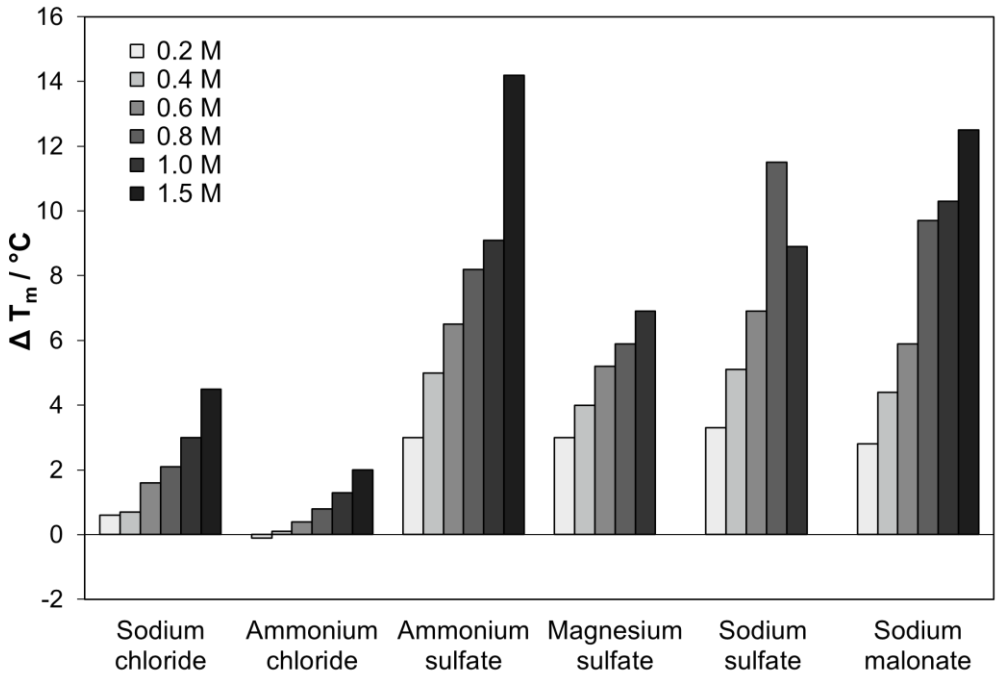

Fig. 5.6: Results from TSA of CPR-DprA with the Durham Salt Screen®[210,255] showing changes in T_m upon addition of salts at increasing concentrations. Magnesium sulfate and sodium sulfate were only tested up to 1.0 M.

The T_m of CPR-DprA was also found to significantly increase with addition of a variety of divalent metal ions. The greatest stabilisation was seen with $NiCl_2$, which caused a distinct shift in the melt curve of CPR-DprA relative to the control, as shown in Fig. 5.7.

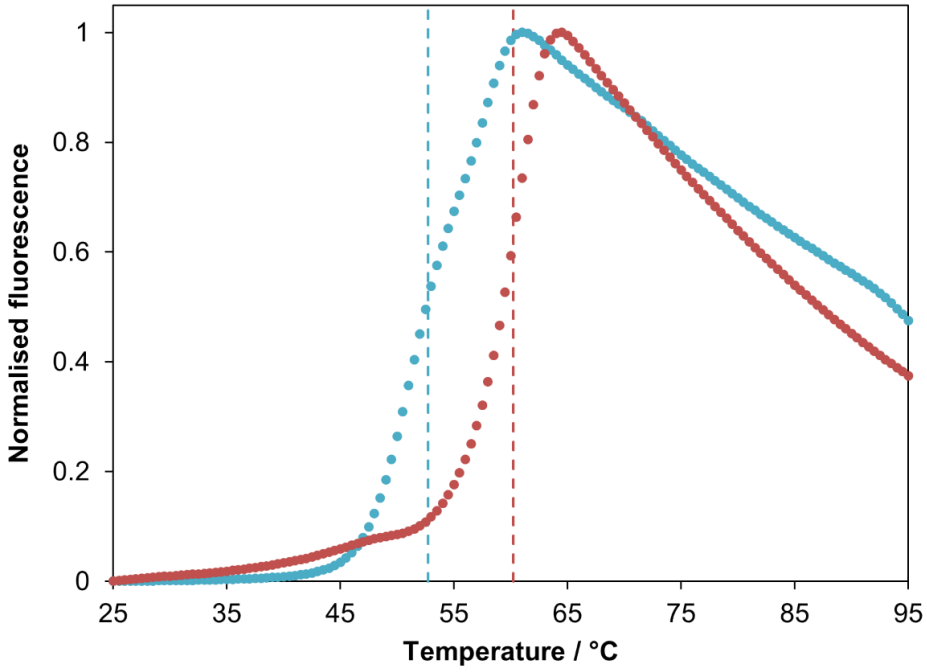

Fig. 5.7: Melt curves showing normalised fluorescence intensities at 590 nm from TSA of CPR-DprA in MilliQ water (blue) and with addition of 0.1 mM $NiCl_2$ (red). Experimental data are shown as individual points. Dashed vertical lines indicate T_m values[255].

The shifted melt curve in the presence of 0.1 mM $NiCl_2$ correlates to a ~ 7 °C increase in T_m. Similar stabilisation was also seen with other divalent metal salts, including $MgCl_2$ and $ZnCl_2$. However, the addition of metal chelators EDTA and EGTA had no effect on the T_m of CPR-DprA relative to the control. This suggests that metal ions are not intrinsic to the structural integrity of CPR-DprA, though are still capable of stabilisation.

5.2.4. Structure determination of CPR-DprA

Crystallisation

Preliminary crystallisation trials were conducted with CPR-DprA at 5.0 mg/ml using a variety of commercially available high-throughput screens. The purity of the protein sample used for these trials is shown in Fig. 5.8.

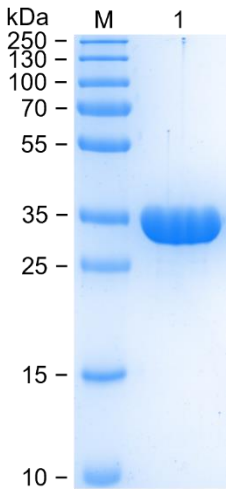

Fig. 5.8: SDS-PAGE analysis of CPR-DprA (MW = 32.3 kDa) purity for preliminary crystallisation trials. M = marker, with MWs of bands on the left; 1 = CPR-DprA sample.

The initial screening experiments identified several promising conditions for optimisation. Four such conditions were found in the Pact *premier*™ HT-96 eco screen[137], each with buffer systems at pH 8.0 – 9.0 and containing 25% polyethylene glycol (PEG) 1500. These four conditions were used as the foundation for a 24-condition optimised screen (OPT1; Table C1).

Another encouraging condition was identified in the JCSG-*plus*™ HT-96 eco screen[137], composed of 0.1 M HEPES pH 7.4 and 10% PEG 6000. This condition was also optimised to produce a second 24-condition screen (OPT2; Table C2) by varying the pH and precipitant concentration, as well as trialling PEG 1500 in place of the original PEG 6000 due to its occurrence in fruitful conditions from the Pact *premier*™ screen.

Potential crystals or 'hits' were identified in conditions from both optimised screens. OPT2 was also found to generate more hits in wells containing PEG 1500 than PEG 6000, indicating that PEG 1500 is a favourable precipitant for CPR-DprA crystallisation. Examples of CPR-DprA crystals from high-throughput trials and optimised screens are shown in Fig. 5.9.

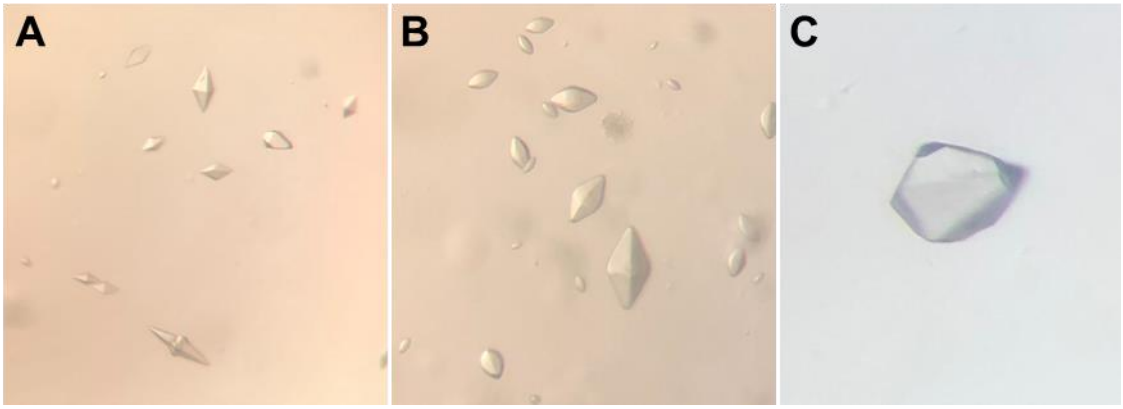

Fig. 5.9: (A) CPR-DprA crystals from initial high-throughput screening grown in 0.1 M PMTP pH 8.0, 25% PEG 1500, from the Pact *premier*™ HT-96 eco screen; approximately 50 µm at widest dimension; **(B)** CPR-DprA crystals from initial high-throughput screening grown in 0.1 M MMT pH 9.0, 25% PEG 1500, from the Pact *premier*™ HT-96 eco screen; approximately 60 µm at widest dimension; **(C)** CPR-DprA crystal grown in 0.1 M MMT pH 7.5, 30% PEG 1500, from OPT1 screen; approximately 120 µm at widest dimension.

Diffraction data collection

CPR-DprA crystals were cryo-protected in 50% glycerol prior to cryo-cooling and transportation to Diamond Light Source (DLS). The crystal in Fig. 5.9C is shown looped for diffraction data collection in Fig. 5.10A, with a resulting diffraction image shown in Fig. 5.10B.

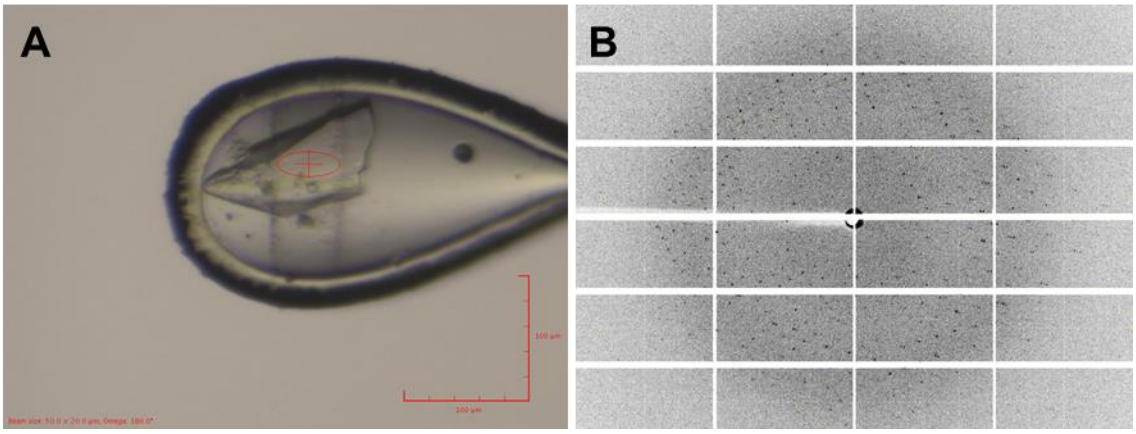

Fig. 5.10: (A) CPR-DprA crystal from Fig. 5.9C mounted on a nylon loop for X-ray diffraction data collection at the I03 beamline, Diamond Light Source (DLS). Beam size (50 µm x 20 µm) and position is indicated with a red oval and cross; 100 µm scale bars are shown in the bottom right corner. **(B)** Resulting diffraction image collected during crystal screening at a wavelength of 0.9762 Å.

In total, 30 CPR-DprA crystals were screened for diffraction, and seven datasets were collected from four well-diffracting crystals. This included two from different parts of the large crystal in Fig. 5.9C/5.10A. The first was collected at 20% transmission. For the second, the transmission was increased to 100% to maximise resolution; however, slight radiation damage was evident as the spot count and expected resolution fell slightly towards the end of collection. Regardless, statistics indicated that this was the best dataset collected for CPR-DprA and merging of multiple datasets was not required. Data collection statistics can be found in Table C3.

Data processing

Automatic data processing by the pipelines on ISPyB was successful, with consistent predictions of either space group $P6_122$ or $P6_322$ and similar resolution limits around 2 Å in most cases. AutoPROC + STARANISO[256,257] predicted a higher resolution than the other pipelines (1.75 Å) coupled with lower completeness (91.7% *vs* ≥ 99.8%), suggesting the data were somewhat anisotropic. To optimise the high-resolution cut-off, data were manually processed using XDS[258] in space group $P6_122$, with a conservative cut-off of 2.1 Å[259].

Model selection for molecular replacement

A total of six DprA crystal structures were identified in the PDB and are outlined in Table 5.1: one for *Sp*DprA, one for *Rp*DprA, and four structures of *Hp*DprA. The availability of these homologous structures made CPR-DprA an encouraging candidate for molecular replacement (MR).

Table 5.1: Details DprA structures deposited in the PDB, identified using the advanced sequence search tool with the CPR-DprA sequence.

PDB code	Species	E value	Sequence identity / %	Query cover / %	Resolution / Å
3uqz	*S. pneumonia*	8.6e-29	41.9	73.2	2.7
3maj	*R. palustris*	3.8e-25	36.1	96.2	2.1
5mll	*H. pylori*	4e-18	38.7	67.2	1.9
4ljr	*H. pylori*	4e-18	38.7	67.2	1.8
4ljl	*H. pylori*	4.2e-18	38.7	67.2	2.2
4ljk	*H. pylori*	4.2e-18	38.7	67.2	2.4

The highest scoring model was the *Sp*DprA protein, which shared the highest sequence similarity (41.9%) with CPR-DprA despite lacking the C-terminal domain (Fig. 5.1). The second highest score was for *Rp*DprA, which had lower overall sequence identity, but notably also covered the C-terminal extradomain and had better resolution than the *Sp*DprA structure (2.1 Å *vs* 2.7 Å). The *Hp*DprA models were of a truncated version of the protein without the C-terminal domain, hence the low query cover.

The *Sp*DprA, *Rp*DprA, and *Hp*DprA sequences were aligned with CPR-DprA using ClustalW[253] and the models (3uqz, 3maj, and 5mll, respectively) were processed using CHAINSAW[260] to prune poorly aligned regions. The resulting homology models were used for MR in Phaser[261]. Whilst no solutions were found using the *Rp*DprA or *Hp*DprA models, the CPR-DprA structure was successfully solved (Table C4) using the *Sp*DprA model (3uqz) with highest sequence similarity. The CPR-DprA crystal structure was found to have a single protein chain in the asymmetric unit (ASU). Phaser also determined the space group as $P6_522$, rather than enantiomorph $P6_122$ predicted during data processing.

5.2.5. Crystal structure of CPR-DprA

The CPR-DprA structure adopts an extended Rossmann fold (eRF) as seen in all other known DprA structures[250,254], comprised of nine β-strands forming a curved sheet through the centre, with short α-helices flanking on either side in a sandwich (Fig. 5.11).

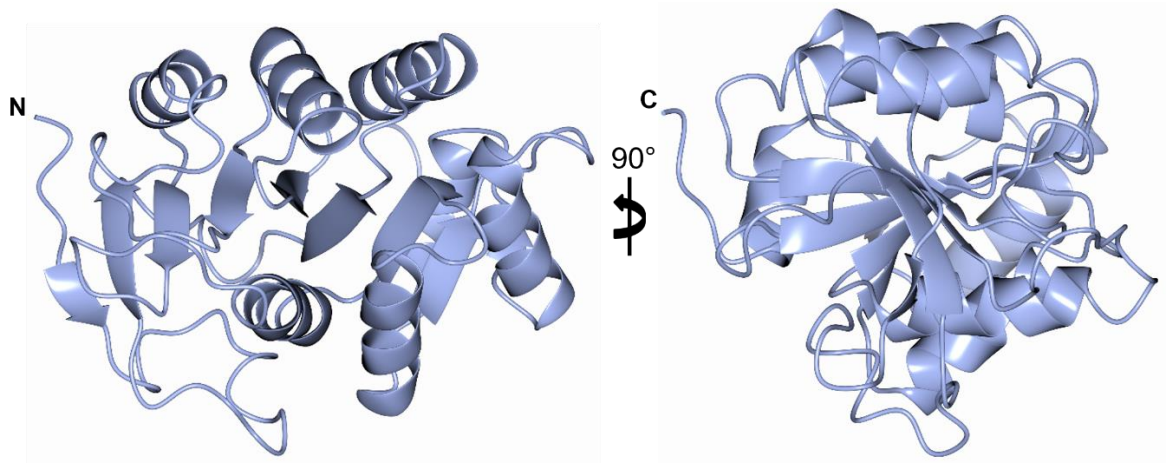

Fig. 5.11: Ribbon diagram of the CPR-DprA monomer (residues 2 to 221) from a side view (left) and front view (right), showing the eRF.

The refined CPR-DprA model has R/R_{free} values of 0.20/0.23; however, the structure only covers 75% of the protein sequence (Leu2 to Pro221) with no discernible electron density for the CTD. It was initially unclear whether the CTD was present but highly flexible, or if a truncated version of the protein missing the extradomain had crystallised.

The Matthews coefficient[262] from Phaser[261] indicated that the measured unit cell containing a full-length CPR-DprA chain would have a solvent content of only ~ 37%, occurring with a probability of only 30 – 40%. When changing the unit cell composition to contain only the portion of the protein identified in the electron density, a much more reasonable solvent content of 52% was predicted, fitting in the 80 – 95% probability band. It is therefore likely that the CPR-DprA underwent degradation during crystallisation, resulting in a truncated protein without the CTD in the crystal structure. This phenomenon has also been observed in previous structural studies of *Hp*DprA, which

found that the full-length protein degraded during the crystallisation process, leaving a truncated protein lacking the CTD[250].

5.2.6. Prediction of C-terminal domain structure using AlphaFold2

Since the CTD of CPR-DprA was missing from the crystal structure, the structure prediction tool AlphaFold2[118,263] was used to generate a model of the full-length protein. The AlphaFold2 CPR-DprA model contained two distinct domains separated by a disordered linker (Fig. 5.12).

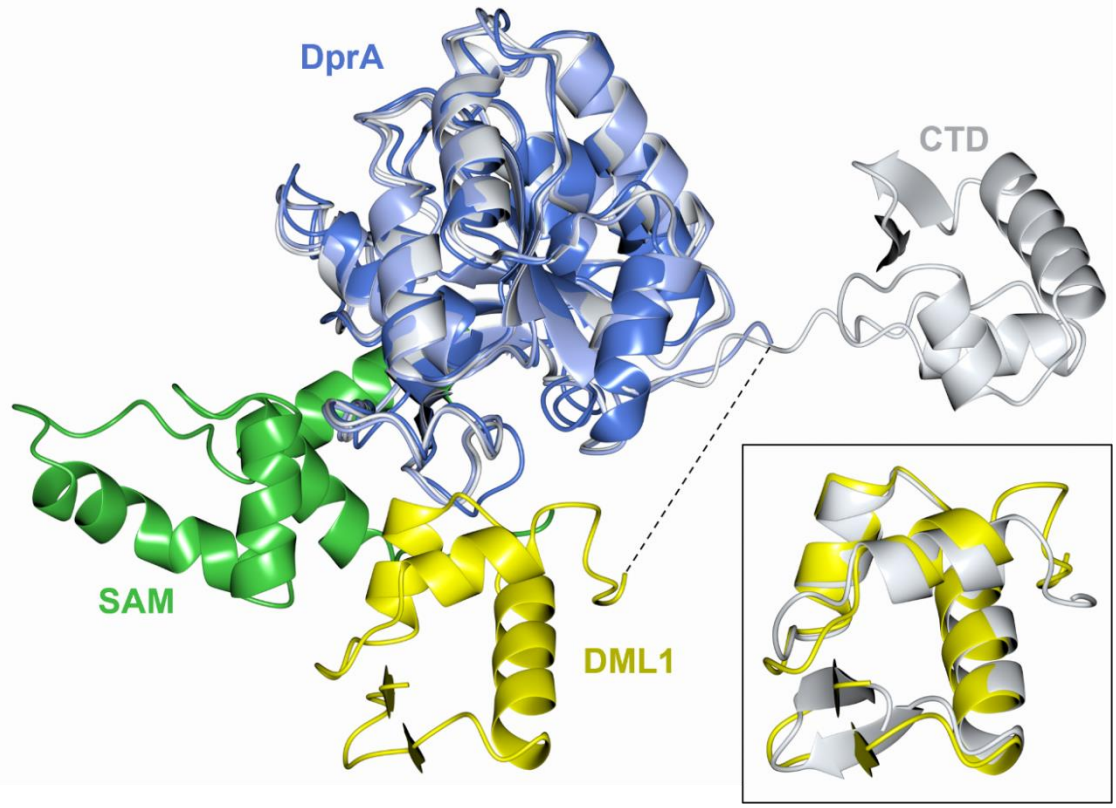

Fig. 5.12: Ribbon diagram of the least-squares superposition of the experimental CPR-DprA core domain (ice blue) with the AlphaFold2[118,263] model (white) and the full-length *Rp*DprA structure (3maj: N-terminal SAM domain in green; DNA_processg_A core DprA domain in royal blue; C-terminal DML1-like domain in yellow). The dashed line represents a loop unresolved in the 3maj electron density[254]. Outlined at the bottom right is the least-squares superposition of the AlphaFold2 predicted CTD of CPR-DprA (white) with the experimentally deduced CTD of *Rp*DprA (yellow).

The eRF of the core DprA domains aligned extremely well between the predicted and experimental structures, with a root mean squared deviation (RMSD) of 0.59 Å (first 220 residues). The AlphaFold2 predicted structure of the CPR-DprA CTD was also found to align well with the experimentally determined DML1-like CTD of *Rp*DprA (RMSD 1.08 Å, 58 residues), despite poor sequence similarity (Fig. 5.2). Whilst the position of the CTD in the CPR-DprA AlphaFold2 model does not align with the *Rp*DprA crystal structure, the linker region connecting the CTD to the core domain in the CPR-DprA had low model confidence (predicted local distance difference test, pLDDT < 70).

Such low pLDDT values are indicative of disorder, flexibility, or regions that are unstructured under typical conditions[264]. It is therefore expected that, whilst the predicted structure of the CTD domain structure is valid, its position relative to the core DprA domain in the AlphaFold2 model is unreliable.

5.2.7. Dimer analysis

The ASU in the CPR-DprA crystal structure was found to contain a single protein chain. Based on SEC analysis and homologous DprA structures[250,254], CPR-DprA was expected to exist as a functional dimer. PISA[265,266] was therefore used to assess potential dimerisation interfaces, identifying a probable dimer formed with a symmetry mate around a crystallographic two-fold axis (Fig. 5.13).

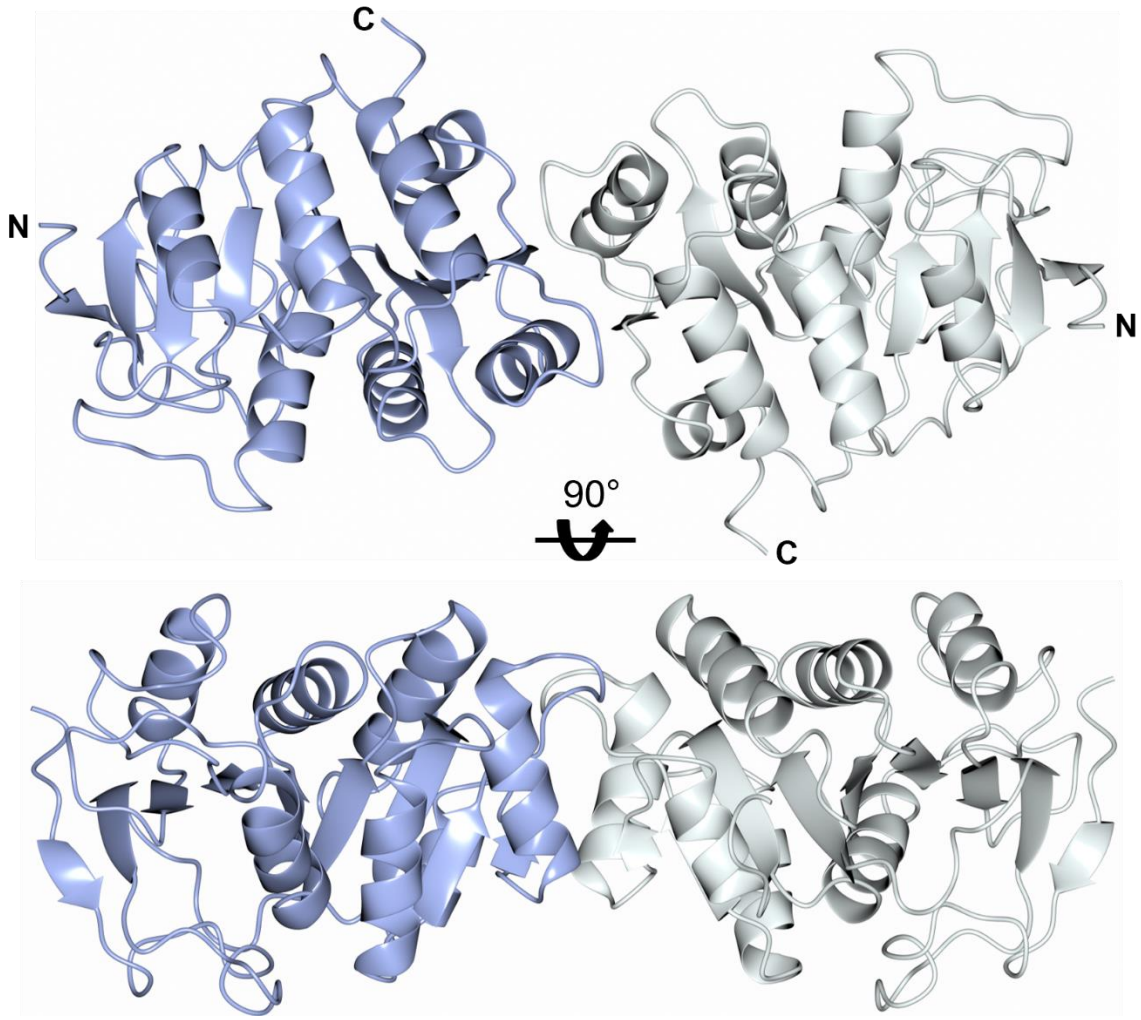

Fig. 5.13: Ribbon diagram of the proposed CPR-DprA functional dimer formed with a crystallographic symmetry mate (-Y, -X, -Z + 1/6). Viewed from above (top) and the side (bottom) showing the end-to-end dimerisation of the eRF. Individual protein chains are coloured in ice blue and white, with N- and C-termini indicated.

Dimerisation of CPR-DprA occurs via the C-terminal end of the eRF that forms the core DprA domain. The dimerisation interface has an area of 952 Å^2 and involves 10.5% of the total residues in

the crystal structure, with a significant total binding energy of -14.0 kcal/mol indicating that it is biologically relevant rather than an artefact of crystal packing. The absence of the CTD in the crystal structure makes it difficult to assess its impact on dimerisation of CPR-DprA. However, the predicted AlphaFold2 model suggests that the CTD folds independently of the eRF, and studies of the *Rp*DprA crystal structure have found its CTD does not inhibit dimerisation. It is therefore unlikely that the CPR-DprA CTD prevents dimerisation.

The dimer interface of CPR-DprA is structurally well conserved between the DprA structures, which each form 'end-to-end' dimers at the eRF. A superposition of the CPR-DprA dimer with homologous DprA assemblies is shown in Fig. 5.14.

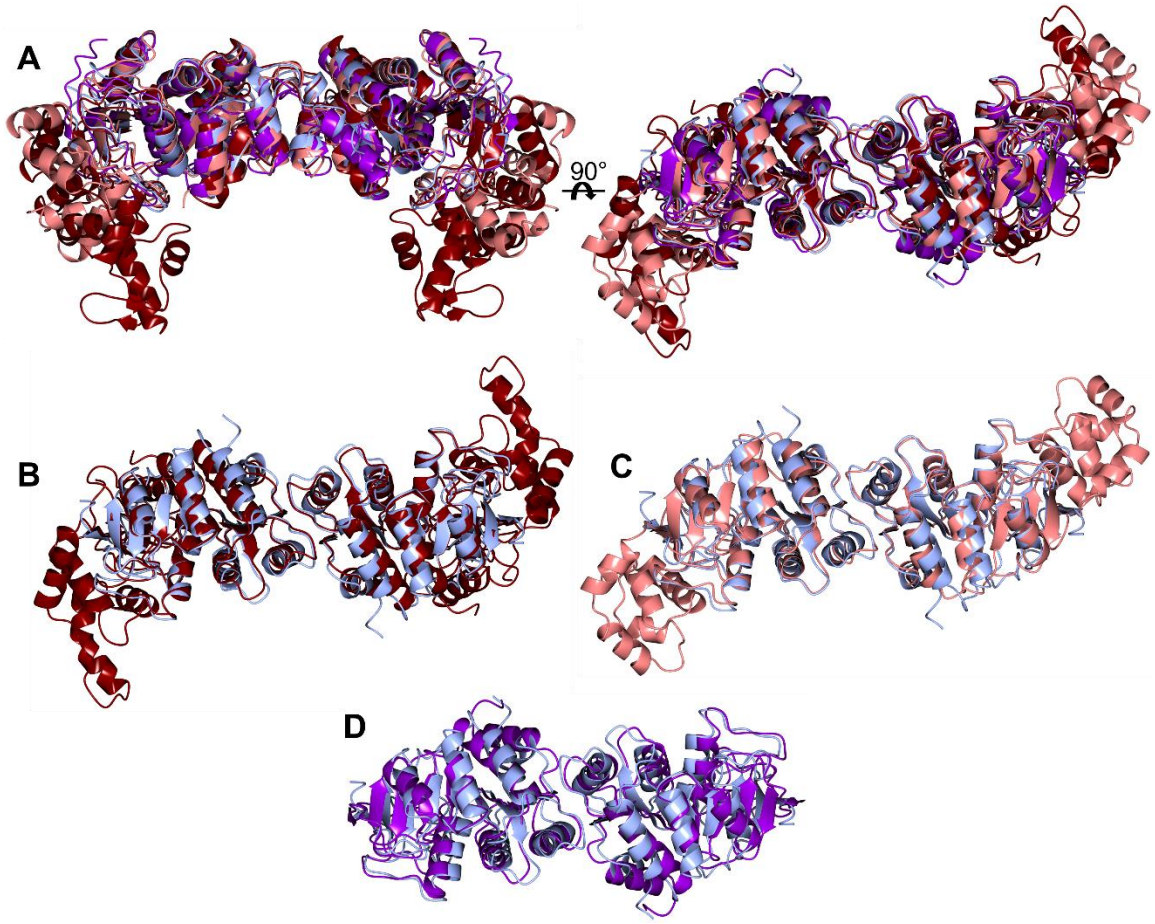

Fig. 5.14: Ribbon diagrams of the least-squares superpositions of DprA dimers from the four homologous structures. CPR-DprA shown in ice blue, *Sp*DprA in red (3uqz; RMSD 1.83 Å / 410 C_α), *Rp*DprA in pale pink (3maj; RMSD 1.32 Å / 410 C_α), and *Hp*DprA in magenta (4ljl; RMSD 2.31 Å / 394 C_α) **(A)** Viewed from the side (left) and the top (right) showing the end-to-end dimerisation of the eRF. **(B)** CPR-DprA and *Sp*DprA. **(C)** CPR-DprA and *Rp*DprA. **(D)** CPR-DprA and *Hp*DprA.

Stabilisation of the *Sp*DprA, *Rp*DprA, and *Hp*DprA dimers involves hydrophobic interactions between several conserved residues, as well as combinations of intermolecular salt bridges and hydrogen bonds[250,254]. Several of these dimer interface residues were found to be conserved in the

CPR-DprA sequence through earlier sequence alignments (Fig. 5.2), and similarly form key interactions, as shown in Fig. 5.15.

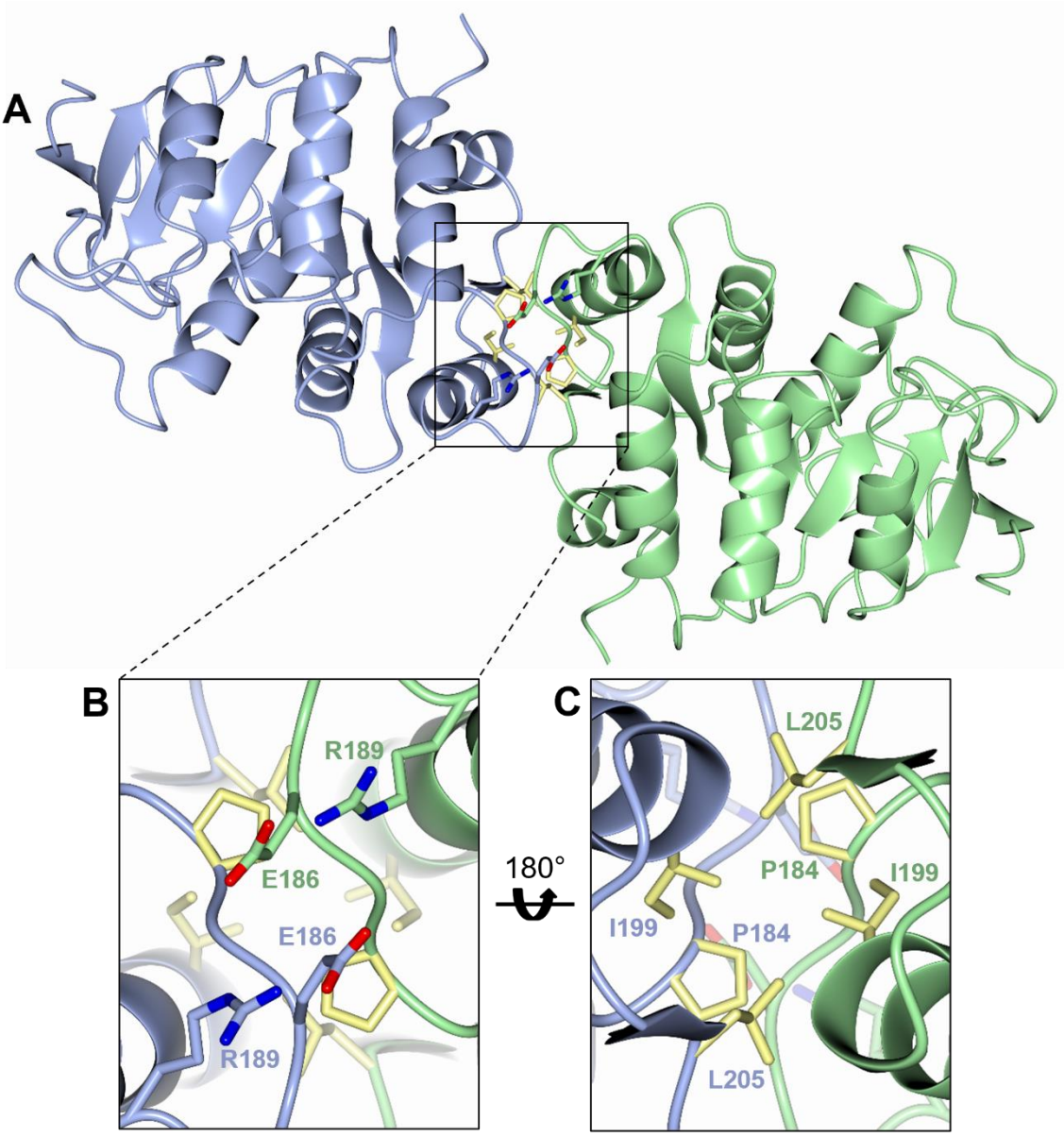

Fig. 5.15: (A) Ribbon diagram of the CPR-DprA dimer highlighting key residues at the dimerisation interface boxed in black. Key residue side chains are shown with cylinder representation, O atoms in red, N atoms in blue, hydrophobic residues in yellow. **(B)** Close-up view of the region boxed in (A) showing residues involved in dimerisation. **(C)** Rotated view of dimerisation interface showing conserved hydrophobic residues.

Dimerisation of CPR-DprA appears to be mediated through hydrophobic interactions between three conserved or positively retained residues: Pro184, Ile199, and Leu205 (Fig. 5.2 and Fig. 5.15). Combined sequence and structural analysis also revealed another pair of key interacting residues at the dimerisation interface, Glu186 and Arg189. Whilst the Arg189 side chain is not well-defined in the density, the positioning of these residues is appropriate for hydrogen bonding. Equivalent residues in *Hp*DprA (Arg185 and Glu188) form intermolecular hydrogen bonds that stabilise the

dimer[250], but these residue identities are switched in the CPR-DprA sequence. This correlated mutation stresses the importance of this structural contact for dimer formation. The structural overlap of the CPR-DprA dimer with homologous assemblies and conservation of several interacting residues at the interface adds confidence to previous assertions that dimerisation is an evolutionarily conserved feature of DprA proteins[254].

5.3. Conclusions

CPR-DprA is one of 24 Virus-X targets and 7 category A targets for which crystal structures have successfully been determined, having progressed smoothly through the biodiscovery pipeline[18]. High yields of stable, soluble protein were expressed and successfully purified by a combination of IMAC and SEC techniques. The high purity and homogeneity of CPR-DprA also facilitated crystallisation in high-throughput trials. Through optimisation of conditions identified during this crystallisation screening, well-diffracting crystals were produced that enabled data collection to a maximum resolution of 2.1 Å.

The availability of several structures of well-studied homologous DprA proteins from *H. pylori, S. pneumoniae*, and *R. palustris* proved instrumental during analysis of CPR-DprA, enabling the novel crystal structure to be solved by MR. Subsequent structural comparisons found that CPR-DprA forms dimers in the same manner as these homologs, largely through hydrophobic interactions between conserved residues. Dimerisation of DprA homologs has been shown to be functionally relevant during the transformation process, and the same can be assumed for CPR-DprA based on the similarities in sequence and structure. Nevertheless, in-depth functional analysis of CPR-DprA is required to assess the role of dimerisation in CPR-DprA function.

The occurrence of DprA proteins in CPR bacteria is of note, as many obligate bacterial species with similarly small genomes do not contain DprA homologs[239]. Since CPR bacteria are incapable of synthesising nucleotides, they are anticipated to scavenge DNA. It is therefore plausible that CPR bacteria are naturally competent, though the CPR-DprA proteins may also carry out broader DNA processing roles. The extensive distribution, conserved sequences, and close functional relationships of the DprA domains suggests a vital and optimised function, and likely an ancient common origin[13].

Chapter 6. Determination of the structure and function of hypothetical protein CPR-C4 from the Candidate Phyla Radiation

6.1. Introduction

6.1.1. Hypothetical proteins

As the availability and species variety of whole genome sequences increases with improvements in DNA isolation and sequencing, so does the number of predicted gene products that are annotated as 'hypothetical proteins'. There is no direct experimental evidence to confirm the translation of these hypothetical proteins *in vivo*, and their biological functions have not been characterised[267]. Determining the functions of these proteins is vital to improving our understanding of the source organisms and their proteomes.

In a newly sequenced bacterial genome, around 70% of gene products can be assigned a predicted function through annotation[158]. The remainder either show some homology to sequences with unknown function (conserved hypothetical) or they have no known homologs at all (hypothetical)[120]. For the hypothetical proteins with no homologs, experimental evidence of their *in vivo* expression is required as this cannot be determined from sequence alone. Even for conserved hypothetical proteins, with homologs often identified across several genomes, bioinformatic analysis is only able to provide a prediction of function using conserved sequence motifs, structural features, and subtle sequence similarity to characterised proteins. True biological function and validation of *in vivo* existence can only be determined by experimentation.

6.1.2. 'Unknown' proteins as key Virus-X targets

Many of the CPR targets within the Virus-X database are hypothetical, or conserved hypothetical proteins, and fall into category C (2.2.2.). These targets are of particular interest, as they may exhibit novel functions that can be harnessed to provide powerful new tools, as was seen with the discovery of the CRISPR-Cas system (Clustered Regularly Interspaced Short Palindromic Repeats and associated protein)[268]. The functions and desirable qualities of these hypothetical proteins are often impossible to predict from nucleotide sequence alone due to a lack of detectable homology. As such, structural determination, using methods including X-ray crystallography, offers a clearer insight into potential function by enabling direct comparisons and structure-based multiple sequence alignments (3D-MSA) that can uncover structural similarities between proteins despite low sequence identity. The 3DM protein superfamily analysis software provides a pipeline for this analysis, fully integrating

sequence data with 3D structures, as well as mined literature data such as mutation and ligand contact information, from all members of a protein superfamily[269].

When selecting C targets for further analysis, priority was given to conserved hypothetical proteins, with unknown function but showing extended conservation across different genomes, increasing the likelihood of biological importantance[18]. One such target was CPR-C4, which was subsequently chosen for characterisation. Its lack of sequence identity to any proteins of known function in public databases led to its categorisation as a high-priority target for structural determination, and assessment of its potentially novel function. Establishing structure-function relationships for hypothetical proteins from the CPR is vital to gain understanding of their little-known source organisms and identify any potential commercial value.

6.2. Results and discussion

6.2.1. CPR-C4 sequence analysis with BLAST

The CPR-C4 gene was identified on the contig discussed in 2.2.1., from a CPR bacterial genome identified at 75 °C and pH 5 in an Icelandic hot spring. It is one of the 70% of genes on this contig that encodes a hypothetical protein, where a potential function could not be established through sequence alone. However, BLAST[151] identified over 400 'hit' bacterial sequences with an Expect value (E) < $1e^{-50}$ against the CPR-C4 amino acid sequence. The closer the E value to zero, the more significant the match between sequences, meaning these hits show strong similarity to the CPR-C4 sequence. CPR-C4 therefore shows extensive conservation across bacteria, with hits from both inside and outside the CPR.

The position of the CPR-C4 gene on the source contig is shown in Fig. 6.1. Its most significant BLAST hit was from an unclassified bacterial phylum, though its flanking genes share most similarity with sequences from the *Candidatus Adlerbacteria* phylum within the CPR.

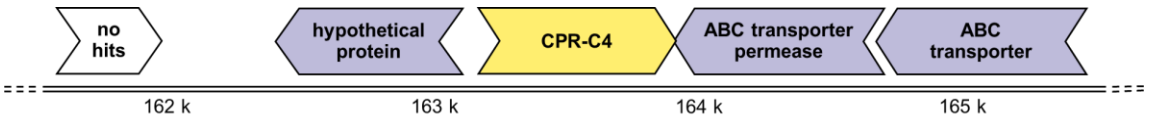

Fig. 6.1: Position of the CPR-C4 gene and its neighbours on the source contig, represented as arrows showing the strand direction of transcription. The predicted function of gene products is stated where annotation has been possible. Genes are colour coded by the phylum classification of their most significant BLAST hit: yellow = unclassified bacteria; purple = *Candidatus Adlerbacteria*; white = unknown. Used under the terms of the Creative Commons CC-BY license[60].

Downstream of CPR-C4 are genes coding for ATP-binding cassette (ABC) transport system proteins, though analogous genes were not identified near CPR-C4 homologs in *Candidatus Kaiserbacteria*

or *Candidatus Adlerbacteria* genomes. CPR-C4 is also transcribed in the opposite strand direction to its neighbours on both sides. The CPR-C4 gene therefore appears unrelated to its flanking genes and its position does not illuminate any potential function, though the sequence analysis with BLAST shows it encodes an important and conserved protein that warrants further investigation.

6.2.2. CPR-C4 sequence analysis with 3DM

To further assess the extent of conservation of CPR-C4 among bacteria, deeper sequence analysis was conducted using the 3DM software. Using the CPR-C4 amino acid sequence, a 3DM subfamily system was created. 3DM information systems are platforms for protein families that collect and integrate data from all members of that family[269]. The CPR-C4 subfamily contains 794 sequences, 25 of which were identified through the metagenomics efforts of the Virus-X consortium. The first 215 of the 221 residues in the CPR-C4 amino acid sequence fell within a conserved 'core region' of this subfamily (Fig. 6.2A).

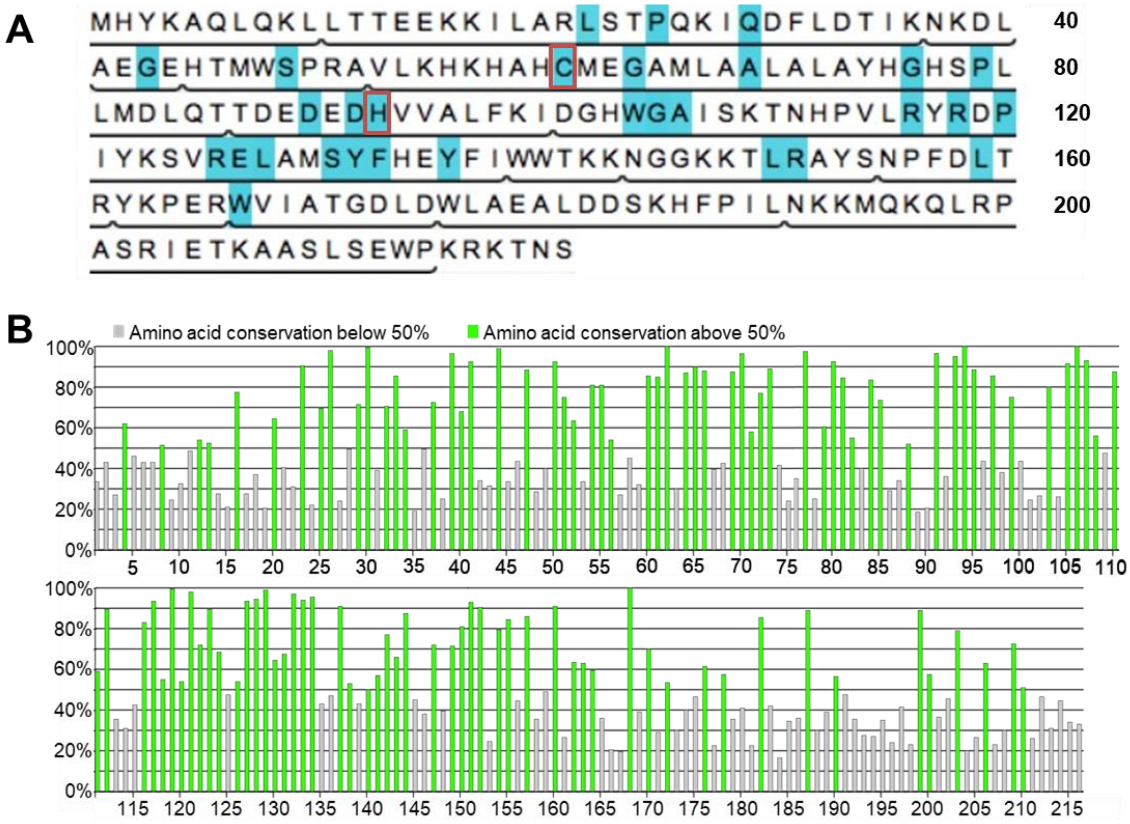

Fig. 6.2: (A) Amino acid sequence of CPR-C4. Residues falling within the conserved core region of the 3DM subfamily are underlined. Residue identities conserved in ≥ 90% of subfamily sequences are highlighted in blue. 99.9% conserved residues C61 and H93 (following CPR-C4 numbering) are boxed in red. **(B)** Percentage conservation of residue identities within the 3DM subfamily core region (following 3DM numbering). Green bars represent ≥ 50% conservation; grey bars indicate < 50% conservation. The 99.9% conserved C61 and H93 in (A) are at positions 62 and 94, respectively. Used under the terms of the Creative Commons CC-BY license[60].

The extent of sequence conservation in this core region can be seen in Fig. 6.2B. Over half (110 / 216) of the residue identities are conserved in ≥ 50% of subfamily sequences, with 15% (32 / 216) conserved in ≥ 90% of cases. Residues towards the edge of the core region are slightly more variable relative to more central residues, though the high level of conservation across the core region is clear. Several residues conserved in nearly all sequences were also identified and subsequently proposed to be important for function. These included a cysteine and histidine (C61 and H93, Fig. 6.2), residues often key to enzymatic activities, that are both 99.9% conserved. Whilst these were clues to potential function, determination of the CPR-C4 structure was necessary for analysis of residue interactions and to assess the influence of these highly conserved residues.

6.2.3. Production and characterisation of CPR-C4

Expression and IMAC purification

To enable effective structural determination, an efficient system for the expression, purification, and characterisation of the CPR-C4 protein had to be established. Initial small-scale tests confirmed that CPR-C4 overexpression could be induced in an *E. coli* host with the addition of 10% L-rhamnose to culture (3.2.2.). However, preliminary large-scale expression and purification experiments resulted in low protein yields, producing less than 1 mg of soluble protein per litre of culture, with large quantities of insoluble protein sequestered in the cell pellet after lysis and centrifugation (Fig. 6.3).

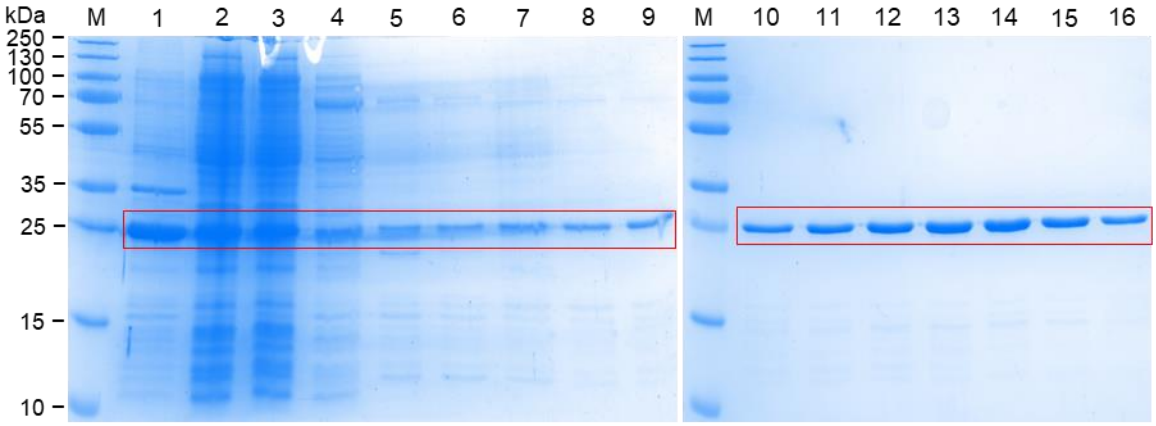

Fig. 6.3: SDS-PAGE analysis of CPR-C4 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 16 = protein fractions. Bands corresponding to CPR-C4 (MW = 27.1 kDa) are indicated with a red box.

There were also several hurdles encountered during purification. CPR-C4 bound tightly to the Ni^{2+} column used for immobilised metal ion affinity chromatography (IMAC), requiring an unusually high concentration of imidazole (> 500 mM) to elute. It was also evident that precipitation occurred with even gentle centrifugation. Despite these issues, IMAC purification resulted in protein of high enough purity for characterisation and preliminary crystallisation screening.

Characterisation

SDS-PAGE and ESI-TOF MS analysis identified a species proposed to be CPR-C4 based on expected MW, with a peak corresponding to a dimer also present in the mass spectrum (Fig. D1). The identity of the species was confirmed as CPR-C4 using trypsin digest MS techniques.

SEC analysis

SEC was used to assess the multimeric state of CPR-C4 in solution, with the resulting chromatogram shown in Fig. 6.4.

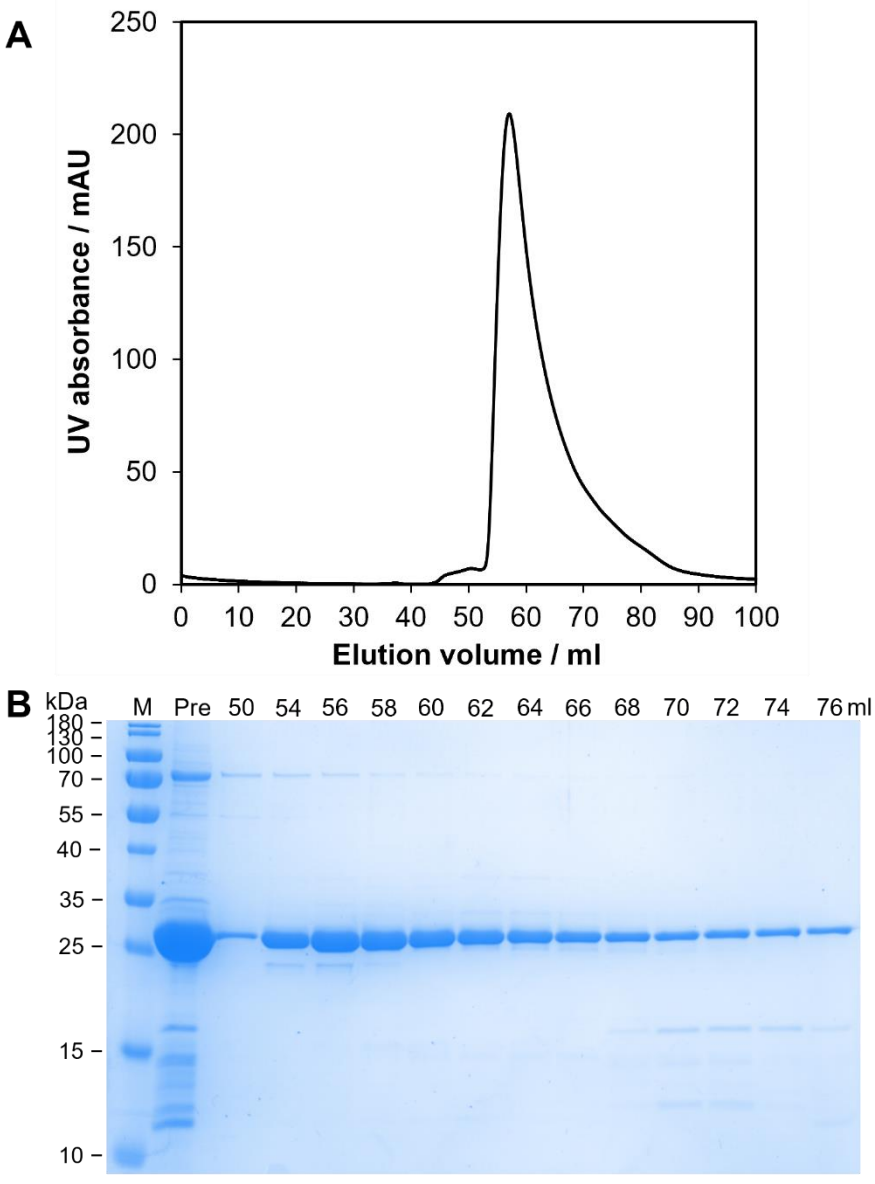

Fig. 6.4: (A) Chromatogram from SEC of CPR-C4. **(B)** SDS-PAGE analysis of elution fractions from SEC. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The protein eluted at a volume corresponding to a 51.7 kDa species, which is approximately twice the monomeric MW of CPR-C4 (27.1 kDa). This implies that CPR-C4 forms dimers, as also evidenced by the ESI-TOF mass spectrometry results (Fig. D1).

Thermal shift analysis

The low yields and precipitation encountered made initial expression and purification of CPR-C4 challenging. TSA was therefore used to identify potentially stabilising conditions to resolve these issues[210,255]. SDS-PAGE was first used to confirm that the protein was free of major impurities (Fig. 6.5).

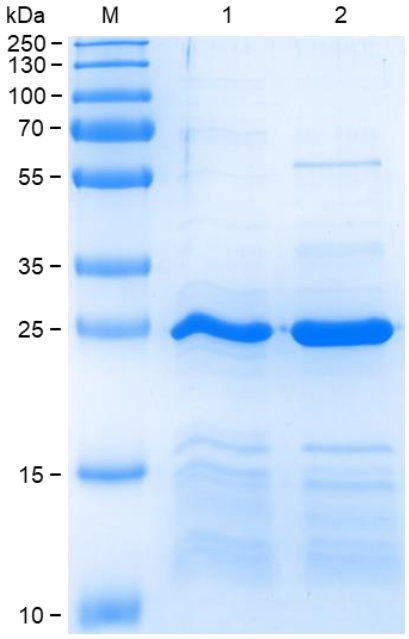

Fig. 6.5: SDS-PAGE analysis of CPR-C4 purity for TSA. M = marker, with MWs in kDa to the left; 1 = SGC buffer; 2 = TSA buffer.

TSA revealed that the reference melt temperature, T_m, of CPR-C4 in MilliQ water was only ~ 29 °C. This provides an explanation for the ready precipitation of CPR-C4 during purification, particularly under ambient temperatures. Given the discovery of the source organism at 75 °C, this considerably low T_m of CPR-C4 is surprising; however, it is not guaranteed that proteins from thermophilic source organisms will be intrinsically thermostable. Molecular chaperones, compatible solutes, and other mechanisms *in vivo* are responsible for a large amount of the stabilisation that allows for life at extreme temperatures[225]. Most of the additives screened resulted in some level of stabilisation of CPR-C4, seen as an increase in T_m relative to the reference in water. As well as establishing conditions to stabilise CPR-C4 during overexpression and purification, this provides an insight into how thermotolerance is achieved in the natural environment of the source organism.

pH screening

The results of TSA with the Durham pH Screen® revealed that CPR-C4 is generally most stable at weakly acidic pH (Fig. 6.6). The greatest stabilisation was seen in citric acid buffers, with an increase

in T_m of 12 °C in 100 mM citric acid (pH 5.5). The high pH buffers above ~ pH 9 resulted in poor fluorescence signals from which accurate T_m values could not be predicted, implying the protein was very unstable under these conditions.

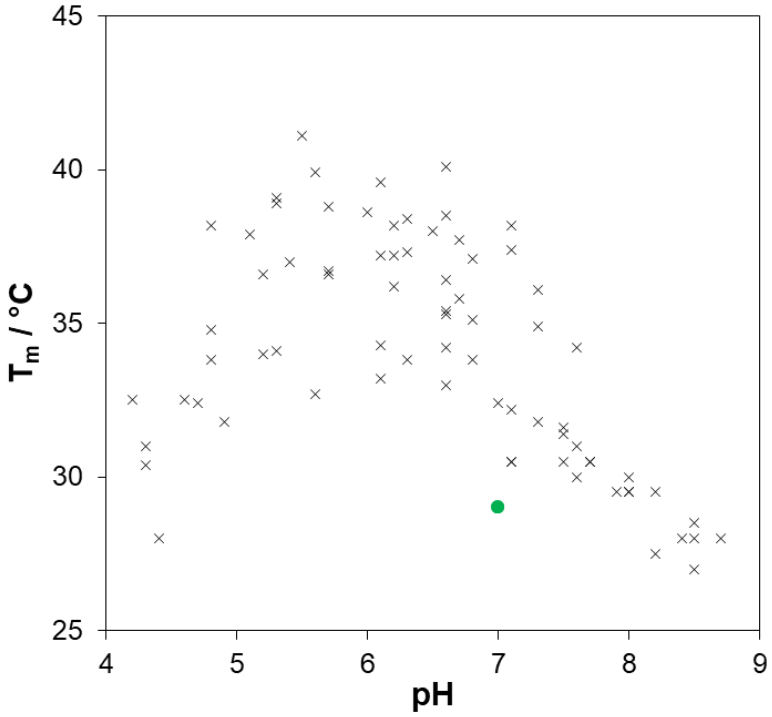

Fig. 6.6: Results from TSA using the Durham pH Screen® showing the effect of pH on the T_m of CPR-C4. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the same pH value are from different buffer systems at identical pH.

It was also noted that CPR-C4 was relatively unstable in the standard HEPES pH 7.5 buffer used during purification, with a T_m of only 32.4 °C, which explains its tendency to precipitate. Whilst the use of acidic buffers such as citric acid for purification would be most stabilising, neutral or weakly alkaline buffers are recommended for IMAC purification using Ni^{2+} affinity. CPR-C4 was found to be moderately stabilised in a pH 7.4 phosphate buffer, with a 4 °C increase in T_m relative to HEPES, providing a suitable alternate buffer for purification.

Salt screening

CPR-C4 was noticeably stabilised by a broad range of inorganic salts, in particular several sodium and ammonium salts that resulted in a T_m increase of ≥ 20 °C when added to a final concentration of 1.5 M (Fig. 6.7).

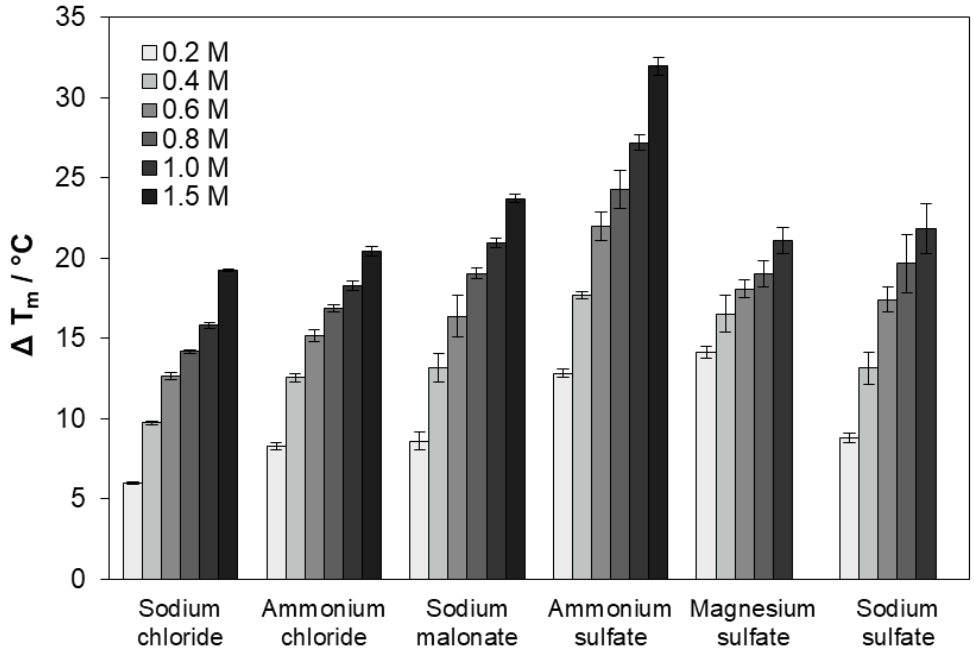

Fig. 6.7: Results from TSA of CPR-C4 with the Durham Salt Screen® showing changes in melt temperature, T_m, upon addition of stabilising salts at increasing concentrations (n = 3, ± st. dev.). Used under the terms of the Creative Commons CC-BY license[60].

For each of these salts, ΔT_m increased with concentration, with only slight variation depending on the salt composition. CPR-C4 is therefore stabilised by high salt concentrations, with seemingly little preference for chemical composition.

Notably, CPR-C4 was massively stabilised by the addition of divalent metal cations. This is most notable with $ZnCl_2$, which caused an increase in T_m of over 30 °C (Fig. 6.8).

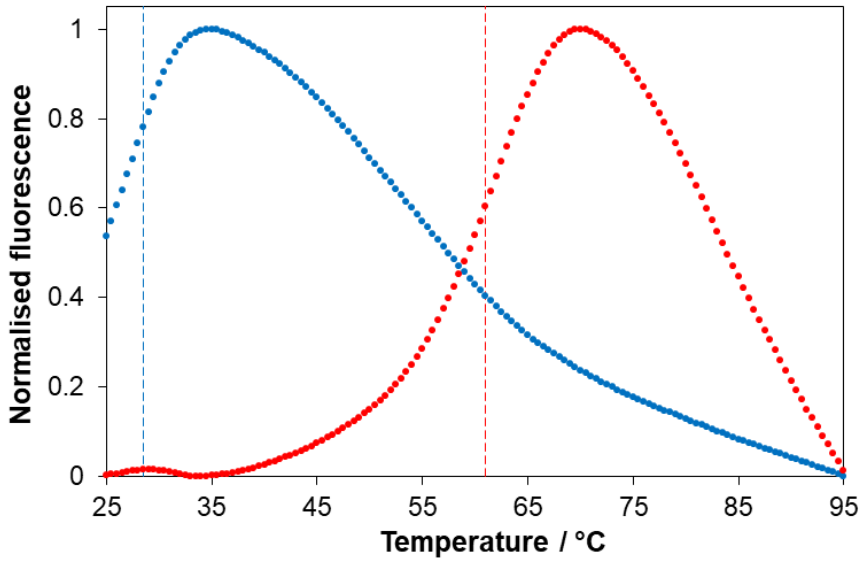

Fig. 6.8: Normalised fluorescence intensities at 590 nm for TSA of CPR-C4 in water (blue) and with 1 mM $ZnCl_2$ (red). Raw experimental data are shown as individual points; dashed vertical lines indicate T_m values of CPR-C4 in water (blue) and with addition of 1 mM $ZnCl_2$ (red). Used under the terms of the Creative Commons CC-BY license[60].

This striking increase in T_m suggested that the stability of CPR-C4 could be dramatically improved with the addition of Zn^{2+}. It also raised the question of whether CPR-C4 could be a zinc metalloprotein, with intrinsic Zn^{2+} required for a structural or catalytic role.

Optimising production of CPR-C4 after TSA

Following the results of TSA, subsequent purification experiments were conducted using a phosphate buffer system, which successfully reduced precipitation. Other adjustments were made to the purification protocol to increase the yield of soluble protein. Previously, CPR-C4 was found to bind tightly to the Ni^{2+} affinity column used during IMAC, eluting over a large volume. This resulted in high volumes of dilute protein, with a subsequent loss of yield during concentration for downstream use. Some protein also remained bound to the column in the presence of 0.5 M imidazole. These protein losses were reduced by increasing the concentration of imidazole in the elution buffer from 0.5 M to 1.0 M and shortening the elution gradient. A comparison of SDS-PAGE gels from the preliminary and optimised purification experiments is shown in Fig. 6.9.

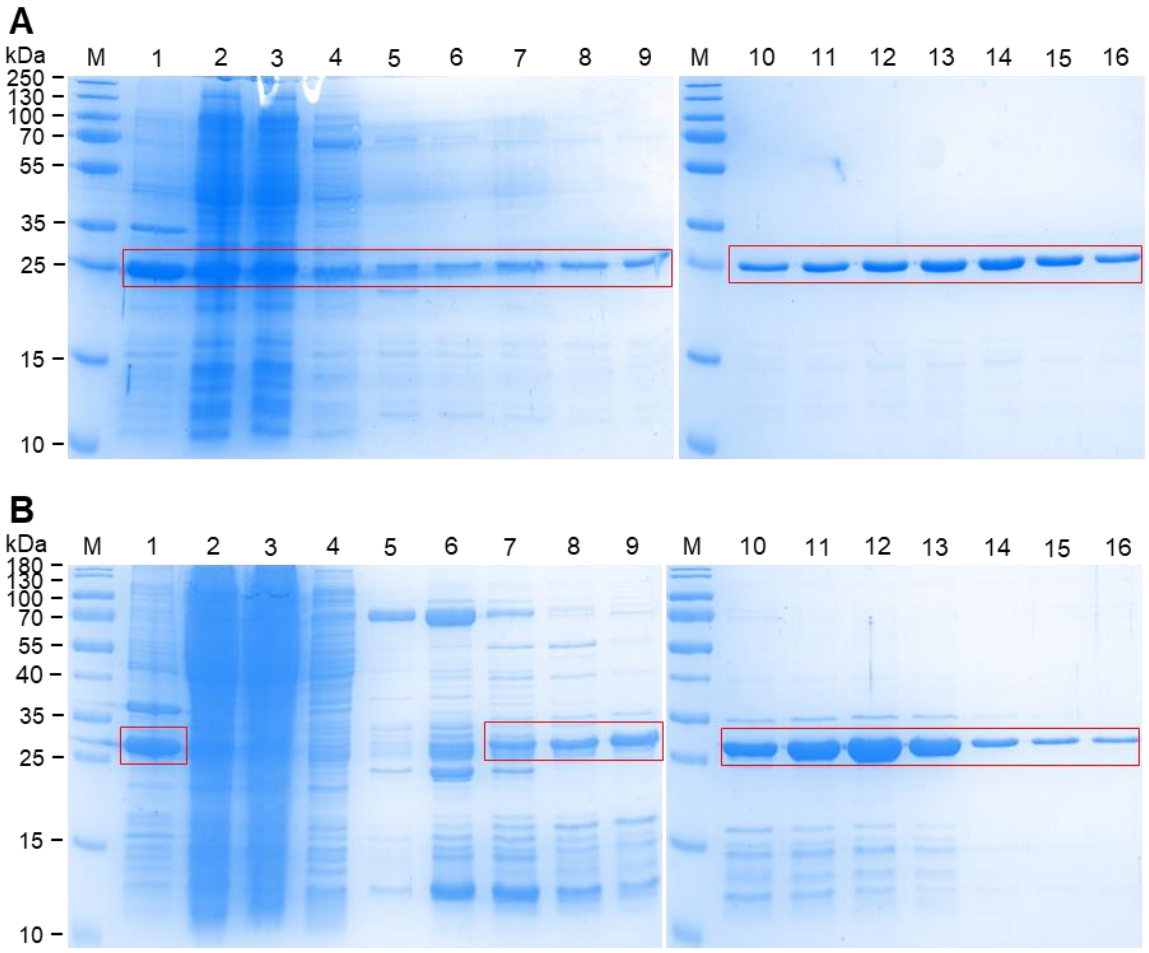

Fig. 6.9: SDS-PAGE analysis of CPR-C4 IMAC purification experiments. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 16 = protein fractions. Bands corresponding to CPR-C4 are indicated with a red box. **(A)** Preliminary IMAC purification with 0.5 M imidazole elution buffer. **(B)** Optimised IMAC purification with 1.0 M imidazole elution buffer.

With the higher final concentration of imidazole and shortened elution gradient, CPR-C4 readily eluted in a smaller volume of more concentrated protein during IMAC. This both increased protein yields and reduced the amount of concentration required for downstream use.

Drawing on the TSA results, changes were also made to the standard expression protocol to improve CPR-C4 production. After the observation of a stark increase in thermotolerance with Zn^{2+}, expression tests were conducted with the addition of $ZnCl_2$ to the culture medium. The resulting SDS-PAGE analysis is shown in Fig. 6.10.

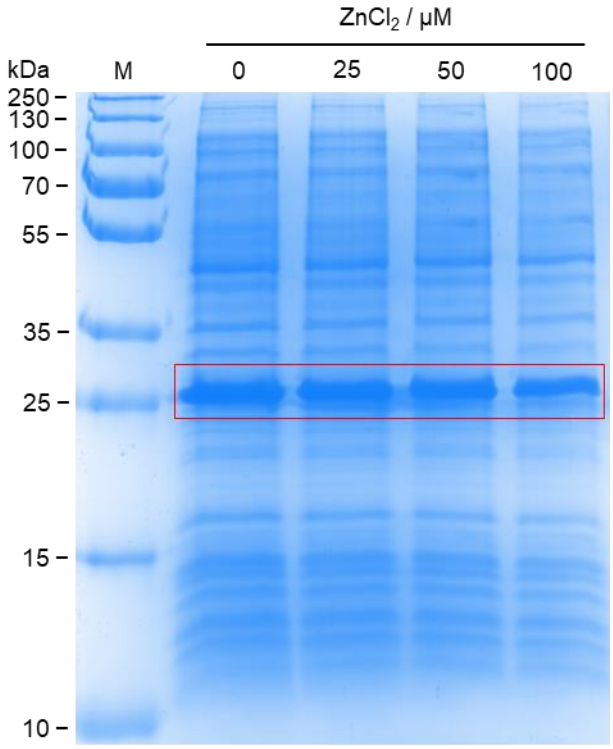

Fig. 6.10: SDS-PAGE of whole cell samples from CPR-C4 expression tests with increasing concentrations of $ZnCl_2$ added to the culture medium. Bands corresponding to CPR-C4 (MW = 27.1 kDa) are indicated with a red box. M = marker, with MWs in kDa to the left.

The final concentration of $ZnCl_2$ used during TSA was 1 mM, though the concentration tested in expression media was limited 100 µM to prevent cytotoxicity in the *E. coli* expression host. The addition of $ZnCl_2$ did not affect the time taken for cultures to reach the optical density (OD) required for induction, or the OD of cultures 16 h after induction, indicating no cytotoxicity even at 100 µM. Based on the size of gel bands in Fig. 6.10, the addition and concentration of $ZnCl_2$ initially appeared to have no effect on the expression levels of CPR-C4. However, full-scale expression experiments with 100 µM $ZnCl_2$ produced > 10 mg of CPR-C4 per litre of culture after purification (Fig. 6.11), a more than 10-fold increase from previous attempts.

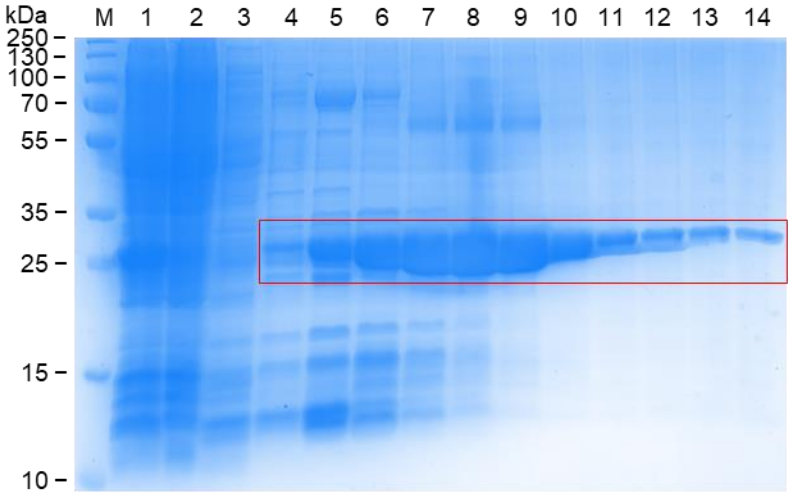

Fig. 6.11: SDS-PAGE analysis of CPR-C4 IMAC purification after overexpression with 100 µM $ZnCl_2$ in media. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 14 = protein fractions. Bands corresponding to CPR-C4 are indicated with a red box.

The increase in soluble protein yield resulting from these small alterations to expression and purification procedures after TSA was critical for downstream experiments, and further demonstrates the stabilising effect of Zn^{2+} on CPR-C4.

6.2.4. Production of CPR-C4 using an alternative expression construct

CPR-C4 was later also produced using a second expression construct: pET28a(+)TEV-*CPRC4*. This construct generated a CPR-C4 fusion protein with an N-terminal His-tag for purification, and a TEV protease recognition sequence that then enabled the tag to be removed.

Expression and IMAC purification

This new construct was transformed into competent T7 Express *E. coli* cells to allow CPR-C4 overexpression by induction with IPTG. Small scale test expressions were successful, with clear overexpression of CPR-C4 on addition of 0.1 M IPTG to cells and no basal expression in noninduced cells (Fig. 6.12).

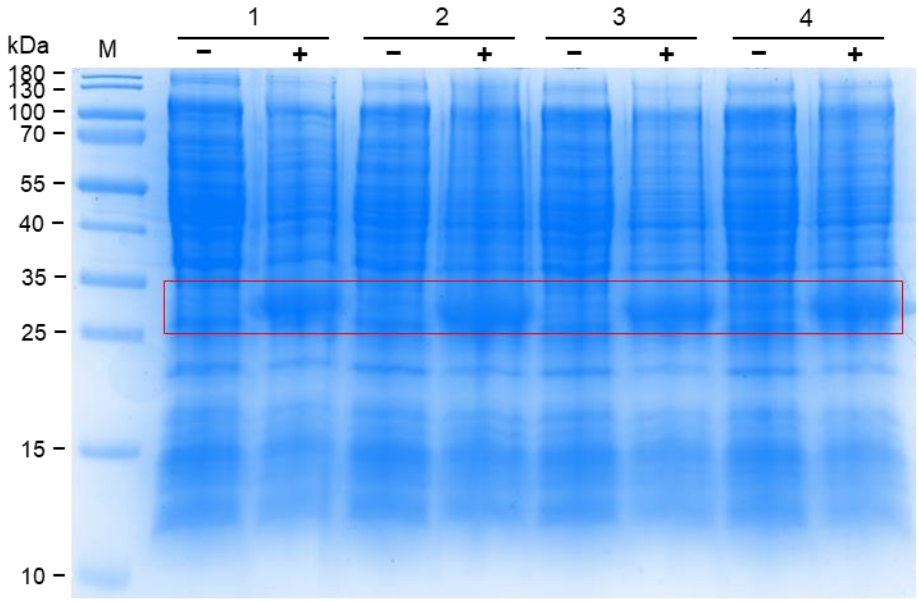

Fig. 6.12: Whole cell SDS-PAGE results of CPR-C4 expression tests with the pET28a(+)TEV-*CPRC4* construct. Noninduced (−) and induced samples with 0.1 M IPTG (+) from four tests 1 - 4 are shown. The red box highlights the positions of bands corresponding to CPR-C4 (MW = 27.1 kDa). M = marker, with MWs in kDa to the left.

Subsequent full-scale expression experiments produced > 10 mg per litre of culture, which was a stark contrast from the < 1 mg produced during original trials with the pJOE5751.1-*CPRC4* construct. It was also found that, unlike with protein from the pJOE5751.1-*CPRC4* construct (6.2.3.), the addition of $ZnCl_2$ to expression media had little to no effect on the quantity of CPR-C4 produced (Fig. 6.13). With ample protein for downstream characterisation and crystallisation, optimisation of production protocols was not necessary in this case.

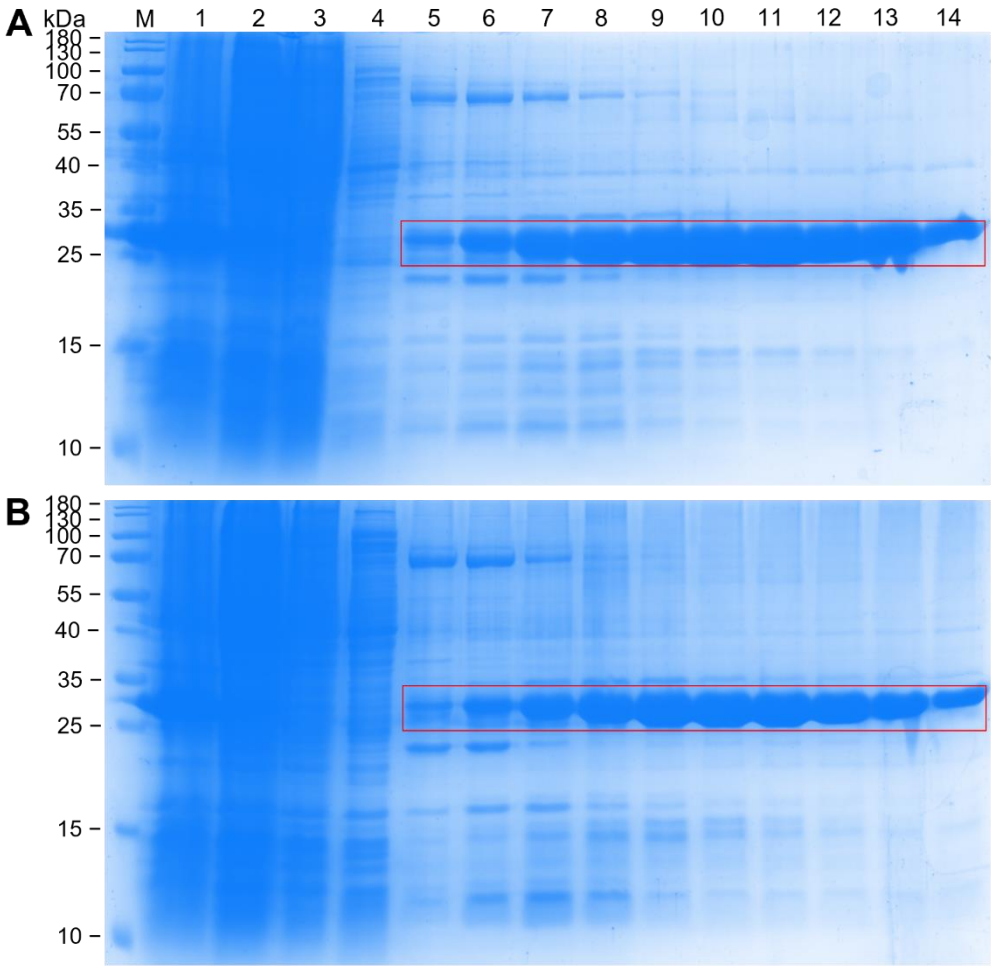

Fig. 6.13: SDS-PAGE analysis of CPR-C4 IMAC purification experiments. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 14 = protein fractions. Bands corresponding to CPR-C4 are indicated with a red box. **(A)** Protein expressed from the pET28a(+)TEV-*CPRC4* construct without addition of $ZnCl_2$ to expression media. **(B)** Protein expressed from the pET28a(+)TEV-*CPRC4* construct with 100 µM $ZnCl_2$ added to expression media.

Following the initial IMAC purification of the tagged protein (Fig. 6.13), the affinity tag was removed using Tobacco Etch Virus (TEV) protease. The tag-free protein then needed to be separated from the tag and any uncleaved protein remaining in the sample. To achieve this, the mixture was run back over the Ni^{2+} affinity column in imidazole-free buffer to elute cleaved protein, followed by high imidazole buffer (500 mM) to remove the tag and any uncleaved protein. However, a large amount of tag-free CPR-C4 remained bound to the column even with addition of 80 mM imidazole (Fig. 6.14), which made purification difficult.

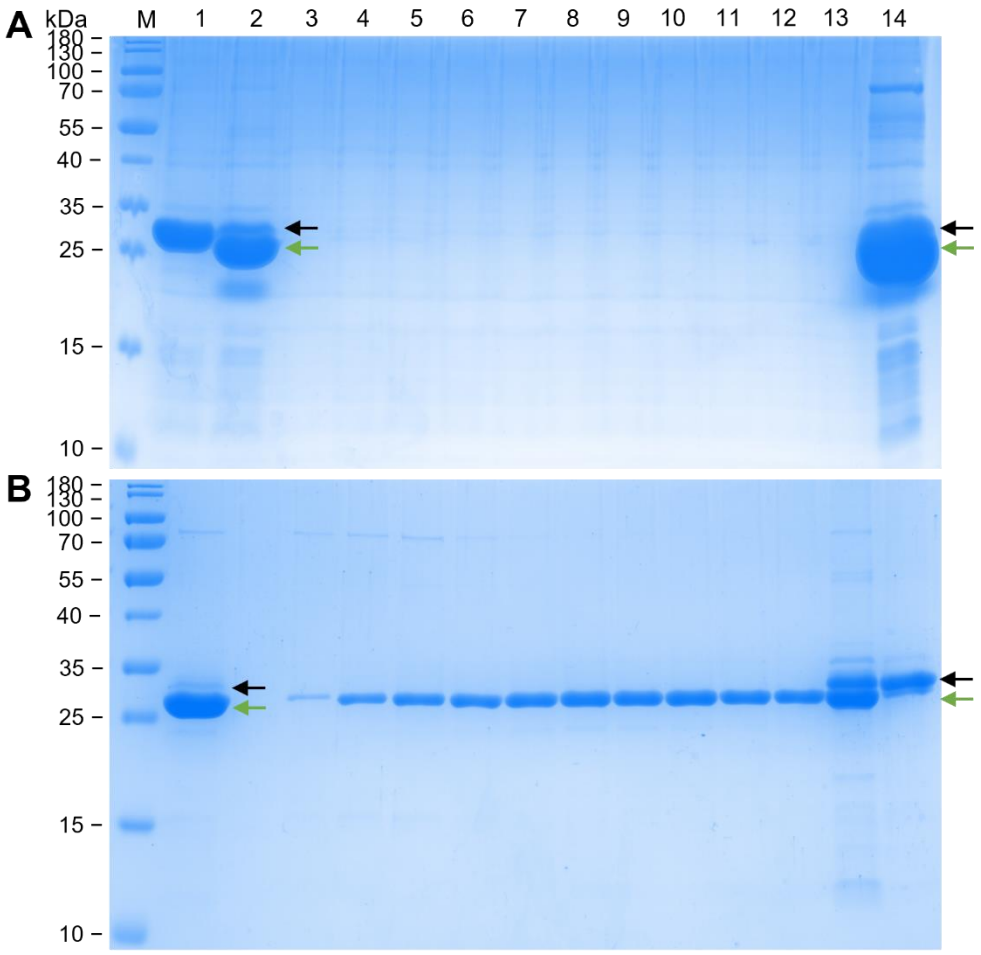

Fig. 6.14: SDS-PAGE analysis of CPR-C4 IMAC purification following tag removal with TEV protease. Successfully cleaved protein is indicated with a green arrow; uncleaved protein is indicated with a black arrow. M = marker, with MWs of bands on the left. **(A)** IMAC with imidazole-free elution buffer: 1 = sample prior to cleavage; 2 = sample after cleavage; 3 – 12 = eluted fractions with 0 mM imidazole; 13 – 14 = eluted fractions with 500 mM imidazole. **(B)** IMAC with 80 mM imidazole: 1 = sample after cleavage; 2 – 11 = eluted fractions with 80 mM imidazole; 12 – 14 = eluted fractions with 500 mM imidazole.

An imidazole gradient was subsequently used to identify an appropriate concentration to fully elute the cleaved protein separately from the uncleaved protein and tag, with the resulting SDS-PAGE analysis shown in Fig. 6.15.

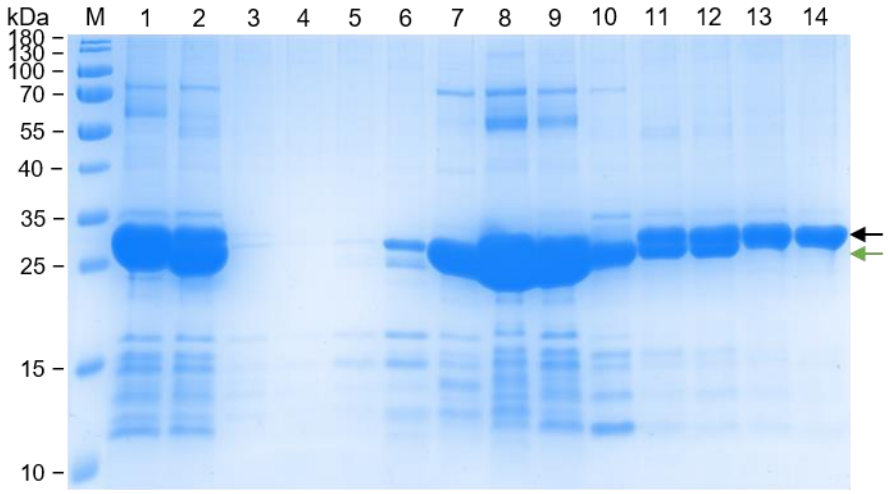

Fig. 6.15: SDS-PAGE analysis of IMAC separation of cleaved and uncleaved CPR-C4 protein. M = marker, with MWs of bands on the left; 1 = sample prior to cleavage; 2 = sample after cleavage; 3 = flow through; 4 = wash; 5 – 14 = protein fractions. The position of bands corresponding to successfully cleaved protein is indicated with a green arrow; uncleaved protein bands are indicated with a black arrow.

This approach was more successful, with most of the cleaved protein eluting independently at 150 – 200 mM imidazole (lanes 6 to 10, Fig. 6.15). However, some remained unseparated from the tagged protein, eluting at 300 – 350 mM imidazole (lanes 11 and 12). This non-specific binding of CPR-C4 to the Ni^{2+} affinity column could be due to its relatively high histidine content (5%), which typically only accounts for 2% of protein residues. It was also later determined that the protein contains an intrinsic metal binding site (6.2.6.) that likely interacts with the immobilised Ni^{2+} during purification.

Characterisation

ESI-TOF MS analysis of the cleaved protein sample from the pET28a(+)TEV-*CPRC4* construct demonstrated that the species present was exactly the MW predicted for tag-free CPR-C4, with a dimer species also present (Fig. D2). Its identity was also confirmed through trypsin digest MS.

SEC analysis

The oligomeric state of the CPR-C4 from the pET28a(+)TEV-*CPRC4* construct was assessed by SEC, with the resulting chromatogram shown in Fig. 6.16.

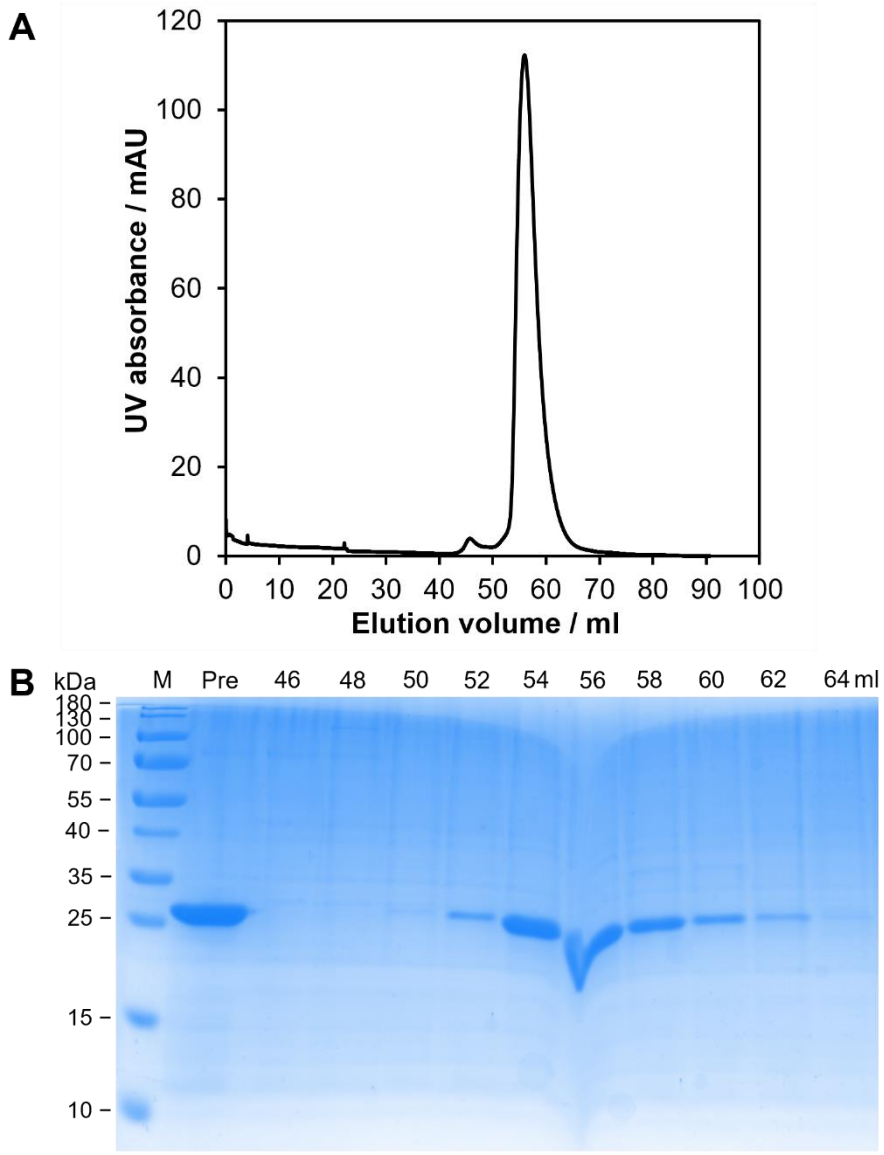

Fig. 6.16: (A) Chromatogram from SEC of CPR-C4 from the pET28a(+)TEV-*CPRC4* expression construct. **(B)** SDS-PAGE analysis of elution fractions from SEC. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The tag-free CPR-C4 was found to elute as a single species at V_e 56.0 ml, corresponding to a 54.6 kDa species, which is the expected MW of the CPR-C4 dimer. As with the protein produced using the original pJOE5751.1-*CPRC4* construct, the ESI-TOF and SEC results therefore suggest that the CPR-C4 protein produced with the pET28a(+)TEV-*CPRC4* construct is dimeric.

6.2.5. Structure determination of CPR-C4 by X-ray crystallography

Crystallisation screening

Determination of the function of CPR-C4 was pursued through structural analysis with X-ray crystallography. Initial high-throughput crystallisation screening using a range of commercially available screens was conducted with CPR-C4 produced using the pJOE5751.1-*CPRC4* construct, at 4.7 mg/ml and of purity shown in Fig. 6.17.

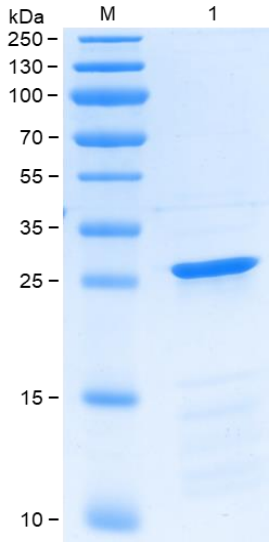

Fig. 6.17: SDS-PAGE analysis showing the purity of CPR-C4 used for initial crystallisation screening experiments. M = marker, with MWs of bands on the left; 1 = CPR-C4 in SGC buffer for crystallisation.

This screening was successful, with CPR-C4 crystallising in two space groups with different crystal morphologies.

Crystal form 1

Crystal growth and optimisation

CPR-C4 crystal form 1 grew quickly across 25% of conditions in the Morpheus® HT-96 Screen (Molecular Dimensions), with dodecahedral morphology. For several of these conditions, crystals could be looped directly from the 96-well plate for diffraction data collection. However, the crystals were small and difficult to mount, requiring optimisation (Fig. 6.18). Morpheus® HT-96 Screen conditions are particularly difficult to adjust as they are comprised of many components and proprietary blended buffer conditions. To improve crystal quality, manual screening using 24-well plates and larger drop volumes was conducted without any change to the Morpheus® Screen conditions (Table D1 and Table D2). These larger drop volumes (3 - 4 µl for 24-well plates *vs* 200 - 300 nl for 96-well plates) typically allow larger crystals to form, which are easier to mount and can give better diffraction[270].

The manual optimisation was a success, producing larger crystals as shown in Fig. 6.18. The form 1 crystals were found to be readily reproducible, though took longer to form in the manual 24-well trays than during high-throughput screening due to the slower vapour diffusion from the larger drop volumes.

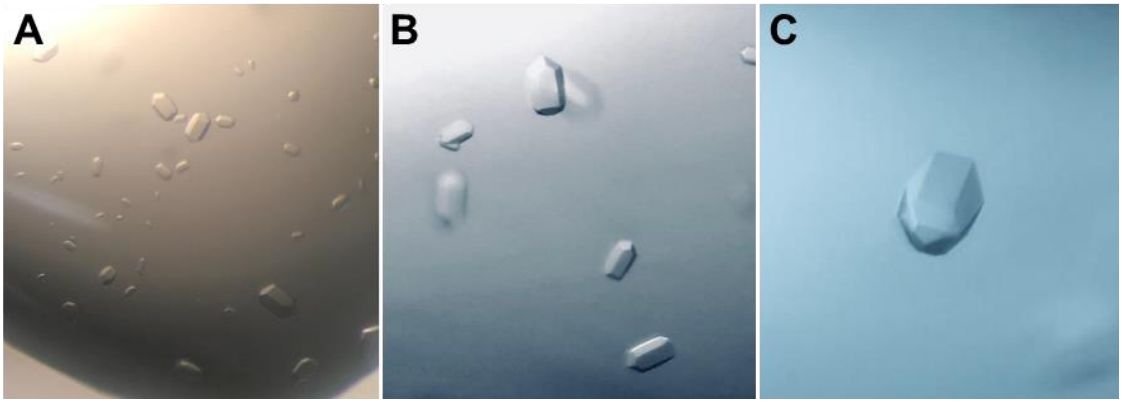

Fig. 6.18: Morphology of CPR-C4 crystal form 1. **(A)** Crystals from initial high-throughput screening with the Morpheus® HT-96 screen. The largest crystals are approximately 70 µm at their widest dimension. **(B)** Optimised crystals from in manual 24-well sitting drop trays with the Morpheus® Screen, approximately 100 µm at their widest dimension. **(C)** Optimised crystal from manual 24-well hanging drop trays using the Morpheus® Screen, approximately 120 µm at its widest dimension.

The form 1 crystals were found to be fragile and frequently attached to well surfaces of sitting drop trays. To resolve this issue, manual screening was also conducted using 24-well hanging drop trays. This alternate set up had no obvious effect on crystal growth but did make manipulation of the crystals considerably easier.

Diffraction data collection

Due to the high concentrations of glycerol and PEG varieties in the Morpheus® HT-96 Screen conditions, no additional cryo-protectant was used prior to cryo-cooling crystals for X-ray diffraction. However, the crystal quality was found to be poor, with significant amounts of ice present on loops. Crystals were subsequently cryo-protected in 50% glycerol prior to freezing, which resulted in reduced ice formation. Four datasets sets were obtained for the crystal shown in Fig. 6.18C, which is shown looped for X-ray diffraction data collection at Diamond Light Source (DLS) in Fig. 6.19.

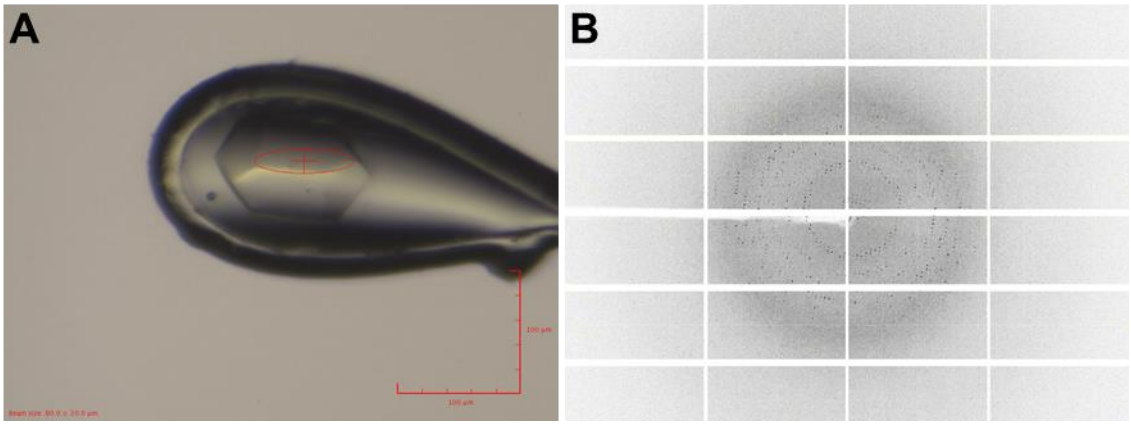

Fig. 6.19: (A) CPR-C4 crystal form 1 mounted on a nylon loop for X-ray diffraction data collection at the I03 beamline, DLS. Beam size (80 µm x 20 µm) and position are indicated with a red oval and cross; 100 µm scale bars are shown in the bottom right corner. **(B)** Resulting diffraction image collected during crystal screening at a wavelength of 1.2824 Å.

Form 1 crystals diffracted to a maximum resolution of 2.6 Å. This resolution is outside the limits of *ab-initio* structure solution programs such as ARCIMBOLDO[271], and the lack of detectable homology prohibited MR methods. X-ray fluorescence scans identified an unambiguous zinc signal from the crystal, which presented the possibility of phasing via single- or multiple-wavelength anomalous dispersion (SAD/MAD) techniques. Zinc is an ideal anomalous scatterer for such phasing methods, with a strong anomalous signal at its K edge (f^2 = 3.9 e) and an absorbance edge (1.284 Å) within the energy range of typical macromolecular crystallography beamlines[142]. A subsequent X-ray fluorescence scan at the zinc edge was used to select appropriate wavelengths to optimising the anomalous signal (Fig. 6.20).

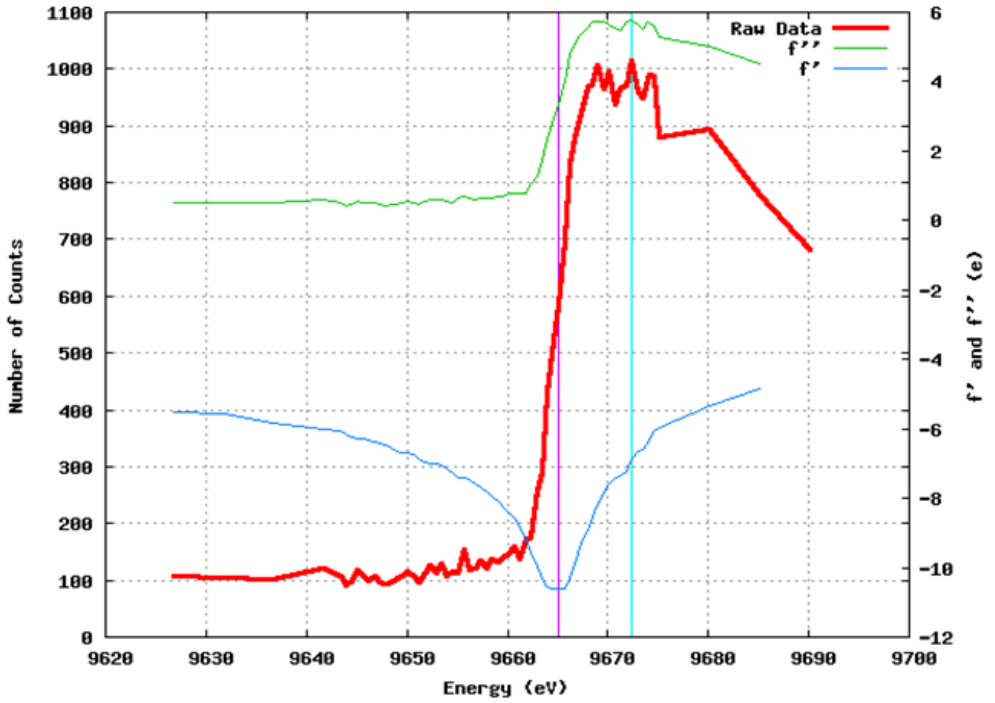

Fig. 6.20: X-ray fluorescence scan at the zinc edge of a CPR-C4 form 1 crystal. The raw data trace is shown in red; the f'' curve is shown in green; the f' curve is shown in blue. The peak energy is indicated with a cyan vertical line; the inflection energy is indicated with a pink vertical line[60].

After collection of a high energy data set at a wavelength of 0.9763 Å with 20% transmission, further datasets were collected at the zinc peak (9668 eV, 1.2824 Å), edge (9659 eV, 1.2836 Å), and inflection point wavelengths (9665 eV, 1.2828 Å) with reduced transmission (10%) to limit radiation damage and enable collection of multiple datasets on one crystal.

Initial data processing and phasing attempts

The processing of diffraction data for CPR-C4 crystal form 1 was performed manually using XDS[258]. This allowed resolution cut-offs to be optimised to maximise the prospect of structure solution for this novel protein from the datasets collected.

The best resolution (2.60 Å) and strongest anomalous signal came from the peak dataset (9668 eV, 1.2824 Å), from which the zinc substructure was solved with the SHELX programme suite[272]. SHELXD[273] successfully identified two zinc sites in the substructure; however, density modification and auto-tracing with SHELXE[274] was ineffective and the data proved to be insufficient for structure solution beyond the zinc substructure.

Many approaches were taken to improve the quality of the CPR-C4 crystals and increase the resolution of diffraction data. These included wider screening with commercially available screens, the use of microseed matrix screening (MMS)[275,276] and classical seeding, as well as varying protein concentration, drop volumes, and experimental set-up (hanging *vs* sitting drop methods). In total, 102 CPR-C4 crystals were screened at DLS for diffraction, with 25 datasets collected. Despite the numerous optimisation attempts, the resolution of diffraction data could not be improved. Further efforts were therefore made to improve data processing with the highest quality dataset in the hopes of obtaining a structure solution.

Structure solution with anisotropic corrections

Automatic data processing by autoPROC + STARANISO[256,257] through ISPyB was able to extend the resolution limit of the peak dataset to 2.32 Å, which ultimately enabled the structure to be solved. The autoprocessed peak data, along with the best native dataset, were fed into the CRANK2 pipeline[277] in the CCP4i2 suite[278], which used the SHELX programmes[272] for automatic phasing by MAD. The Matthews coefficient calculated for the crystal system pointed to three molecules in the asymmetric unit (ASU) as being most probable. However, the presence of two zinc sites identified through SHELXD, along with the extremely delicate nature of the crystals and implied high solvent content, strongly suggested a composition of two molecules per ASU. This proved to be correct: CPR-C4 form 1 was solved in space group $P3_221$, with two Zn^{2+} ions in the ASU and two protein chains generating a dimer with two-fold rotational non-crystallographic symmetry (rNCS). The rNCS was exploited through density modification with Parrot[279], using NCS restraints[280] and a forced high solvent content of 69% based on the Matthews coefficient for two molecules in the ASU. This was enormously beneficial, improving the map and allowing both chains to be built. The density for chain A, which was built first, was notably better defined, with the side chains of His-tag residues interpretable. Data collection, processing, and refinement statistics are reported in Table D4.

Crystal form 2

Crystal growth and optimisation

Over a six-month period, crystals grew in a new form with cubic rod morphology in a preliminary screening experiment (Fig. 6.21A). These crystals were small, and so attempts were made to optimise them through manual out-screening of the crystallisation conditions by varying the pH and reagent concentrations (Table D3).

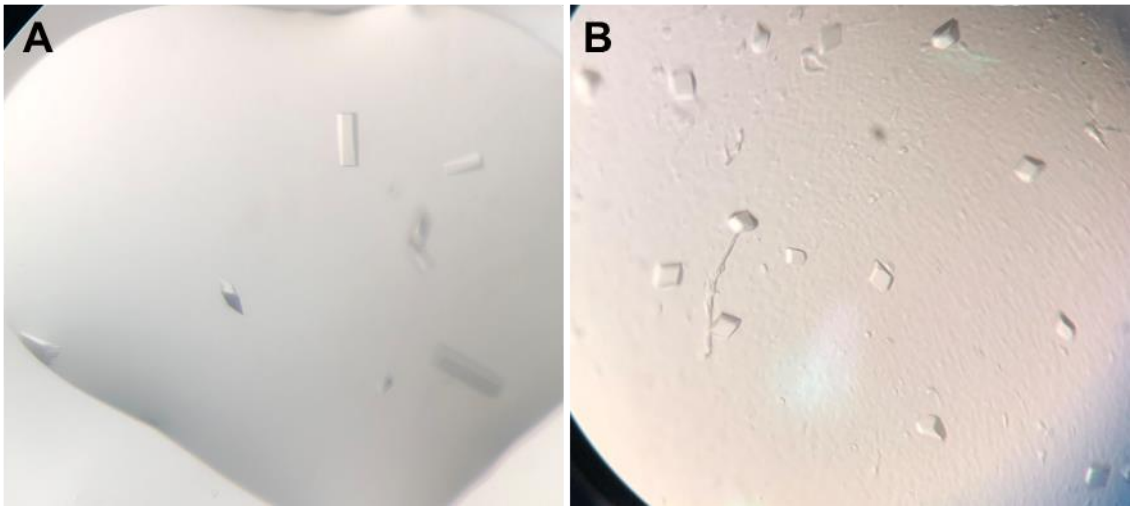

Fig. 6.21: Morphology of CPR-C4 crystal form 2. **(A)** Crystals from initial high-throughput screening with the JCSG-plus™ HT-96 ECO Screen; the largest crystals are approximately 50 µm at their widest dimension. **(B)** Crystals from manual out-screening attempts in 24-well sitting drop trays, approximately 30 µm at their widest dimension.

The form 2 crystals grew slowly and were difficult to reproduce, only forming in one of 24 conditions during out-screening (Fig. 6.21B; Table D3). Fortunately, they were more robust with a suspected lower solvent content than the form 1 crystals and could be looped and cryo-cooled for diffraction data collection despite their small size.

Diffraction data collection

Due to the smaller size of the form 2 crystals, data collection was carried out using the microfocus beamline (I24) at DLS (Fig. 6.22). It was initially presumed that these crystals would also contain zinc, and so a strategy was set up for the collection of a MAD dataset. An X-ray fluorescence scan at the zinc edge did identify a zinc signal, although much weaker than for crystal form 1.

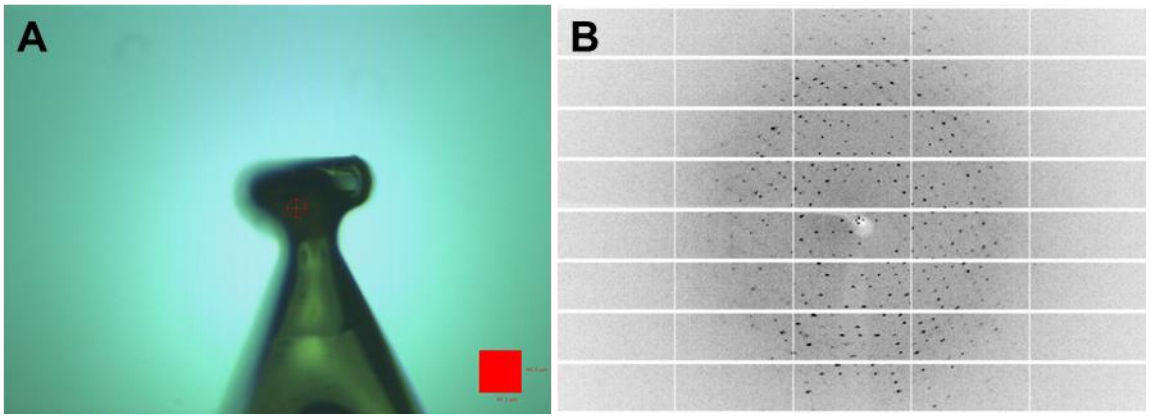

Fig. 6.22: (A) CPR-C4 crystal form 2 mounted on a Kapton loop for X-ray diffraction data collection at the I24 beamline, DLS. Beam size and position is indicated with a red circle and cross; 46.1 µm x 46.4 µm scale square is shown in the bottom right corner. **(B)** Resulting diffraction image collected during crystal screening at a wavelength of 1.2819 Å.

Due to a technical error during data collection, only 360 images (0.10° oscillation) were initially collected at the native (0.9686 Å), inflection point (1.2810 Å), and peak (1.2819Å) wavelengths, rather than the intended 3600 images. Autoprocessing was successful, with each pipeline on ISPyB selecting space group I222 and predicting a maximum resolution of 2.1 Å with xia2 DIALS[281], but with unsurprisingly low completeness (~ 75%) and low multiplicity (< 2). A second MAD dataset was therefore collected from the same crystal, though signs of radiation damage were apparent, and only the first full dataset collected was of high enough quality for downstream processing. Regardless, this was the best crystal form 2 dataset collected, and issues with reproducing crystals prohibited subsequent data collection. The xia2 DIALS pipeline again selected space group I222 and predicted a resolution limit of 2.36 Å with 100% completeness.

Downstream processing was done manually using the xia2 DIALS user interface (DUI), with a high-resolution cut-off of 2.58 Å and space group I222. However, it was later noted that the overall R_{merge} value was unusually high (0.505). The resolution was subsequently cut back to 2.68 Å, though this only slightly improved the R_{merge} (0.453). R_{merge} is dependent on multiplicity, which in this case is high (9.7). R_{pim}, which is independent of multiplicity, was therefore used as a better indicator of the dataset quality (R_{pim} = 0.153)[282].

Phasing and structure solution

As with crystal form 1, the resolution was below that typically required for success with *ab-initio* programmes such as ARCIMBOLDO[271]. There was also a much weaker anomalous signal in comparison to crystal form 1, and initial attempts to solve the structure by zinc phasing were unsuccessful. However, once the structure of CPR-C4 was solved in the original crystal form 1, the structure for this second crystal form in space group I222 could readily be solved by MR. Unlike form 1, no Zn^{2+} ions and only one polypeptide chain were present per ASU.

Crystal form 3

Crystal growth

Subsequent to solving the structure of CPR-C4 in crystal form 1 and 2, crystallisation was also conducted using protein without the His-tag, produced using the pET28a(+)TEV-*CPRC4* construct (6.2.4.). This provided the opportunity to assess the influence of the non-native His-tag, as there have been instances of such tags interfering with crystallisation of proteins[221,283,284].

Crystallisation experiments were conducted after removal of the His-tag and subsequent purification by IMAC and SEC. The purity of CPR-C4 used in crystallisation screening is shown in Fig. 6.23.

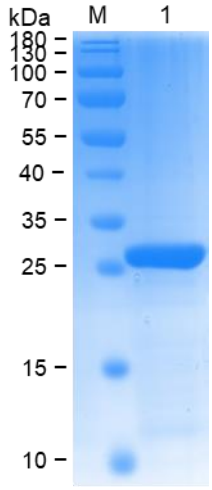

Fig. 6.23: SDS-PAGE analysis showing the purity of CPR-C4 produced using the pET28a(+)TEV-*CPRC4* construct used for initial crystallisation screening experiments. M = marker, with MWs of bands on the left; 1 = CPR-C4 for crystallisation.

The high-throughput screening resulted in a third crystal form, obtained from a 6 mg/ml protein solution with 0.2 M lithium sulfate, 0.1 M Tris pH 8.5, 15% w/v PEG 4000. These new crystals were plate-like and grew rapidly in a variety of clusters, as shown in Fig. 6.24.

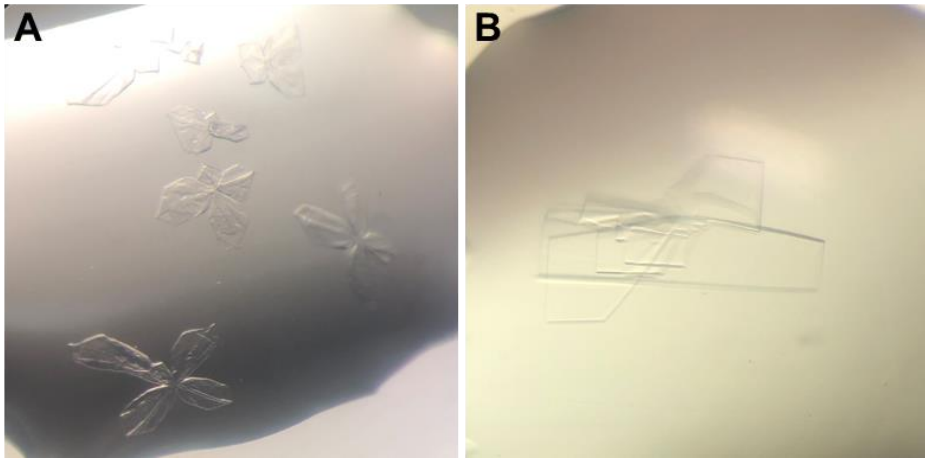

Fig. 6.24: Morphology of CPR-C4 crystal form 3. **(A)** Crystals from initial high-throughput screening with the Morpheus® HT-96 Screen, forming flower-like clusters. **(B)** Cluster of plates from initial high-throughput screening.

These stacks of thin plates were not ideal for X-ray diffraction analysis, as such crystals are often difficult to handle. However, the crystals were large enough to allow suitable single plates to be isolated for harvesting and cryo-cooling.

Diffraction data collection

The new crystals from the pET28a(+)TEV-*CPRC4* construct were successfully looped and cryo-protected for diffraction data collection at DLS (Fig. 6.25).

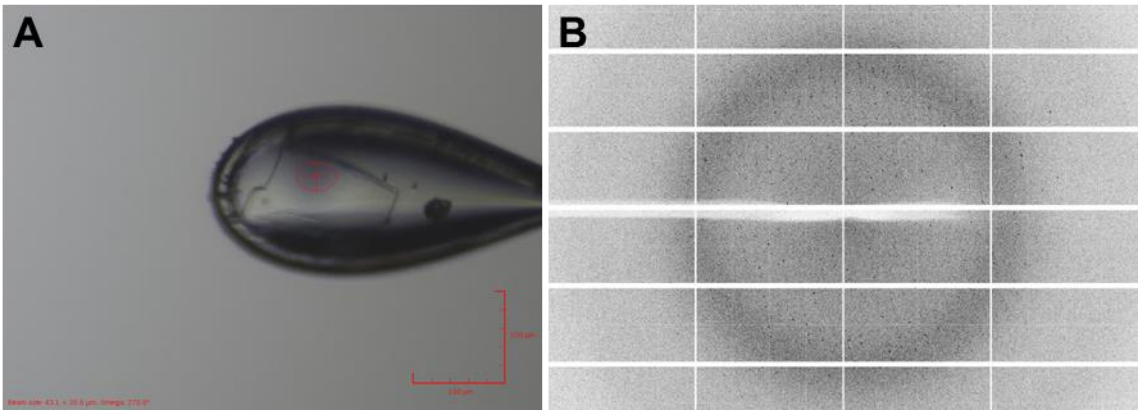

Fig. 6.25: (A) CPR-C4 crystal form 3 mounted on a nylon loop for X-ray diffraction data collection at the I04 beamline, DLS. Beam size (43.1 µm x 30.0 µm) and position is indicated with a red oval and cross; 100 µm scale bars are shown in the bottom right corner. **(B)** Resulting diffraction image collected during crystal screening at a wavelength of 0.9795 Å.

An X-ray fluorescence spectrum scan and UV fluorescence scan at the zinc edge identified that zinc was not present in the form 3 crystals. Collection of a MAD dataset, as was collected from the zinc-containing form 1 crystals, was therefore redundant and data was instead collected at the native wavelength (0.9795 Å) with higher dosage to maximize the resolution[285].

Autoprocessing was successful, with all but one pipeline selecting space group $P2_12_12_1$. The diffraction data were manually processed using XDS[258] and a high-resolution cut-off of 2.25 Å was determined based on $CC_{1/2} > 0.3$ and $I/\sigma(I) > 1.0$ in the outer shell. This was a considerable improvement on the form 1 and form 2 data (2.60 Å and 2.68 Å, respectively).

Phasing and structure solution

With the structure of CPR-C4 already solved in the previous crystal forms, MR with Phaser[261] using form 1 as a homology model was performed, in space group $P2_12_12_1$ and with two protein chains per ASU. The improved resolution of the form 3 structure subsequently enabled the geometries of the other models to be improved, reducing the numbers of Ramachandran outliers from 7 to 0 in form 1, and from 3 to 1 in form 2.

6.2.6. Crystal structures of CPR-C4

Overall structure of the CPR-C4 homodimer

In all three crystal forms, the structure of CPR-C4 is a homodimer displaying a mixed αβ fold, with short α-helices surrounding a small central region of curved anti-parallel β-sheet. Crystal forms 1 and 3 contain two protein chains in the ASU, forming the functional dimer. Whilst form 2 only has a single chain in the ASU, the dimer is formed with a neighbouring symmetry mate with crystallographic symmetry about a 2-fold axis. The three structures in the different crystal forms superimpose very well, with no significant deviation (Fig. 6.26, Table 6.1).

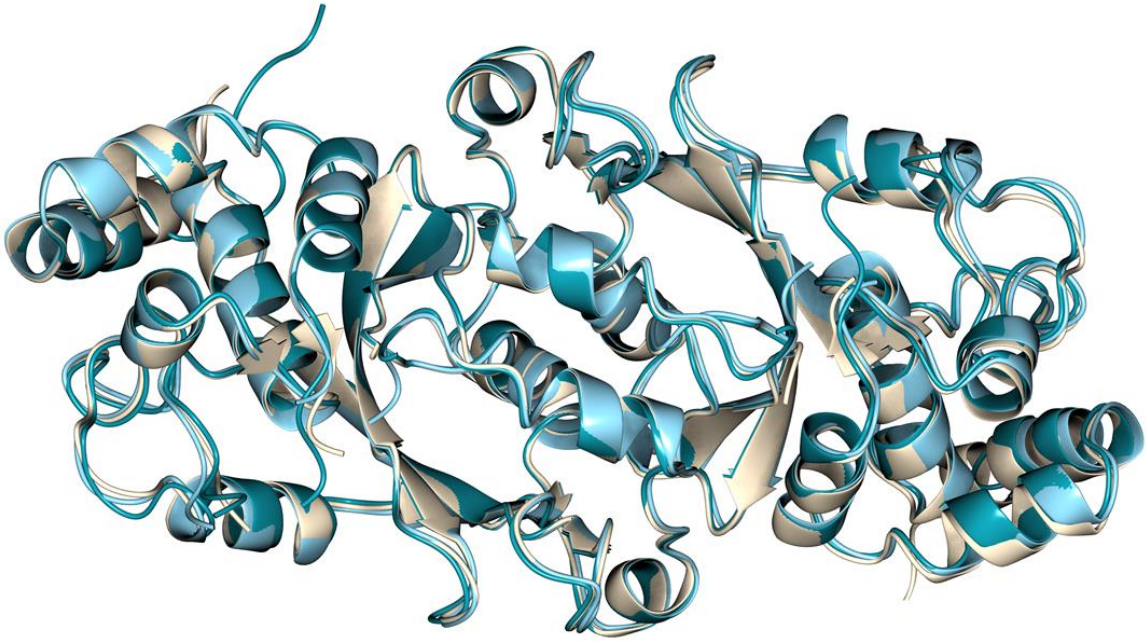

Fig. 6.26: Ribbon diagram of the least-squares superposition of CPR-C4 dimers from the three crystal forms: form 1 shown in teal, form 2 in light blue, and form 3 in white; Zn^{2+} ions found in form 1 chain A only have been omitted for clarity. Used under the terms of the Creative Commons CC-BY license[60].

Table 6.1: Matrix of C_α RMSD values in Å between the CPR-C4 chains from the three crystal forms; numbers of residues aligned are shown in parentheses[60].

	7OB6 chain A	**7OB6 chain B**	**7OB7**	**7PJO chain A**	**7PJO chain B**
7OB6 chain A		**0.52 (213)**	**0.37 (209)**	**0.42 (211)**	**0.64 (213)**
7OB6 chain B	**0.52 (213)**		**0.27 (209)**	**0.22 (211)**	**0.25 (211)**
7OB7	**0.37 (209)**	**0.27 (209)**		**0.31 (209)**	**0.31 (209)**
7PJO chain A	**0.42 (211)**	**0.22 (211)**	**0.31 (209)**		**0.27 (215)**
7PJO chain B	**0.64 (213)**	**0.25 (211)**	**0.31 (209)**	**0.27 (215)**	

It is clear from the structural superposition that there are no major deviations in overall fold between the three crystal forms, with tightly conserved secondary structure features and only slight variations in flexible loop regions. A particularly disordered loop between residues 38 and 45 could not be built, even in the highest resolution form 3 structure.

Dimer interface analysis

To evaluate whether the CPR-C4 dimer observed in the crystal structures is the biological unit of the protein responsible for function, interface, and assembly analysis with PISA[265] was conducted. The

dimer showed an interface area of 1669.1 Å^2 per monomer (15%) with a binding energy of -26.2 kcal mol^{-1} (crystal form 1), which is strong evidence that the interface is part of the natural biological assembly of the protein. This is supported by SEC analysis (Fig. 6.4 and Fig. 6.16). The functional unit of CPR-C4 *in vivo* is therefore highly likely to be a dimer.

6.2.7. Structural analysis of the CPR-C4 Zn^{2+} binding sites

Zn^{2+} ions are known to play vital roles in numerous metalloproteins, with both structural and catalytic functions[286]. To assess the potential role of the Zn^{2+} identified in the CPR-C4 structure, the binding site residue identities and geometries were analysed.

The positions of the Zn^{2+} ions in CPR-C4 were independently determined from anomalous dispersion in the course of solving the structure. Surprisingly, the two Zn^{2+} ions identified in crystal form 1 were both found to interact with the same chain (chain A) at two distinct sites (Fig. 6.27). Chain A also shows slightly greater structural deviation from the other crystal forms than chain B where Zn^{2+} is absent (Table 6.1), which suggests that Zn^{2+} binding induces marginal differences in structure.

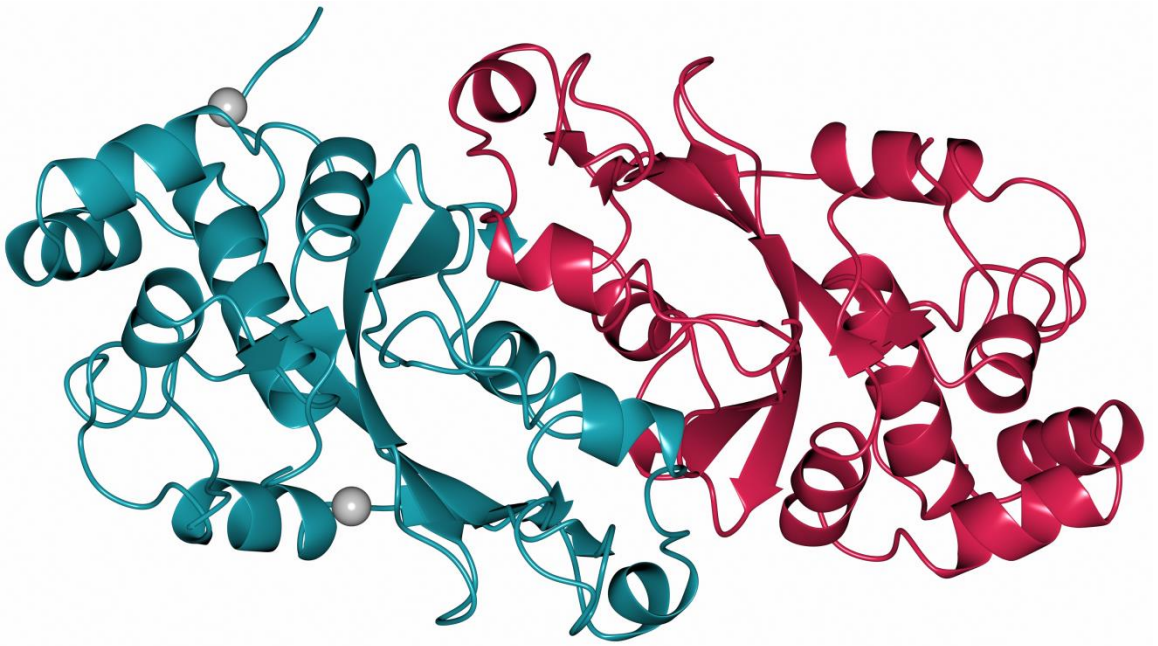

Fig. 6.27: Ribbon diagram of the CPR-C4 dimer (crystal form 1) showing the two molecules in the ASU perpendicular to the rNCS axis; chain A is shown in teal, chain B in crimson; Zn^{2+} ions are shown as grey spheres. Used under the terms of the Creative Commons CC-BY license[60].

The Zn^{2+} ions are both bound in a tetrahedral geometry, as is most common for Zn^{2+} coordination[287]. The first is coordinated by three aspartate side chains (D83, D92, D182) and a water molecule (Fig. 6.28A), and the second is coordinated by two histidine residues from the N-terminal His-tag (H-7 and H-5), another histidine (H75), and a glutamate residue (E179) from a neighbouring symmetry mate (Fig. 6.28B), forming a crystal contact. Since this second Zn^{2+} site is formed through the non-native His-tag, it was not considered to be important for the structure or function of CPR-C4.

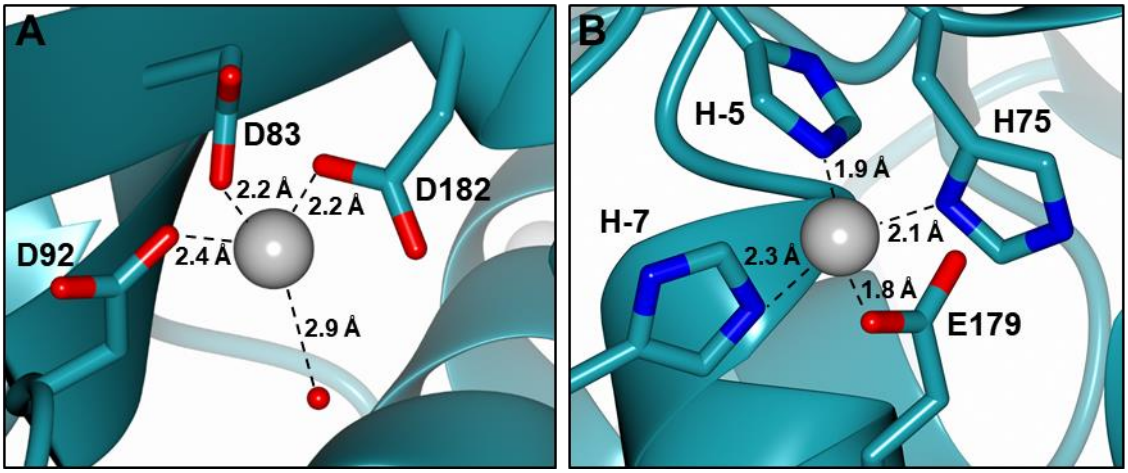

Fig. 6.28: Close-up views of the CPR-C4 Zn^{2+} binding sites. Distances in Å are indicated with dashed lines; side chains are shown with cylinder representation, O atoms in red, N atoms in blue; the Zn^{2+} ions are shown as grey spheres with 0.5 van der Waals radii. **(A)** The tri-aspartate Zn^{2+} binding site, showing tetrahedral coordination to three aspartate side chains and a water molecule; the coordinated water molecule is shown as a red sphere[60]. **(B)** The Zn^{2+} site formed through the non-native His-tag, showing tetrahedral coordination to three histidine side chains from chain A and a glutamate side chain from a neighbouring symmetry mate with symmetry operator -X+Y, -X, Z+1/3.

The presence of a well-defined water molecule in the tri-aspartate Zn^{2+} site is indicative of a catalytic over a structural Zn^{2+} site, where the metal coordination sphere is typically comprised entirely of protein side chains[287]. However, the arrangement and identities of the coordinating residues are unusual. In the binding sites of classic catalytic Zn metalloproteins, the first two residues are only one to three amino acids apart in the primary sequence, with much greater variation in the separation between the second and third coordinating residue[286]. In the case of CPR-C4, the first two residues (D83 and D92) are separated by eight residues, and the third (D182) is relatively distant. The tri-aspartate arrangement is also atypical, with histidine accounting for most residues coordinated to Zn^{2+} in catalytic sites, and cysteine in structural sites. When considering the makeup of the binding site alone, CPR-C4 therefore appears to contain an unusual but possible catalytic Zn^{2+} site.

Despite the lack of Zn^{2+} in crystal forms 2 and 3, the key aspartate residues superimpose well across all three structures and the geometry of the binding site is maintained even in the absence of Zn^{2+} (Fig. 6.29).

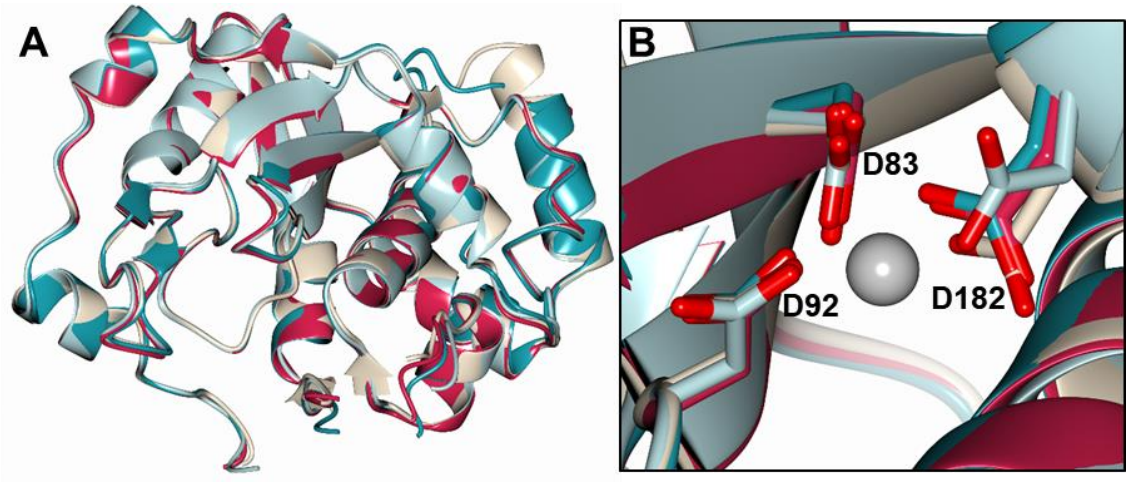

Fig. 6.29: (A) Least-squares superposition of CPR-C4 chains: form 1 chain A shown in teal, form 1 chain B in crimson, form 2 in light blue, and form 3 chain A in white; Zn^{2+} ions found in form 1 chain A only have been omitted for clarity. **(B)** Least-squares superposition of the tri-aspartate Zn^{2+} site across the CPR-C4 chains coloured as in (A); Zn^{2+} ions from form 1 chain A are shown as grey spheres with 0.5 van der Waals radii. Used under the terms of the Creative Commons CC-BY license[60].

The unusually tight binding of CPR-C4 to Ni^{2+} affinity columns in the absence of a His-tag (6.2.3.) suggests that this site can bind Ni^{2+}. It is also possible this is a cation binding site that could contain other metal ions from buffer or crystallisation solutions, such as Na^{+} and K^{+}. However, Fig. 6.30 highlights that no other metal cations are bound in the absence of Zn^{2+}, implying this is not a promiscuous site.

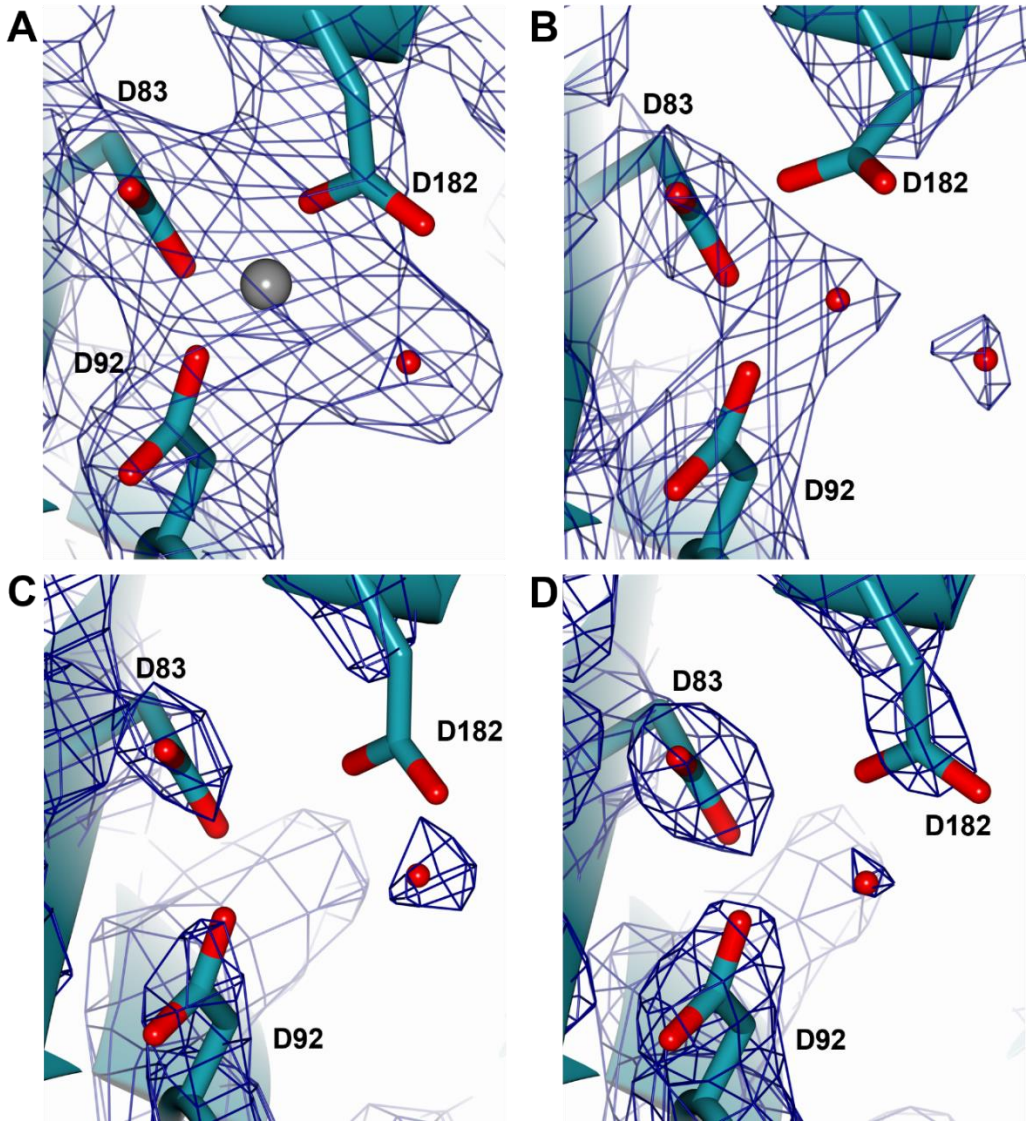

Fig. 6.30: Comparison of electron density at the tri-aspartate Zn^{2+} binding site between the CPR-C4 crystal forms. In all panels the main chain is shown with cylinder representation in teal, O atoms in red, N atoms in blue; water molecules are shown as red spheres. The $2F_o$-F_c map is shown in dark blue with chickenwire representation at contour level $\sigma = 1.5$. This high sigma was chosen to make clear the position of Zn^{2+} and coordinating residues in form 1 chain A. At $\sigma = 1.0$, D182 is well fitted in density across all three crystal forms. **(A)** Electron density for crystal form 1 chain A showing the location of bound Zn^{2+}, depicted as a grey sphere, with coordinating water molecule and aspartate residues. **(B)** Electron density for form 1 chain B. **(C)** Electron density for form 2. **(D)** Electron density for form 3 chain A. Used under the terms of the Creative Commons CC-BY license[60].

The tetrahedral ligand geometry is also strongly preferable for Zn^{2+} binding over other transition metal cations or Na^+/K^+[288,289]. Additionally, the Zn^{2+} identified in the form 1 structure was incorporated into the protein without being supplemented in expression media, buffers, or crystallisation reagents, and TSA showed greater stabilisation upon the addition of Zn^{2+} than with other metal ions (6.2.3.). This evidence collectively supports the observation that this is an intrinsic Zn^{2+} binding site in CPR-C4.

6.2.8. Structure analysis of CPR-C4 with 3DM

Identification of structural homologs

Whilst homologs could not be identified from the CPR-C4 sequence alone, the availability of the CPR-C4 structure enabled 3D comparisons to be conducted using 3DM[269]. Surprisingly, only two proteins were identified through structural alignments as having significant structural homology to CPR-C4: the human vasohibins VASH1 and VASH2. These protein isoforms are carboxypeptidases responsible for cleaving the C-terminal tyrosine of α-tubulin during microtubule regulation, in concert with the α-helical Small Vasohibin Binding Protein (SVBP)[290–293]. VASH1 and VASH2 are part of the transglutaminase-like cysteine protease superfamily and share a sequence identity of 87.8%, though only show 17.6% and 15.4% sequence identity to CPR-C4, respectively. Despite this low sequence identity, the vasohibin structures align well with CPR-C4, with RMSDs of 1.91 Å/114 C_α (VASH1, PDB accession code 6J8F) and 1.93 Å/120 C_α (VASH2, 6J4P).

3DM superfamily

With the availability of the CPR-C4 structure and identification of structural homologs, the 3DM subfamily system for CPR-C4 was extended to include the subfamilies of the newly identified human VASH1/2 structural homologs. These additional subfamilies were denoted 6J8FB and 6J4PA, after the PDB accession codes of their template structures (6J8F chain B and 6J4P chain A)[290]. Collectively, these three subfamilies formed a 3DM superfamily system, containing 657 and 817 sequences homologous to the human VASH1 and VASH2 proteins, respectively, along with 967 sequences for CPR-C4. This superfamily of 2441 sequences enabled deeper analysis of the structural and previously inaccessible sequence relationships between the human vasohibins and bacterial CPR-C4.

Combined sequence and structural analysis

Despite the low sequence similarity, structural alignments showed that the vasohibins share the core mixed αβ fold seen in CPR-C4 (Fig. 6.31). 3DM utilised a structure-based multiple sequence alignment (3D-MSA) approach to delineate the residue positions forming the conserved 'core' regions in the CPR-C4 and VASH1/2 structures. One of the powerful tools provided by 3DM is a synchronised numbering system, which assigned all structurally equivalent residues in the superfamily sequences the same '3D number'. This enabled direct comparison of corresponding residues between sequences, linking sequence alignment data to the template structures.

The 3DM alignment tool also allowed each residue position to be assessed individually, identifying positions where a particular residue identity was highly conserved throughout the superfamily. In the process of generating the 3DM superfamily system, many types of protein-related data, including protein-ligand contacts, residue mutations and protein variants, were extracted from sequences, structures, and the accompanying literature, and collated[269]. This meant that residues known to form ligand contacts in structures collated by the 3DM system could be identified and compared to equivalent residues in the CPR-C4 sequence.

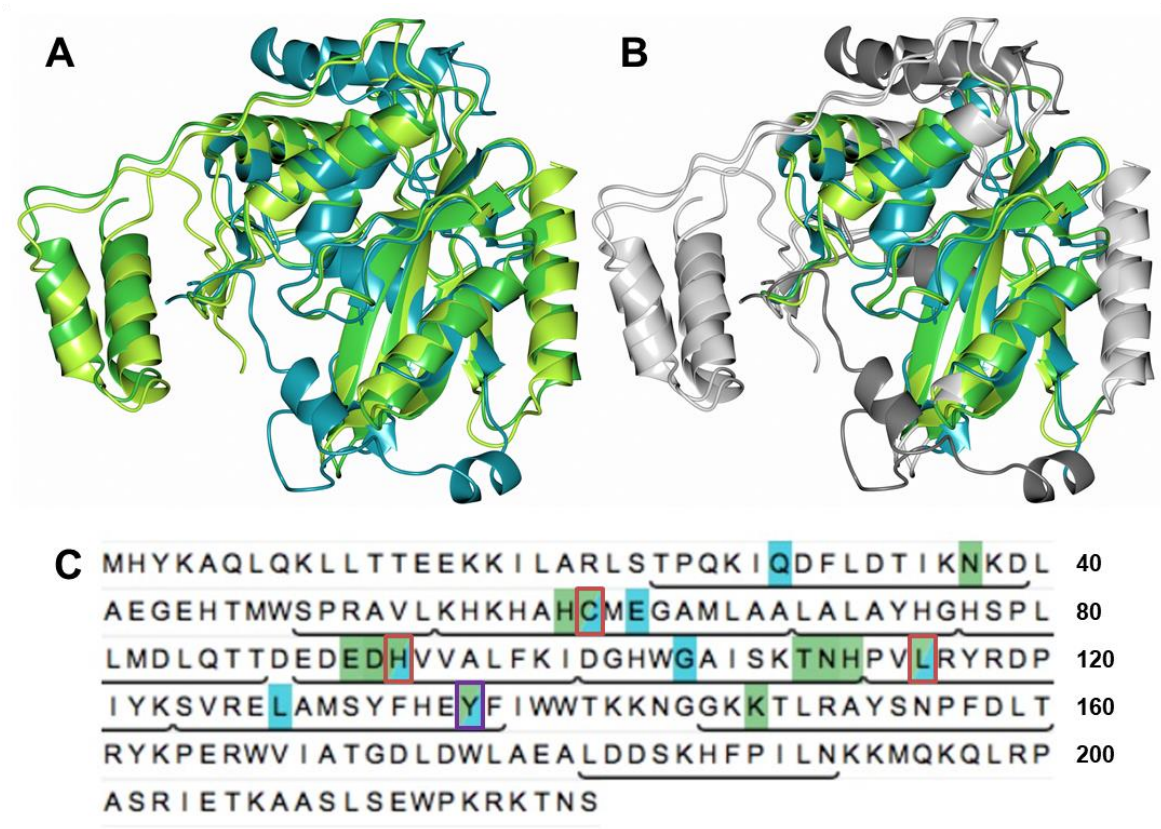

Fig. 6.31: (A) Ribbon diagram showing the structural alignment of CPR-C4 (form 1 chain A, teal) with *Homo sapiens* VASH1 (6J8F chain B, green) and VASH2 (6J4P chain A, yellow green) through least-squares superposition. **(B)** The 3DM core region, sharing significant structural similarity across CPR-C4 and human VASH1/2, is shown in the same colour representation as (A); the remaining 'variable regions' comprised largely of peripheral helices are shown in dark grey for CPR-C4 and in light grey for VASH1/2. **(C)** Amino acid sequence of CPR-C4, with the 3DM core region underlined. Residues conserved in ≥ 95% of 3DM superfamily members are shown in blue; residues that form ≥ 8 ligand contacts in the superfamily are shown in green. The three residues boxed in red form the conserved catalytic triad (C61, H93, L115, following CPR-C4 sequence numbering). The residue boxed in purple is a conserved tyrosine involved in Leu(carbonyl) positioning in VASH1/2. Used under the terms of the Creative Commons CC-BY license[60].

Of the 221 residues in the CPR-C4 sequence, 130 form the 3DM core region that is structurally conserved throughout the extended superfamily, as shown in Fig. 6.31B and C. In VASH1/2, this core region contains the binding site of the α-tubulin substrate[290], with 15 residue positions shown to form ligand contacts in the literature. Only five of the 15 residue identities at these positions were conserved in the CPR-C4 sequence; however, three of these are known to form the non-canonical catalytic triad of VASH1/2: C61, H93 and L115 (following the CPR-C4 sequence, Fig. 6.31C). These residues have been shown to be responsible for the protease activity of the vasohibins through mutation studies[290], and their conservation in the CPR-C4 sequence provides an enticing clue to the potential function of CPR-C4.

Analysis of the catalytic triad

The conservation of the vasohibin catalytic triad residues, in sequence and 3D structure, was assessed to determine if they were maintained throughout the superfamily. For clarity, these residues are referred to by their placement in the CPR-C4 sequence: C61, H93 and L115, with their positioning in the CPR-C4 structure shown in Fig. 6.32A. The residues forming this unconventional catalytic triad identified in VASH1/2 structurally superpose well across the three template structures (Fig. 6.32B) and are highly conserved throughout all three subfamilies (Fig. 6.33).

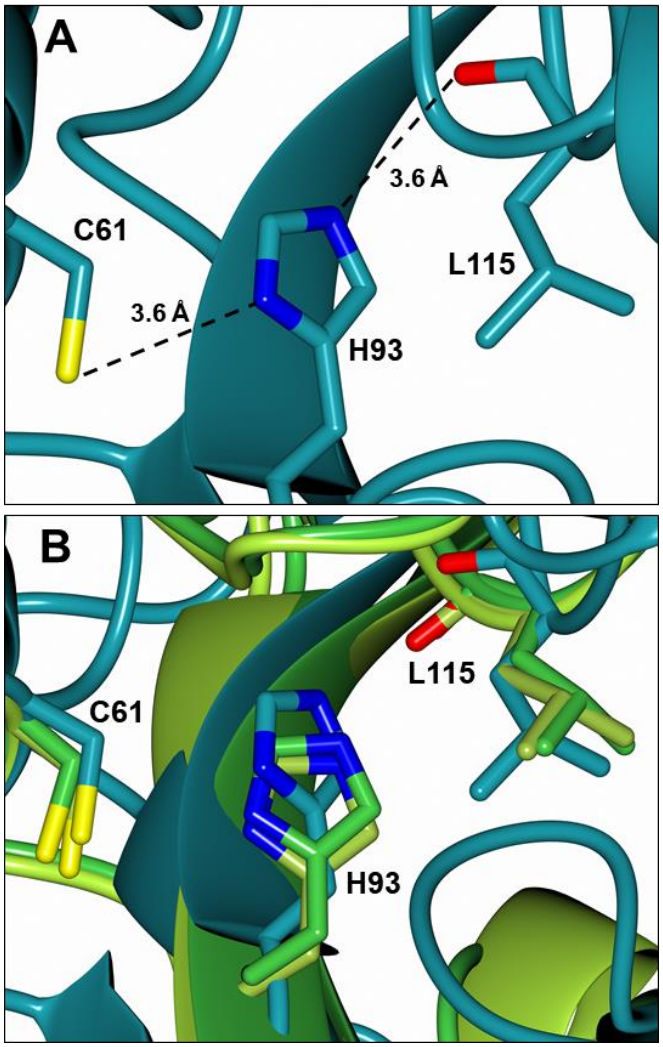

Fig. 6.32: (A) Close-up view of the catalytic triad residues in the CPR-C4 crystal structure (form 1 chain A). The main chain is shown with ribbon representation in teal; side chains of the catalytic triad residues and the backbone carbonyl of L115 are shown with stick representation, O atom in red, N atoms in blue, S atom in yellow; interatomic distances in Å are indicated with dashed lines. The distance between H93 and L115 in forms 2 and 3 is 3.7 Å and 3.6 Å, respectively. Average B-factors for these residues are 92.9 $Å^2$, 89.4 $Å^2$, and 84.9 $Å^2$ for C61, H93, and L115, respectively. **(B)** Catalytic triad residues in CPR-C4 (form 1 chain A, teal) aligned with equivalent 3D residues from human VASH1 (6J8F chain B, green) and VASH2 (6J4P chain A, yellow green) through least-squares superposition. Used under the terms of the Creative Commons CC-BY license[60].

Prior to structure solution, C61 and H93 had been identified as 99.9% conserved in the CPR-C4 subfamily during sequence analysis (6.2.2.). The conservation of these residues was now found to extend to the 3DM superfamily, with leucine also being the predominant residue identity at position 115 (Fig. 6.33). The importance of these residues in VASH1/2 is clear, with C61, H93 and L115 also found to form 9, 10, and 8 ligand contacts in the vasohibin superfamilies, respectively, during 3DM literature mining.

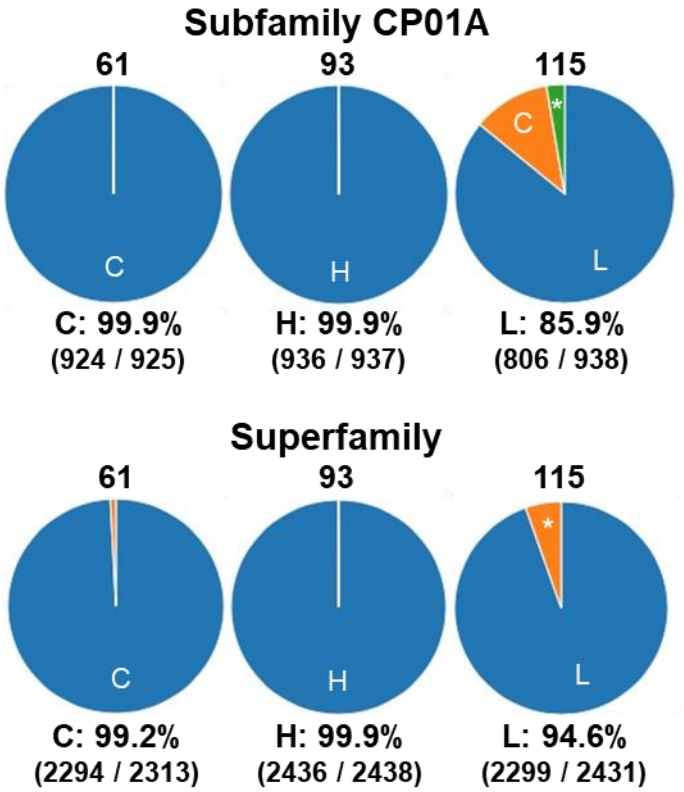

Fig. 6.33: Conservation of the catalytic triad residue identities across the CPR-C4 subfamily and the extended superfamily comprising the CPR-C4, VASH1 and VASH2 subfamilies; * indicates all other residue identities. Used under the terms of the Creative Commons CC-BY license[60].

In the vasohibins, it has been shown that the leucine backbone carbonyl group forms a hydrogen bond with the imidazole ring of histidine, orientating histidine for deprotonation of cysteine and the formation of a thiolate-imidazolium ion pair for catalysis[290]. The occurrence of leucine here is unusual, with the side chains of asparagine, glutamine, or aspartate much more typical in this position[294]. The capability of other residues to fulfil this role explains why leucine is less well conserved than its essential cysteine and histidine counterparts in the 3DM superfamily.

Based on the mechanism of protease activity in the vasohibins, it was proposed that the backbone carbonyl of L115 in CPR-C4 forms a hydrogen bond with an imidazolic nitrogen of H93, which in turn activates the cysteine thiol[290]. The distance between the imidazole ring of H93 and the backbone carbonyl of L115 in CPR-C4 is 3.6 Å (Fig. 6.32A), which is larger than in the vasohibins (2.8 Å and 2.7 Å for VASH1 and VASH2, respectively). However, these side chain conformations are dynamic

and will change during catalysis, making the formation of a hydrogen bond entirely plausible. Evidence for the formation of this hydrogen bond in CPR-C4 also comes from a conserved tyrosine residue, Y136 (Fig. 6.31C). This tyrosine is known to help position the key leucine carbonyl for hydrogen bonding with histidine in the vasohibins[290,291], and is also present in CPR-C4, with 92.5% conservation in the subfamily and 95.8% conservation in the extended superfamily (Fig. 6.34).

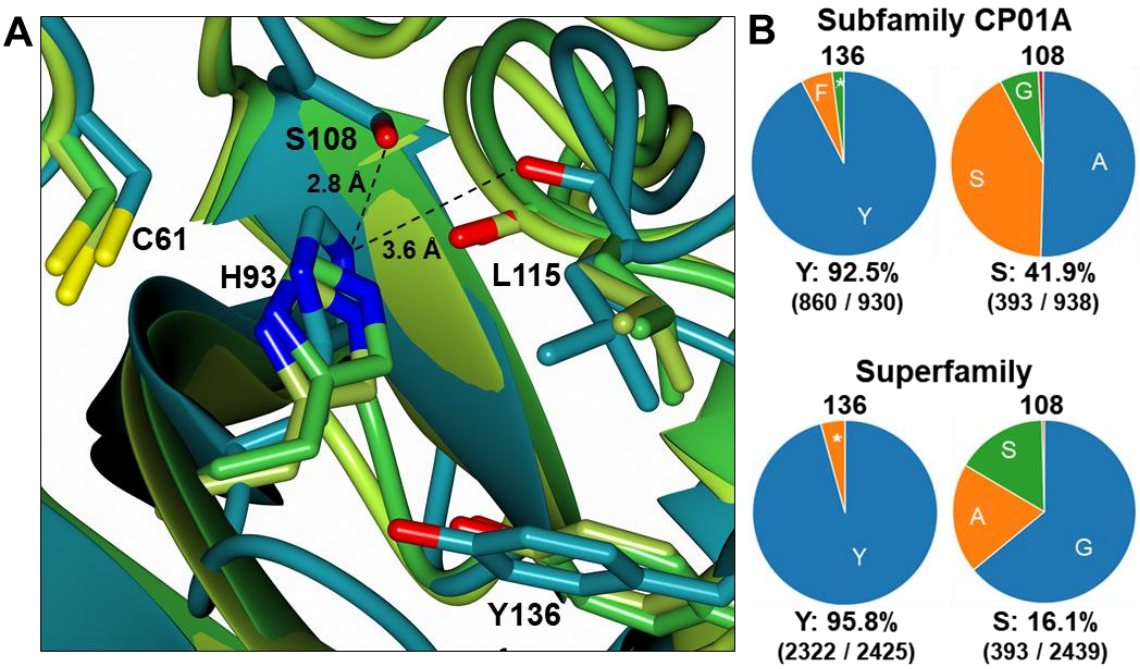

Fig. 6.34: (A) Catalytic triad residues in CPR-C4 (form 1 chain A, teal) aligned with equivalent 3D residues from human VASH1 (6J8F chain B, green) and VASH2 (6J4P chain A, yellow green) through least-squares superposition. The positions of the conserved tyrosine residue (Y136) and proximal serine residue (S108) are indicated. The main chain is shown with ribbon representation; specified side chains and the backbone carbonyl group of L115 are shown with cylinder representation, O atoms in red, N atoms in blue, S atoms in yellow; interatomic distances in Å are indicated with dashed lines. **(B)** Conservation of the tyrosine and serine residues proximal to the CPR-C4 catalytic triad across the CPR-C4 subfamily and the extended superfamily comprising the CPR-C4, VASH1 and VASH2 subfamilies; * indicates all other residue identities[60].

Intriguingly, a serine residue (S108) was also identified near H93, at a more appropriate distance for hydrogen bonding (2.8 Å, Fig. 6.34A). It is possible that this serine could form the third position in the catalytic triad rather than the leucine carbonyl. However, serine is only present in 41.9% of the CPR-C4 subfamily sequences, with alanine in over 50% (473 / 938), and glycine in all vasohibin family sequences. S108 is therefore unlikely to be crucial to function, though may play a role in activation for potential catalysis in some CPR-C4 subfamily proteins. It is also interesting to note that the vasohibin catalytic triad was initially believed to contain a Cys-His-Ser catalytic triad, but this serine was determined to be too far from the key histidine for hydrogen bonding and was instead found to be involved in substrate recognition[290]. S108 in CPR-C4 could similarly be involved in substrate recognition.

Whilst the conservation of the unusual catalytic triad in both sequence and structure is a crucial indicator of shared function, there are key differences between CPR-C4 and the vasohibins. VASH1/2 are known to require a small binding protein (SVBP) for activity, but no potential binding partner has been identified for CPR-C4 on the source contig. CPR-C4 is also a homodimer and lacks the helices known to form intermolecular contacts between monomeric VASH1/2 and SVBP; the superpositions in Fig. 6.31B highlight the differences in these peripheral helices between the CPR-C4 and vasohibin structures. Additionally, the conserved catalytic triad in the vasohibins is located deep in a negatively charged substrate binding pocket, capable of accommodating positively charged α-tubulin[290,293]. In CPR-C4, the proposed C61-H93-L115 is found at the base of a pocket lined with positively charged residues (Fig. 6.35), implying a negatively charged substrate. More crucially, eukaryotic α-tubulin is not present in the source bacterium[295], and so cannot be the CPR-C4 substrate.

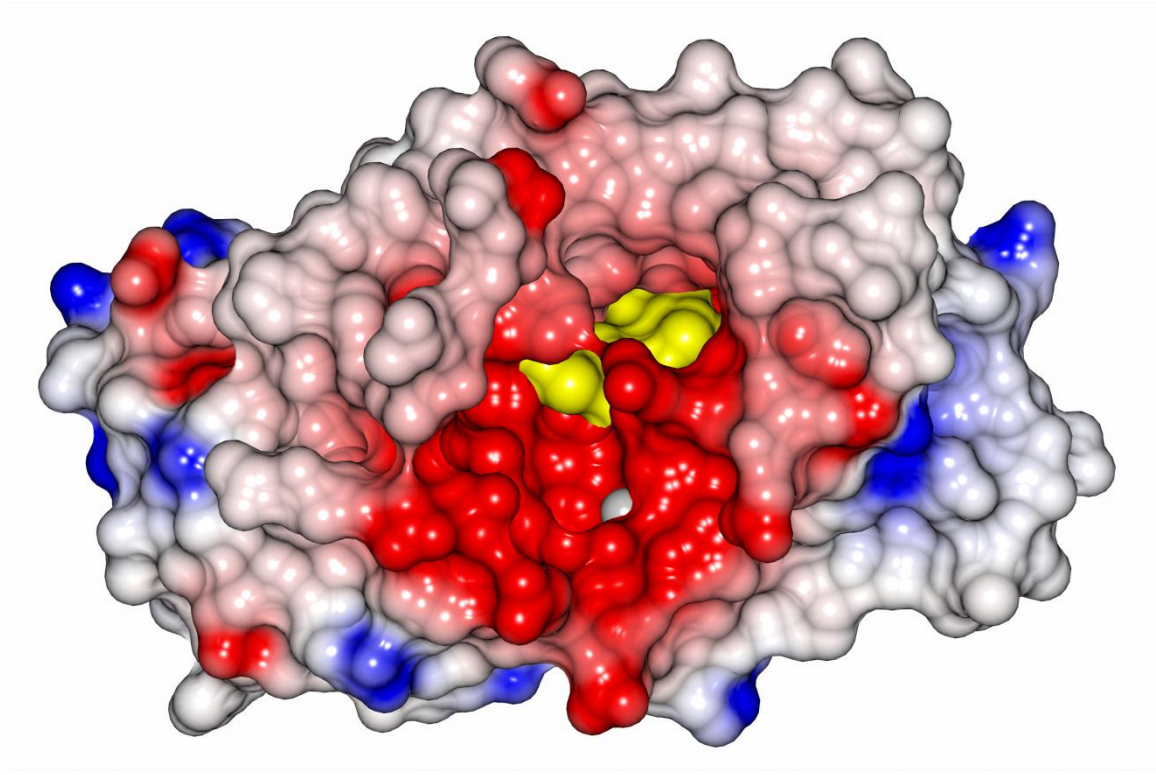

Fig. 6.35: Electrostatic surface representation of CPR-C4 (form 1 chain A) showing the proposed substrate binding pocket of CPR-C4, lined with positively charged residues; red shows positive potential, blue shows negative potential; the surfaces of the catalytic triad residues are shown in yellow; Zn^{2+} ion represented as a grey sphere with 1.0 van der Waals radius[60].

The analysis of the 3D-MSAs employed by 3DM therefore points to CPR-C4 being a cysteine protease, utilising the same non-canonical Cys-His-Leu(carbonyl) catalytic triad as its VASH1/2 structural homologs, albeit with a very different substrate.

Analysis of the Zn^{2+} binding site

With the prediction that CPR-C4 is a cysteine protease, the role of the Zn^{2+} identified in the form 1 crystal structure was evaluated. The extent of aspartate conservation at the binding site residue

positions was assessed across the CPR-C4 subfamily and the wider superfamily (Fig. 6.36). All three positions are within the structurally conserved 3DM core region (Fig 6.31B). Following the CPR-C4 sequence, these positions are denoted 83, 92, and 182.

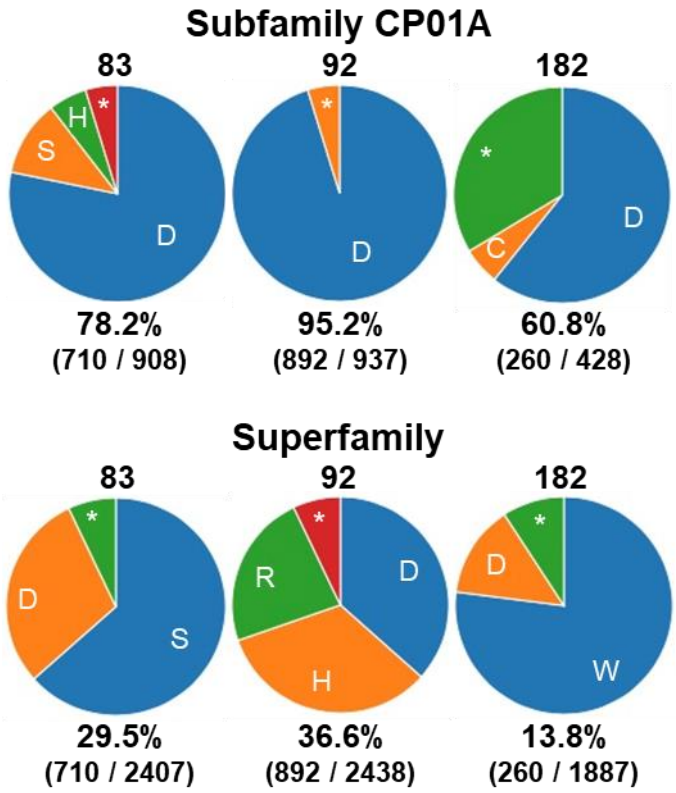

Fig. 6.36: Conservation of equivalent residues of the CPR-C4 tri-aspartate Zn^{2+} binding site across the CPR-C4 subfamily and the extended superfamily containing the CPR-C4, VASH1 and VASH2 subfamilies; * indicates all other residue identities. Used under the terms of the Creative Commons CC-BY license[60].

Aspartate is the predominant residue at all three positions across the CPR-C4 subfamily. However, less than half of the sequences in the subfamily (428 / 967) contain a residue aligned at position 182. Whilst there is clearly some conservation of aspartate in these residue positions, the tri-aspartate Zn^{2+} binding site seen in CPR-C4 is therefore not a requirement for proteins in the subfamily.

The conservation of aspartate at these positions also does not extend to the wider superfamily. All superfamily instances of aspartate are found in sequences from the CPR-C4 subfamily, with none in the vasohibin subfamilies. At position 83 the major residue in the superfamily is serine (63.4%), with similar proportions of aspartate, histidine, and arginine at position 92, and tryptophan at position 182 in the majority of sequences (77.0%). Residues 83, 92, and 182 from CPR-C4, and structurally equivalent residues in VASH1/2, are superposed in Fig. 6.37.

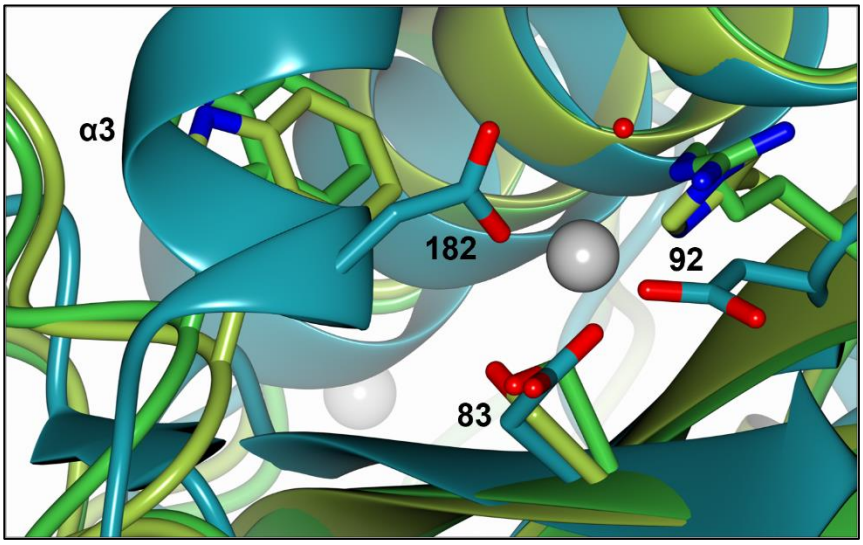

Fig. 6.37: Zn^{2+} binding site from the CPR-C4 structure (form 1 chain A, teal) aligned with equivalent residues in human VASH1 (6J8F chain B, green) and VASH2 (6J4P chain A, yellow green) through least-squares superposition; the main chain is shown with ribbon representation; residue side chains are shown with cylinder representation, O atoms in red, N atoms in blue; Zn^{2+} ions are shown as a grey spheres with 0.5 van der Waals radii and the coordinated water molecule is shown as a red sphere (note: these are only present in the CPR-C4 form 1 chain A structure)[60].

Whilst 83 and 92 superpose reasonably well across the three structures, the coordination of D182 to Zn^{2+} in CPR-C4 appears to pull the chain out of alignment with the vasohibin structures at the start of the α3 helix (Fig. 6.37). This evidence collectively points to a lack of conservation of the CPR-C4 Zn^{2+} binding site in the vasohibin subfamilies.

It was previously noted that the nature of the CPR-C4 Zn^{2+} site is at first glance indicative of a catalytic over a structural site (6.2.7.). However, literature mining with 3DM found no instances of residues forming metal ion contacts throughout the vasohibin subfamilies, and the 3DM residue analysis makes it abundantly clear that the Zn^{2+} binding site is not conserved across the superfamily. Since Zn^{2+} or other metal ions are not required for protease activity in the vasohibins, it is unlikely that the Zn^{2+} in CPR-C4 has a catalytic role or is a requirement for activity. The TSA experiments (6.2.3.) instead support the notion that Zn^{2+} plays a stabilising structural role, significantly increasing the thermotolerance of CPR-C4.

6.2.9. Activity analysis of CPR-C4

Protease activity assays

3D-MSA analysis with 3DM predicted that CPR-C4 is a cysteine protease, employing a Cys-His-Leu(carbonyl) catalytic triad to cleave an unidentified substrate. CPR-C4 activity was anticipated to be highly specific based on the unusual active site geometry, and what is known of the VASH homologs. Whilst the 3DM analysis was useful for gaining insight into the function of CPR-C4, experimental evidence was required to confirm that the protein shows protease activity.

Consequently, biochemical assays were carried out using a commercially available kit designed to detect protease activity. This kit utilises a casein derivative heavily labelled with green fluorescent BODIPY-FL dye, as a substitute for the unknown substrate of CPR-C4. Fluorescence of the uncleaved BODIPY-FL casein substrate in solution is self-quenched, but degradation by a protease releases fluorescently labelled peptide fragments, causing an increase in fluorescence at wavelengths detectable with a standard fluorescein filter[296].

If a protease is present in the reaction mixture, increasing its concentration should result in a measurable increase in fluorescence. This was found to be the case with CPR-C4, with a distinct increase in fluorescence emission measured at increasing concentrations of CPR-C4 relative to a fixed concentration of casein substrate (~ 0.4 µM), as shown in Fig. 6.38A. The CPR-C4 protein used in these assays was highly pure to minimise the likelihood of any detected activity being the result of a contaminant protease.

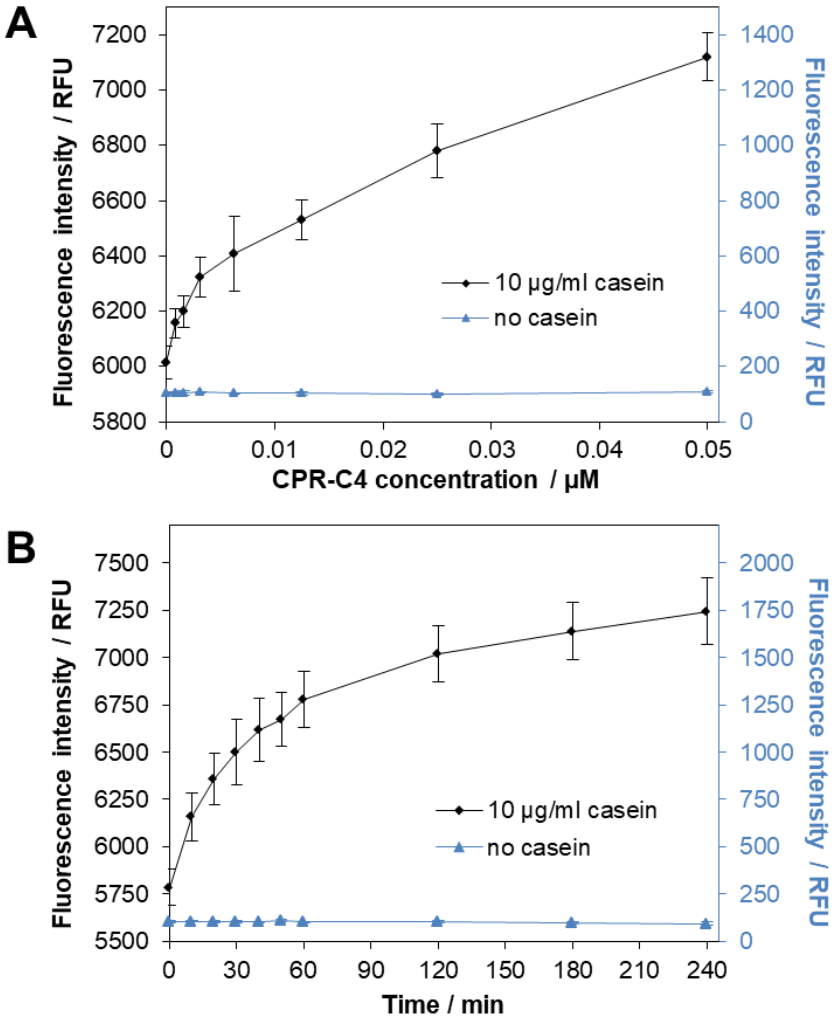

Fig. 6.38: (A) Protease activity data for increasing concentrations of CPR-C4 in the presence (n = 8, ± st. dev.) and absence (n = 3, ± st. dev.) of BODIPY-FL casein substrate, measured as fluorescence intensity in relative fluorescence units (RFU) after 3 h at 30 °C; ex/em = 528 ± 20 nm / 485 ± 20 nm. **(B)** Time course of the increase in fluorescence intensity for 0.025 µM CPR-C4 in the presence (n = 8, ± st. dev.) and absence (n = 3, ± st. dev.) of BODIPY-FL casein substrate. Used under the terms of the Creative Commons CC-BY license[60].

Additions of CPR-C4 in the nM range (to a maximum of 0.05 μM) resulted in a significant increase in fluorescence intensity relative to non-proteolytic controls (CPR-C4 without casein, and casein without CPR-C4). This indicates the presence of fluorescently labelled peptide fragments in the reaction mixture, resulting from proteolytic cleavage of the BODIPY-FL casein substrate by CPR-C4 rather than any intrinsic fluorescent properties of the enzyme or substrate. The fluorescence intensity was also found to increase over time at fixed concentrations of CPR-C4, as the quenched BODIPY-FL casein substrate is cleaved into fluorescent fragments. This is demonstrated by the time course shown in Fig. 6.38B. The data presented are reminiscent of those reported using the same assay performed with the cysteine protease calpain[296], which adds credence to the proteolytic activity of CPR-C4. This experimental data validates that CPR-C4 is a protease, as projected from the 3D-MSA analysis.

Thiol blocking assays

Although the results in Fig. 6.38 show that CPR-C4 has protease activity, they do not prove that the activity is directly due to a catalytic cysteine, as predicted from comparisons to human VASH1/2. Cysteine residues are easily oxidised and are subject to a variety of post-translational modifications[297,298], so the protease activity assays were conducted in the presence of a reducing agent (TCEP) to control the state of the cysteine. ESI-TOF MS analysis was routinely performed during protein preparation to monitor any potential modification, including oxidations. No evidence of covalent post-translational modifications to CPR-C4 were detected in the MS analysis, or in the electron density maps of the three crystal structures.

To determine whether cysteine is responsible for the protease activity of CPR-C4, experiments were also conducted with TCEP in conjunction with iodoacetamide, which alkylates free thiol groups of reduced cysteine residues and renders them inactive[299]. Alkylation with iodoacetamide results in a 57 kDa adduct that can be detected through mass spectrometry. Standard protocols for cysteine modification with iodoacetamide and other alkylating agents typically require sample reduction at ~ 60 °C prior to addition of alkylating agent for mass spectrometric peptide mapping. Whilst these denaturing conditions are necessary for peptide mapping techniques, temperatures were limited to 30 °C to ensure CPR-C4 remained folded, and any reduction in protease activity was due to cysteine alkylation rather than denaturation. However, ESI-TOF MS analysis of CPR-C4 after reduction and alkylation showed that the cysteine remained unmodified, even with sufficiently high concentrations of reducing agent and iodoacetamide (10 mM and 20 mM, respectively), long incubation periods, and the use of freshly prepared iodoacetamide solutions carefully protected from light to prevent degradation. The susceptibility of cysteine residues to modification is influenced by a host of factors, including solvent accessibility, pK_a, and neighbouring residue chemistry[300]. It is possible that the key cysteine in CPR-C4, which sits deep in a pocket of charged residues, is inaccessible and escapes modification with iodoacetamide. An alternative method to confirm that C61 is also responsible for

the protease activity of CPR-C4 is site-directed mutagenesis. Such experiments have been conducted previously on the active site cysteine in VASH1 and were found to abolish the activity of the enzyme[293], showing the cysteine is required for protease activity.

A substantial amount of further experimental work is required to fully characterise the activity of CPR-C4 and is largely hindered by the lack of a native substrate. However, the experimentally demonstrated protease activity, in conjunction with structural alignment of catalytic triad residues in CPR-C4 and VASH1/2, and the near 100% conservation of the catalytic triad residues based on 3D-MSAs, is convincing evidence that these residues form the active site of CPR-C4.

Potential self-cleavage of CPR-C4

A common feature of canonical cysteine proteases is expression as inactive or less active precursors called zymogens, which are then cleaved to form the fully mature, active protease[301,302]. During analysis of CPR-C4, it was noticed that the full-length protein underwent degradation during some SEC experiments, which was detected by SDS-PAGE analysis as shown in Fig. 6.39.

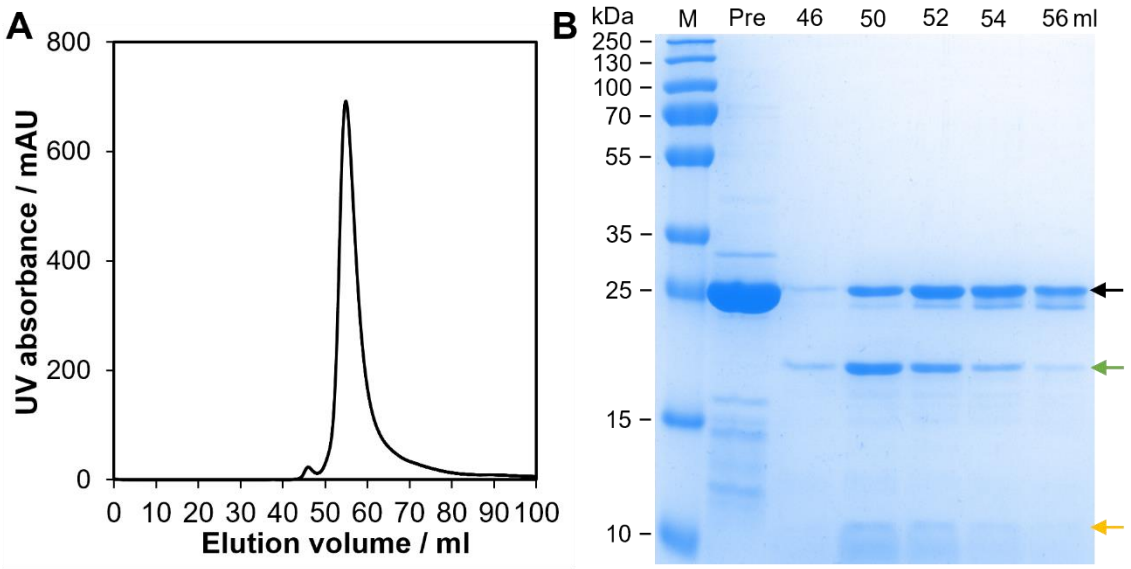

Fig. 6.39: (A) Chromatogram from SEC of CPR-C4. The major peak at 56.8 ml corresponds to a 51.7 kDa species, concluded to be dimeric (54.2 kDa) CPR-C4. **(B)** SDS-PAGE analysis of elution fractions from SEC. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above. The position of bands corresponding to full-length CPR-C4 is indicated with a black arrow; bands corresponding to suspected cleavage products that are absent in the sample prior to SEC are indicated with green and yellow arrows.

SDS-PAGE identified a distinct species with MW ~ 20 kDa, and a further smearing of protein bands is visible at ~ 10 kDa, both of which were not present in the sample prior to SEC and were confirmed to be CPR-C4 through trypsin digest MS. On each occasion that degradation was identified, the size of the major product was consistent, and no further degradation occurred over time, implying a single specific cleavage rather than general degradation by a contaminant protease. Other than the trypsin

used in the digest, no contaminant proteases were detected in the samples, implying that the CPR-C4 underwent self-cleavage.

This proposed self-cleavage of CPR-C4 was ostensibly random. Despite efforts to closely follow a strict protocol during purification, the cleavage occurred only in 20% of purification experiments. Numerous attempts were made to determine a potential cause of the erratic and unpredictable cleavage. It is well understood that pH is an important factor in the activation of cysteine proteases, with auto-processing of zymogens to mature proteases known to occur under the influence of pH, typically at acidic pHs[302–305]. To determine if acidic pH could induce the proposed self-cleavage of CPR-C4, samples were incubated at a range of acidic pH values and analysed by SDS-PAGE, with the results shown in Fig. 6.40.

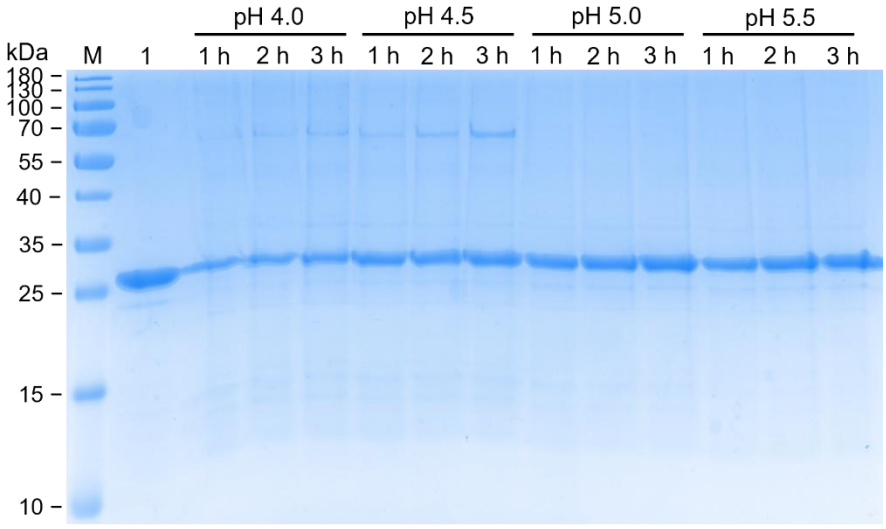

Fig. 6.40: SDS-PAGE analysis of purified CPR-C4 after incubation at 30 °C in sodium acetate buffer at varying acidic pH values. M = marker, with MWs of bands on the left; 1 = sample in pH 7.4 phosphate purification buffer.

The pH change from the weakly alkaline purification buffer (pH 7.4, lane 1 Fig. 6.40) to acidic pH did not induce cleavage of the full-length CPR-C4. Additional experiments with a variety of buffer compositions, temperatures, incubation periods, and additives such as reducing agents and metal chelators, were also ineffective. In cases where cleavage did not happen spontaneously during purification, it could not successfully be induced. A method of cleavage and activation could therefore not yet be established, though it remains feasible that full-length CPR-C4 undergoes self-cleavage to form to a fully mature enzyme, similar to canonical cysteine proteases[302]. Further work is required to uncover the mechanism of activation and determine if the truncated cleavage product shows higher protease activity than full-length CPR-C4.

6.2.10. Phylogenetic analysis of CPR-C4

The discovery of the structural and functional similarities between bacterial CPR-C4 and the human vasohibins was surprising, particularly when considering the low sequence identity, and pointed towards an evolutionary relationship undetectable at the sequence level alone. These proteins could be distantly related by sequence divergence from a common ancestor, or the similarities could have arisen through structural and functional convergence. In order to assess the evolutionary relationship between CPR-C4 and the human VASH proteins in greater detail, a phylogenetic tree shown in Fig. 6.41 was generated, incorporating 92 protein sequences (Table D5) with significant similarity to CPR-C4 (49 sequences) and VASH1/2 (43 sequences). These representative sequences were identified through separate BLAST[151] searches of the CPR-C4 and VASH1/2 protein sequences using an E value cut-off of 0.05 and were selected to cover a broad range of taxa and E values.

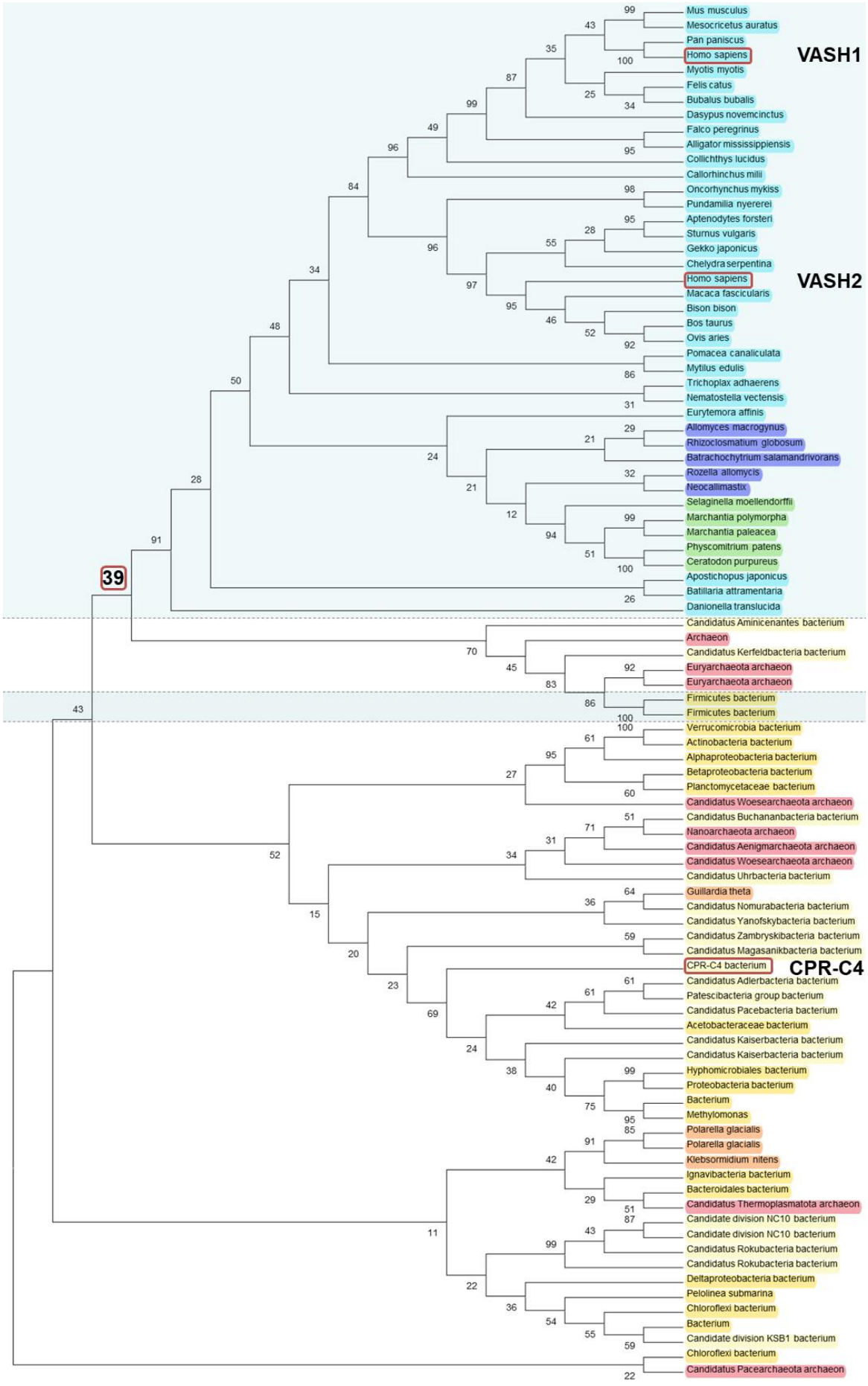

Fig. 6.41: Phylogenetic analysis of CPR-C4 and human VASH1/2 proteins showing a bootstrap consensus tree generated in MEGAX[306] using protein sequences related to CPR-C4 and human VASH1/2 after

alignment in ClustalW[253] (Table D5). Sequences from BLAST[151] searches of the human VASH1/2 proteins are indicated with a pale blue background to distinguish from sequences resulting from BLAST of CPR-C4. Branches are labelled with the species name of the host organism where known, and colour coded according to taxonomic classification: Animalia are coloured blue, Plantae in green, Fungi in dark blue, Algae in orange, Archaea in red, Bacteria in yellow, with bacteria from the CPR in pale yellow. The locations of human VASH1/2 and CPR-C4 within the tree are indicated with labelled red boxes. The evolutionary history was inferred using the Maximum Likelihood (ML) method and Jones-Taylor-Thornton (JTT) matrix-based model using MEGAX[306–308]. Bootstrap values from 500 replicates are shown next to each branch and indicate the percentage of replicate trees in which the associated taxa clustered together. The bootstrap value of 39% at the node linking CPR-C4-related sequences to human VASH1/2-related sequences is highlighted in grey. Used under the terms of the Creative Commons CC-BY license[60].

The sequences within the tree segregate almost perfectly into two halves, with those relating to human VASH1/2 congregating in the top half in Fig. 6.41. These sequences further group based on their taxonomic classification, with clusters of Animalia and Fungi species, shown in light and dark blue, and Plantae species in green. Only two prokaryotic sequences, both from *Firmicutes* bacteria, were identified as showing some similarity to human VASH1/2 within the constraints of the BLAST[151] search. These are the only VASH1/2-related sequences that show some crossover into the lower half of the tree, which is otherwise comprised of the sequences linked to CPR-C4. Bacterial (non-CPR and CPR in yellow and light yellow, respectively), Archaeal (in red) and Algal (in orange) sequences were all identified in the CPR-C4 BLAST search, though importantly no eukaryotic sequences.

The separation of sequences in the tree, both regarding their taxonomy and their relationship to either VASH1/2 or CPR-C4, means there is no obvious candidate for a 'common ancestor' from which the sequences diverged. The taxa at the node linking these two groups of sequences also only clustered together in 195 / 500 repeats (39%, Fig. 6.41). This together implies a weak, if any, evolutionary relationship between CPR-C4 and the human vasohibins that is not identifiable by sequence analysis alone. However, it appears extremely unlikely that the similarities in core fold, protease function, and extremely unusual catalytic triad all arose 'by chance'. The overall structural and functional resemblances between the proteins instead strongly support the notion that bacterial CPR-C4 and the human vasohibins are indeed related by evolutionary divergence from a currently unknown common ancestor and are part of an important cysteine protease family present in species from across the tree of life.

6.3. Conclusions

The mixed αβ-fold structure of the CPR-C4 homodimer was successfully determined in $P3_221$ using Zn-MAD phasing techniques to a resolution of 2.60 Å. A further two crystal forms were also solved

by MR in space groups I222 and $P2_12_12_1$ at resolutions of 2.68 Å and 2.25 Å, respectively. The three crystal structures of CPR-C4 are the first reported structures of a protein from a thermophilic CPR bacterium[60] and revealed the function of a key hypothetical protein conserved throughout the CPR. A major reason that such hypothetical proteins with unknown functions dominate the genomes of CPR organisms is that they cannot be cultured using currently available methods. The metagenomics identification and subsequent heterologous expression approach used here removes the need for cultivation, enabling the discovery of previously inaccessible CPR genomes and broadening our knowledge of their contents. As the numbers of available CPR genomes increase, and with the progression of phylogenetic methods, the roles of other conserved novel gene products will be uncovered, expanding our knowledge of this little understood branch of the tree of life.

The analysis of CPR-C4 using a 3D-MSA approach uncovered distant structural homology to the human vasohibins, VASH1 and VASH2, leading to its functional annotation as a cysteine protease with an unusual Cys-His-Leu(carbonyl) catalytic triad. The 3DM protein family system for CPR-C4 proved to be an important *in-silico* tool for prediction of this protease activity, which was uncovered through the remote sequence parallels detected in alignments. The 3D-MSA used by 3DM are typically of higher quality than classical multiple sequence alignments and are also key to the success of the recent structure prediction tools, AlphaFold2[118,263] and RoseTTAFold[118,263,309]. Models of the CPR-C4 structure generated with these tools (Fig. D3) were of high enough quality to retroactively solve the CPR-C4 crystal structure using MR methods. AlphaFold2 and RoseTTAFold therefore clearly have the potential to aid experimental structure solutions of other hypothetical proteins within the CPR and beyond, where suitable homology models are not available in experimental structure databases such as the PDB. The powerful link between protein structure and function, as demonstrated here with CPR-C4, shows the importance of structure determination for establishing the function of hypothetical proteins, where sequence alone is insufficient.

The lack of knowledge regarding the source organism of CPR-C4 makes it difficult to speculate on a potential function for this protease *in vivo*. The recurrence of CPR-C4-type proteases with high sequence conservation throughout the candidate phyla hints at an important function *in vivo* that cannot be confirmed without greater understanding of the lifestyle of CPR organisms. It is well understood that CPR species have little capacity for the *de novo* synthesis of many essential metabolites, including amino acids, and instead form symbiotic partnerships with hosts from across the tree of life[42,43,150]. CPR-C4 could be involved in providing essential amino acid metabolites both to the CPR organism and its symbiotic partner through the degradation of peptide substrates. Uncovering the functions of more of the hypothetical proteins that constitute 70% of the source contig is necessary to further our understanding of the CPR and expose the specific role of CPR-C4 and its homologs. Proteases account for around 60% of marketed enzymes worldwide, and the potential specificity and thermotolerant properties of CPR-C4 introduce additional innovation value. This warrants further investigation of CPR-C4 as an enticing candidate for biotechnological applications in future.

Chapter 7. Production and characterisation of a DnaK/ClpB bi-chaperone system from the Candidate Phyla Radiation

7.1. Introduction

7.1.1. Protein folding and molecular chaperones

To function correctly, most proteins must fold into specific three-dimensional structures encoded within their amino acid sequence[310–312]. Whilst protein folding *in vivo* can occur spontaneously, many newly synthesised proteins require assistance from a complex network of molecular machines to acquire and maintain functionally active conformations. These machines, known as molecular chaperone proteins, perform a variety of functions with the overall aim of maintaining protein quality control, known as 'proteostasis'[313,314].

Molecular chaperones are required to promote correct folding of nascent polypeptide chains as they emerge from the ribosome into an exceedingly crowded cytosolic environment that can otherwise trigger misfolding and aggregation[311,313]. This is achieved through recognition by chaperones of features that distinguish unfolded proteins from folded proteins, such as hydrophobic residues and unstructured backbone regions[154,315]. Binding of chaperones to these substrates shields the surfaces of the polypeptide chains from unproductive interactions during the folding process, preventing aggregation[314]. Chaperones involved in folding of newly synthesised chains do so through cycles of substrate binding and release correlated to ATP hydrolysis and assisted by a variety of co-chaperone proteins[316,317].

Chaperones also function to maintain proteins in their functionally active forms, as protein structures are dynamic and conformationally flexible, with their functional state comprised of an ensemble of related structures[313]. As a result, they are typically only partially stable, and even small environmental changes can result in unfolding due to the small energy barriers between native and misfolded states[315]. Protein unfolding and aggregation is therefore a major concern under stress conditions, such as heat shock. Molecular chaperones function as a protective system to combat the adverse effects of protein misfolding during stress conditions, and many members of these systems are also known as heat shock proteins (HSPs) due to their upregulation under thermal stress[313].

7.1.2. DnaK/ClpB bi-chaperone system

A variety of molecular chaperone systems are known with family members present throughout the tree of life[314,318,319]. One particularly well-studied bacterial chaperone is DnaK, with eukaryotic homologs known as Hsp70 proteins. *E. coli* DnaK (*Ec*DnaK) has served as the prototype for study, with functions including *de novo* protein folding, preventing aggregation of heat-denatured proteins, solubilisation of protein aggregates, and regulation of the heat shock response[311,314]. DnaK functions in a cycle of substrate binding and release through ATP-hydrolysis, mediated by the co-chaperone DnaJ (Hsp40) and the nucleotide exchange factor GrpE[316,320,321]. The actions of DnaK depend on its interactions with these partner proteins, and the ATPase activity of DnaK has been shown to increase 50-fold in the presence of DnaJ and GrpE[321].

Additional association of DnaK with another chaperone, ClpB (Hsp100), is also required for disaggregation of protein aggregates[322]. ClpB is part of the AAA+ superfamily (ATPase associated with diverse cellular activities)[323–326] and cooperates with the DnaK chaperone system to solubilise and refold aggregated proteins for cells to survive severe heat stress[327,328].

7.1.3. Applications of chaperones in biotechnology

The biomedical and biotechnological applications of chaperone proteins are extensive. Protein misfolding and aggregation are well-recognised problems during the overexpression of recombinant proteins in heterologous systems, such as *E. coli*[162,168,219]. The copious quantities of protein produced can overwhelm cellular folding machinery, generating inclusion bodies of biologically inactive aggregates[162]. There has been extensive research into strategies for improving the production of soluble recombinant proteins, and effective refolding of aggregated material[168,315,329,330]. The capacity for molecular chaperones to assist both *de novo* folding and disaggregation of protein aggregates makes them attractive tools for this purpose.

Co-overexpression of ClpB and components of the DnaK chaperone system alongside proteins of interest has been demonstrated to increase target solubility, as well as improve recovery from aggregates[168,329,330]. This is typically achieved through the insertion of chaperone-encoding genes alongside the target gene in an expression vector, allowing concurrent overexpression. Bacterial host cells can alternatively be transformed with a secondary plasmid containing chaperone-encoding genes. This improves the intracellular folding environment of the expression host by increasing the cytosolic concentrations of chaperones, facilitating proper folding, and lowering the amount of aggregation for many recombinant proteins[315,329]. The use of *E. coli* strains harbouring additional chaperone genes proved to be greatly beneficial during the expression of viral targets from deep-sea metagenomes by Virus-X partners[18]. However, co-expression of molecular chaperones does not always improve soluble protein yields for aggregation-prone targets. Success is dependent on the target protein, and the levels and combinations of chaperones used. Since chaperone systems are

comprised of several collaborating components, it is usually necessary to combine these in a single host strain[22,23].

As well as encouraging correct folding from the outset during recombinant overexpression, the DnaK/ClpB bi-chaperone system can successfully solubilise and refold target proteins from aggregated material *in vivo* and *in vitro*[315]. This presents a cleaner and more efficient alternative to chemical additives for the solubilisation and refolding of protein aggregates[330].

As discussed in 2.2.3., a bi-chaperone system was identified from the metagenome of a CPR organism, formed of four components denoted as CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB. Characterisation of this CPR-DnaK/ClpB bi-chaperone system is necessary to reveal any unique features that could make its components particularly suitable for commercialisation. Greater understanding of chaperone modes of action will also enable the development of artificial systems that can mimic and enhance their valuable functions[154,315].

7.2. Results and discussion

7.2.1. Target identification and BLAST analysis

Genes predicted to encode the components of a DnaK/ClpB bi-chaperone system were identified on a single contig from a CPR genome. The positions of the genes on the contig are shown in Fig. 7.1.

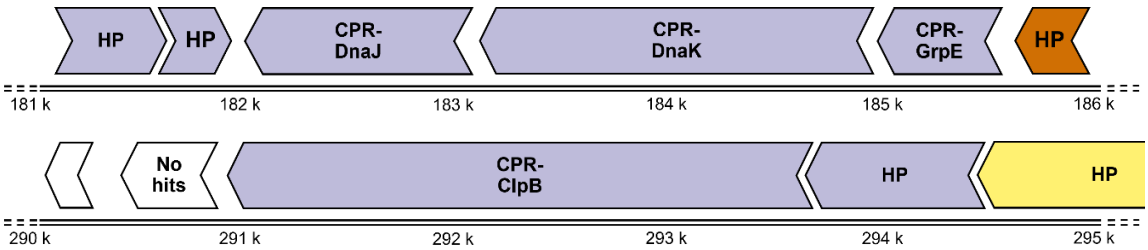

Fig. 7.1: Positions of the chaperone genes and their neighbours on the source contig, represented as arrows showing the strand direction of transcription. The predicted function of gene products is stated where annotation has been possible; HP = hypothetical protein. Genes are colour coded by the phylum classification of their most significant BLAST hit: yellow = unclassified bacteria; purple = *Candidatus Adlerbacteria*; orange = *Candidatus Kaiserbacteria*; white = unknown.

The predicted molecular chaperone CPR-DnaK, co-chaperone CPR-DnaJ, and nucleotide exchange factor CPR-GrpE are all adjacent on the contig, transcribed in the same direction and likely concurrently. The predicted ATP-dependent disaggregase, CPR-ClpB, was found further downstream on a separate region of the contig, surrounded by hypothetical proteins with unknown function.

BLAST[151] searches against the chaperone protein sequences revealed that the hits with highest sequence similarity for all four targets were from candidate bacterial phyla within the CPR; details of the BLAST search results are outlined in Table 7.1.

Table 7.1: Details of the top BLAST hits for the CPR bi-chaperone system targets, and sequence identities (seq. ID) with well-studied *E. coli* homologs.

Target	Description	Taxonomy	E value	Seq. ID	Seq. ID to *E. coli*
CPR-GrpE	Nucleotide exchange factor, GrpE	*Candidatus Adlerbacteria*	$7e^{-66}$	57.8%	35.6%
CPR-DnaJ	Molecular chaperone, DnaJ	*Candidatus Adlerbacteria*	$5e^{-127}$	78.2%	40.4%
CPR-DnaK	Molecular chaperone, DnaK	*Patescibacteria group*	0.0	85.9%	58.4%
CPR-ClpB	ATP-dependent chaperone, ClpB	*Candidatus Adlerbacteria*	0.0	85.7%	49.2%

Through BLAST searches, functions could confidently be assigned to these CPR proteins, and they were therefore classified as category A targets, indicating known function. All four CPR proteins were identified as key targets for characterisation due to their potential biotechnology applications. Since all four proteins were successfully overexpressed in small-scale tests, full-scale expression was conducted to enable biophysical characterisation and crystallisation experiments.

7.2.2. Production and characterisation of CPR-GrpE

Expression and IMAC purification

During full-scale expression, CPR-GrpE was found to overexpress well, yielding ~ 20 mg protein per litre of culture. The SDS-PAGE results from IMAC purification are shown in Fig. 7.2.

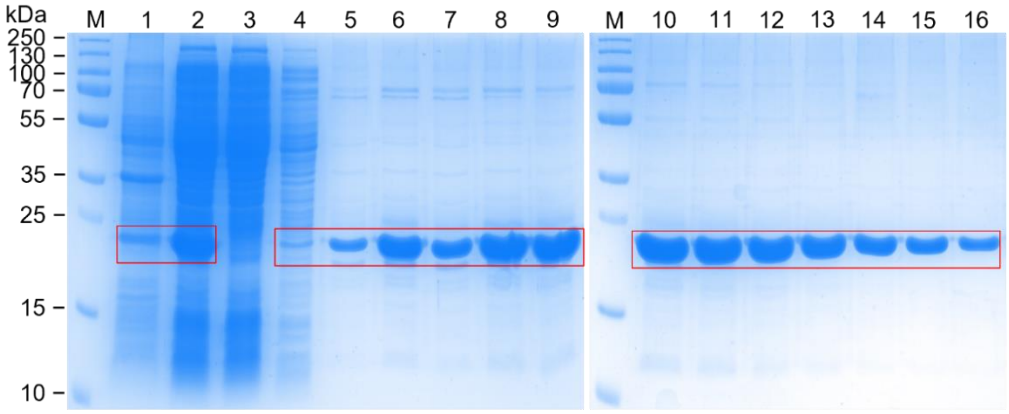

Fig. 7.2: SDS-PAGE analysis of CPR-GrpE IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 16 = protein fractions. Bands corresponding to CPR-GrpE (MW = 20.8 kDa) are indicated with a red box.

The majority of CPR-GrpE produced was found to be soluble, with much greater quantities detected in the lysate than the cell pellet, reflected by the band strengths in Fig. 7.2. IMAC purification was also highly successful, with minimal loss of protein in the flow through and wash, and reasonably pure target protein in eluted fractions.

SEC analysis

The purity of CPR-GrpE after IMAC purification was sufficient for downstream characterisation (Fig. 7.2). However, SEC was still necessary to evaluate the oligomeric state of CPR-GrpE and remove any aggregated protein prior to crystallisation. The resulting chromatogram and SDS-PAGE analysis of eluted fractions are shown in Fig. 7.3.

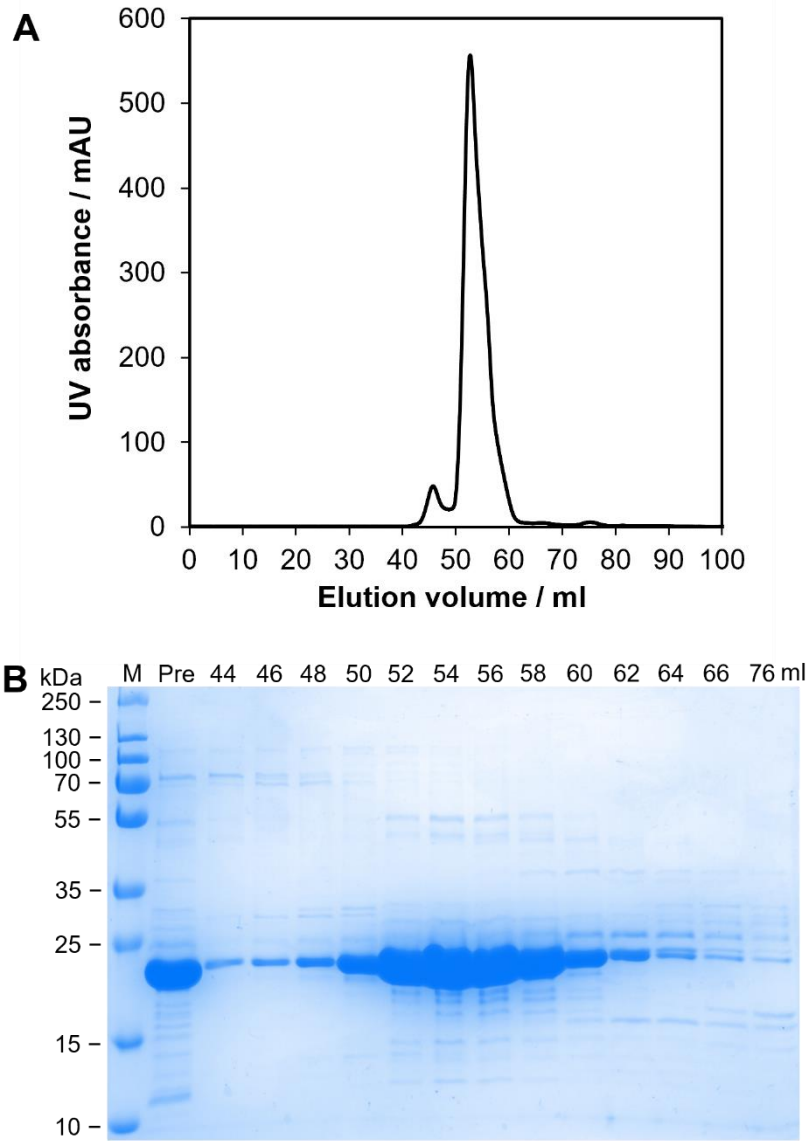

Fig. 7.3: (A) Chromatogram from SEC of CPR-GrpE. **(B)** SDS-PAGE analysis of elution fractions from SEC of CPR-GrpE (MW = 20.8 kDa). M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The minor peak at 45.7 ml is close to the void volume (V_o) of the column (44.6 ml) and likely corresponds to aggregated protein. The major peak in the chromatogram with V_e = 52.7 ml (Fig. 7.3A) is well resolved and sharp, indicating a largely homogenous protein sample with a MW around 67.1 kDa. This species was confirmed to be CPR-GrpE based on expected MW by SDS-PAGE (Fig. 7.3B). The estimated MW based on V_e was closest to that of the CPR-GrpE trimer (MW = 62.4 kDa). However, it was noted that GrpE homologs are dimeric, with a long α-helical tail region that gives a deceptively high Stokes radius[320,331,332]. It is likely that CPR-GrpE forms a similar dimer and so elutes at a lower V_e than typical globular proteins of the same MW.

Characterisation

An ESI mass spectrum was successfully obtained for CPR-GrpE, with a MW within 3 Da of the expected MW, accounting for N-terminal methionine excision (NME)[224] (Appendix E). The identity of CPR-GrpE was also confirmed by trypsin digest MS.

CPR-GrpE was also analysed using circular dichroism (CD) to determine if the protein was folded. The resulting CD spectrum is shown in Fig. 7.4.

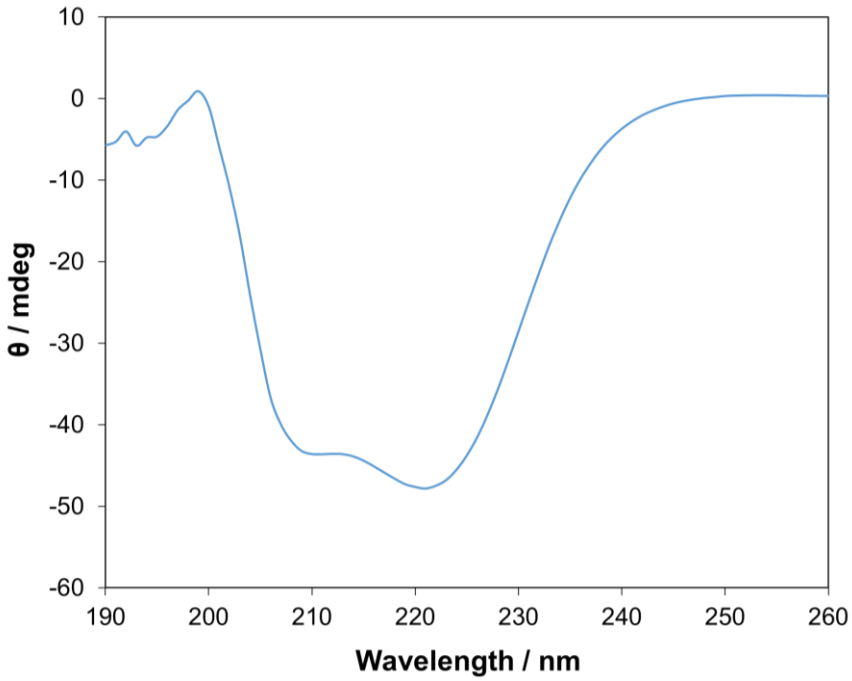

Fig. 7.4: Circular dichroism spectrum for CPR-GrpE, with CD measured in ellipticity (θ).

The spectrum indicates that CPR-GrpE is folded, and the two negative peaks at 208 nm and 222 nm are characteristic of a protein with significant regions of α-helices[333,334]. This correlates with the structures of homologous GrpE dimers, which are known to contain long α-helical tails[320,331,332].

Thermal shift analysis

The results of TSA with CPR-GrpE were atypical, with two thermal transitions generating two melt temperatures (T_{m1} and T_{m2}) seen under most conditions. The output table generated during data processing by NAMI[210,255] is shown in Fig. 7.5A and presents the T_m values for each condition relative to a reference in MilliQ water from wells A1 and A2. On closer inspection of the associated melt curves, it was found that TSA with CPR-GrpE generates an unusually shaped curve with two peaks (Fig. 7.5B).

A

	1	2	3	4	5	6	7	8	9	10	11	12
A	41.4 54.3	41.8 54.5	N/S	50.1	46.2 56.9	44.9 55.4	41.6 55.1	54.6	43.9 60.7	41.7 55.5	41.9 56.7	41.7 55.6
B	58.3	41.6 54.7	41.4 56.6	55.0	42.8 56.8	41.7 54.8	41.5 54.9	41.7 54.4	43.9 57.1	41.9 54.8	41.9 55.8	41.2 55.5
C	41.2 55.4	41.7 54.5	41.8 56.1	41.5 54.7	40.3 55.8	39.4 54.4	40.3 55.8	40.6 54.4	37.7 51.6	40.9 52.7	53.3	41.2 54.3
D	41.8 56.0	41.1 54.8	41.7 54.4	41.6 55.1	41.6 54.9	54.6	43.7	54.8	41.6 54.2	54.6	54.5	41.2 54.5
E	43.0 61.3	41.3 55.2	41.2 56.2	41.9 54.7	41.2 55.1	41.9 54.5	43.2 56.3	54.5	41.8 55.9	54.8	56.0	41.4 54.6
F	54.4	54.8	53.1	41.7 55.5	57.1	41.8 54.9	57.3	41.6 54.7	43.9 55.7	54.9	42.7 57.2	41.6 54.9
G	57.0	43.3 56.8	41.3 56.4	41.6 55.2	41.6 55.2	42.2 54.6	34.2 86.5	36.5	52.9	41.5	55.0	41.6 54.5
H	41.7 56.3	42.2 54.5	41.2 58.1	41.4 54.9	40.9 55.3	41.7 54.6	41.8 56.7	41.6 55.0	42.2 56.8	42.1 54.7	40.8 55.1	40.6 54.7

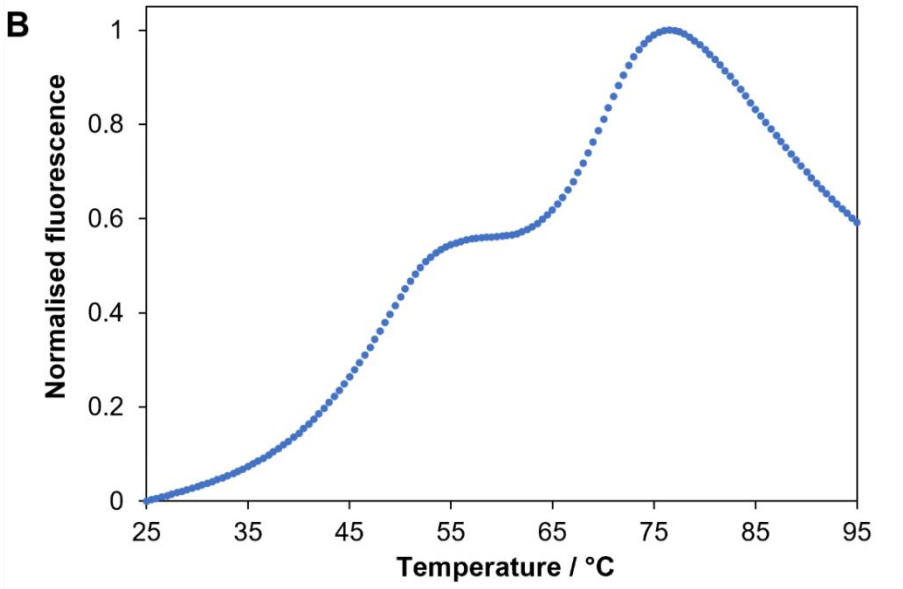

Fig. 7.5: (A) Results from the 96-well plate TSA of CPR-GrpE with the Durham Osmolyte Screen®. Colours are indicative of differences in the melt temperature (T_m) from that in water (wells A1 and A2): a colour change from colourless through light blue to dark blue indicates an increase in T_m. Green indicates either two stages in the melting process or variation in data. N/S means no signal due to denatured protein. **(B)** Melt curve showing normalised fluorescence intensities at 590 nm of CPR-GrpE in 1.0 M sodium malonate from TSA with the Durham Salt Screen®. Experimental data are shown as individual points.

Across the three Durham Screens®, the reference T_m values for CPR-GrpE were ~ 41.5 °C (T_{m1}) and ~ 54.5 °C (T_{m2}). Whilst the T_m values did vary depending on the conditions, the double peak of the melt curves made it difficult to confidently assess the stabilising or destabilising effects of different additives.

Some useful information could be gathered from TSA results with the Durham Salt Screen®. Increasing concentrations of a variety of common salts, such as sodium sulfate and ammonium sulfate, generally increased the stability of CPR-GrpE, reflected by an increase in T_m values. Particularly high T_m values were seen in 1.5 M sodium malonate, which caused > 7 °C (48 °C) and > 16 °C (71 °C) shifts in T_{m1} and T_{m2}, respectively. The melt curves for CPR-GrpE with sodium malonate are shown in Fig. 7.6, which demonstrates the stark rise in T_{m2} as the salt concentration is increased.

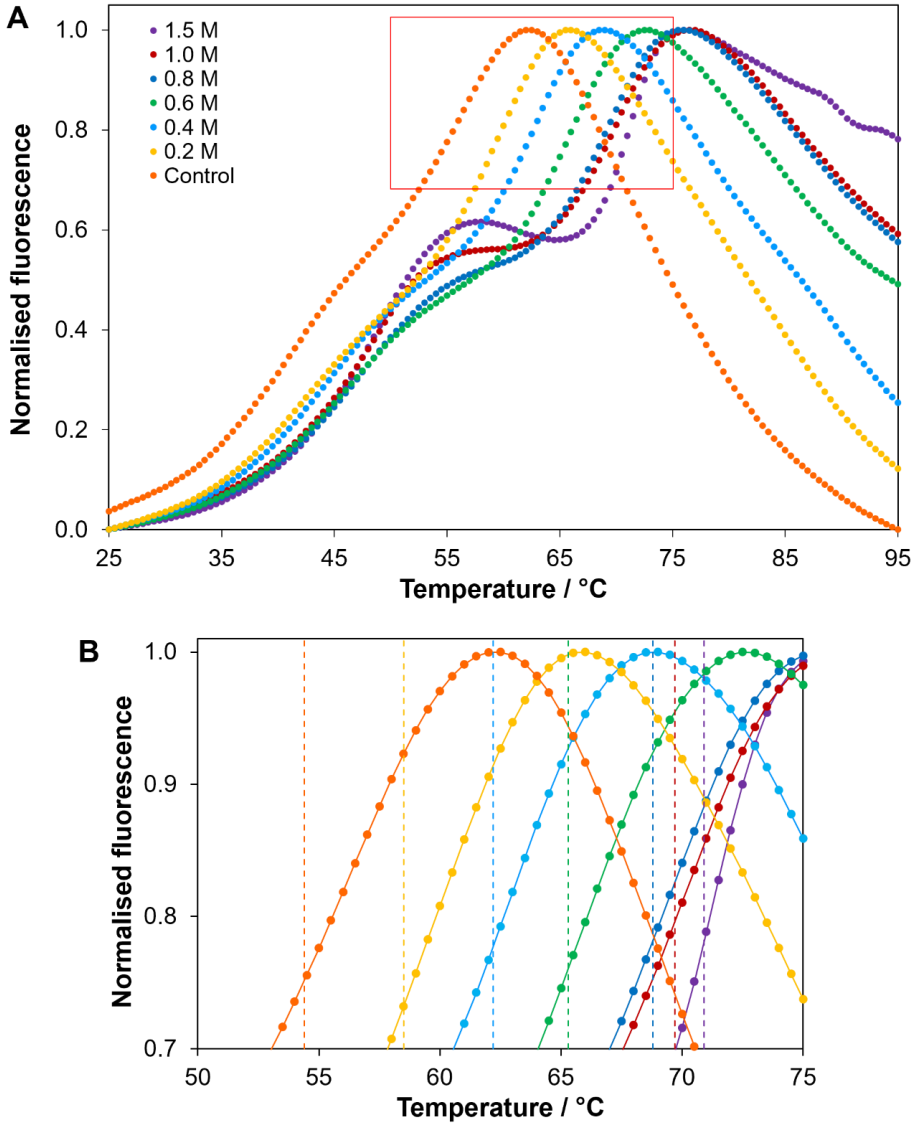

Fig. 7.6: (A) Melt curves showing normalised fluorescence intensities at 590 nm from TSA of CPR-GrpE with increasing concentrations of sodium malonate against a control in MilliQ water. Experimental data are shown as individual points. **(B)** Closer view of the region outlined in red from (A), highlighting the shift in T_{m2} with increasing concentrations of sodium malonate. Dashed vertical lines indicate T_{m2} values.

Similar increases in T_{m2} values were also seen with increasing concentrations of sodium sulfate and ammonium sulfate. CPR-GrpE is therefore stabilised at high salt concentrations, with a particular preference for sodium malonate. It was also noted that as salt concentration was increased, the first peak in the melt curve became more pronounced (Fig. 7.6A). It could be that the stabilising conditions promote the thermal transition represented by the first peak.

Significance of the double melt curve

It was initially thought that the first thermal transition in the melt curve was due to dissociation of the CPR-GrpE dimer. The second transition was then proposed to be the result of denaturation. However, the *E. coli* GrpE (*Ec*GrpE) dimer has been shown to also undergo two thermal transitions, with T_{m1} ~ 50 °C and T_{m2} ~ 75 °C[335]. In *Ec*GrpE, the first thermal transition is independent of GrpE concentration, implying it is not the result of dimer dissociation. Instead, this transition has been assigned to unfolding of the long α-helical tail of the GrpE dimer structure[331,335]. The second thermal transition with T_m ~ 75 °C is due to unfolding of the remaining structure. GrpE from the hyperthermophile *Thermus thermophilus* (*Tt*GrpE) similarly has two thermal transitions, with T_m values of ~ 90 °C and ~ 100 °C, which again are proposed to correspond to helix unfolding, and denaturation, respectively[332]. The first thermal transition for these GrpE homologs is initiated at temperatures representing heat shock conditions for their respective organisms: ~ 45 °C for mesophilic *E. coli*, and ~ 85 °C for hyperthermophilic *T. thermophilus*[332,336]. GrpE has therefore been suggested to function as a thermosensor, with the initial partial unfolding acting as a regulatory element for the DnaK chaperone system by indicating heat shock[320,332].

The two thermal transitions observed for CPR-GrpE likely result from the same structural changes as its *E. coli* and *T. thermophilus* counterparts. Considering that CPR-GrpE was identified from a metagenome isolated at 75 °C, it could be expected that its thermal transition temperatures would resemble those of *T. thermophilus*, which thrives at 70 – 75 °C[110]. However, the highest temperatures observed for CPR-GrpE during screening were only ~ 48 °C and ~ 71 °C for T_{m1} and T_{m2}, respectively (1.5 M sodium malonate, Fig. 7.5), which is even lower than for mesophilic *Ec*GrpE. It is, however, possible that the unknown conditions in which CPR-GrpE operates *in vivo* are significantly more stabilising than are replicated in the TSA screens, which would result in higher T_m values that better match those recorded for GrpE from other thermophilic species. Assuming this stabilisation is possible, a dual melt temperature may allow CPR-GrpE to function as a thermosensor in the high temperature environment in which the sample was collected.

Crystallisation

Crystallisation screening

Preliminary crystallisation trials were conducted for CPR-GrpE using commercially available high-throughput screens. Fig. 7.7 illustrates the purity of the CPR-GrpE sample used for crystallisation screening after IMAC and SEC purification.

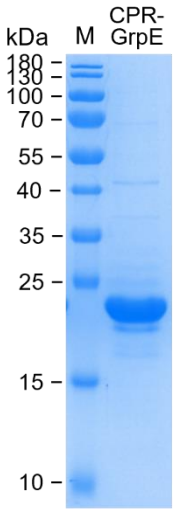

Fig. 7.7: SDS-PAGE showing purity of the chaperone proteins after IMAC and SEC for preliminary crystallisation trials.

Initial screening of CPR-GrpE at 4.9 mg/ml did not reveal any promising crystallisation conditions, with most drops remaining clear after 2 months. Screening was therefore repeated with higher concentrations of CPR-GrpE, at 6.4 mg/ml, 8.6 mg/ml, and 10.7 mg/ml. In this second round of screening, potential crystalline material was identified in 2.0 M ammonium phosphate monobasic, 0.1 M Tris pH 8.5, from the Structure Screen 1+2 HT-96. This condition was manually outscreened by varying the pH of the Tris buffer system and the ammonium phosphate concentration in 24-well sitting drop trays (Appendix E). The crystals produced in these manual trays were very thin and flat (Fig. 7.8A), proving difficult to harvest and diffracting poorly (~ 12 Å), and so required further optimisation.

Crystal optimisation with Microseed Matrix Screening

To improve the quality of the CPR-GrpE crystals, Microseed Matrix Screening (MMS)[275,276] experiments were conducted. Seed stocks were produced using crystals from the manual outscreening experiments, which were then used to nucleate crystal formation. Fig. 7.8 shows the results of MMS on the CPR-GrpE crystal morphology.

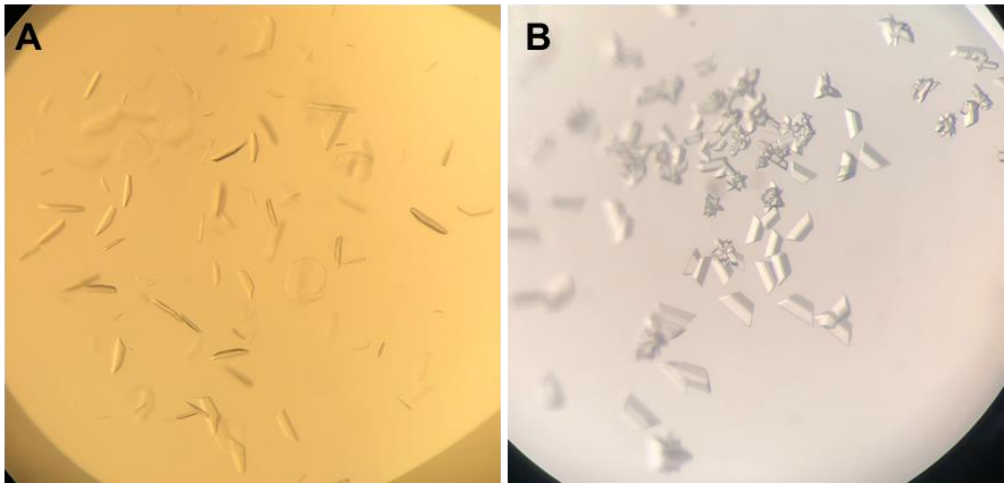

Fig. 7.8: Optimisation of CPR-GrpE crystals through MMS. **(A)** Original crystals identified through manual outscreening experiments and used to produce seed stock for MMS. **(B)** Crystals produced through MMS using seed stock produced from crystals in (A).

Whilst the preliminary crystallisation screening produced crystals in only one conditions in the Structure Screen 1+2 HT-96, MMS resulted in crystals in six conditions. The new crystals after MMS were also a more desirable shape, progressing from flat, plate-like crystals to the trapezoidal prisms shown in Fig. 7.8B.

These new crystals grew rapidly, with the images in Fig. 7.8B and Fig. 7.9A captured less than 24 h after MMS experiment set-up. However, within a further 24 h the crystals had visibly degraded, with the edges becoming feathered and uneven, as shown in Fig. 7.9B.

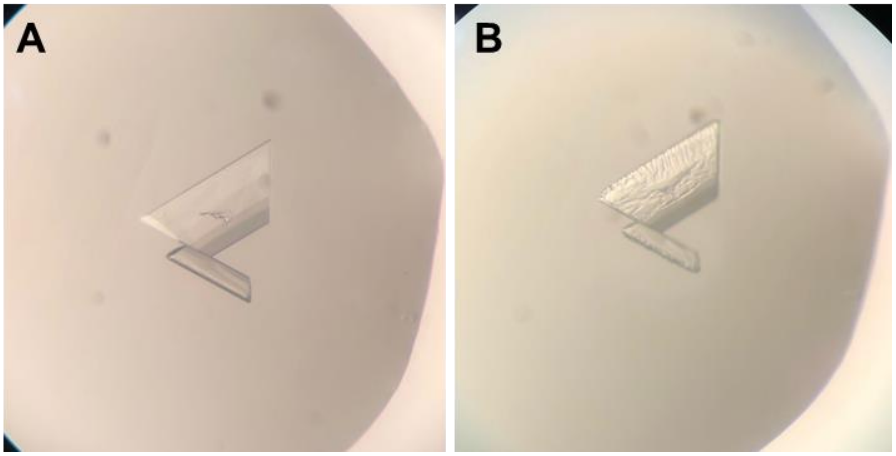

Fig. 7.9: CPR-GrpE crystals grown using MMS. **(A)** 24 h after set-up. **(B)** 48 h after set-up.

The rapid nature of this crystal degradation means that CPR-GrpE crystals should be harvested within ~ 24 h to maximise crystal quality for diffraction data collection. Conducting crystallisation experiments at lower temperatures or with the addition of glycerol could also slow the crystal growth and subsequent degradation. Whilst diffraction data could not be collected on the degraded crystals in Fig. 7.9, the results from MMS are encouraging, and will hopefully lead to CPR-GrpE structure determination in future.

7.2.3. Production and characterisation of CPR-DnaJ

Expression and IMAC purification

The expression levels of CPR-DnaJ in full-scale cultures were extremely high, with approximately 30 mg of soluble protein obtained per litre of culture (Fig. 7.10).

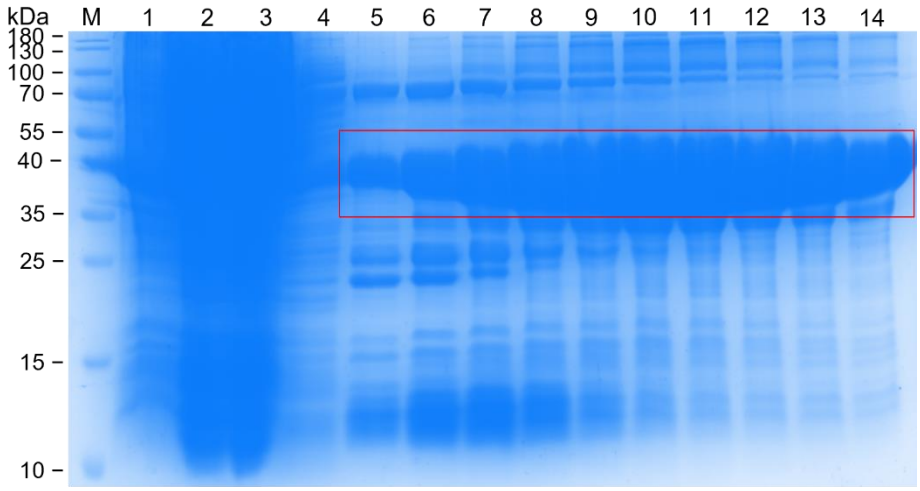

Fig. 7.10: SDS-PAGE analysis of CPR-DnaJ IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 14 = protein fractions. Bands corresponding to CPR-DnaJ (MW = 40.4 kDa) are indicated with a red box.

The lanes showing cell pellet, lysate, and flow through samples in Fig. 7.10 are heavily overloaded due to the high quantities of CPR-DnaJ present. Despite the losses in the flow through and wash, ample soluble CPR-DnaJ was recovered, though significant amounts of other contaminant proteins remain after IMAC.

SEC analysis

The purity of CPR-DnaJ after IMAC was reasonably poor (Fig. 7.10), and SEC was required for further purification. The results of SEC and corresponding SDS-PAGE analysis are shown in Fig. 7.11.

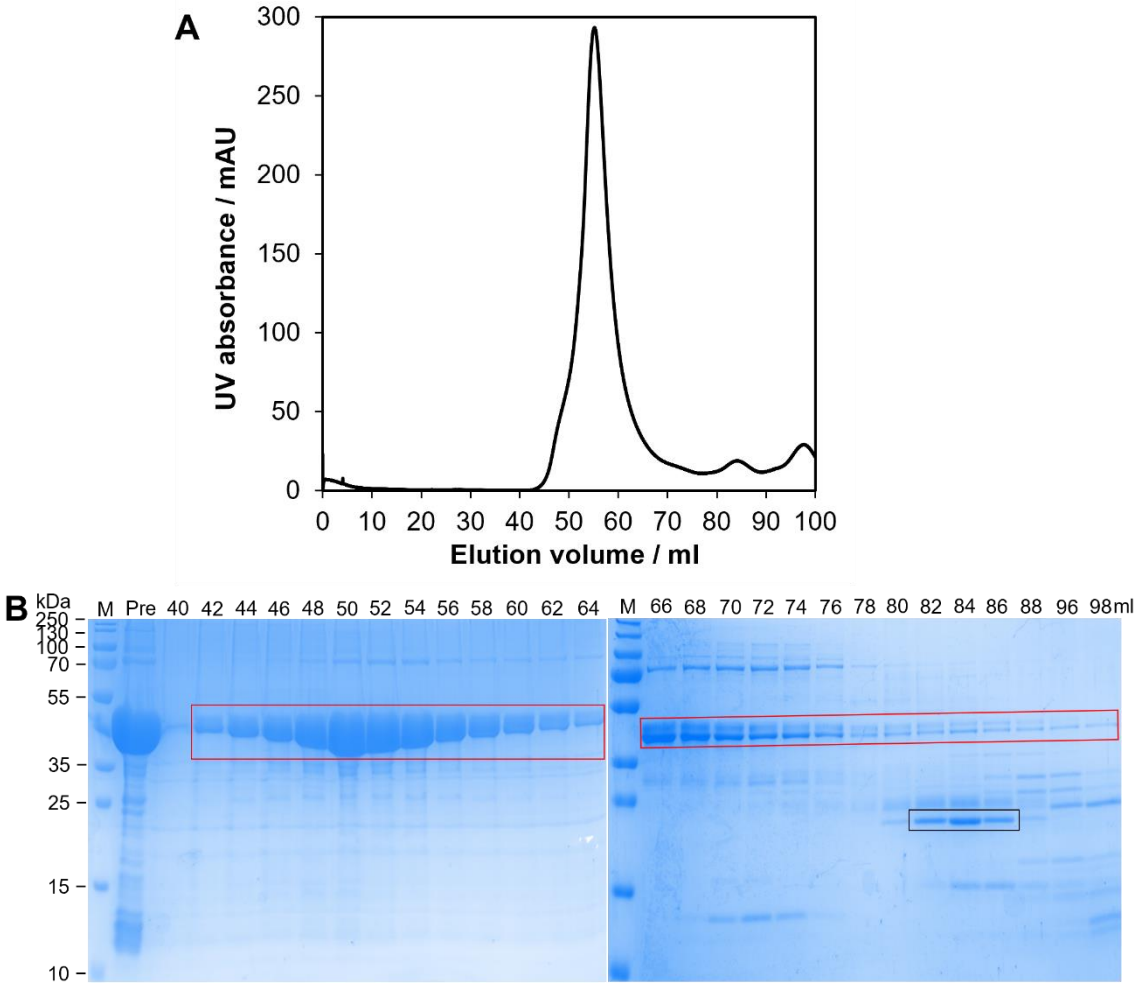

Fig. 7.11: (A) Chromatogram from SEC of CPR-DnaJ. **(B)** SDS-PAGE analysis of elution fractions from SEC of CPR-DnaJ (MW = 40.4 kDa). The position of bands corresponding to CPR-DnaJ are indicated with a red box. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The chromatogram from SEC shown in Fig. 7.11A identifies a single predominant species with V_e = 55.1 ml, which was confirmed to be CPR-DnaJ based on expected MW by SDS-PAGE analysis (Fig. 7.11B). Based on homologous DnaJ proteins, CPR-DnaJ is predicted to be monomeric (40.4 kDa), though its V_e corresponds to a > 500 kDa species, which is > 12 times its MW. The CPR-DnaJ in this sample is therefore heavily aggregated.

Close observation of the SDS-PAGE results in Fig. 7.11B also revealed two bands in the region of the 35 kDa to 55 kDa region of the gel; this is particularly pronounced in lanes 15 to 20. This suggests that CPR-DnaJ copurified with another protein of similar MW during SEC. The small peak at 84.1 ml corresponds to a 44.5 kDa species, which is close to the MW of the CPR-DnaJ monomer (40.4 kDa). However, SDS-PAGE revealed that the pertinent species, outlined in black in Fig. 7.11B, was < 25 kDa and so could not be the full-length CPR-DnaJ protein. The small peak at 97.6 ml corresponds to a 14.3 kDa species, which is likely an *E. coli* contaminant.

Characterisation

Due to the precipitation encountered during purification, CD was also used to determine if CPR-DnaJ was folded; the resulting spectrum is shown in Fig. 7.12.

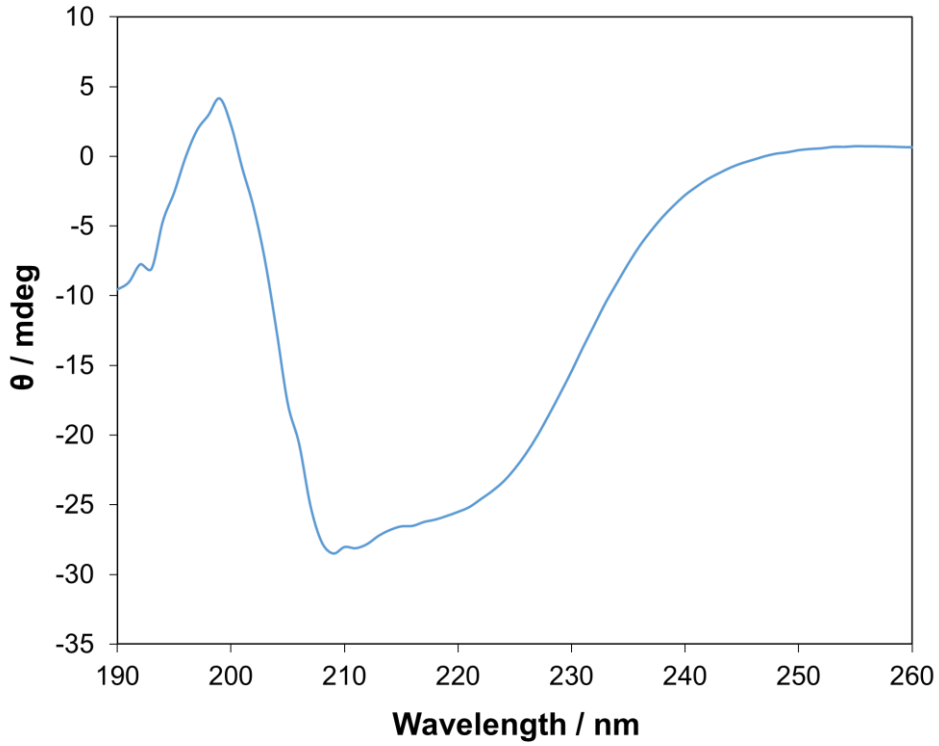

Fig. 7.12: Circular dichroism spectrum for CPR-DnaJ, with CD measured in ellipticity (θ).

The CD spectrum for CPR-DnaJ is slightly ragged, which could be due to sample heterogeneity. Regardless, the spectrum does show that the protein is folded in solution, with characteristics of both α-helical and antiparallel β-sheet regions. The negative peaks at 208 nm and 222 nm, typical of a predominantly α-helical protein, are not as well defined as for CPR-GrpE (Fig. 7.4), likely due to contribution from β-sheet that has characteristic negative peaks at 218 nm and around 195 nm[333,334].

Trypsin digest MS was used to confirm the identity of CPR-DnaJ (40.4 kDa) in the sample after SEC. However, common *E. coli* contaminants RecA (38.0 kDa) and DNA directed RNA polymerase subunit α (36.5 kDa) were also identified in the sample[337,338]. The close MWs of these proteins makes separation through SEC methods problematic, explaining the presence of multiple bands in Fig. 7.11B. An additional purification step would be required to separate CPR-DnaJ from these similar MW proteins, such as IEC. However, the protein precipitated heavily during purification, and it was decided to instead conduct TSA with the remaining sample to identify stabilising conditions.

Thermal shift analysis

TSA of CPR-DnaJ resulted in suboptimal melt curves that prevented accurate estimation of T_m values. This meant no comparisons could be drawn to find stabilising conditions. An example melt curve is shown in Fig. 7.13.

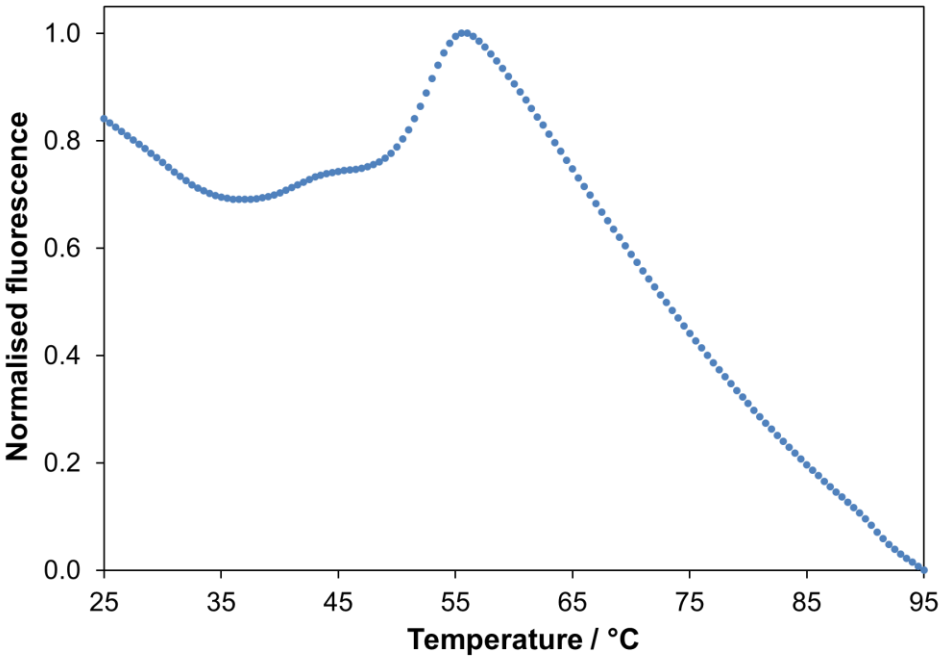

Fig. 7.13: Melt curve showing normalised fluorescence intensities at 590 nm for CPR-DnaJ in MilliQ water, from the results of TSA with the Durham pH Screen®[210,255]. Experimental data are shown as individual points.

The background fluorescence was high, and no distinct thermal transitions could be identified. It could be that CPR-DnaJ has exposed hydrophobic regions, making it difficult to track thermal denaturation with the standard fluorescent dye used for TSA. This could be resolved in future by exploiting the intrinsic fluorescence of the tyrosine residues in the CPR-DnaJ sequence, or thermal denaturation could be tracked using alternative methods such as CD or differential scanning calorimetry (DSC).

Due to the issues with instability and precipitation, work on the target was postponed at this stage to focus on the other CPR targets. Purification in an alternative buffer, such as Tris or phosphate, could alleviate these issues in future to enable further characterisation and crystallisation.

7.2.4. Production and characterisation of CPR-DnaK

Expression and IMAC purification

CPR-DnaK also overexpressed well in full-scale cultures, with ~ 25 mg of soluble protein retrieved per litre of culture by IMAC purification (Fig. 7.14).

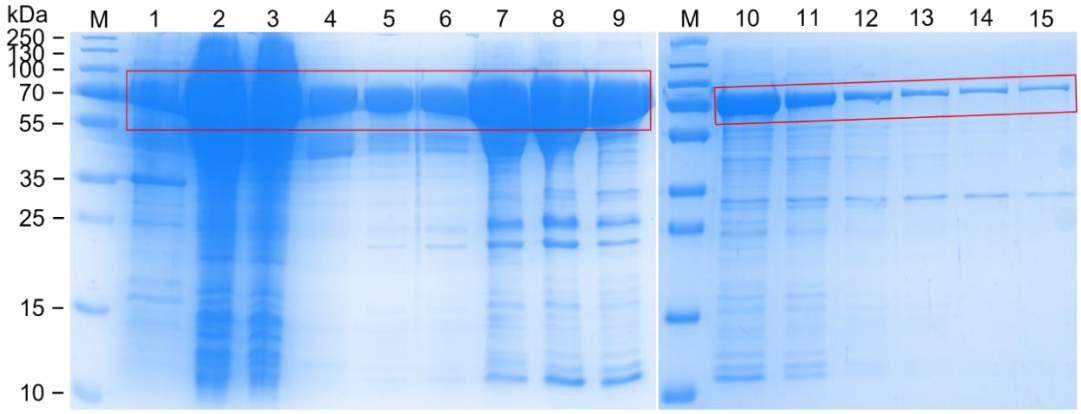

Fig. 7.14: SDS-PAGE analysis of CPR-DnaK IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 15 = protein fractions. Bands corresponding to CPR-DnaK (MW = 69.6 kDa) are indicated with a red box.

Although insoluble CPR-DnaK was detected in the cell pellet, a far stronger band of soluble protein was observed in the lysate. A significant quantity of CPR-DnaK was also lost in the flow through and wash during IMAC purification; nevertheless, plentiful target protein was retrieved for further purification by SEC and downstream characterisation.

SEC analysis

Several contaminant proteins remained in the CPR-DnaK sample after IMAC purification (Fig. 7.14). SEC was therefore conducted to further purify the target, as well as assess its oligomeric state. The results of SEC and accompanying SDS-PAGE analysis are shown in Fig. 7.15.

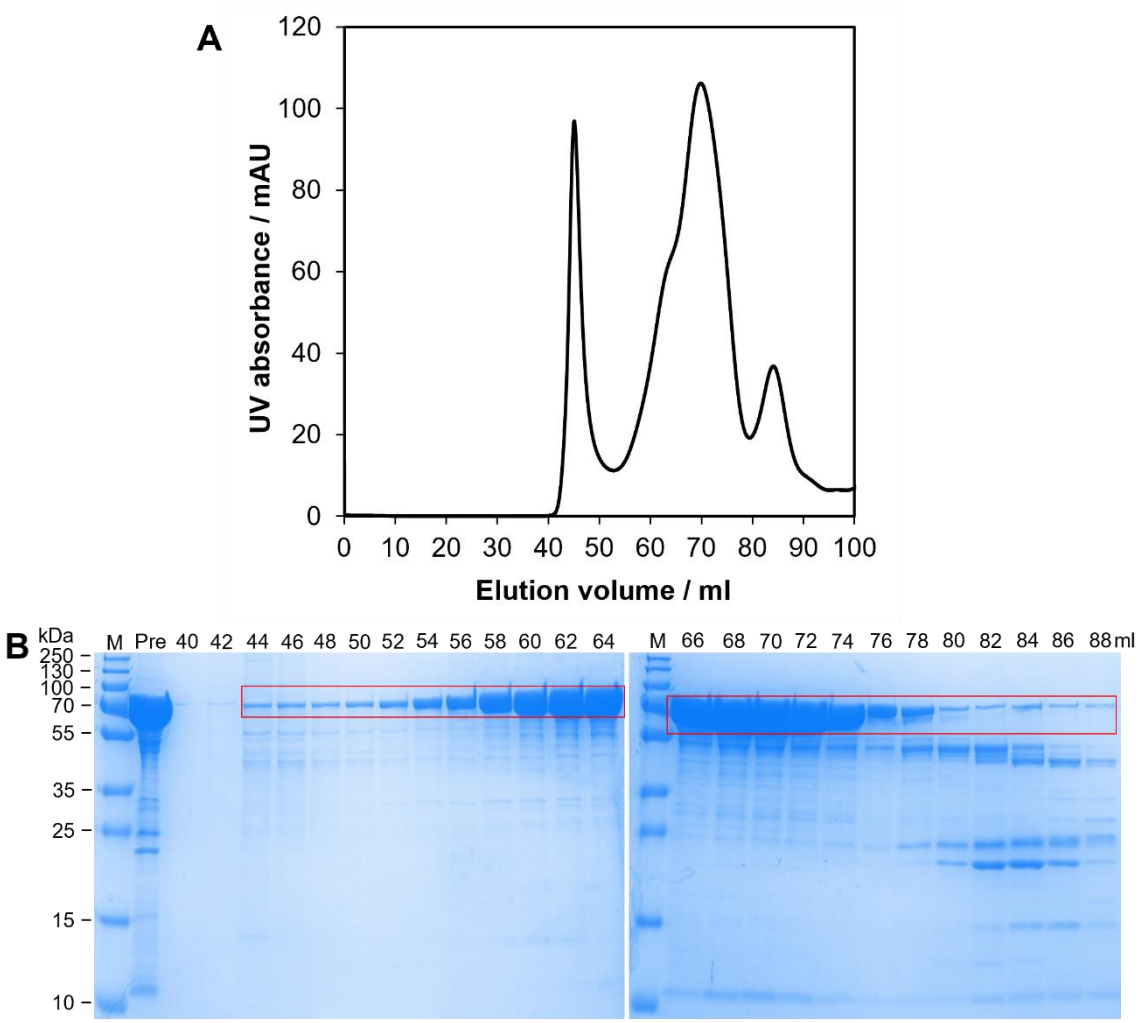

Fig. 7.15: (A) Chromatogram from SEC of CPR-DnaK. **(B)** SDS-PAGE analysis of elution fractions from SEC. M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The peak at 45.1 ml corresponds to a species with MW > 1 MDa, which is most likely aggregated protein. The peak at 84.1 ml corresponds to a 44.6 kDa species, which could be fragmented CPR-DnaK or a lower MW contaminant. The major peak at 69.9 ml corresponds to a 147.1 kDa species, concluded to be dimeric CPR-DnaK (139.2 kDa). However, the broad, unsymmetrical shape of the major peak indicates a somewhat heterogeneous sample. This could be the result of contaminant proteins with similar MWs to CPR-DnaK, as are evident from SDS-PAGE analysis (Fig. 7.15B).

Characterisation

The identity of the species highlighted in Fig. 7.15B was confirmed to be CPR-DnaK through trypsin digest MS. CD was also performed to determine if purified CPR-DnaK was folded (Fig. 7.16).

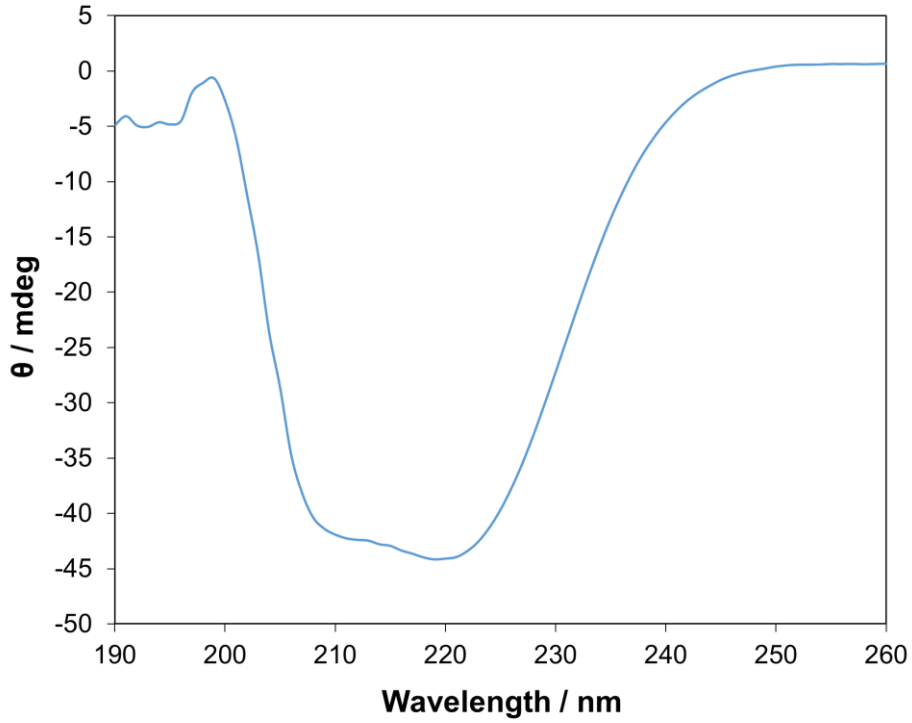

Fig. 7.16: Circular dichroism spectrum for CPR-DnaK, with CD measured in ellipticity (θ).

The CD spectrum of CPR-DnaK is typical of a folded protein, with clear regions of secondary structure determined from the shapes of the curves in characteristic wavelength regions[333,334]. Based on the combination of negative peaks in the 208 nm to 222 nm region, and negative ellipticity around 195 nm, CPR-DnaK contains a mixture of α-helical and β-sheet secondary structure.

Thermal shift analysis

pH screening

The TSA melt curves of CPR-DnaK are more typical than those of CPR-GrpE and CPR-DnaJ, with a single sharp melt transition observed at T_m ~ 47.4 in water. Interestingly, most conditions screened resulted in a decrease in T_m relative to this reference, as demonstrated by the results of TSA with the Durham pH screen® in Fig. 7.17.

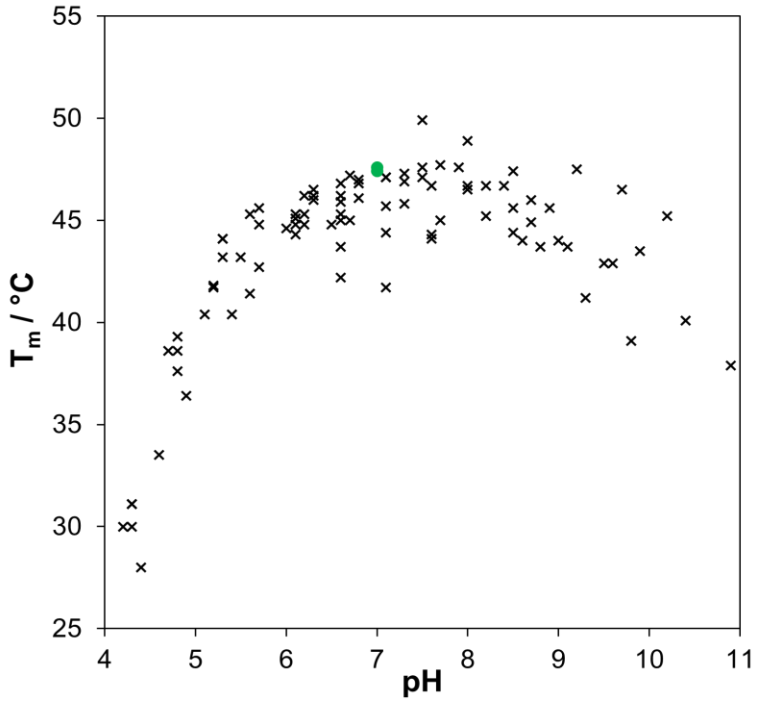

Fig. 7.17: Results from TSA using the Durham pH Screen®[210,255] showing the effect of pH on the T_m of CPR-DnaK. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the same pH value are from different buffer systems at identical pH.

Most buffers across the pH spectrum resulted in a decrease in T_m relative to the control, with extremes of pH causing the greatest destabilisation, particularly acidic conditions under pH 5.5. The pH appears to have little effect at middling values, though the buffer composition does cause variation in T_m at identical pH values. This is particularly noticeable at pH 7.1, with the T_m ranging from 41.7 °C in 100 mM bis-Tris propane, to 47.1 °C in 100 mM bis-Tris and 100 mM MOPS buffer. HEPES buffer at pH 7.5, as used for purification of CPR-DnaK, had very little effect on T_m, with a minor decrease of ~ 0.3 °C. However, 100 mM imidazole caused a ~ 3 °C reduction in T_m; imidazole was therefore removed from CPR-DnaK purification buffers immediately after IMAC in later experiments.

Salt screening

TSA with the Durham Salt Screen® revealed the effects of salt concentration and composition on the thermostability of CPR-DnaK. The changes in T_m with increasing concentrations of a variety of salts are shown in Fig. 7.18.

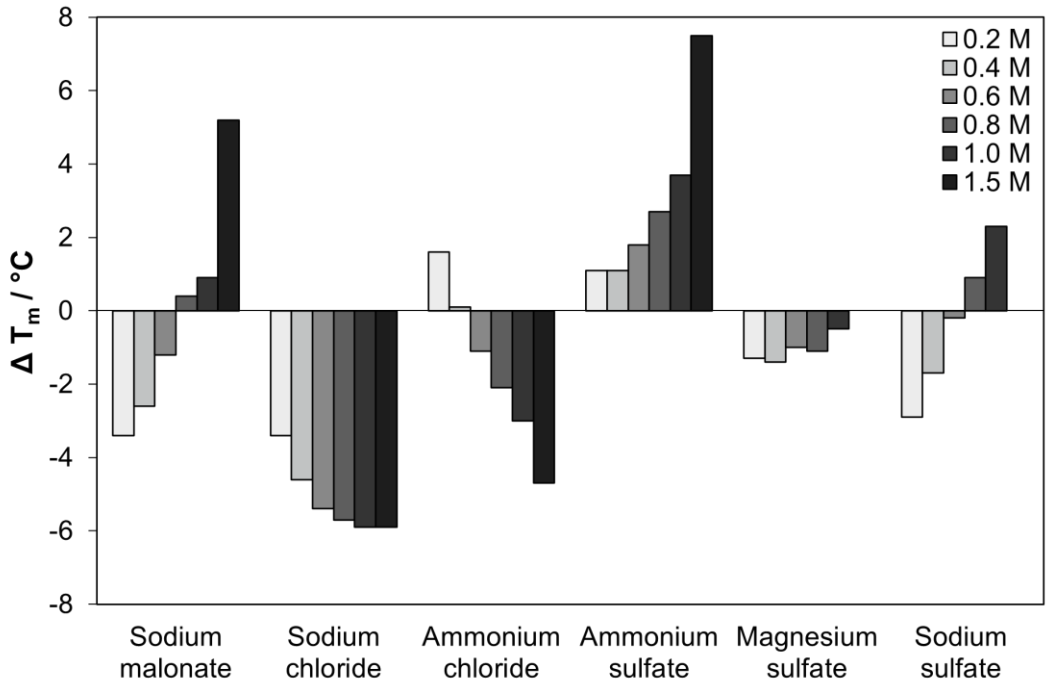

Fig. 7.18: Results from TSA of CPR-DnaK with the Durham Salt Screen®[210,255] showing changes in T_m upon addition of salts at increasing concentrations. Magnesium sulfate and sodium sulfate were only tested up to 1.0 M.

Of the generic salts tested, only ammonium sulfate was found to be stabilising across the full range of concentrations. Interestingly, sodium malonate and sodium sulfate were stabilising at high concentrations (≥ 0.8 M) but destabilising at lower concentrations (≤ 0.6 M). Conversely, ammonium chloride was slightly stabilising at low concentrations, becoming increasingly destabilising above 0.4 M. Sodium chloride was destabilising at all concentrations tested, and increasingly so at higher concentrations. These results together could suggest that, whilst ammonium and sulfate ions are stabilising, sodium and chloride ions destabilise CPR-DnaK. This would clarify the switch from destabilising to stabilising with increasing concentrations of ammonium chloride and sodium sulfate, which are both composed of one apparently stabilising and one destabilising ion. The results of sodium chloride addition are particularly significant since this is used as standard in buffers throughout purification and characterisation. For future assays with CPR-DnaK, ammonium sulfate could be used as an alternative salt in buffers.

Osmolyte screening

The Durham Osmolyte Screen® also revealed several stabilising conditions that increased the T_m of CPR-DnaK, including the natural osmolytes glycerol, trimethylamine N-oxide (TMAO), and glycine[339]. The addition of L-glutamic acid was also noteworthy, causing a decrease in T_m of 17 °C relative to the control in water, demonstrated by a distinct shift in the melt curve as shown in Fig. 7.19.

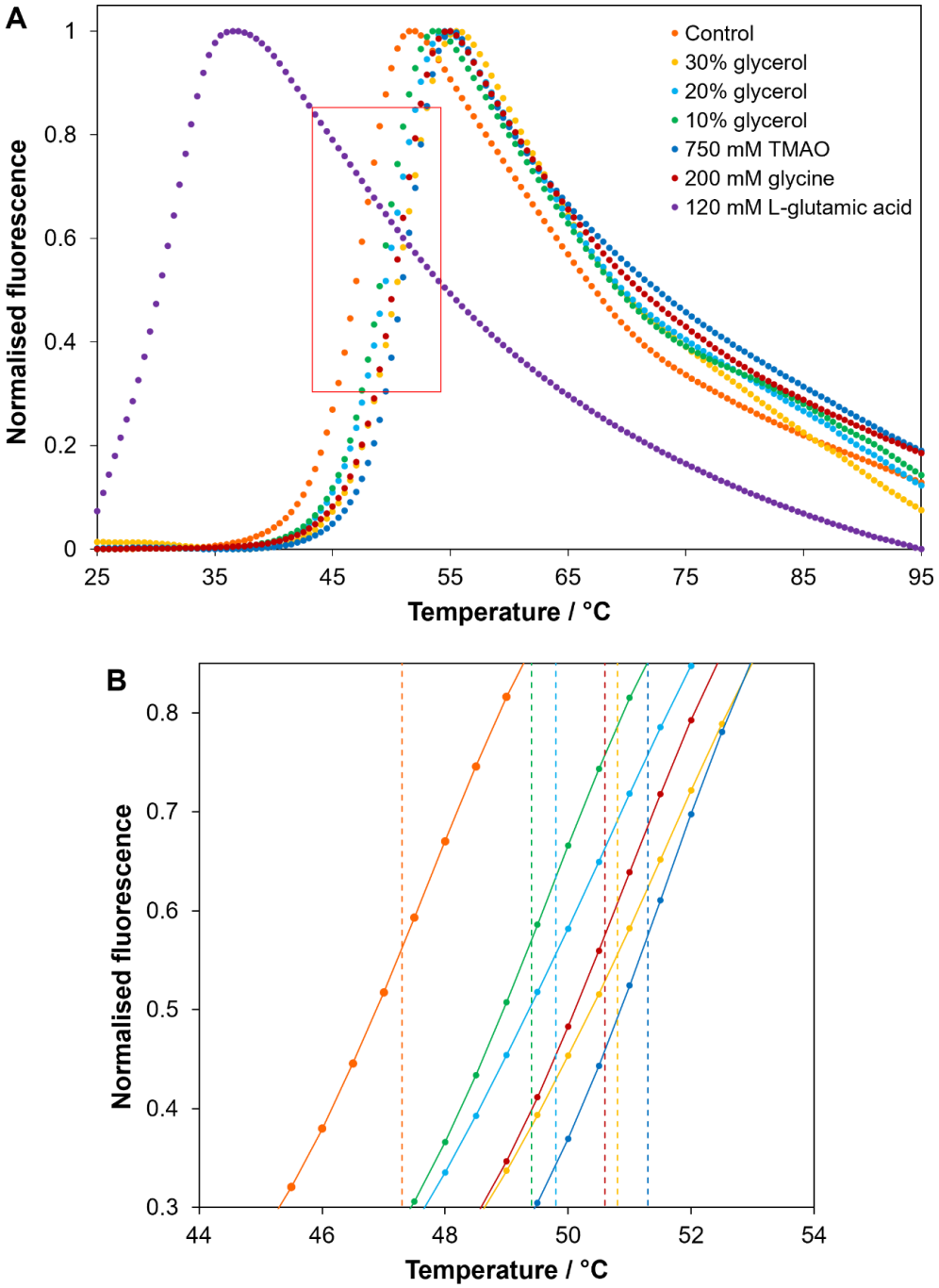

Fig. 7.19: (A) Melt curves showing normalised fluorescence intensities at 590 nm from TSA of CPR-DnaK with various osmolytes against a control in MilliQ water. Experimental data are shown as individual points. **(B)** Closer view of the region outlined in red from (A), highlighting the shift in T_m with addition of osmolytes. Dashed vertical lines indicate calculated T_m values. For clarity, the melt curve of destabilising L-glutamic acid is not shown.

The greatest stabilisation was seen with TMAO, which caused a ~ 4 °C increase in T_m. The extent of stabilisation seen with glycerol increased with concentration, though even the lowest concentration assessed (10%) caused a significant ~ 2.5 °C increase in T_m. Since it was readily available, 10% glycerol was subsequently added to storage buffer to improve the stability of CPR-DnaK.

The T_m values observed during TSA with CPR-DnaK are still well below the 75 °C that the source metagenome was isolated; however, the increased thermostability resulting from addition of some common salts and natural osmolytes gives an indication of how CPR-DnaK could be stabilised *in vivo*.

7.2.5. Production and characterisation of CPR-ClpB

Expression and IMAC purification

Full-scale overexpression of CPR-ClpB was extremely successful, producing a maximum of ~ 40 mg of soluble target protein per litre of culture. The results of SDS-PAGE after IMAC purification are shown in Fig. 7.20.

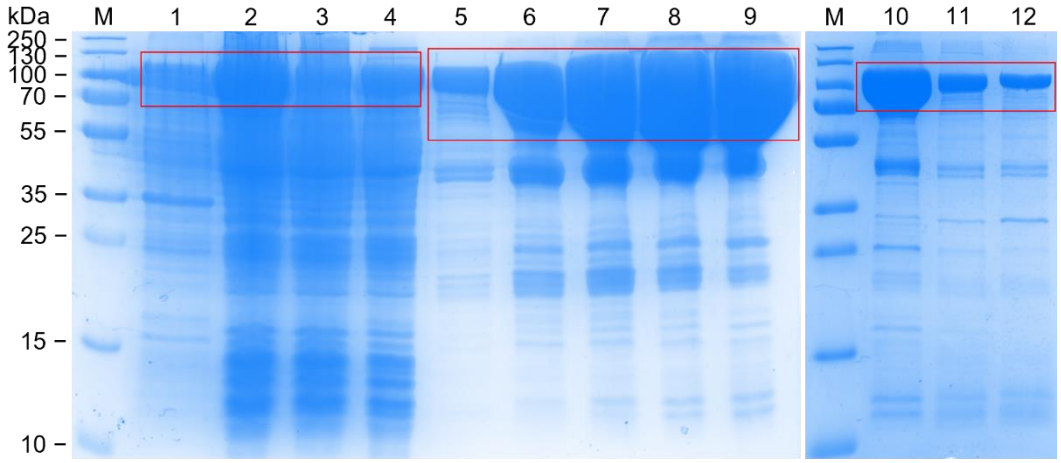

Fig. 7.20: SDS-PAGE analysis of CPR-ClpB IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 12 = protein fractions. Bands corresponding to CPR-ClpB (MW = 102.0 kDa) are indicated with a red box.

Despite some insoluble protein being detected in the cell pellet, the vast majority of CPR-ClpB produced was found to be soluble. Some CPR-ClpB was lost through column loading and washing during IMAC, though the soluble protein obtained was still more than sufficient for downstream uses.

Due to the high MW of CPR-ClpB, later SDS-PAGE analysis was conducted using an 8% polyacrylamide gel rather than the standard 12% used for other targets.

SEC analysis

Further purification was required to remove remaining contaminant proteins from the CPR-ClpB sample after IMAC (Fig. 7.20). Subsequent SEC analysis improved the purity of CPR-ClpB, as can be seen in Fig. 7.21.

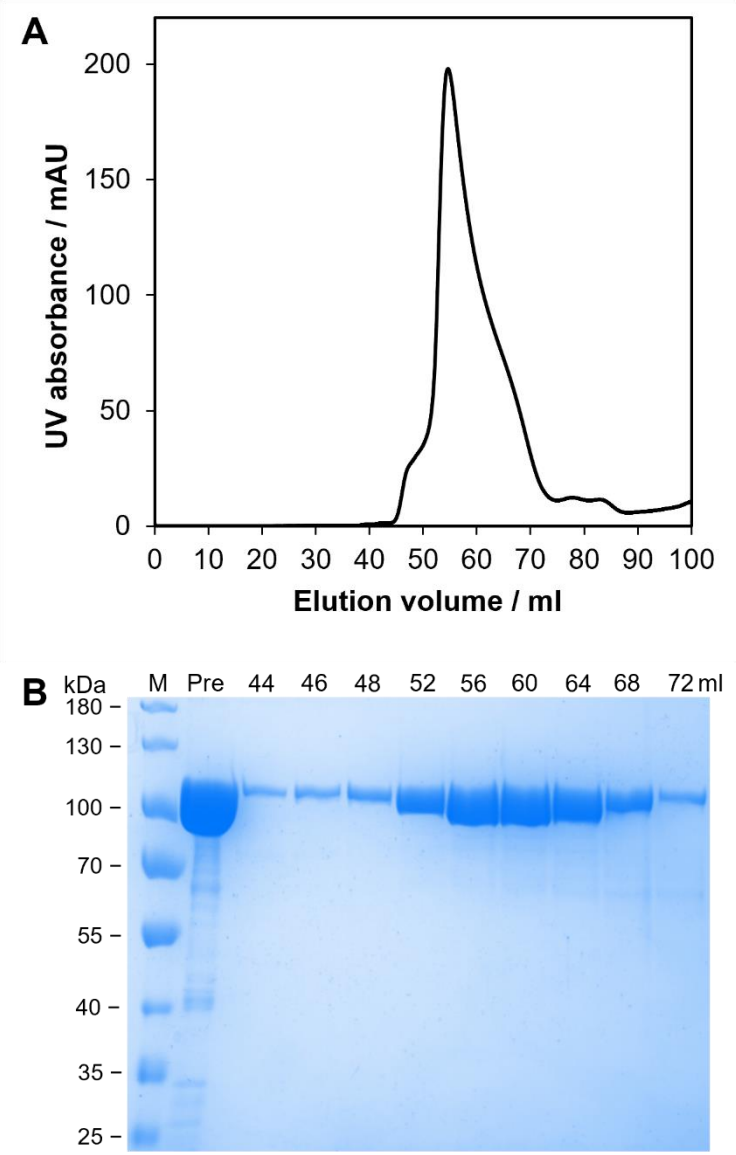

Fig. 7.21: (A) Chromatogram from SEC of CPR-ClpB. **(B)** SDS-PAGE analysis of elution fractions from SEC with CPR-ClpB (MW = 102 kDa). M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The shoulder on the left side of the major peak at ~ 46.5 ml correlates to a > 1 MDa species, presumed to be aggregated CPR-ClpB. The major peak at 54.7 ml corresponds to a 528.3 kDa species, which is slightly higher that the CPR-ClpB pentamer (MW = 510.0 kDa). Homologous ClpB proteins are known to assemble into hexameric complexes, and so it is proposed that CPR-ClpB will do likewise. However, the MWs of a CPR-ClpB pentamer and hexamer are at the top end of the fractionation range of the column used (10 kDa to 600 kDa[340]). SEC studies with a Superose 6 Increase column that can resolve proteins of higher MWs (up to 5 MDa[340]) would be beneficial to properly analyse the size of the CPR-ClpB complex and separate it from aggregated protein or other multimers.

Characterisation

The MW of CPR-ClpB was beyond the attainable limit of the ESI-TOF MS technologies of the departmental service in Durham. Trypsin digest MS was therefore used to confirm the identity of the species isolated through SEC (Fig. 7.21) as CPR-ClpB.

The CPR-ClpB sample was also analysed by CD to assess whether the protein was folded, with the resultant spectrum shown in Fig. 7.22.

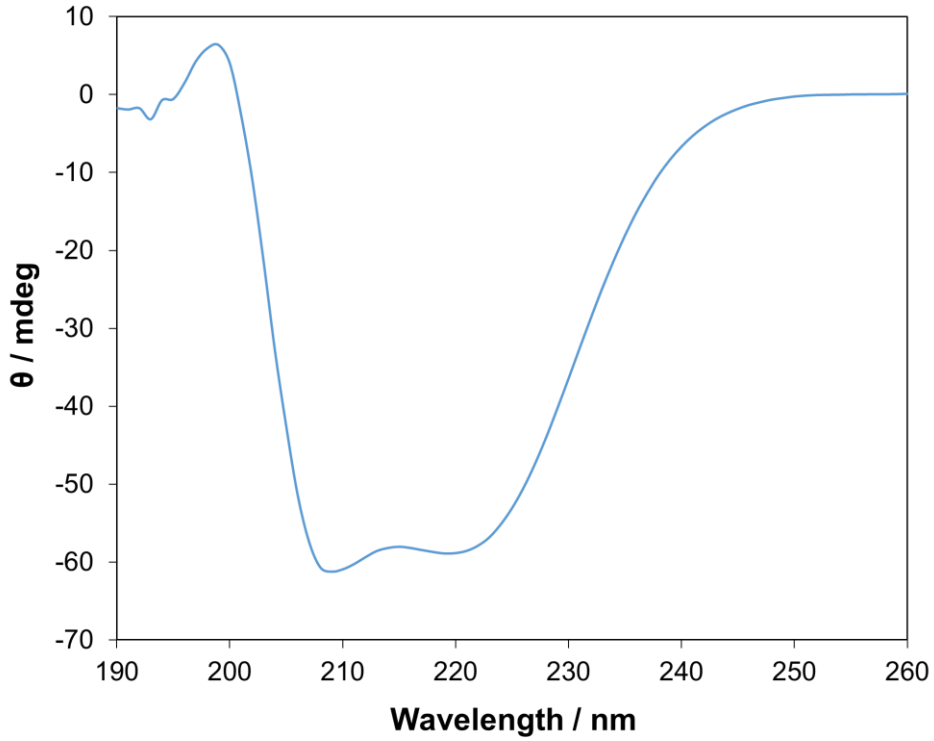

Fig. 7.22: Circular dichroism spectrum for CPR-ClpB, with CD measured in ellipticity (θ).

The CD analysis confirmed that the target protein was folded in solution. The two distinct negative peaks at 208 nm and 222 nm also imply that CPR-ClpB has significant α-helical secondary structure[333,334].

Thermal shift analysis

pH screening

CPR-ClpB displayed a characteristic melt curve during TSA, with a T_m of ~ 47.4 across the three Durham Screens®. Interestingly, it was destabilised by most conditions tested during pH screening. The effect of pH on the T_m of CPR-ClpB is shown in Fig. 7.23.

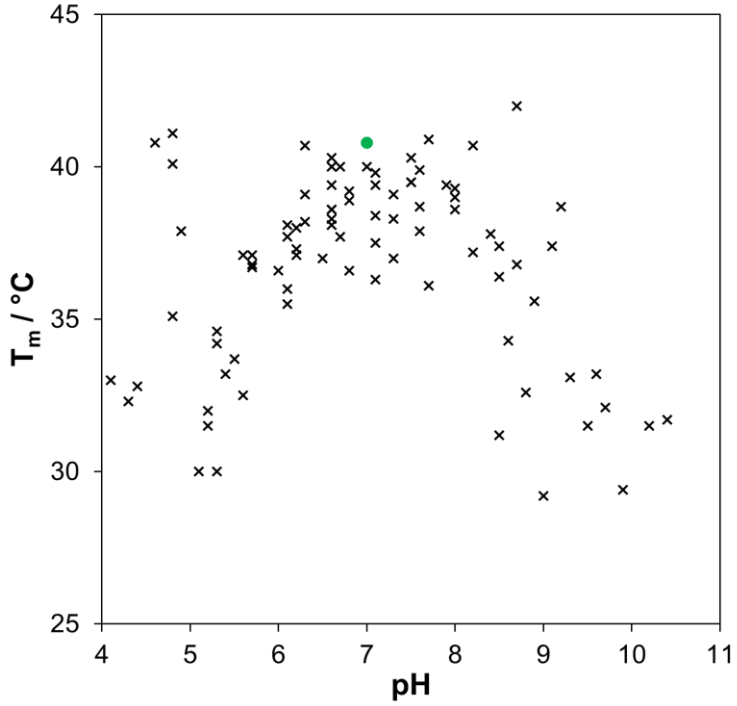

Fig. 7.23: Results from TSA using the Durham pH Screen®[210,255] showing the effect of pH on the T_m of CPR-ClpB. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the same pH value are from different buffer systems at identical pH.

As could be expected, CPR-ClpB was generally most stable at middling pH values between pH 6 and 8, with a significant decrease in T_m at more extreme pH values. The buffer composition does also have a notable effect; for example, at pH 7.1, there is a ~ 5 °C range in T_m depending on the buffer composition. The pH 7.5 HEPES buffer used as standard during purification was found to be marginally destabilising, with a ~ 1 °C reduction in T_m relative to the control. However, it is worth noting that the buffer concentration used in the TSA screen is significantly higher than for purification (100 mM *vs* 20 mM). Since TSA results provided no clear alternative candidates for a weakly alkaline buffer suitable for IMAC, HEPES was maintained in purification buffers. As with CPR-DnaK, imidazole was also found to destabilise CPR-ClpB, and so was removed quickly after IMAC purification.

Salt screening

CPR-ClpB was found to be destabilised by malonate and chloride salts, and stabilised by sulfate ions, as can be seen in Fig. 7.24.

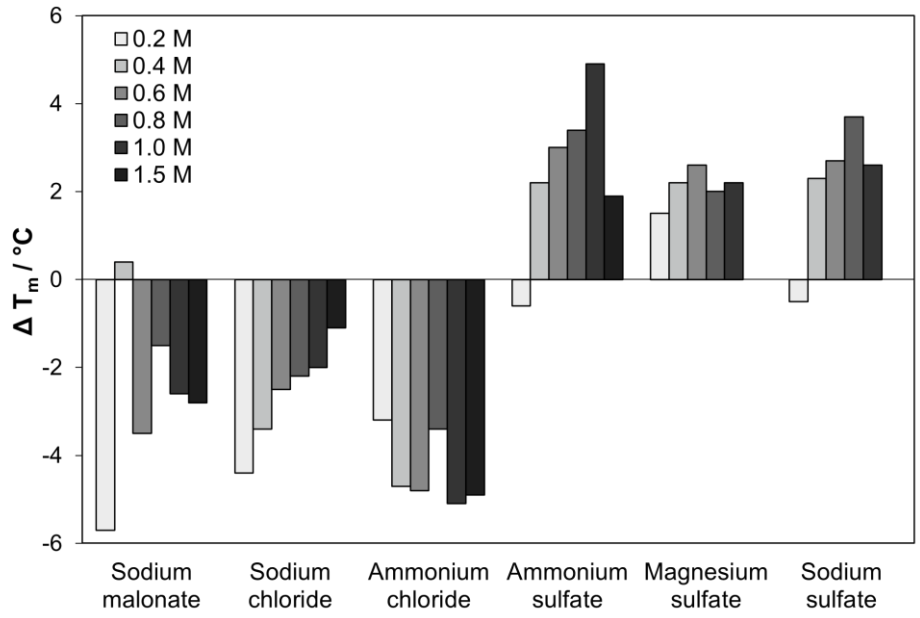

Fig. 7.24: Results from TSA of CPR-ClpB with the Durham Salt Screen® showing changes in T_m upon addition of salts at increasing concentrations. Magnesium and sodium sulfate were only tested up to 1.0 M.

Sodium chloride, which is present in purification buffers, caused a notable reduction in T_m. The extent of this destabilisation increased as the concentration of NaCl was decreased, with a ~ 4 °C reduction of T_m at 200 mM. An alternative stabilising salt, such as magnesium sulfate, could replace NaCl in future if necessary.

Osmolyte screening

TSA of CPR-ClpB with the Durham Osmolyte Screen® revealed that the protein was stabilised by number of common osmolytes, with a particular preference for glycerol and high concentrations of sugars and sugar alcohols (Fig. 7.25).

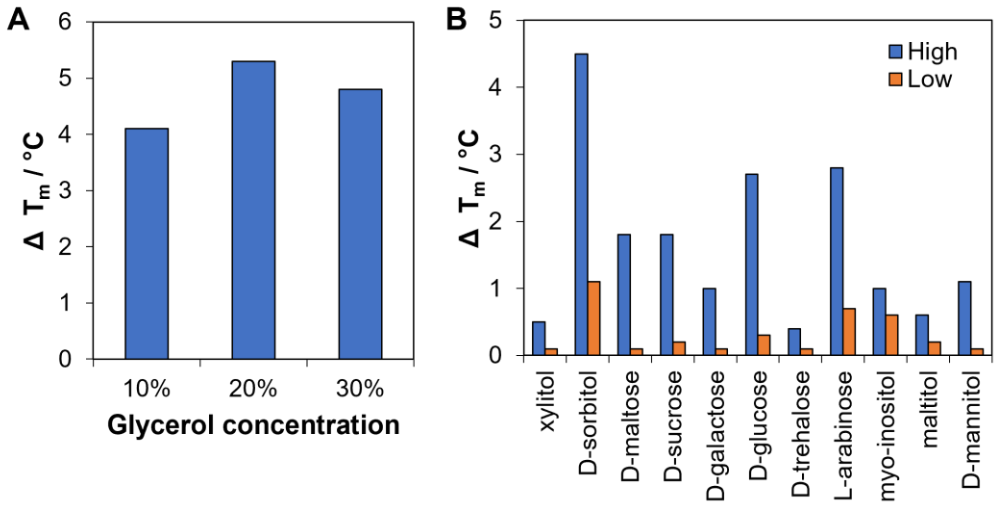

Fig. 7.25: (A) Results of glycerol addition on the T_m of CPR-ClpB relative to a control in MilliQ water, from TSA with the Durham Osmolyte Screen®[210,255]. **(B)** Results of the addition of sugars and sugar alcohols on the T_m of CPR-ClpB relative to the control, from TSA with the Durham Osmolyte Screen®.

Glycerol addition resulted in a > 4 °C increase in T_m at all three concentrations (Fig. 7.25A), and CPR-ClpB was subsequently stored in buffer containing 10% glycerol. The sugars and sugar alcohols shown in Fig. 7.25B were tested at a 'high' and a 'low' concentration, depending on their solubility and effective concentration ranges (Table A3)[210]. In each case, the higher of the two concentrations tested was considerably more stabilising, with 1 M D-sorbitol resulting in the largest increase in T_m of > 4.5 °C. CPR-ClpB was also significantly stabilised by TMAO, with an increase in T_m of > 7 °C at 750 mM and > 2 °C at 100 mM. The broad stabilisation of CPR-ClpB with a wide range of sugars, sugar alcohols and other common natural osmolytes suggests that CPR-ClpB could be stabilised by a variety of 'chemical chaperones' *in vivo*[339].

Crystallisation

Attempts to crystallise CPR-ClpB were unsuccessful, with precipitation occurring in most screened conditions even at very low (> 2 mg/ml) protein concentration. This is likely due to the large size (612 kDa) of the predicted hexameric CPR-ClpB complex, since with larger protein complexes are typically more flexible and less stable, which can discourage effective packing into ordered diffracting crystals[341].

Previous studies of homologous ClpB proteins have involved a combination of X-ray crystallography of isolated ClpB monomers, along with single particle cryo-Electron Microscopy (cryo-EM) of the hexameric complexes[324–326,328,342–345]. Attempts could be made to isolate the CPR-ClpB monomer for crystallisation. However, the putative CPR-ClpB hexamer is of more interest as the predicted functional unit of the protein and is an ideal candidate for structural determination by cryo-EM due to its large size[346]. Work towards structure determination of CPR-ClpB using single particle cryo-EM will be addressed in Chapter 8.

7.3. Conclusions

All four components of the CPR-DnaK/ClpB bi-chaperone system overexpressed in large quantities using the standard expression protocol outlined in 3.1. Insoluble expression was evident in each case; however, the levels of soluble expression were substantially higher, and ample quantities of soluble protein were obtained for all four targets. Optimisation of the expression protocol was therefore not considered necessary.

IMAC purification proved to be successful for capturing the target proteins, though additional purification with SEC was required prior to characterisation and crystallisation. Through the use of calibrated gel filtration columns, these SEC experiments also allowed the multimeric states of the proteins to be assessed. Attempts to obtain ESI-mass spectra for the three larger proteins (CPR-DnaJ, CPR-DnaK and CPR-ClpB) were unsuccessful, as the MWs of these proteins were above the attainable limit of available facilities. The proteins were therefore analysed by trypsin digest MS,

which confirmed their identities. Trypsin digest MS also revealed some *E. coli* contaminant proteins in the chaperone protein samples, though these were only present in minor quantities, and so further purification was not considered necessary. Additional purification steps, such as IEC, may prove necessary to facilitate crystallisation in future, particularly for CPR-DnaJ and CPR-DnaK.

TSA was conducted with the CPR chaperone proteins to assess their intrinsic thermostability and identify stabilising conditions to improve protein production and assist crystallisation experiments. The preliminary TSA results revealed that the chaperone proteins are less thermostable than anticipated when considering they are predicted to manage the heat shock response of their thermophilic source organism. It cannot be ruled out that the source organism did not originate from the intended sampling site, as contamination is always a possibility during metagenomics studies. However, due to time and resource limitations, these assays were only conducted once at this stage. TSA should be optimised and repeated to obtain more reliable results from which conclusions on thermostability, and potentially activity, could be drawn. As they stand, the results offer some insight into improving stability and reveal avenues for future investigation.

The CPR chaperone targets were given high priority due to their potential applications in the biotechnology sector. This groundwork has provided a strong starting point from which to optimise structural determination and functional characterisation of these proteins, leading to the eventual study of the bi-chaperone system as a whole and interactions between the constituent proteins. Since homologous chaperone proteins are known to interact and function cooperatively[316], the CPR-DnaK/ClpB bi-chaperone system is an excellent target for cryo-EM analysis, which could uncover the detailed workings and interactions of the various components during heat shock. Combined activity analysis will also provide a more accurate indication of thermostability and reveal potential applications for this CPR chaperone system.

Chapter 8. Single particle cryo-electron microscopy studies of the CPR-ClpB disaggregase

8.1. Introduction

8.1.1. Single particle cryo-electron microscopy

Cryo-electron microscopy (cryo-EM) uses 2D electron microscopy images of proteins or biological macromolecules embedded in a thin layer of vitreous ice to reconstruct their 3D structures[347,348]. This technique has seen a meteoric rise in popularity in the structural biology field, with numbers of cryo-EM structures deposited in the PDB rapidly increasing year on year. As of April 2022, the Electron Microscopy Data Bank (EMDB) contains nearly 20,000 entries (19,812, 27th April 2022), with 4483 entries from 2021 alone, compared to only 223 in 2010[349]. Single particle cryo-EM is extremely versatile and has been used to reveal the structures of a variety of biomolecules, including multi-component complexes, ribosomes, membrane proteins, and whole viruses[350].

Early cryo-EM structures were of poor resolution, earning it the moniker 'blobology' from some scientists[347,351]. However, a 'resolution revolution' resulting from advancements in hardware, such as faster cameras and direct electron detection, motion correction, as well as automation of data collection and processing, has led to an explosion in high-resolution structures[131,350,352–356]. Such technical improvements have even led to structures with resolution below 1.3 Å[130,132]. With the progression of cryo-EM to atomic resolution capacity, it is proving to be an increasingly attractive technique among structural biologists. Cryo-EM has arguably replaced X-ray crystallography as the preferred method for notoriously challenging membrane proteins, particularly with developments in lipid nanodiscs that provide a stabilising pseudo-native environment to facilitate structural analysis[130,357–360].

8.1.2. Negative staining

Obtaining a high-resolution structure through cryo-EM often demands comprehensive sample preparation, with iterative rounds of screening and optimisation required[350,361,362]. Negative stain EM offers relatively simple and inexpensive sample pre-screening at the start of the process, providing a useful indication of sample quality with immediate feedback[350,363,364]. This technique involves imaging a sample of interest embedded in a layer of dried heavy metal staining solution, typically uranyl acetate or formate, which significantly increases the contrast of the contained specimen relative to the background[364–366]. Whilst it does not resolve the high-resolution details or give internal structural information, it can quickly reveal overall domain level molecular structure. It also offers

general information on sample quality, such as homogeneity and aggregation[367]. This in turn assists sample preparation for high-resolution structure determination with cryo-EM[350,363].

8.1.3. Cryo-EM studies of ClpB homologs

Bacterial ClpB proteins and their eukaryotic Hsp100 homologs have been widely studied due to their importance in maintaining proteostasis[324–326,328,342–345]. A total of 96 PDB entries were identified as having significant similarity to CPR-ClpB using the sequence search tool[117]. These ClpB/Hsp100 proteins consistently form homohexameric, spiral-like structures, with a central channel through which protein substrates are threaded during disaggregation, in a dynamic process driven by ATP hydrolysis[368–370]. An example structure of the Hsp104 protein from the thermophilic fungus *Chaetomium thermophilum*, sharing 47.7% sequence identity with CPR-ClpB, is shown in Fig. 8.1.

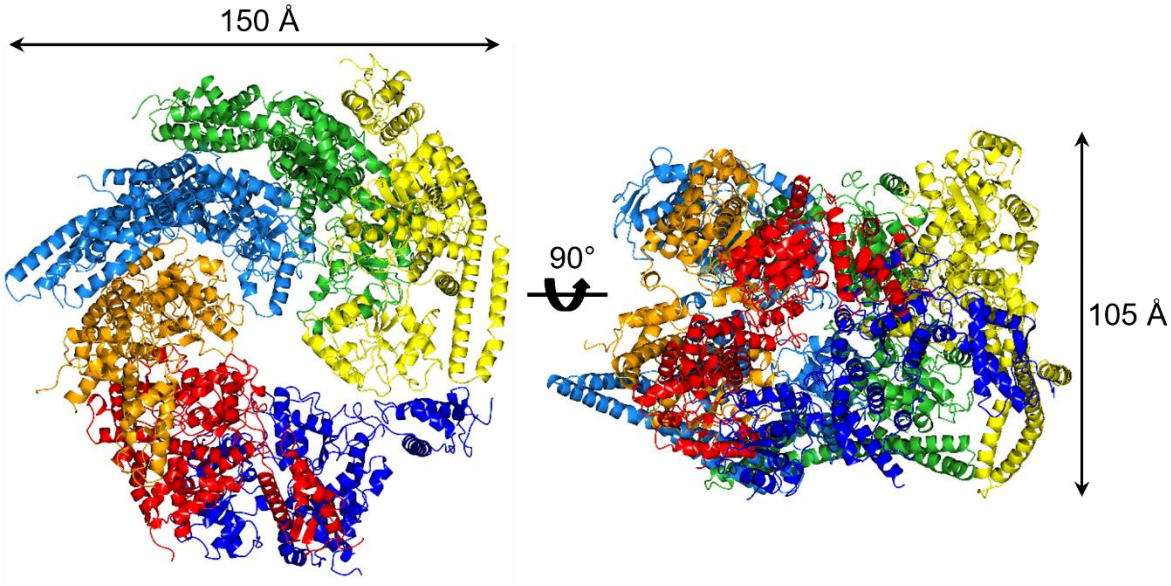

Fig. 8.1: Cryo-EM structure of the Hsp104 homohexamer from *C. thermophilum* (PDB accession code 7cg3; map resolution range of 4.5 - 7.4 Å)[371], in a staggered ring conformation with ADP. Left = top view; right = side view. Ribbon diagrams are coloured by chain and were generated using CCP4mg[372]. Approximate dimensions are shown in Å.

ClpB/Hsp100 hexamers have significant conformational flexibility, with several conformations detected upon binding of different nucleotides[370,371,373], and even multiple conformations for a single nucleotide state[373]. The dynamic hexamers can form near-symmetrical ring-like structures, as well as more open, extended spirals during substrate threading. Due to the conservation of this flexible homohexameric structure in ClpB/Hsp100 proteins, CPR-ClpB is expected to form a similarly dynamic oligomer with multiple conformation states. This, along with the large size of the CPR-ClpB hexamer (612 kDa), makes cryo-EM the most suitable method for structural analysis.

8.2. Results and discussion

8.2.1. Negative staining of CPR-ClpB samples

Prior to cryo-EM studies, a sample containing the CPR-ClpB hexamer from SEC was analysed by negative stain transmission election microscopy (TEM) by the Electron Microscopy Research Services at Newcastle University. The initial concentration (1 mg/ml) was found to be too high, though with dilution (0.02 mg/ml), individual particles could be identified from the resulting micrographs (Fig. 8.2).

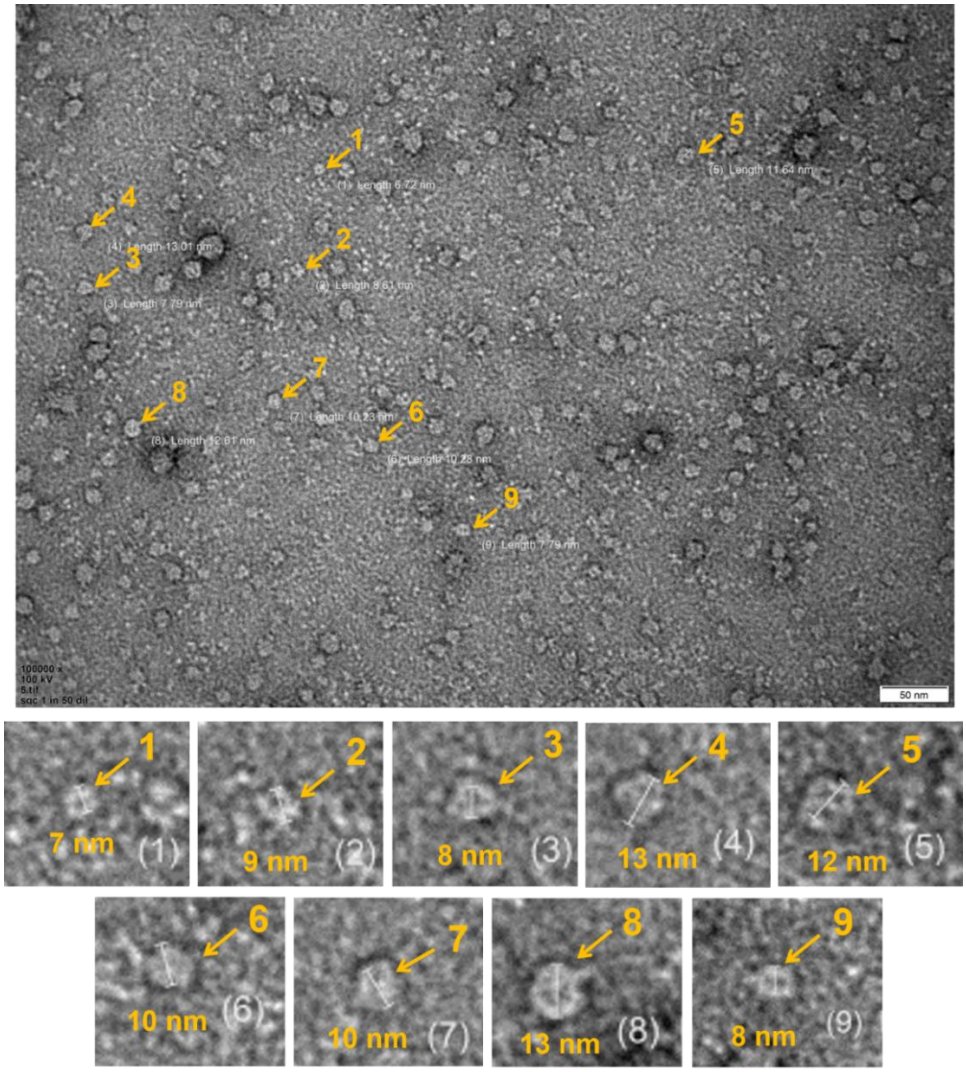

Fig. 8.2: Top: negative stain TEM micrograph of CPR-ClpB sample. Scale bar at the bottom right corner shows 50 nm/500 Å. Bottom: closer views of individual particles highlighted with yellow arrows, and approximate particle diameters.

Based on homologous ClpB/Hsp100 structures, it was expected that the diameter of the ClpB ring-like hexamer would be ~ 150 Å (15 nm). Greater discrepancy in height is expected due to conformational flexibility, though it is likely in the range of 80 Å to 110 Å. Whilst there is some heterogeneity and variation in size evident in Fig. 8.2, particles fitting the overall expected shape and estimated dimensions of the CPR-ClpB hexamer can readily be identified from the negative stain micrograph.

8.2.2. Single particle cryo-EM

Grid screening

Cryo-EM experiments for CPR-ClpB were performed by Dr Daniel Maskell at the Astbury Centre, Leeds. After preparation of duplicate grids at varying concentrations of CPR-ClpB, the grids were screened to identify the most suitable concentration and select regions with suitable ice thickness and particle distribution for data collection (Fig. 8.3).

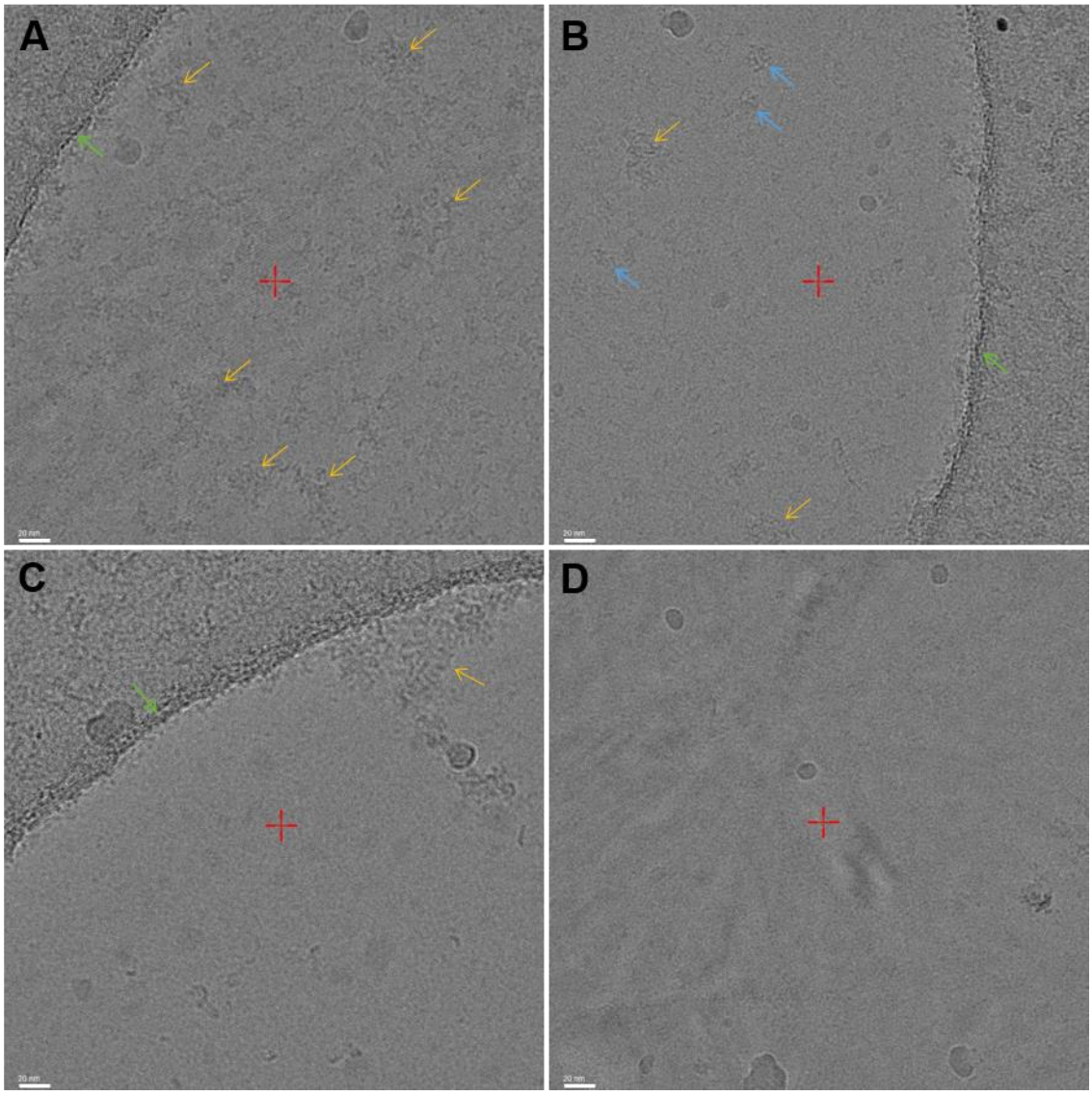

Fig. 8.3: Screening images for CPR-ClpB grids at varying concentrations. Green arrows indicate foil hole edges. Yellow arrows indicate regions of heavy aggregation. Blue arrows indicate potential single CPR-ClpB hexamer species. Scale bars at the bottom left corner of each image show 20 nm/200 Å. **(A)** 1 mg/ml CPR-ClpB. **(B)** 0.5 mg/ml. **(C)** 0.25 mg/ml. **(D)** 0.2 mg/ml with ATPγS.

Sample heterogeneity and aggregation was evident in micrographs at all concentrations of CPR-ClpB. At 1 mg/ml, heavy aggregation in both grids made it exceedingly difficult to distinguish any individual species (Fig. 8.3A). Dilution to 0.5 mg/ml alleviated this aggregation somewhat, with several appropriately sized particles for the CPR-ClpB hexamer also visible (Fig. 8.3B). Screening

of the first grid at 0.25 mg/ml CPR-ClpB revealed some aggregation around the edges of foil holes (Fig. 8.3C), and unsuitably thick ice that could limit high-resolution structural information[374,375]. The duplicate grid contained regions of ice with more acceptable thickness, though many foil holes within the grid squares were dry due to heavy blotting. The lower concentration of sample also appeared too dilute, with fewer distinguishable species than at 0.5 mg/ml. Data were therefore collected from a grid with 0.5 mg/ml ClpB (Fig. 8.3B), since this was the most suitable in terms of both ice quality and sample concentration.

Grids were also prepared for CPR-ClpB with the non-hydrolysable ATP analogue adenosine-5′-o-(3-thio-triphosphate) (ATPγS) to allow structural comparisons in the presence and absence of a nucleotide. However, a rippling effect in the ice (Fig. 8.3D) made these grids unsuitable for imaging.

Data collection and processing

A total of 813 movies were collected from foil holes across two grid squares on a grid prepared with 0.5 mg/ml CPR-ClpB (Fig. 8.3B). During inspection of motion-corrected micrographs, the contrast transfer function (CTF) fit values appeared to cluster based on expected resolution (Fig. 8.4). The explanation for this is unclear, though there are several possible reasons, including microscope issues during data collection, poor motion correction, crystalline ice on the grid, or radiation damaged sample. The poorer quality micrographs (> 10 Å resolution) were excluded at this stage, with particles subsequently picked from the remaining 316 micrographs.

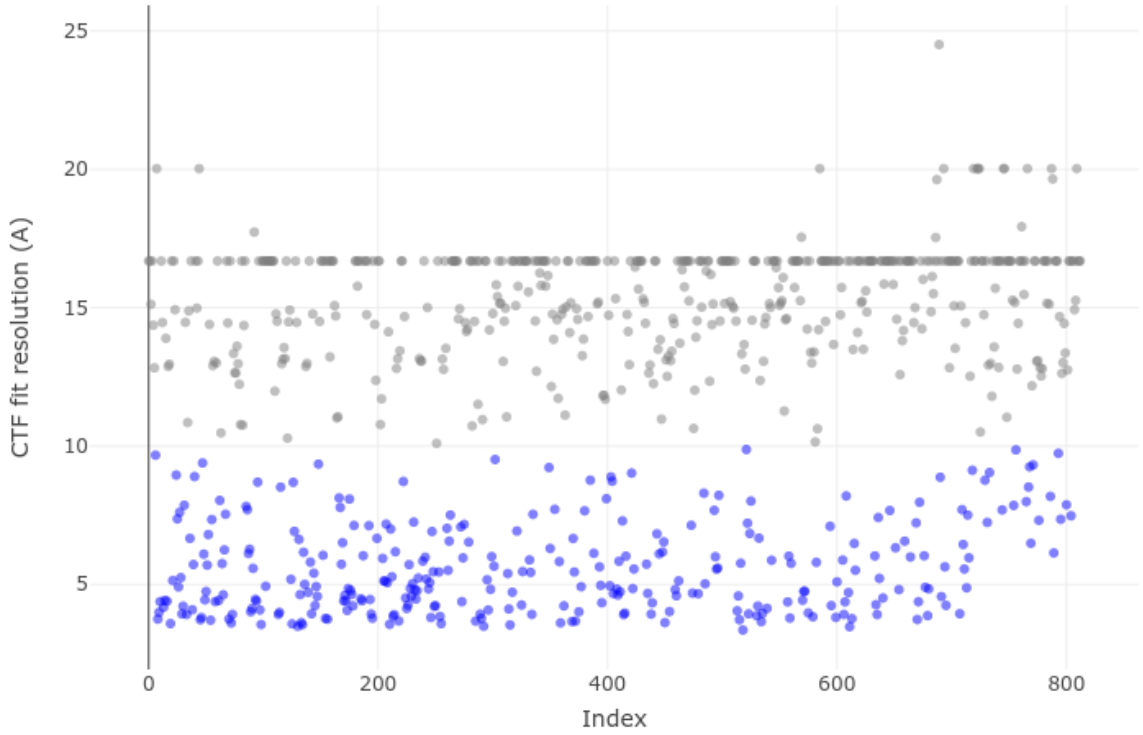

Fig. 8.4: Plot of CTF fit resolution in Å for each of the 813 motion corrected micrographs. Blue dots (≤ 10 Å) represent micrographs selected for particle picking. Grey dots (> 10 Å) represent excluded micrographs.

In the initial round of particle picking, 262,598 particles were identified using the Blob Picker tool in cryoSPARC[376], based on the expected dimensions of the CPR-ClpB hexamer. This number was approximately halved through removing obvious 'junk' particles, such as those in regions of heavy aggregated material, and implementing selection based on high normalised cross-correlation (NCC) scores and local power (NCC score > 0.1; 380 < local power < 660), which are indicative of a good agreement with expected particle size and significant signal, respectively[377,378].

A large number (200) of 2D classes were generated from the 127,239 remaining particles to account for sample heterogeneity and multiple predicted conformations of CPR-ClpB. From this, 25 encouraging classes containing a total of 111,343 particles were assigned as templates for a second round of particle picking. The resulting 106,087 particles after inspection and selection (NCC score > 0.35; 561 < local power < 826) were used to generate further sets of 50, 100, and 200 2D classes, with the most promising classes observed in the largest set (200). The 27,532 particles constituting 43 reasonable classes were used for *ab-initio* reconstruction of five separate coarse-resolution maps, again to account for anticipated heterogeneity. The most promising map is shown in Fig. 8.5.

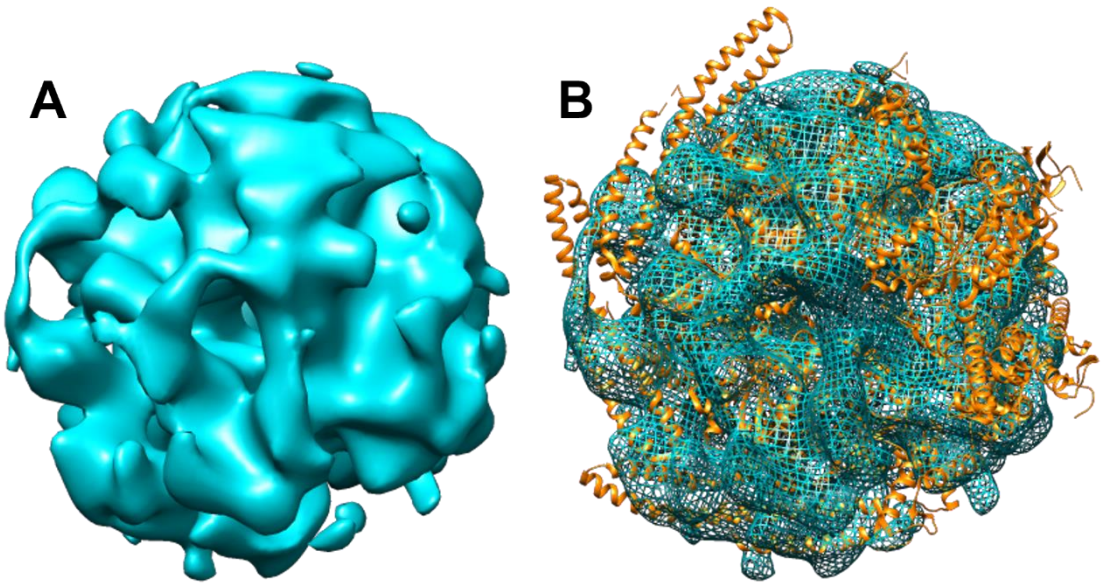

Fig. 8.5: (A) CPR-ClpB map surface from *ab-initio* reconstruction and refinement at contour level 0.0013. **(B)** Map surface in mesh (blue) overlayed with the crystal structure of Hsp104 from *C. thermophilum* (orange; resolution 2.7 Å)[371] using the 'fit model to map' tool in UCSF Chimera[379].

The overall dimensions of the map surface in Fig. 8.5 are appropriate for the CPR-ClpB hexamer. However, the coarse resolution of the map does not show distinct subunits or a central pore through a hexameric ring or spiral. Optimisation of data processing, with iterative rounds of particle picking, and template classification could improve the resulting maps to an extent. However, it was decided that optimising sample preparation for collection of better-quality data was the most practical solution.

8.2.3. Optimisation of CPR-ClpB samples

SEC analysis

Over the course of several SEC experiments with CPR-ClpB following an identical protocol on the same gel filtration column, large variations in elution volume, V_e, were observed (Fig. 8.6).

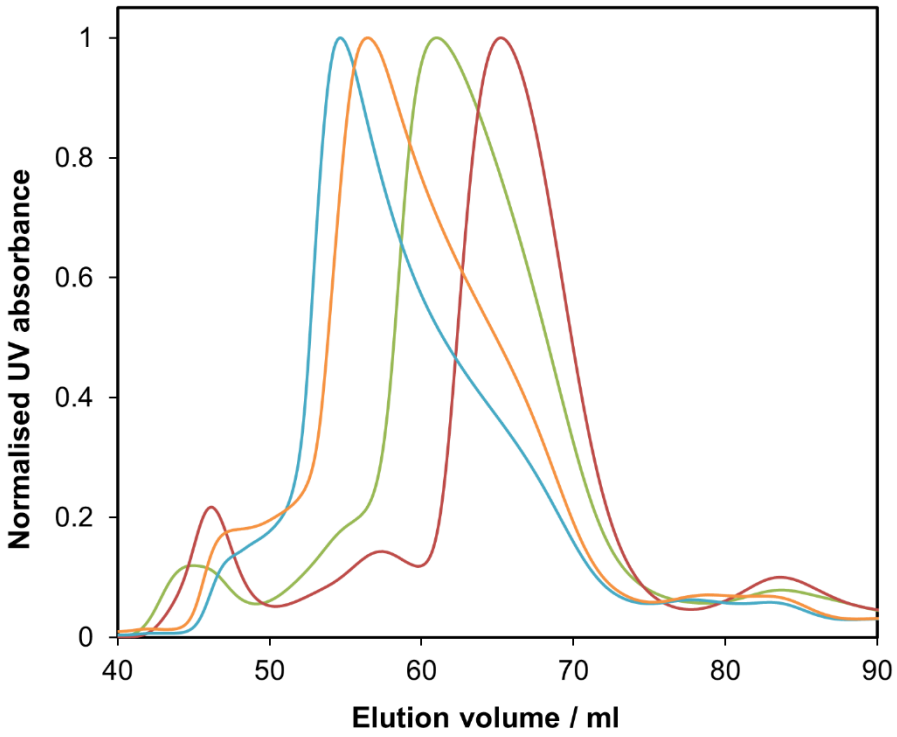

Fig. 8.6: Chromatograms from four separate SEC experiments with CPR-ClpB in SGC buffer, showing significant variation in V_e.

Early experiments found CPR-ClpB eluted at a V_e corresponding to a MW between that of a pentamer and hexamer (Fig. 8.6 blue curve; V_e = 54.7 ml, MW = 529 kDa), and the target was proposed to be hexameric based on the oligomeric states of ClpB/Hsp100 homologs. However, later experiments following the same procedure saw a large increase in retention time, with the major species eluting at volumes corresponding more closely to a trimer (Fig. 8.6 green curve; V_e = 61.0 ml, MW = 310 kDa) or dimer (Fig. 8.6 red curve; V_e = 65.4 ml, MW = 215 kDa). Peaks around 45 - 47 ml correspond to a MW > 1 MDa and are likely due to aggregated protein. SEC was also attempted with alternative standard buffers, Tris and phosphate, though these consistently resulted in a V_e ~ 65 ml (MW = 220 – 230 kDa), correlating most closely to a CPR-ClpB dimer (204 kDa). It is therefore likely that CPR-ClpB can form several distinct oligomers, with dissociation of these oligomers also implied by the broad trailing peaks in Fig. 8.6. This instability of the CPR-ClpB hexamer provides an explanation for the heterogeneity observed in cryo-EM micrographs (Fig. 8.3).

Glutaraldehyde crosslinking with nucleotides

Optimisation was required to promote formation of the functional CPR-ClpB hexamer and stabilise it for structural analysis. Previous structural studies found that the hexameric form of ClpB homologs is stabilised by nucleotides including ADP, ATP, and non-hydrolysable ATP equivalents such as ATPγS[380,381]. CPR-ClpB was therefore incubated with a variety of nucleotides and covalently cross-linked with glutaraldehyde, in an effort to stabilise the hexamer. SEC was then conducted to determine the oligomeric state of CPR-ClpB; the resulting chromatograms are shown in Fig. 8.7.

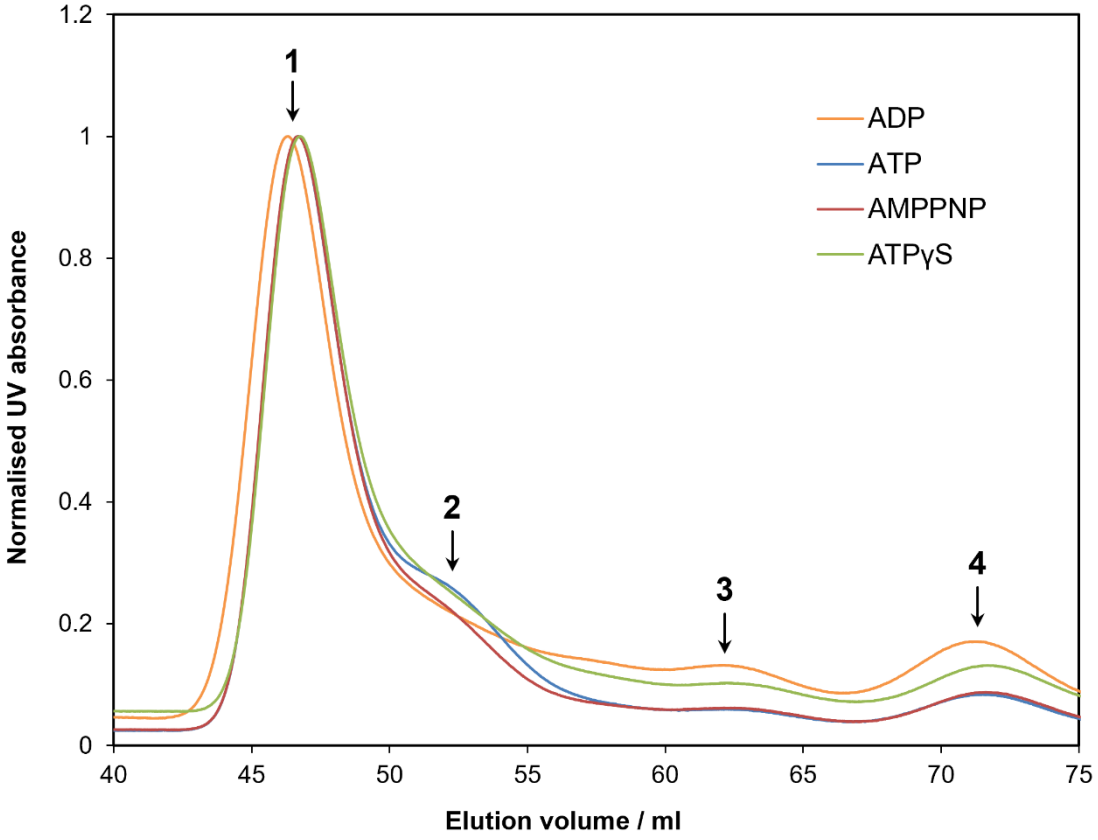

Fig. 8.7: Chromatograms from SEC of CPR-ClpB after cross-linking with nucleotides. Peaks of interest are indicated with arrows. AMP-PNP = adenylyl-imidodiphosphate.

The chromatograms from SEC following crosslinking with nucleotides (Fig. 8.7) were notably different to those from standard CPR-ClpB SEC experiments without crosslinking (Fig. 8.6). With all four nucleotides tested, the V_e of the major species (peak 1, ~ 47 ml) shifted significantly towards the void volume, V_o, of the column, with an estimated MW > 1 MDa. However, small quantities of several other CPR-ClpB species were also identified. The major peak has a distinct shoulder to the right (peak 2, ~ 52 ml; ~ 660 kDa species) in chromatograms with ATP and its analogues, though less pronounced with ADP. This species is likely to be the CPR-ClpB hexamer. Peaks 3 and 4 (3: V_e = 62 ml, ~ 290 kDa species; 4: V_e = 72 ml, ~ 120 kDa species) correspond most closely to the CPR-ClpB trimer and monomer, respectively, based on estimated MW.

Negative staining with crosslinked CPR-ClpB samples

To assess the contents of the distinct species identified from SEC after crosslinking, negative stain TEM was conducted on samples from each peak in Fig. 8.7 for the four nucleotides. Negative stain micrographs of sample from peak 1 are shown in Fig. 8.8.

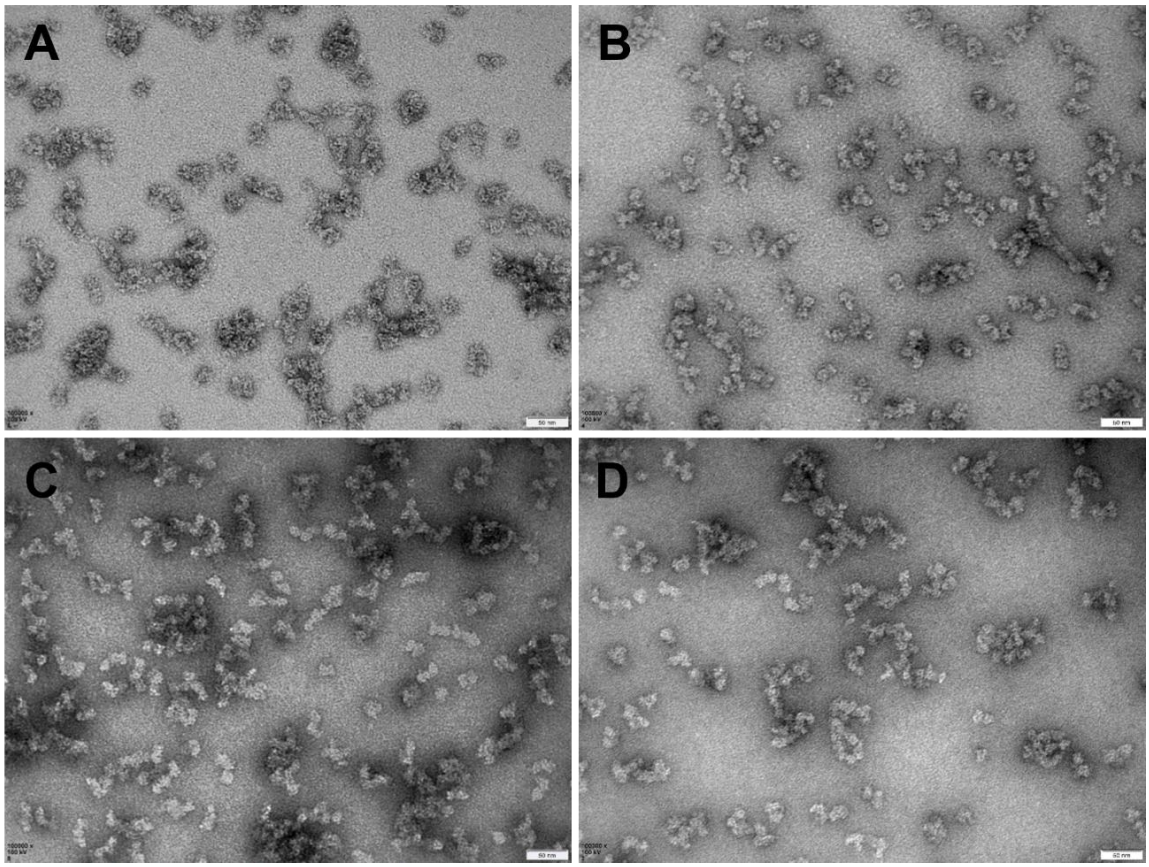

Fig. 8.8: Negative stain TEM micrographs of crosslinked CPR-ClpB samples with various nucleotides from peak 1 in Fig. 8.7. Scale bars at the bottom left corner of each image show 50 nm/500 Å. **(A)** ADP. **(B)** ATP. **(C)** AMP-PNP. **(D)** ATPγS.

All four micrographs in Fig. 8.8 contain aggregated CPR-ClpB, with few, if any, individual species of appropriate size and shape for the predicted hexamer. Crosslinking therefore resulted in extensive aggregation, as expected from the V_e during SEC.

Micrographs from negative stain experiments on samples expected to contain hexameric CPR-ClpB (Fig. 8.7 peak 2) are shown in Fig. 8.9.

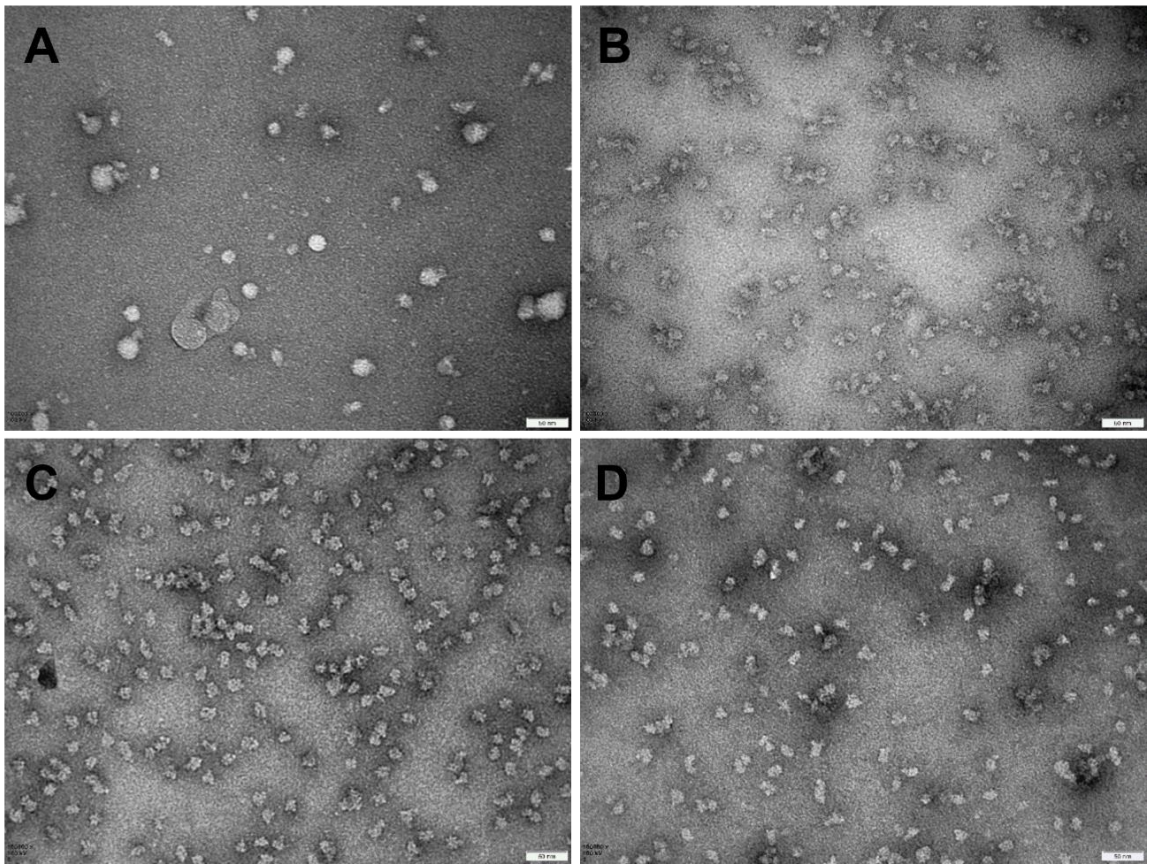

Fig. 8.9: Negative stain TEM micrographs of crosslinked CPR-ClpB samples with various nucleotides from peak 2 in Fig. 8.7. Scale bars at the bottom left corner of each image show 50 nm/500 Å. **(A)** ADP. **(B)** ATP. **(C)** AMP-PNP. **(D)** ATPγS.

Plausible hexameric species could not be identified in micrographs of the CPR-ClpB sample with ADP (Fig. 8.9A). However, the micrographs with ATP, ATPγS, and AMP-PNP were more promising, with a good distribution of potential hexameric species of appropriate size. The dynamic nature of the hexamer is also evident, with heterogeneity in the shape of particles that could correspond to a variety of hexamer conformations. The absence of such particles in the ADP-containing sample correlates with the lack of a distinct peak in the chromatogram in Fig. 8.7 relative to samples with ATP and its analogues.

The results of negative stain TEM experiments with the putative CPR-ClpB trimer (Fig. 8.7 peak 3) are shown in Fig. 8.10.

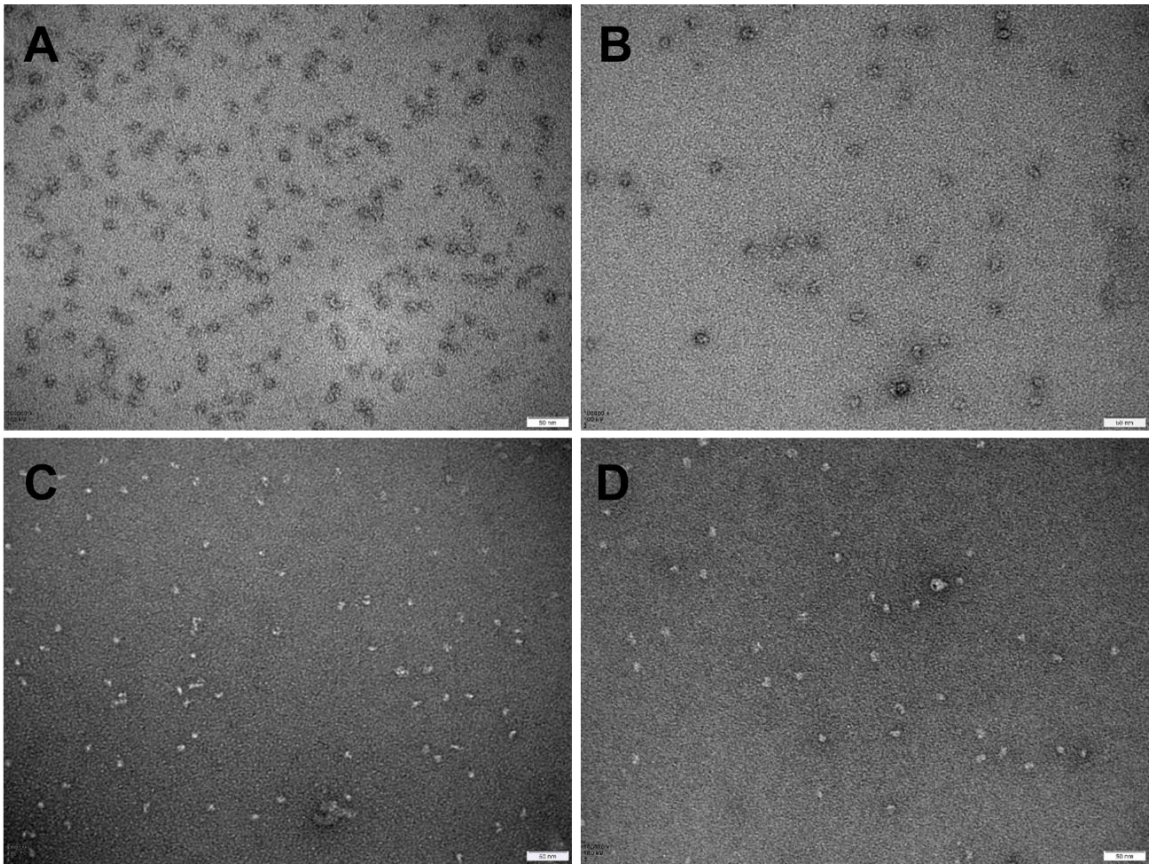

Fig. 8.10: Negative stain TEM micrographs of crosslinked CPR-ClpB samples with various nucleotides from peak 3 in Fig. 8.7. Scale bars at the bottom left corner of each image show 50 nm/500 Å. **(A)** ADP. **(B)** ATP. **(C)** AMP-PNP. **(D)** ATPγS.

A relatively pure, monodisperse species is clear in the CPR-ClpB with ADP micrograph (Fig. 8.10A). As anticipated, this species is smaller than the putative hexamers in Fig. 8.9, and likely corresponds to the CPR-ClpB trimer based on the estimated MW from SEC. Similar particles are present in the micrograph with ATP though are more sparsely distributed (Fig. 8.10B), and they are not visible with ATPγS or AMP-PNP (Fig. 8.10C and D, respectively). In conjunction with SEC analysis, the results shown in Fig. 8.9 and Fig. 8.10 suggest that ATP and its non-hydrolysable analogues promote formation of a CPR-ClpB hexamer, whilst ADP encourages trimer formation.

The micrographs from negative stain TEM with sample proposed to contain the CPR-ClpB monomer (Fig. 8.7 peak 4) are shown in Fig. 8.11.

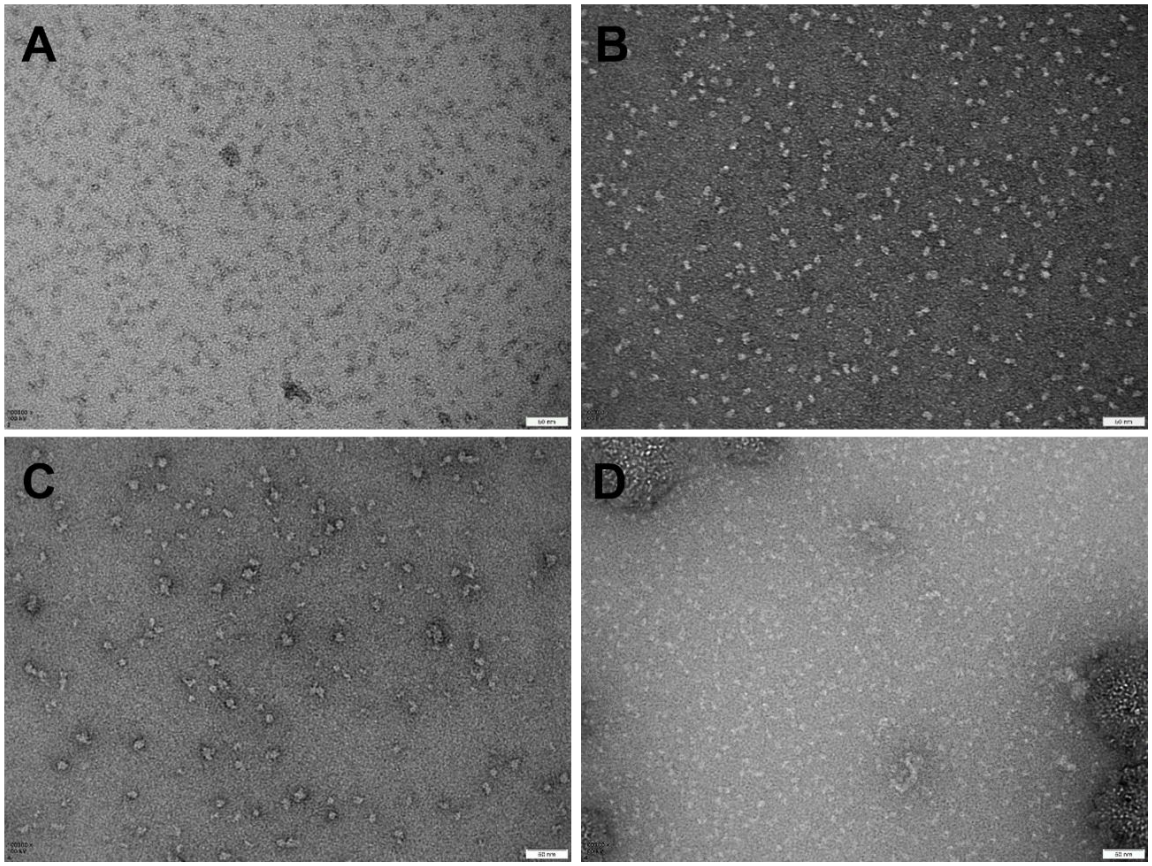

Fig. 8.11: Negative stain TEM micrographs of crosslinked CPR-ClpB samples with various nucleotides from peak 4 in Fig. 8.7. Scale bars at the bottom left corner of each image show 50 nm/500 Å. **(A)** ADP. **(B)** ATP. **(C)** AMP-PNP. **(D)** ATPγS.

The micrographs in Fig. 8.11 each contain large populations of a smaller CPR-ClpB species than the proposed hexamers and trimers from Fig. 8.9 and Fig. 8.10, respectively. This again correlates with the expected size of the species from SEC (Fig. 8.7 peak 4), adding confidence to its classification as the CPR-ClpB monomer.

Glutaraldehyde crosslinking therefore resulted in several distinct CPR-ClpB species, including a likely CPR-ClpB hexamer with ATP and its analogues (Fig. 8.9). Each of the CPR-ClpB samples in Fig. 8.8 – 8.11 were notably more homogenous than the un-crosslinked sample from earlier negative staining experiments (Fig. 8.2). The crosslinking did, however, cause most of the protein to aggregate (Fig. 8.7 and Fig. 8.8). Nevertheless, sufficient non-aggregated CPR-ClpB hexamer could be resolved by SEC for downstream single particle cryo-EM analysis.

Cryo-EM analysis of crosslinked CPR-ClpB sample

Negative stain TEM micrographs of CPR-ClpB from peak 2 (Fig. 8.7) with ATP and its analogues showed relatively monodisperse samples containing promising hexamers (Fig. 8.9). However, species aggregation and dissociation were observed in grids made using similarly prepared CPR-ClpB samples after vitrification (Fig. 8.12).

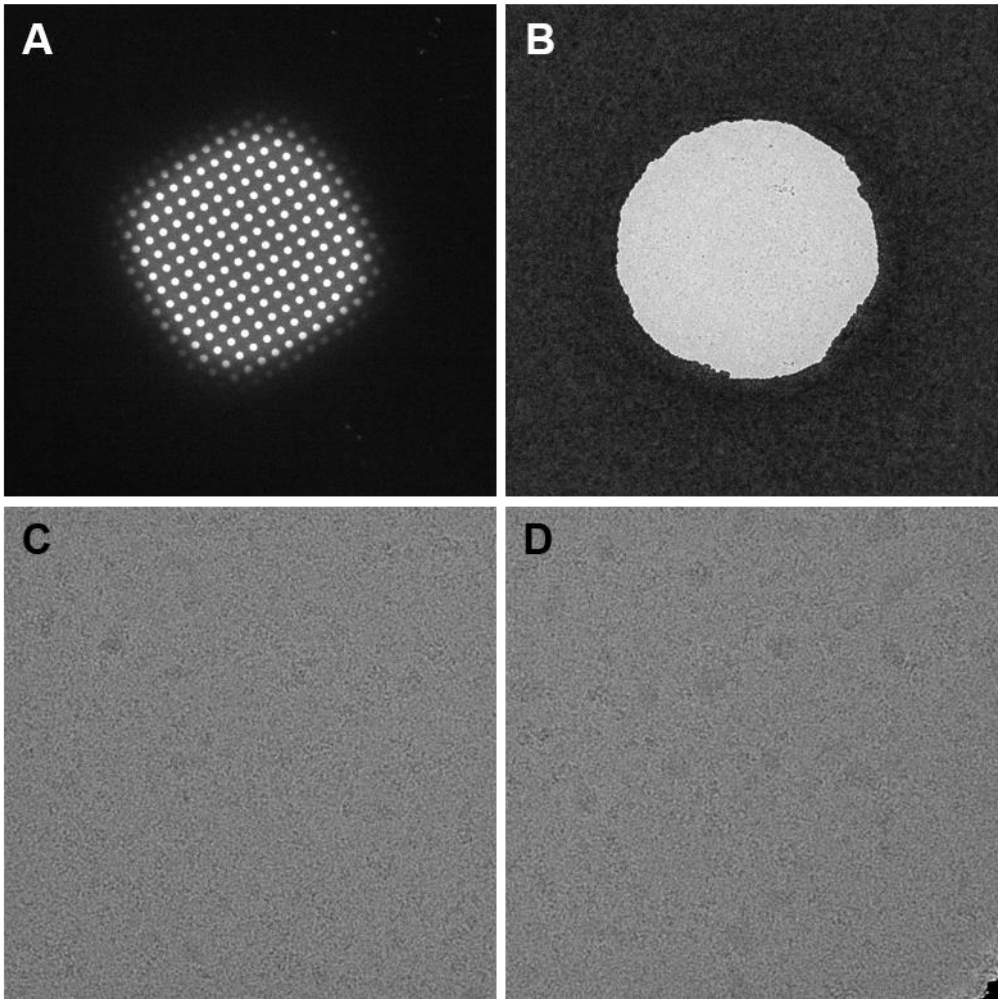

Fig. 8.12: Screening of grids prepared with the crosslinked CPR-ClpB and ATPγS sample. **(A)** Grid square selected for data collection. **(B)** Representative foil hole from grid square in (A) with suitable ice thickness. **(C)** and **(D)** Magnified views of the foil hole in (B).

CPR-ClpB appears prone to aggregation and disintegration during the vitrification process. These are common issues encountered during grid preparation[361], and are likely the result of unfavourable interactions between the CPR-ClpB particles and the air-water interface[361,382,383]. Work is now ongoing to improve the quality of the CPR-ClpB sample and grid preparation, including the use of detergents and alternative grid materials.

8.3. Conclusions

Preliminary negative-stain experiments with CPR-ClpB were promising; however, instability and aggregation was evident from cryo-EM grids during screening and data collection. It is common for initial grid preparation of new cryo-EM targets to be unsuccessful, and optimisation can require a variety of approaches to identify the causes of issues from a multitude of variables[350,361]

Aggregation and dissociation can often be dealt with by optimising the sample in the early stages, ahead of grid preparation[350]. Glutaraldehyde crosslinking combined with SEC was therefore

attempted to stabilise and isolate the CPR-ClpB hexamer, though resulted in heavy aggregation. A starting point would be less aggressive crosslinking with a lower percentage glutaraldehyde (0.0025%). During cryo-EM studies of a large octameric exocyst complex, this was found to effectively preserve particles on grids where 0.1% glutaraldehyde resulted in aggregation and disintegration[361,384]. The use of detergents and surfactants could also help reduce unwanted aggregation[361], and the results of TSA could be used to guide simple alterations to buffer composition, such as pH or ionic strength, or the addition of stabilising additives. However, optimisation of CPR-ClpB samples is hampered by limited access to the apparatus required for both negative staining and cryo-EM experiments. As the availability of microscopes and other necessary equipment increases, this process will become considerably smoother.

Whilst the Virus-X project chiefly applied X-ray crystallography for target structure determination, this technique is not suitable for all proteins, including large and conformationally flexible proteins such as CPR-ClpB. Though the work towards a structure is ongoing, CPR-ClpB is the first Virus-X target to have been studied with cryo-EM and remains a promising candidate for this continually advancing technique. The groundwork discussed here will hopefully lead to a high-quality structure from which comparisons can be drawn to homologous ClpB/Hsp100 proteins. This would also provide a solid foundation for further cryo-EM studies of the CPR-DnaK/ClpB bi-chaperone system, to reveal the interactions that underpin the mechanisms by which aggregated proteins are solubilised and refolded *in vivo*.

Chapter 9. Production and characterisation of targets from a thermophilic giant bacteriophage

9.1. Introduction

9.1.1. Exploring viral diversity through metagenomics

Whilst the focus of the work discussed in this thesis is bacterial targets from the CPR, the primary motivation of the Virus-X project was to explore viral genomes[18,19]. Viruses are infectious agents that exploit the metabolic machinery of a host organism in order to replicate themselves, with populations estimated to outnumber the stars in the observable universe[385–388]. Where there is life, there are viruses, occupying every conceivable ecological niche and playing key roles in many biogeochemical and ecological processes[386,389–391]. Yet, as of February 2019, fewer than 5000 viral species had been formally documented[392], leaving a vast amount of unexplored sequence space within the 'virosphere'[18,393,394]. Previous challenges in detecting and isolating viruses have limited their study in the past, though the arrival of metagenomics and advances in NGS have made viral genomes much more accessible[9,393,395–398].

Enzymes identified from merely a handful of viral genomes are responsible for many major advancements of contemporary molecular biology and biotechnology[18,397,399]. These viruses include the *E. coli* phages lambda, T4, and T7, which provided enzymes that were instrumental in the development of molecular cloning technologies[153,399,400]. The reservoir of unexplored genetic diversity within the virosphere can therefore be expected to hold great innovation potential, and exploring new sources of viral enzymes will undoubtedly lead to further progression of the biotechnology, industrial, and biomedical fields[393,401–403]. Analysis of viral gene pools is also necessary to deepen our understanding of ecosystem dynamics, virus-host interplay, and the influence of viruses on evolution[404–409].

9.1.2. Discovery and implications of giant viruses

After their discovery at the end of the 19th century, virus particles were originally distinguished from cellular organisms by their comparatively small size[410,411]. Any 'microorganism' that could pass through a sterilising filter or could not be visualised by light microscopy was considered a virus[412]. A more formal set of criteria was later proposed and refined which stated that viruses are biological entities possessing a DNA or RNA genome, but that are unable to synthesise ATP, do not encode the apparatus for protein translation, and do not divide[412,413]. All living cellular organisms, including host-dependent intracellular parasites, are able to perform these tasks[411].

The discovery of the giant (400 nm), amoeba-infecting *Mimivirus* in 2003 challenged the previously steadfast criteria for viruses[411,414]. These criteria were in fact a hindrance to the discovery of the *Mimivirus*, or 'Mimicking Microbe' virus, which was thought to be a bacterium for a decade after its initial discovery[414]. *Mimivirus* has an exceptionally large genome (~ 1.2 Mbp) that is vastly more complex than traditional viruses, and larger than the genomes of some bacteria[415–417].

The numbers of giant viral genomes being identified are now continually increasing thanks to the mNGS approach[393,418]. Their large genomes harbour many genes typical of cellular genomes, such as those encoding translation machinery, and even diverse ribosomal proteins[393,419]. Protein production is considered to be absent in viruses, and a lack of translation system components was previously a common definition for viruses. Giant viruses therefore cast doubt over traditional definitions, blurring the division between cellular life and viruses[411,420–422].

9.1.3. Jumbophages

Another group of larger viruses are the jumbophages: tailed bacteriophages with > 200 kbp genomes and markedly large virions up to 450 nm in diameter[423–425]. Jumbophages have been isolated from a broad range of environments and show remarkable genetic diversity, presenting an excellent source of novel enzymes for exploration[423]. It has been suggested that these large phages evolved from smaller phages by acquiring new genetic material, allowing them to control protein production upon infection and so reducing their dependence on their bacterial hosts[423]. Further exploration of jumbophages and other giant viruses could therefore provide insight into the origins of cellular life, and the evolutionary progression from cell-dependent entities to autonomous cellular lifeforms[423].

9.1.4. The '*Ubervirus*'

During the bioprospecting efforts of the Virus-X project, a phage with an unusually large dsDNA genome was discovered in a hot spring sampling site in Iceland. The genome of this thermophilic jumbophage, nick-named the '*Ubervirus*', is referred to as 'Genome X' (GX) and encodes many hypothetical proteins with no significant homology to any characterised proteins. Many of the predicted gene products scattered across the length of the *Ubervirus* genome show sequence similarity to 'classical' proteins of T4-like phages, namely structural components of the complex virion, bolstering the claim that the genome is phage in origin.

Whilst much more analysis is required of the *Ubervirus* and its host, it is likely to share many characteristics with other giant viruses. In particular, the *Ubervirus* genome appears closely related to that of a giant virus previously described in a large study of viral metagenomes[393]. This giant phage was identified in a bioreactor sample and has a closed genome of nearly 600 kbp and > 1100 genes[393]. Both the *Ubervirus* and this previously reported giant phage were found to encode some ribosomal proteins and components of the translational machinery[393], further fuelling discussions regarding the definition of a virus.

The vast majority of genes products from the *Ubervirus* genome were annotated as hypothetical proteins with unknown taxonomic classification, highlighting a wealth of unexplored enzymes. Several of these gene products were selected as key targets for characterisation by Virus-X consortium members. Production of these targets followed the general pipeline used for the CPR targets (3.1). The selected GX proteins have entirely unknown structures and functions, and so present opportunities for the discovery of novel and unique functionalities with innovation potential.

9.2. Results and discussion

9.2.1 Sequence analysis of GX targets

Five GX targets (GX-hyp-2, GX-hyp-4, GX-hyp-6, GX-hyp-16, and GX-hyp-28) were selected from the 589 kbp *Ubervirus* contig containing 1058 genes. All five selected genes encoded hypothetical proteins lacking detectable homology with proteins of known structure, and so were allocated to category C. High priority for structure determination was therefore given to these targets as they could have potentially distinct structures, with accordingly novel functions. The positions of these target genes on the source contig are shown in Fig. 9.1.

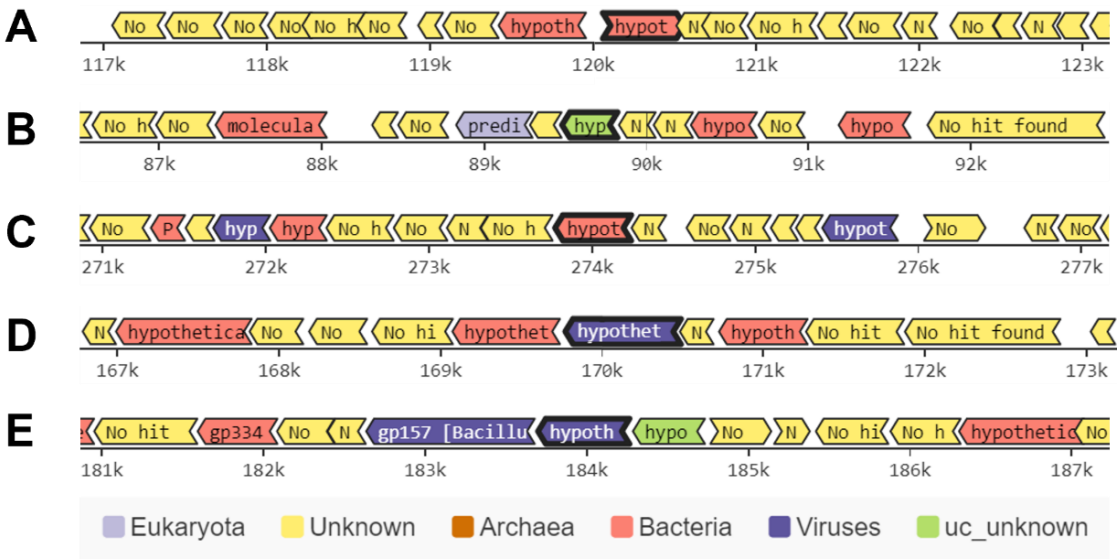

Fig. 9.1: Positions of the GX target genes and their neighbours on the *Ubervirus* source contig from the Virus-X sequence database[18]. Predicted genes are represented as arrows showing the strand direction of transcription and colour coded by the taxonomic classification of their most significant BLAST hit. Selected target genes are outlined in black: **(A)** GX-hyp-2; **(B)** GX-hyp-4; **(C)** GX-hyp-6; **(D)** GX-hyp-16; **(E)** GX-hyp-28.

The GX target sequences were analysed using BLAST[151]. Over 100 'hit' sequences with significant similarity to each of GX-hyp-2 and GX-hyp-6 were identified, occurring in viruses and various bacterial and archaeal species. Proteins sharing sequence similarity to GX-hyp-16 were also found

to be common in bacteria and some viruses. The GX-hyp-4 and GX-hyp-28 sequences returned fewer hits (< 50) among sequenced viruses and bacteria, suggesting they are less widespread.

All BLAST hit sequences identified for the five GX targets were annotated as hypothetical proteins, revealing no clues as to potential function. Of note, however, was the detection of a conserved domain from a superfamily of unknown function (Pfam07505; DUF5131) in several GX-hyp-16 hits. This family of bacterial and phage proteins share three highly conserved cysteine residues in the format Cx_6CxxC, as well as many highly conserved residues. This motif was subsequently identified in the GX-hyp-16 sequence, and so this target can be considered a member of the DUF5131 protein family.

9.2.2. Cloning and preliminary characterisation

The MW, estimated molar extinction coefficient (ε) at 280 nm, and isoelectric point (pI) of the selected GX targets were calculated to assist production and characterisation. These are detailed in Table 9.1.

Table 9.1: Parameters for the GX targets. The molar extinction coefficient, ε, is that at 280 nm.

Target	MW / kDa	ε / M^{-1} cm^{-1}	Theoretical pI
GX-hyp-2	20.7	26930	9.5
GX-hyp-4	14.8	20970	9.7
GX-hyp-6	19.3	9970	6.1
GX-hyp-16	29.4	74940	6.5
GX-hyp-28	20.5	15930	6.2

As with the CPR targets, the GX targets were cloned commercially into the pJOE5751.1 plasmid for overexpression in *E. coli* (3.2.1.). The targets were expressed as fusion proteins with a short and non-removable N-terminal His-tag (MTMITHHHHHHGS) to facilitate purification by IMAC. The MWs in Table 9.1 assume N-terminal methionine excision (NME) of all targets due to the tag sequence and NME processes in the *E. coli* expression host[223,224].

9.2.3. Expression trials of GX targets

Agarose gel electrophoresis and Sanger sequencing of purified plasmids from transformed cells confirmed that transformations of the pJOE5751.1-*GX* constructs into competent *E. coli* cells were successful. Small-scale expression trials were then conducted for each putative GX protein. Single colonies were selected from transformation plates for induction tests with and without the addition of 0.2 % L-rhamnose in culture to allow assessment of basal expression levels. Fig. 9.2 shows the

results of whole cell SDS-PAGE of non-induced (−) and induced (+) *E. coli* cultures transformed with the pJOE5751.1-*GX* plasmids.

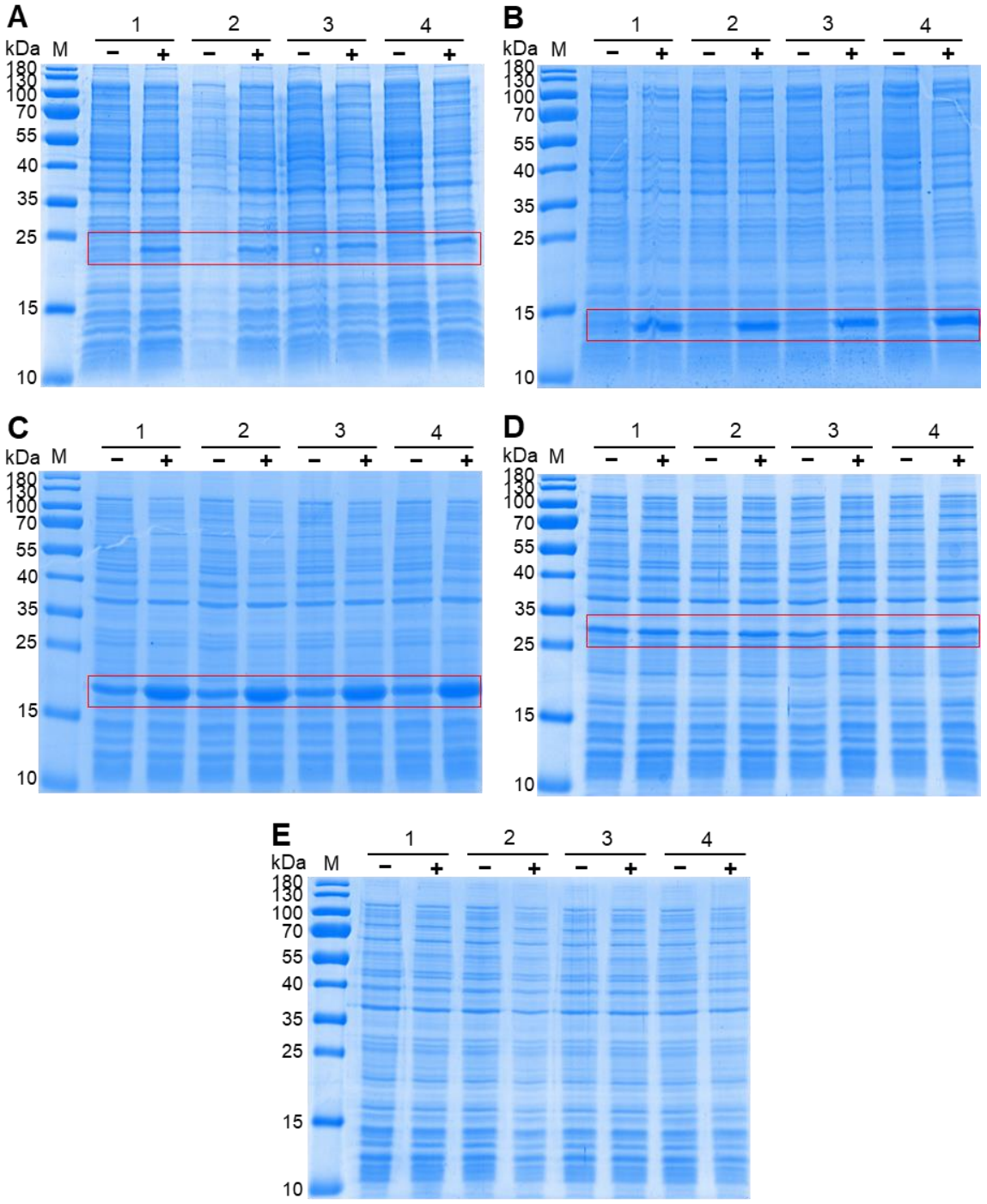

Fig. 9.2: Whole-cell SDS-PAGE results from small-scale expression tests of GX targets. Noninduced (−) and induced samples with 0.2% L-rhamnose (+) from four tests 1 - 4 are shown. M = marker, with MWs in kDa to the left. Positions of bands corresponding to target protein expression are indicated with a red box, where applicable. **(A)** GX-hyp-2, MW = 20.7 kDa; **(B)** GX-hyp-4, MW = 14.8 kDa; (**C)** GX-hyp-6, MW = 19.3 kDa; (**D)** GX-hyp-16, MW = 29.4 kDa; (**E)** GX-hyp-28, MW = 20.5 kDa.

As shown in Fig. 9.2, four of the five tests were successful (GX-hyp-2/4/6/16), with obvious target overexpression upon induction with 0.2 % L-rhamnose across all four colonies. The strength of gel bands from the SDS-PAGE results also enabled qualitative judgement of basal and induced expression levels, which are outlined in Table 9.2:

Table 9.2: Results of small-scale test expressions of GX targets: − signifies no detectable expression, + represents weak expression, ++ for moderate expression, and + + + for high expression.

Target	MW / kDa	Basal expression	Induced expression
GX-hyp-2	20.7	−	+
GX-hyp-4	14.8	−	+ +
GX-hyp-6	19.3	+ +	+ + +
GX-hyp-16	29.4	+	+
GX-hyp-28	20.5	−	−

GX-hyp-2 and GX-hyp-4 (Fig. 9.2A and B, respectively) showed conclusive overexpression upon induction with L-rhamnose, with negligible basal expression in non-induced cultures. GX-hyp-6 (Fig. 9.2C) does show substantially increased expression upon addition of L-rhamnose, though there is significant basal expression in non-induced cultures. A similar situation is seen for GX-hyp-16 (Fig. 9.2D), with similar expression levels in non-induced and induced cultures. Whilst this could prove problematic if the recombinant protein were toxic to host cells, GX-hyp-6 and GX-hyp-16 expression had no obvious effect on cell growth, and so basal expression was disregarded.

Of the five GX targets, only GX-hyp-28 (Fig. 9.2E) showed no detectable expression, even after repeated tests. When returning to this protein in future, higher concentrations of rhamnose, variations in growth conditions, and alternative expression host strains could be trialled. A different expression vector utilising a T7 hybrid system could also be of benefit. It is also possible that, due to its hypothetical nature, the *GX-hyp-28* gene product is not a naturally expressed protein.

In order to characterise as many targets as possible in the time available, only targets that overexpressed well in small-scale trials (GX-hyp-4 and GX-hyp-6) were carried forward to full-scale expression at this stage. Full-scale expression was also conducted for GX-hyp-16, since the presence of a conserved cysteine motif (9.2.1.) made this a particularly curious target.

9.2.4. GX-hyp-4

Expression and IMAC purification

GX-hyp-4 overexpressed well during large-scale expression experiments; however, SDS-PAGE revealed that most of the protein was insoluble and contained within the cell pellet (Fig. 9.3).

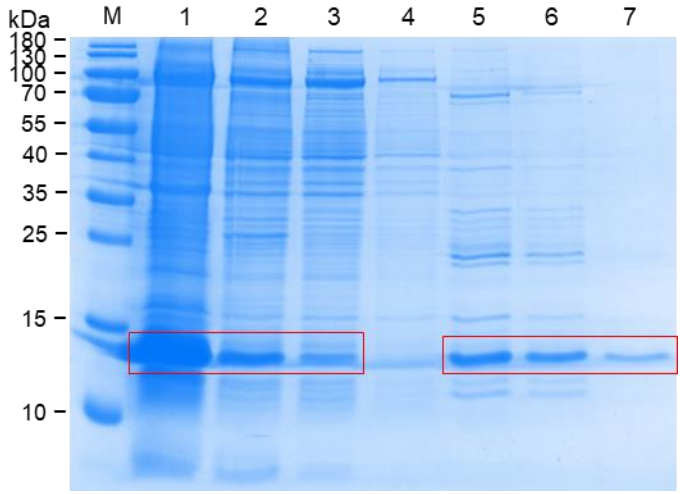

Fig. 9.3: SDS-PAGE analysis of GX-hyp-4 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 7 = protein fractions. Bands corresponding to GX-hyp-4 (MW = 14.8 kDa) are indicated with a red box.

A small quantity of soluble GX-hyp-4 was retrieved from the cell lysate through IMAC. However, the protein was unstable and precipitated rapidly, prohibiting further purification by SEC or MS analysis. Due to these issues, work on GX-hyp-4 was paused to focus on the other GX targets. In future, growth conditions should be optimised to increase the soluble protein yield so that TSA can be conducted. TSA could reveal stabilising conditions that enable further characterisation and crystallisation screening of GX-hyp-4.

9.2.5. GX-hyp-6

Expression and IMAC purification

GX-hyp-6 expressed well in full-scale experiments, yielding around 10 mg of soluble protein per litre of culture, and with reasonably high purity achieved after IMAC (Fig. 9.4).

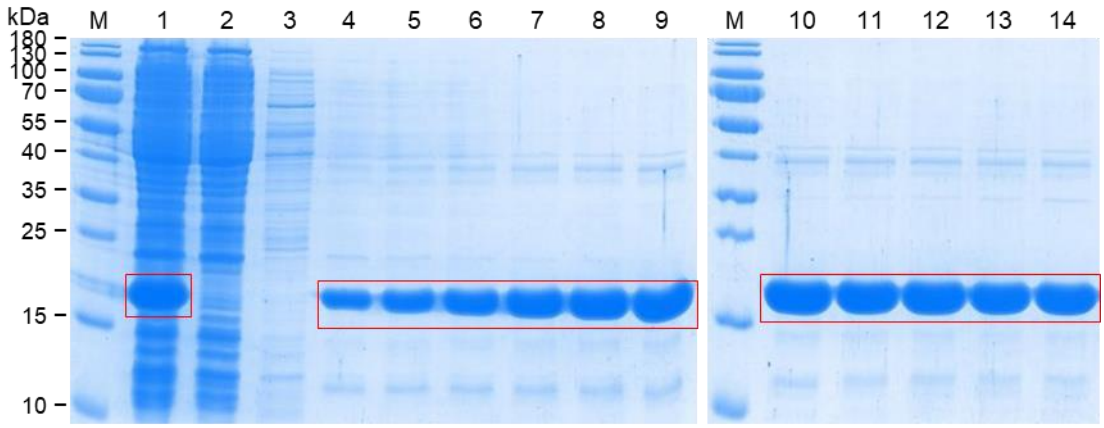

Fig. 9.4: SDS-PAGE analysis of GX-hyp-6 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell lysate; 2 = flow through; 3 = wash; 4 - 14 = protein fractions. Bands corresponding to GX-hyp-6 (MW = 19.3 kDa) are indicated with a red box.

Levels of soluble expression far outweighed a small amount of insoluble expression, and the quantities of GX-hyp-6 were more than sufficient for downstream characterisation. Minor contaminants remained in the GX-hyp-6 sample following IMAC (Fig. 9.4), though it was suitably pure for MS and TSA.

Characterisation

The identity of the purified species was confirmed to be GX-hyp-6 by ESI-TOF MS (Fig. G1). The MW of the molecular ion (19278 Da) is within 1 Da of the expected MW of GX-hyp-6 (19277 Da). As predicted, the mass spectrum also indicates N-terminal methionine excision (NME), with a loss of 131 Da from the MW of the full-length protein.

SEC analysis

SEC was conducted to further purify GX-hyp-6 for crystallisation, and to determine its oligomeric state. The resulting chromatogram and SDS-PAGE analysis of eluted fractions are shown in Fig. 9.5.

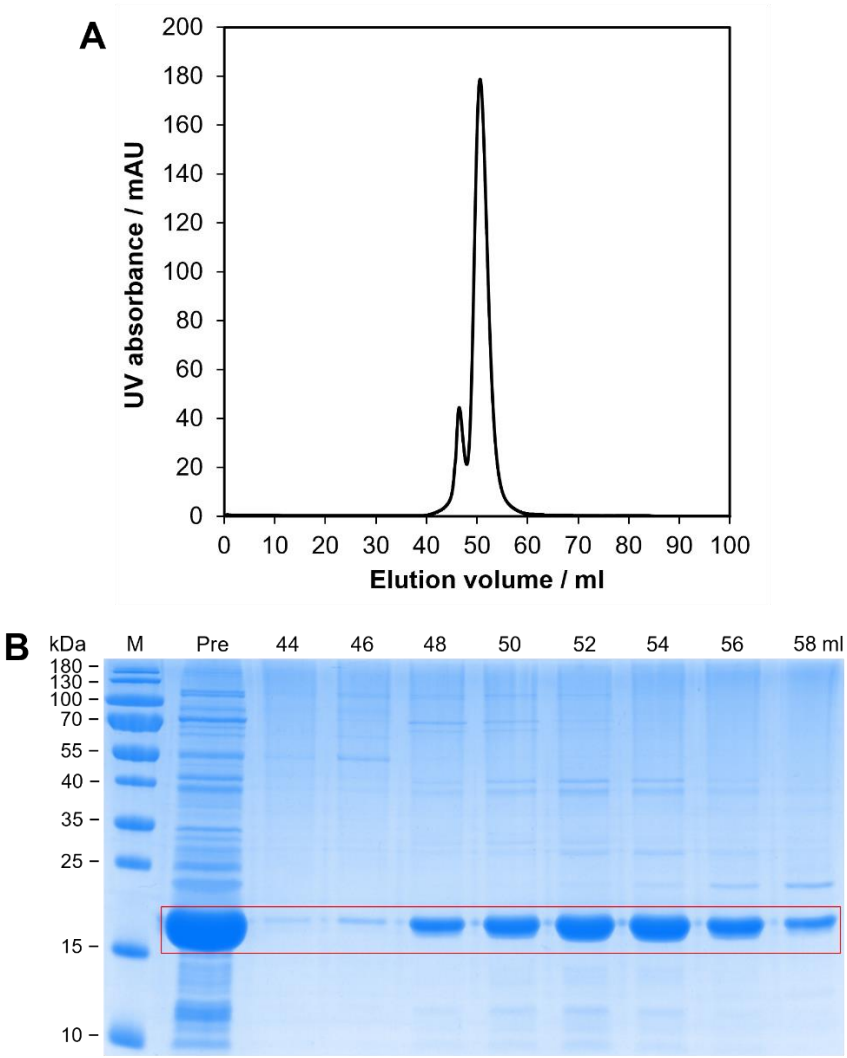

Fig. 9.5: (A) Chromatogram from SEC of GX-hyp-6. **(B)** SDS-PAGE analysis of elution fractions from SEC of GX-hyp-6 (MW = 19.3 kDa). M = marker, with MWs of bands on the left; Pre = sample prior to SEC; protein fractions are shown with corresponding elution volumes above.

The major species in the sample eluted as a sharp, well-resolved peak with a V_e of 50.6 ml, corresponding to an approximate MW of 76.6 kDa (Fig. 9.5A). This species was confirmed to be GX-hyp-6 based on expected MW by SDS-PAGE (Fig. 9.5B). The V_e corresponded to an expected MW close to that of a GX-hyp-6 tetramer (77.2 kDa). The minor species eluting at V_e 46.5 ml has an estimated MW of 99.6 kDa, close to a GX-hyp-6 pentamer (96.5 kDa), though could also be a contaminant protein in the sample. From the SEC results, it is concluded that GX-hyp-6 is likely tetrameric.

Thermal shift analysis

TSA of GX-hyp-6 with the Durham Screens® enabled the intrinsic thermostability of the target to be evaluated, whilst also searching for stabilising conditions[210,255]. GX-hyp-6 was found to be stable, with a T_m of 53.9 in water.

pH screening

Results of TSA with the Durham pH Screen® for GX-hyp-6 revealed a striking preference for acidic buffers, as displayed in Fig. 9.6.

A

	1	2	3	4	5	6	7	8	9	10	11	12
A	53.2	53.6	36.6	67.6	73.5	74.2	31.2 77.0	79.2	77.8	76.1	79.0	78.2
B	75.2	77.4	73.6	72.4	77.5	75.1	78.4	79.1	77.2	74.1	69.1	60.4
C	71.3	63.8	57.1	69.9	68.9	55.8	80.6	77.3	72.2	70.3	65.0	56.4
D	73.2	70.5	61.3	72.3	63.4	56.7	76.1	76.7	57.5	76.1	72.5	54.8
E	61.0	51.9	50.7	74.6	70.2	59.6	70.8	54.1	52.7	73.9	62.9	53.9
F	74.3	65.3	50.3	65.5	52.6	46.3	56.5	44.2	39.0	57.2	46.2	42.9
G	45.7	41.0	37.0	52.6	40.4	39.0	48.5	41.0	38.0	37.0	33.0	32.0
H	50.6	47.7	45.1	36.0	30.0	30.0	36.0	32.0	31.0	31.0	30.0	30.0

B

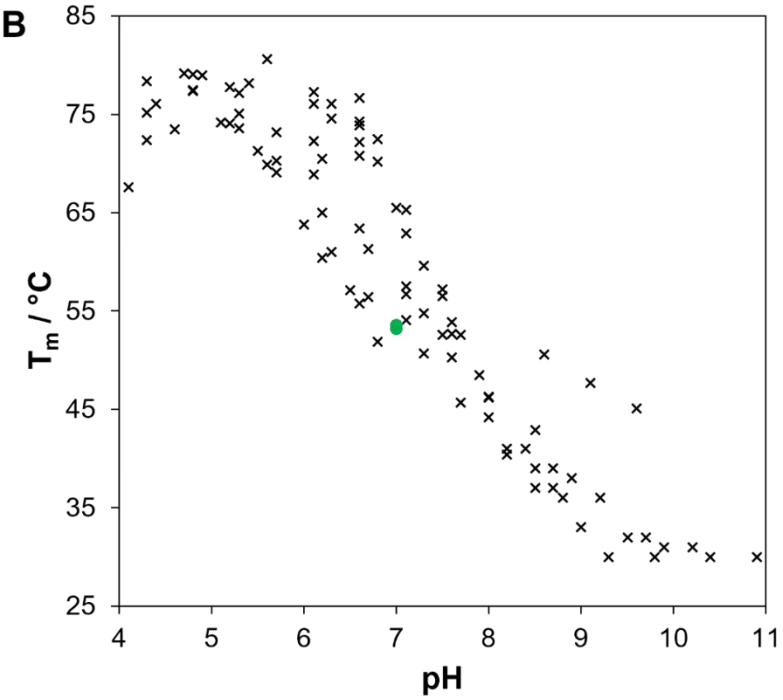

Fig. 9.6: (A) Results from the 96-well plate TSA of GX-hyp-6 with the Durham pH Screen®[210,255]. Colours are indicative of differences in the melt temperature (T_m) from that in water (wells A1 and A2): a colour change from colourless through light blue to dark blue indicates an increase in T_m. Green indicates either two stages in the melting process or variation in data. The pH of conditions increases across the screen: rows A to E contain acidic buffers, whilst rows F to H contain alkaline buffers (Table A2). **(B)** Results from TSA using the Durham pH Screen® showing the effect of pH on the T_m of GX-hyp-6. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at similar or identical pH values are from different buffer systems.

The range of T_m values for GX-hyp-6 across the screen spans > 50 °C, from a lowest T_m of 30 °C in several alkaline conditions to 80.6 °C in pH 5.6 MES buffer. The stability of the GX-hyp-6 appears strongly linked to pH, with optimal stability under acidic conditions and a general decrease in T_m as pH is increased above ~ 5. This suggests that GX-hyp-6 is adapted to function at low pH, correlating with the fact that the *Ubervirus* metagenome was isolated in an acidic environment (pH 5).

Salt screening

TSA with the Durham Salt Screen® found that GX-hyp-6 is strongly stabilised by ammonium chloride and sulfate salts, as demonstrated in Fig. 9.7.

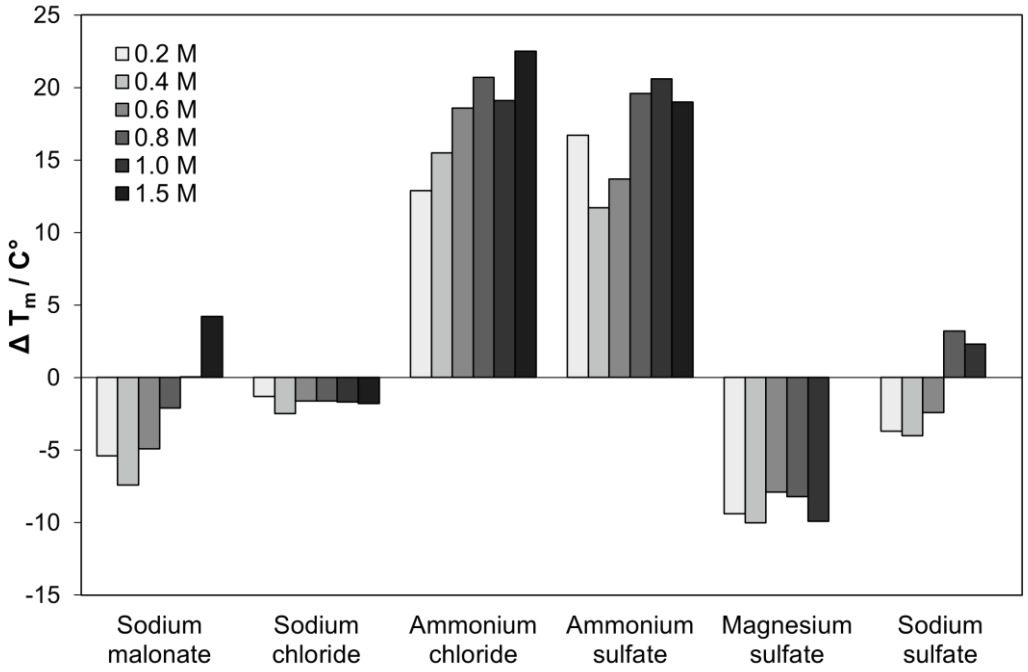

Fig. 9.7: Results from TSA of GX-hyp-6 with the Durham Salt Screen®[210,255] showing changes in T_m upon addition of salts at increasing concentrations.

All concentrations of the ammonium salts tested resulted in an increase in T_m of > 10 °C, up to ~ 22 °C (T_m = 76.8 °C) with 1.5 M ammonium chloride. Ammonium salts could therefore be useful additives for stabilising the protein during downstream functional assays and crystallisation experiments. Conversely, sodium and magnesium salts tended to destabilise GX-hyp-6, with the addition of magnesium sulfate causing a reduction in T_m of nearly 10 °C from the reference in water across the full range of concentrations tested. Sodium chloride also caused a slight decrease in T_m of ~ 2 °C at all concentrations. However, due to the relatively stable nature of GX-hyp-6, any destabilisation from sodium chloride present in purification buffers was of little consequence, and its removal from GX-hyp-6 purification buffers was not considered necessary.

Crystallisation

The quantities of GX-hyp-6 following IMAC and SEC purification were sufficient for high-throughput crystallisation trials. A concentration of 4.5 mg/ml was used for these preliminary screening experiments, with the sample purity displayed in Fig. 9.8.

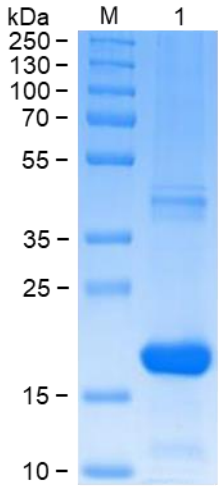

Fig. 9.8: SDS-PAGE analysis showing purity of GX-hyp-6 (MW = 19.3 kDa) for preliminary crystallisation trials. M = marker, with MWs of bands on the left; 1 = GX-hyp-6 sample.

This initial set of screening experiments gave little in the way of promising conditions for optimisation, producing a variety of clear and precipitated drops after extended incubation. Nevertheless, a variety of strategies to promote GX-hyp-6 crystal production could be tested in future. Additional purification, such as with IEC, could facilitate crystallisation by removing heterogeneities that remain in the sample after SEC. Further crystallisation trials with a wider range of screens and a variety of protein concentrations could also be beneficial. The insights gained from TSA results could also be implemented; for example, using low pH buffers and stabilising additives such as ammonium chloride in crystallisation experiments.

9.2.6. GX-hyp-16

Expression and IMAC purification

Full-scale expression experiments produced up to 15 mg of GX-hyp-16 per litre of culture. This was a stark contrast to the low expression levels seen in small-scale test (Fig. 9.2D), likely resulting from differences in growth parameters such as greater aeration achievable in larger shaking cultures[199]. SDS-PAGE analysis of fractions from expression and purification of GX-hyp-16 are shown in Fig. 9.9.

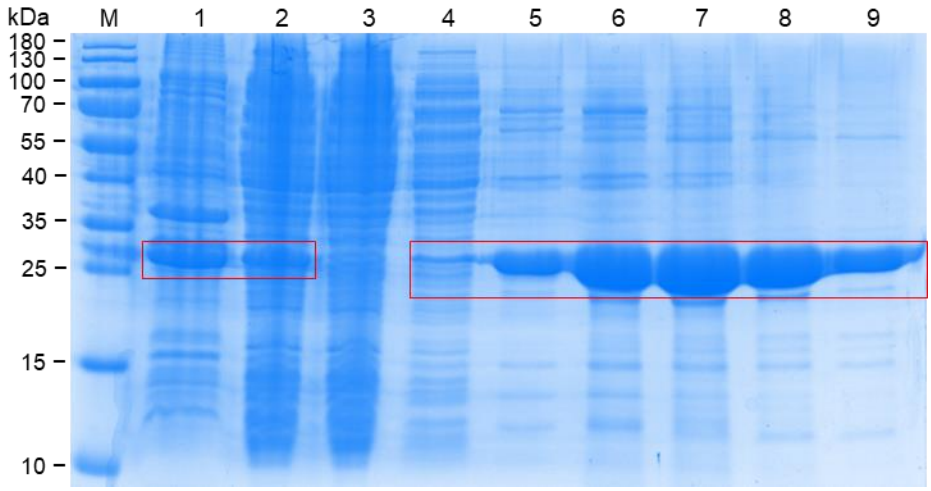

Fig. 9.9: SDS-PAGE analysis of GX-hyp-16 IMAC purification. M = marker, with MWs of bands on the left; 1 = cell pellet; 2 = cell lysate; 3 = flow through; 4 = wash; 5 - 9 = protein fractions. Bands corresponding to GX-hyp-16 (MW = 29.4 kDa) are indicated with a red box.

Fractions containing GX-hyp-16 were immediately apparent due to their brown colour. Despite small amounts of contaminant proteins in the sample, SDS-PAGE indicated that GX-hyp-16 was the overwhelming species after IMAC, and of high enough purity for biophysical characterisation. However, the protein precipitated heavily after only 1 h, both at 4 °C and RT, and even at low concentrations (< 1 mg/ml). This indicated that GX-hyp-16 was highly unstable, which made further analysis challenging.

Characterisation

GX-hyp-16 was found to be stable enough in MilliQ water to obtain an ESI-TOF mass spectrum (Fig. G2). The molecular ion mass (29403 Da) is within 10 Da of the expected MW for GX-hyp-16 (29413 Da), accounting for the anticipated NME.

SEC analysis

Minor contaminant proteins remained in the GX-hyp-16 sample after IMAC purification (Fig. 9.9), and the sample was expected to contain aggregated protein due to the heavy precipitation encountered throughout production and characterisation. SEC was therefore conducted to further purify the target prior to crystallisation trials, as well as assess its oligomeric state. The results of SEC and accompanying SDS-PAGE analysis are shown in Fig. 9.10.

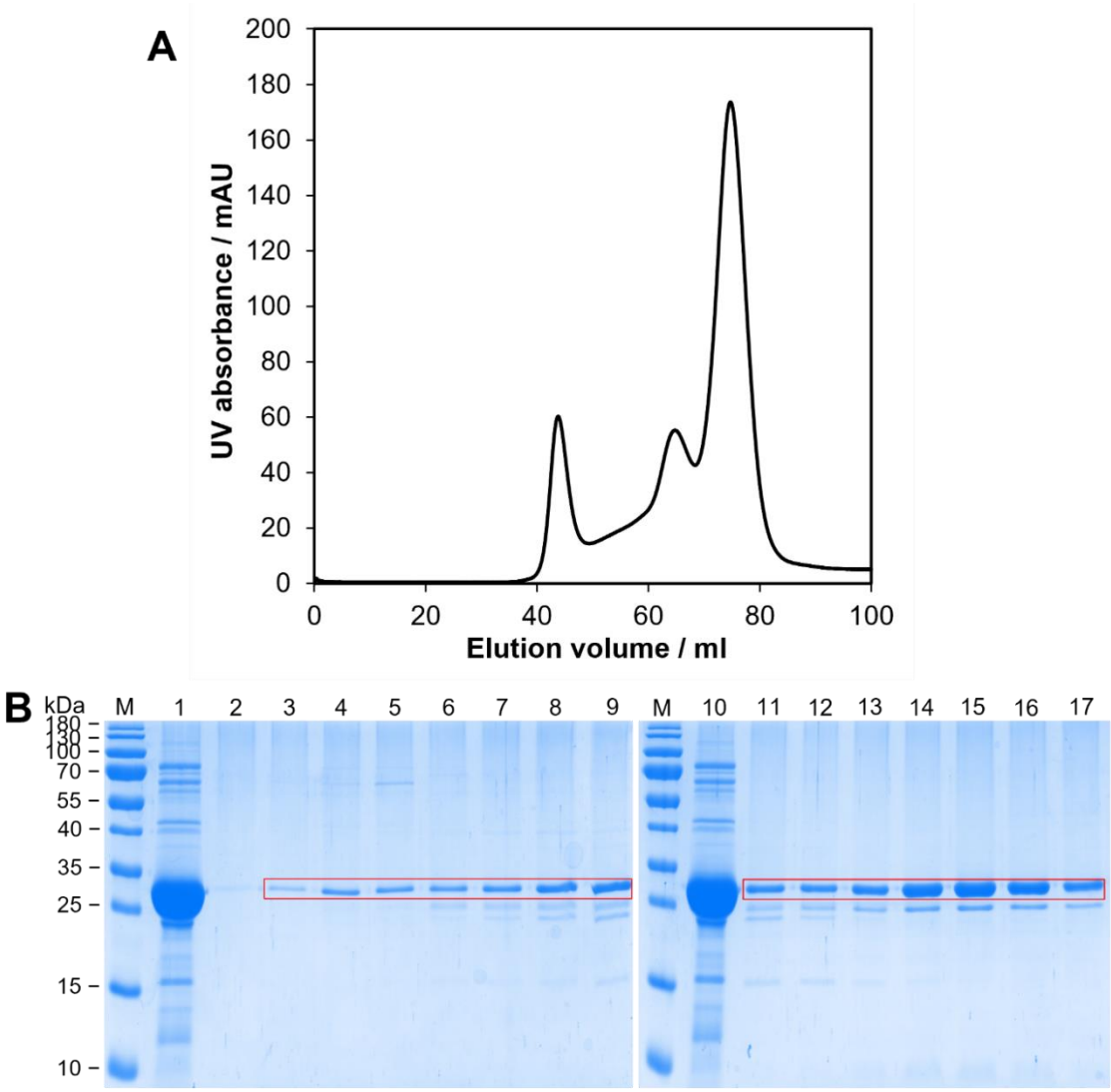

Fig. 9.10: (A) Chromatogram from SEC of GX-hyp-16. **(B)** SDS-PAGE analysis of elution fractions from SEC of GX-hyp-16 (MW = 29.4 kDa). M = marker, with MWs of bands on the left; 1/10 = sample prior to SEC; 2 - 9 and 11 - 17 = protein fractions.

The chromatogram in Fig. 9.10A resolves three distinct species, showing the GX-hyp-16 sample from IMAC sample was considerably heterogeneous. The shoulder at 64.8 ml is ~ 31.1 kDa, which could be the GX-hyp-16 monomer. The minor peak at 43.8 ml is at the void volume, V_o, and corresponds to aggregated protein. However, the V_e of the major peak (74.7 ml) was unexpectedly high, correlating with a ~ 16.6 kDa species, which is roughly half the MW of GX-hyp-16. Despite this, the species was shown to be the expected MW of GX-hyp-16 (29.4 kDa) through SDS-PAGE analysis (Fig. 9.10B). The misleadingly high retention time is likely due to unexpected interactions between GX-hyp-16 and the SEC column matrix.

GX-hyp-16 also co-purified with a contaminant protein of MW ~ 25 kDa. This is likely one of several native *E. coli* proteins commonly found to co-purify during IMAC, such as Crp (23.6 kDa), CAT (25.5 kDa), and YadF (25.0 kDa)[337], or a truncated form of GX-hyp-16. Due to the similarity in MWs

prohibiting separation by SEC, additional purification is necessary to remove this contaminant from GX-hyp-16.

Thermal shift assays

TSA was conducted to find favourable and stabilising conditions for GX-hyp-16, to increase yields of stable protein for crystallisation experiments[210,255]. The reference T_m of GX-hyp-16 in water was consistently 32.0 °C in both the Durham pH and Salt Screens®, and the highest T_m encountered during TSA was only 40.2 °C. This relative instability is unexpected when considering the *Ubervirus* metagenome was identified in a high temperature (75 °C) environment. However, these TSA experiments do not replicate the natural environment of the protein *in vivo*, where additional factors such as chaperone proteins and compatible solutes likely provide significant stabilising effects[225]. It is also possible that the protein is unstable due to improper folding during recombinant overexpression in *E. coli*[162]. Even so, the TSA results did offer some insight into how the protein could be stabilised during production.

pH screening

Several conditions from the Durham pH Screen® resulted in denatured GX-hyp-16 protein at the lowest temperature analysed (25 °C), or generated melt curves with poorly defined transitions, leading to no signal or poorly estimated T_m values for ~ 20% of conditions. Nevertheless, several buffer systems were identified that produced a slight increase in T_m relative to water and the HEPES buffer used during purification. An overview of the results is shown in Fig. 9.11.

A

	1	2	3	4	5	6	7	8	9	10	11	12
A	32.0	32.0	N/S	N/S	N/S	30.0	N/S	41.0 78.0	N/S	N/S	N/S	30.0
B	38.0 74.4	N/S	31.0	41.8 68.2	N/S	29.0	41.0 78.0	41.0 78.0	29.0	30.0	33.2	37.1
C	33.0	37.0	38.3	34.3	35.0	36.5	30.0 35.6	31.0	33.0	31.0	33.8	35.9
D	32.0	33.0	35.6	33.0	36.0	37.7	32.0	32.0	33.0	32.0	33.0	34.4
E	37.1	36.7	37.0	33.0	34.0	36.9	33.0	33.0	33.0	31.0	33.9	33.4
F	33.0	35.3	35.1	33.0	33.0	34.0	34.8	34.6	33.2	33.0	33.0	33.0
G	33.4	33.0	32.0	34.3	33.3	33.0	33.0	33.0	32.0	33.0	30.0	29.0
H	31.0	30.0	30.0	30.0	29.0	N/S	32.0	30.0	29.0	29.0	41.0 78.0	29.0

B

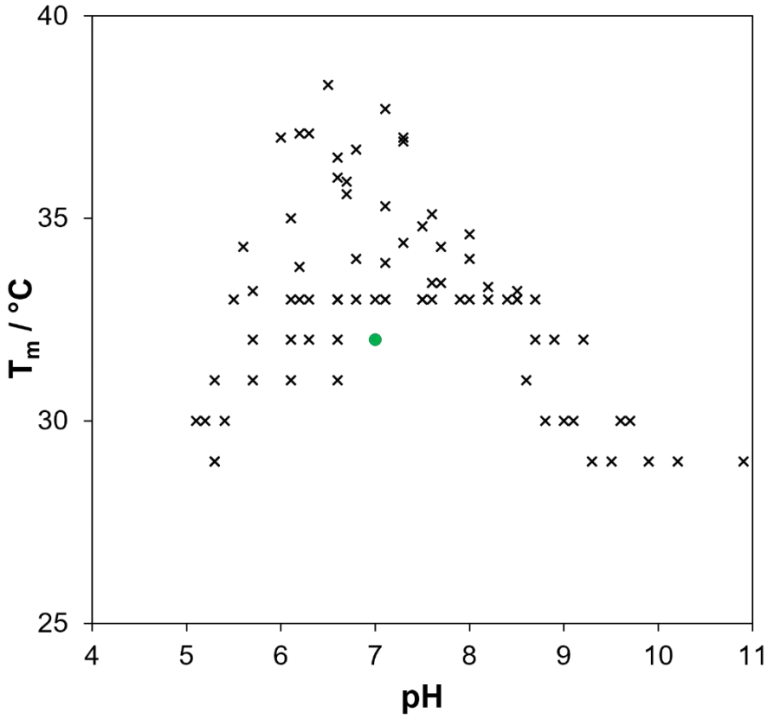

Fig. 9.11: (A) Results from the 96-well plate TSA of GX-hyp-16 with the Durham pH Screen®[210,255]. Colours are indicative of differences in the T_m from that in water (wells A1 and A2): a colour change from colourless through light blue to dark blue indicates an increase in calculated T_m. Green indicates either two stages in the melting process or variation in data. N/S means no signal due to denatured protein. The pH of conditions increases across the screen: rows A to E contain acidic buffers, whilst rows F to H contain alkaline buffers (Table A2). **(B)** Results from TSA using the Durham pH Screen® showing the effect of pH on the T_m of GX-hyp-16. Values corresponding to the reference T_m in water are shown as green circles. Results with no signal or multiple calculated T_m values have been excluded for clarity. Points at the similar or identical pH value are from different buffer systems.

GX-hyp-16 was particularly unstable at the extremes of pH tested, with all conditions at pH ≤ 5 causing denaturation even at 25 °C. In general, GX-hyp-16 seemed to favour weakly acidic conditions, though with a strong preference for particular buffer compositions. For example, there was a > 5 °C range in T_m for buffers at pH 6.6, from 31 °C in MOPS to 36.5 °C in succinic acid (Fig. 9.11A, E10 and C6, respectively). The condition generating the highest T_m (38.3 °C) contained citric

acid at pH 6.5 (Fig. 9.11A C3). This could be a useful buffer system for storage of GX-hyp-16; however, neutral or weakly alkaline buffers are recommended for Ni^{2+} affinity IMAC, so the use of citric acid could negatively impact purification. Of the buffers tested in the pH 7 – 8 range, GX-hyp-16 was most stabilised by pH 7.4 phosphate buffer (T_m 37.0 °C; Fig. 9.11A E3), which caused a ~ 4 °C increase in T_m relative to the condition best reflecting the original purification buffers (pH 7.5 HEPES, Fig. 9.11A F5). A phosphate buffer system is therefore a suitable replacement for HEPES during IMAC purification.

Salt screening

TSA with the Durham Salt Screen® revealed that GX-hyp-16 is stabilised considerably by high concentrations of several common sodium salts (Fig 9.12).

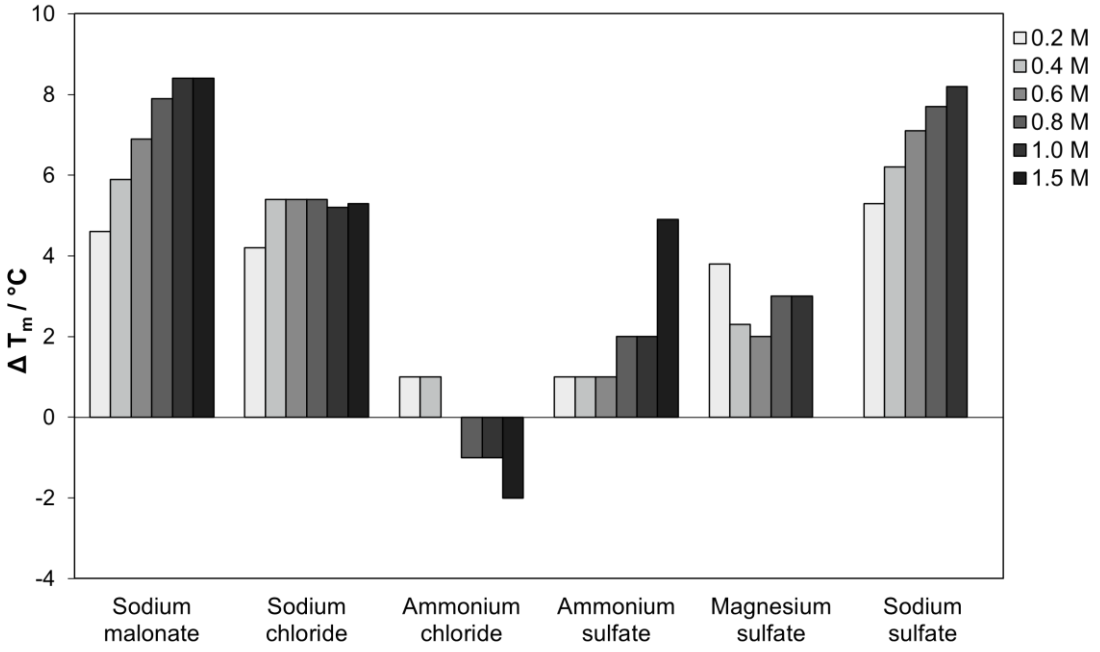

Fig. 9.12: Results from TSA of GX-hyp-16 with the Durham Salt Screen®[210,255] showing changes in T_m upon addition of salts at increasing concentrations. Magnesium sulfate and sodium sulfate were only tested up to 1.0 M.

Sodium malonate, sodium chloride, and sodium sulfate all generated > 4 °C increases in T_m across the concentration range tested. The extent of stabilisation generally increased with salt concentration, with a substantial > 8 °C increase in T_m at ≥ 1.0 M sodium malonate and sodium sulfate. Including a high concentration of one of these sodium salts in GX-hyp-16 buffers could therefore enhance protein stability, which will be particularly beneficial for crystallisation experiments.

Optimised purification

Following TSA, the production protocols for GX-hyp-16 were adapted to improve its stability and reduce precipitation. In place of the SGC buffer (pH 7.4 HEPES, 300 mM NaCl) used in initial rounds of purification, pH 7.4 phosphate buffer with a higher concentration (500 mM) NaCl was

trialled. This optimised buffer dramatically improved the stability of the protein, with no precipitation observed during purification.

9.3. Conclusions

The work presented here forms the foundation for structure determination of several novel protein targets from an uncultivated giant phage, designated the *Ubervirus*. Four of the five GX targets selected by the Virus-X consortium were successfully overexpressed in *E. coli*, with three also produced in large-scale cultures so far and one undergoing crystallisation screening. However, instability of these recombinantly expressed proteins has presented a challenge, with heavy precipitation of both GX-hyp-4 and GX-hyp-16 occurring during purification experiments. In the case of GX-hyp-16, TSA proved to be invaluable, identifying favourable buffer components to stabilise the protein. Further work is now required to optimise production of other GX targets, working towards crystallisation and structural analysis.

At the beginning of the Virus-X project, structure determination of these hypothetical proteins would likely have been non-trivial, requiring more involved phasing strategies than standard MR. However, with the advent of structure prediction tools AlphaFold2[118,263] and RoseTTAFold[309], experimental structure solutions will likely be more straightforward for such targets. The models generated by AlphaFold2 and RoseTTAFold may even be of sufficient quality to begin hypothesising potential functions of the GX targets, and other hypothetical proteins from the *Ubervirus*, to be fully characterised after structural determination.

The *Ubervirus* metagenome was an exciting discovery by the Virus-X consortium. However, some doubt remains over its origins. Whilst the *Ubervirus* was believed to have been isolated from a high temperature environment, its close relationship to sequences from milder conditions is contradictory. Additional efforts to confirm that it originated from the intended geothermal sampling site were ineffective, and the possibility remains that the genome was a contamination[54]. Regardless of its origin, the *Ubervirus* genome remains a valuable source of novel enzymes, and further study of its contents will also help to fill considerable gaps in our understanding of viral ecology and evolution.

Chapter 10. Conclusions

10.1. Contribution of CPR targets to the Virus-X project

The advancement of CPR targets through successive stages of the Virus-X biodiscovery pipeline followed a funnel-like shape, with a decreasing number of targets moving from each stage to the next, as anticipated from previous structural genomics projects[18,145]. The approach taken involved halting work on challenging targets at each stage, and prioritising the most amenable targets first. The main bottlenecks encountered were the soluble expression and crystallisation of CPR proteins, as was also the case for the wider Virus-X project[18]. However, the substantial number of targets at the start of the pipeline ensured that several could progress to each consecutive stage, and the success of the approach is clear from Fig. 10.1.

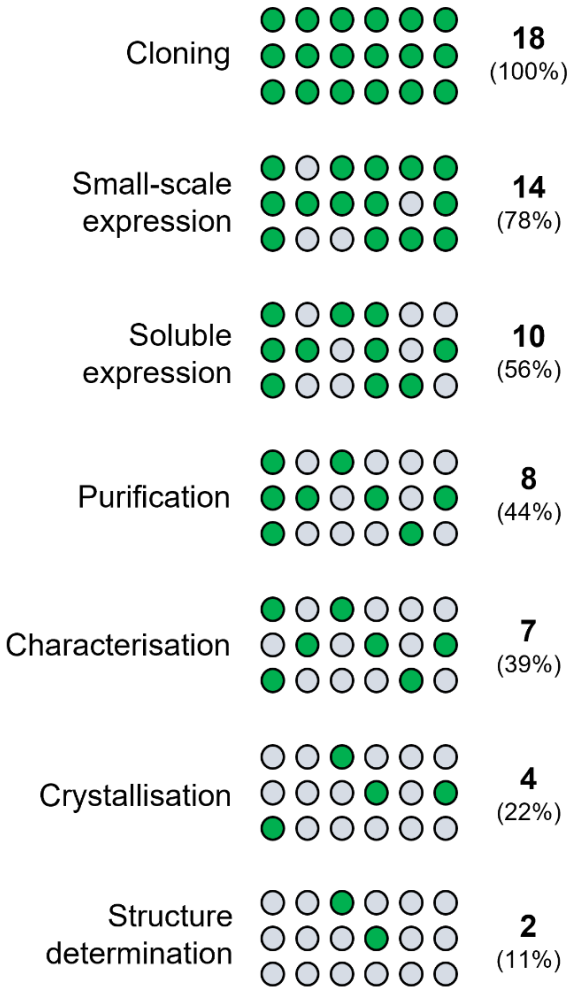

Fig. 10.1: Numbers of CPR targets progressing through the stages of production to crystal structure determination. Green circles represent successful targets at each stage; grey circles represent targets requiring additional work to continue through the pipeline.

Of the 18 original targets selected for production, two successfully advanced through the full pipeline to novel crystal structures (CPR-DprA and CPR-C4), and promising conditions for crystal production have been identified for a further two targets (CPR-exo-1 and CPR-GrpE). Additionally, CPR-ClpB has been identified as a suitable candidate for cryo-EM, and work towards its structure using this rapidly evolving method is ongoing.

The CPR targets discussed in this thesis represent only a fraction of the total Virus-X targets. Regardless, the production and characterisation of the CPR targets contributed significantly to the overall output of the project, with CPR-DprA and CPR-C4 providing two of the 26 total novel crystal structures determined. An overview of the results at each stage of the biodiscovery pipeline across the whole Virus-X project is shown in Fig. 10.2.

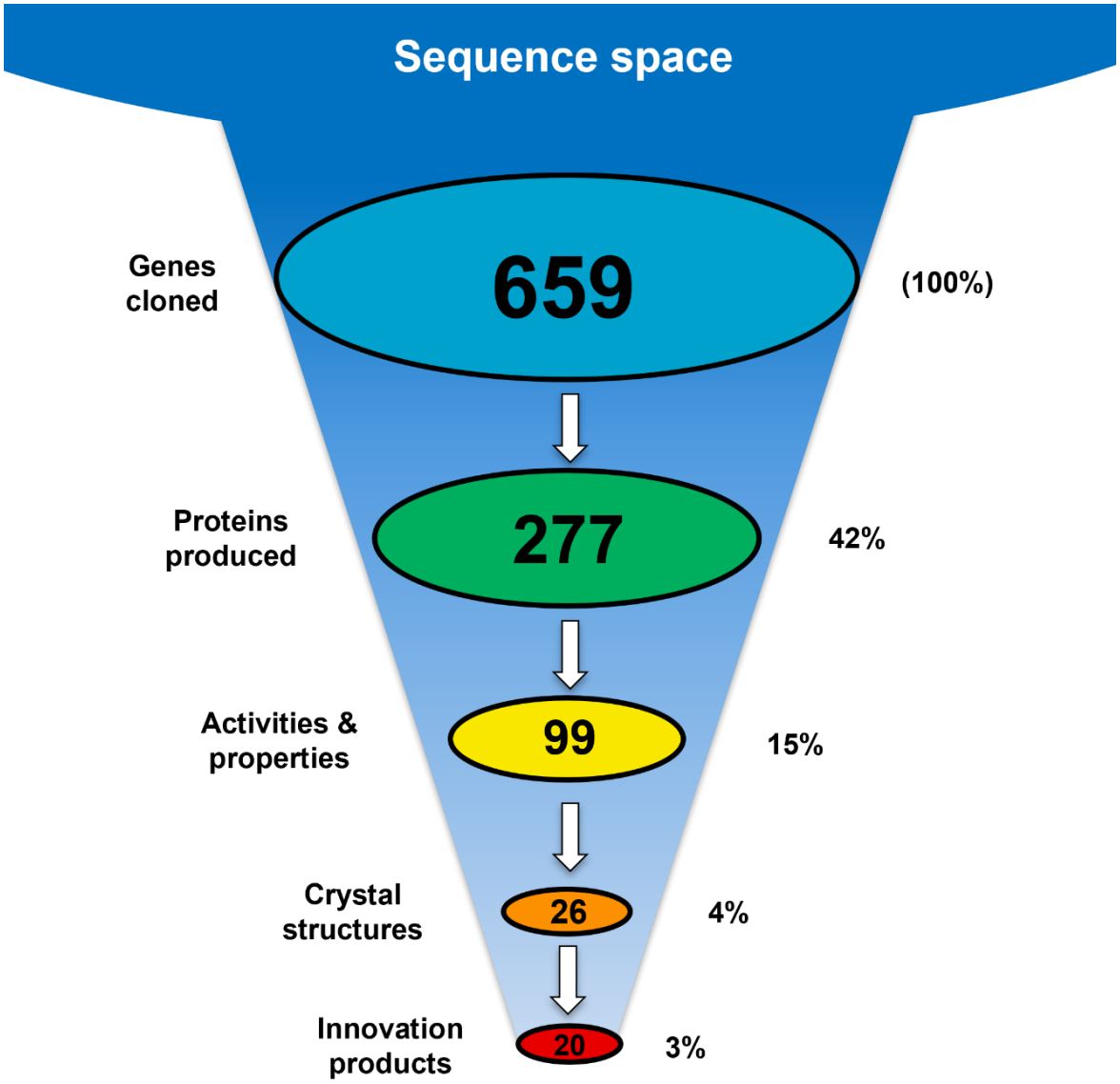

Fig. 10.2: Overall output of the Virus-X project, from sampling total biological sequence diversity, through to commercial products. The numbers of successful targets at each stage are shown, as well as percentages of the total genes cloned[18].

Success rates for the CPR targets were notably higher than those for the overall Virus-X project. Soluble expression and structure determination was achieved for 56% (10/18) and 11% (2/18) of CPR targets, respectively, compared to only 42% (277/659) and 4% (26/659) across all Virus-X targets. Despite the limited time frame, the CPR target success rates are also markedly greater than for many structural genomics projects. 78% of the cloned CPR targets were successfully expressed in trials, in comparison to only 60% from a summary of 17 structural genomics (SG) projects up to June 2004[426]. The proportion of structural solutions from these SG projects is also considerably lower than for the CPR targets (2.6% *vs* 11%)[426], though the scale of target input was obviously vastly different (> 34,000 *vs* 18 genes cloned). Drawing comparisons to more recent results of SG projects is challenging, since such efforts are now considerably reduced, and published results focus on the number of structures deposited in the PDB rather than the ratios of input targets to structures[427,428]. Regardless, it is clear that the outputs from CPR target study are substantial and had a significant impact on the Virus-X project.

10.2. Protein production

10.2.1. CPR targets

The use of standardised expression and purification protocols was an efficient way of producing and evaluating many targets concurrently, and in a short time frame. The *E. coli* expression host and L-rhamnose-induced vectors proved very effective for the recombinant expression of the bacterial CPR proteins, with 14 out of 18 expressing well in small-scale trials (Fig. 10.1). However, it appears that the use of a highly tuneable L-rhamnose system was unnecessary for these targets since frequent basal expression did not detriment the *E. coli* host.

The quick purification of targets was facilitated by the incorporation of a small N-terminal His-tag. The use of non-removable His-tags saved time by avoiding the need for proteolytic cleavage and additional purification steps, though at the risk of the tag influencing protein structure and activity[192]. Even amongst the relatively small number of targets under investigation here, a protein was encountered where the His-tag had a notable impact on structure (CPR-C4), introducing a non-native metal binding site. The high priority of this target necessitated the design of an additional construct to enable His-tag-removal, which ultimately led to an improved crystal structure with better resolution and model geometry[60]. Whilst the time and resources saved by retaining the tags was beneficial for the large number of CPR targets, this result highlights that additional care may be required further down the line as a compromise.

The two-step purification method of IMAC followed by SEC resulted in sufficiently high target purity for characterisation, and structural studies, in most cases. The addition of a further polishing step, such as IEC, may improve the chances of crystallisation for several targets, including CPR-DnaJ and CPR-DnaK, and help to optimise crystal quality for CPR-exo-1 and CPR-GrpE. Further

purification of CPR-ClpB will also help to reduce the sample heterogeneity evident from cryo-EM micrographs, for improved data processing and structure determination.

10.2.2. GX targets

The five viral GX targets presented more of a challenge than the bacterial CPR targets. Instability was a recurring issue, with GX-hyp-4 and GX-hyp-16 precipitating heavily during purification experiments. Protein insolubility was also a prominent issue with other viral targets during the Virus-X project, and seemingly an inherent issue of expression in a bacterial host system[18].

The *Ubervirus* genome encodes components of the translational machinery and ribosomal proteins. It is possible that this jumbophage combines these components with those of a host, creating a unique environment for the correct production and folding of proteins. There could also be unknown post-translational modifications that *E. coli* expression hosts lack the necessary folding environment and translational systems to carry out effectively. Considerable work may be required to enable successful production and experimental structure determination, made more challenging by our limited understanding of viral proteomes. This emphasises the need for continued study of viruses to better access their unexplored genome contents.

10.3. CPR-DprA

CPR-DprA, a predicted DNA processing protein, was expressed and purified following standard protocols, and formed high-quality diffracting crystals. The X-ray crystal structure of the core DprA domain of CPR-DprA was determined by MR to a resolution of 2.1 Å, with R/R_{free} = 0.20/0.23. Consistent with homologs, this domain in CPR-DprA adopts an extended Rossmann fold, which forms homodimers through conserved hydrophobic interactions and hydrogen bonding. Though a C-terminal extradomain (CTD) with predicted DNA binding capabilities is missing from the crystal structure, a model generated using the AlphaFold2 structural prediction tool was found to align well with the experimentally determined CTD of DprA from *R. palustris*, regardless of low sequence identity.

Sequence and structural similarities between CPR-DprA and DprA homologs imply that CPR-DprA will be able to bind ssDNA, though this assertion should be explored experimentally. Electrophoretic mobility shift assays (EMSAs) could determine DNA binding capabilities and preferences of CPR-DprA, whilst co-crystallisation with DNA will allow probing of DNA-binding modes. Deeper examination of the CPR-DprA CTD is also necessary, given that functional diversity in DprA proteins seems to be primarily conferred by these additional domains[245,250,254]. The discovery of a new and specific DNA-binding functionality would give this target commercial prospects. The role of DprA proteins in metabolically limited CPR organisms, particularly for essential exogenous DNA uptake[429,430], should also be explored.

10.4. CPR-C4

A particularly notable outcome of this work was the structural determination and subsequent functional annotation of CPR-C4 as a novel protease with remarkable structural homology to the human vasohibins, despite low sequence similarity. During initial target selection, this hypothetical category C target was highlighted as being of particular interest due to its wide occurrence throughout the bacterial domain, despite a lack of sequence similarity to any characterised proteins. Its structure was determined in three crystal forms, with the original structure in $P3_221$ (2.7 Å, R/R_{free} = 0.20/0.23) solved by utilising intrinsic Zn^{2+} for MAD phasing. Structural comparisons with the 3DM software revealed the unexpected similarity between CPR-C4 and the vasohibin proteases, with subsequent 3D-MSAs also uncovering near total conservation of a highly unusual cysteine-histidine-leucine catalytic triad throughout vasohibin and CPR-C4-related sequences. CPR-C4 was therefore predicted to display protease activity similar to that of the vasohibins, which was then confirmed experimentally. This analysis of CPR-C4 exemplifies the effectiveness of structure-based studies to establish the function of hypothetical proteins, where sequence alone is insufficient. As a result of this work, the function of a family containing hundreds of hypothetical protein sequences from the CPR and beyond has been revealed.

Further work is now needed to investigate the specific role of CPR-C4 and expose any unique functionalities. Optimal conditions for protease activity should be determined, and the increased thermostability with Zn^{2+} noted during TSA should be explored in more depth. Uncovering its native substrate will also be a crucial step towards characterising CPR-C4, as well as establishing prospective applications for the protease. Synthetic peptide libraries could be used to sample potential substrates across a large amount of sequence space and assess substrate specificity. Continued analysis of CPR genomes to increase our knowledge of their proteomes and metabolisms will also serve to clarify the role of CPR-C4 *in vivo*.

10.5. CPR-DnaK/ClpB bi-chaperone system

All four components of the CPR bi-chaperone system (CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB) were successfully overexpressed and purified for characterisation and structural studies. Promising crystallisation conditions have been found for CPR-GrpE, and CPR-ClpB was identified as an excellent target for cryo-EM studies, with preliminary data collected and optimisation underway. This work forms a foundation from which to optimise target production, to increase the likelihood of structure determination and functional characterisation of both the individual components and the full cooperating chaperone system. The use of cryo-EM will be particularly valuable here, as it could reveal multiple conformations to give a dynamic picture of the chaperone system that can help to unravel its mechanism of action. These studies will become easier as access to cryo-EM equipment and facilities widens, and sample preparation throughput increases.

The DnaK/ClpB bi-chaperone system identified from a thermophilic CPR organism has tangible biotechnology applications, namely in recombinant expression to stabilise proteins and support the folding process. The innovation potential of these proteins is demonstrated by the fact that a patent was filed in 2020 for a similar set of novel thermostable heat-shock proteins discovered during the Virus-X project[18]. Structure determination for the CPR chaperone targets and comparisons to homologous structures, in combination with functional assays, could reveal distinct and valuable traits that make them preferable over existing chaperone enzymes. An important task now is to determine the optimal conditions under which this system could be used for during recombinant expression, as well as the range of physio-chemical conditions it can tolerate. If the system can be shown to function under extreme temperatures, it could be valuable for use in thermophilic expression hosts.

10.6. Thermophilic potential of CPR and GX targets

TSA proved to be a valuable tool for identifying stabilising conditions for target proteins, and was pivotal for improving production of CPR-C4 and GX-hyp-16. TSA also enabled a rapid qualitative assessment of the intrinsic thermostability of target proteins. Since the CPR and GX genes were derived from a high temperature environment (75 °C), it was anticipated that the encoded proteins would be thermophilic. However, the majority of target proteins were considerably less thermostable than originally expected. It is possible that the *Ubervirus* and CPR source organism did not come from the proposed geothermal sample site and are instead the result of DNA contamination, as is a common issue with genome isolation through metagenomics[54]. It is also possible that the hot spring could be contaminated with non-thermophilic organisms through the flow of groundwater into the hot spring[36], though our current inability to isolate or cultivate the CPR source organism prevents assessment of its preferred growth temperature. Nevertheless, the CPR and *Ubervirus* genomes are filled with unexplored and potentially valuable enzymes, and continued analysis of their proteomes may also reveal clues as to their optimum growth environment.

10.7. The impact of advanced structural prediction tools

During the course of this project, the powerful structure prediction tools AlphaFold2[118,263] and RoseTTAFold[309] were released and have already revolutionised the structural biology field. Were the Virus-X project to begin now, the biodiscovery pipeline would likely include automatic structural prediction of gene products using these tools. This would assist selection of interesting targets for subsequent production, structure determination, and in-depth functional analysis. The 54 million gene products identified by the Virus-X project are more than could be produced and characterised on any imaginable timescale, and so an approach centred around structure prediction could rapidly expand the rate at which sequence diversity can be explored. Accelerating functional predictions for the swathes of hypothetical proteins within uncultivated genomes would assist the search for novel and unique enzymes. It should, however, be remembered that these tools are not a magic bullet for

all proteins, and structure determination by X-ray crystallography, and increasingly by cryo-EM, will remain a necessary part of bioprospecting.

It would be interesting to implement structure prediction for the entire proteome of the CPR organism from which the targets discussed here were selected, as well as other available CPR genomes. This could more rapidly broaden our knowledge of these poorly understood organisms, likely finding proteins that complete metabolic pathways thought to be missing or broken. It could also reveal other features that distinguish the CPR from non-CPR bacteria, helping to establish their evolutionary separation. A similar approach would also be useful for the *Ubervirus* genome, to further study the blurred distinction between giant viruses and cellular life.

10.8. Shining a light on microbial dark matter

The Virus-X project has been a successful endeavour to explore new genetic territories and sequence diversity, including that of the enigmatic microbial dark matter. Targets from the Candidate Phyla Radiation saw particular success, with 14 of the 18 CPR targets cloned for production expressed using a heterologous system (78%), leading to large quantities of soluble and pure protein for eight novel proteins (44%). Within the limited time frame of this project, seven of these targets have been characterised (39%) and promising crystallisation conditions have been identified for four (22%). The structures of two novel CPR targets, CPR-DprA and CPR-C4, have been determined through X-ray crystallography (11%), which enabled more in-depth assessment of their features than could be achieved through sequence analysis alone. The DNA processing enzyme, CPR-DprA, contains an extended Rossmann fold that is highly conserved amongst DprA proteins throughout the bacterial domain. Additionally, CPR-DprA contains a C-terminal extradomain with a conserved winged helix-like fold predicted by AlphaFold2, despite significant sequence divergence from homologs. The hypothetical protein CPR-C4 was discovered to have a mixed αβ fold with a striking resemblance to that of the human vasohibin proteases, again in spite of low sequence identity, suggesting an evolutionary relationship undetectable at the sequence level. CPR-C4 could subsequently be characterised as a novel cysteine protease containing an unusual cysteine-histidine-leucine(carbonyl) catalytic triad using structure-based multiple sequence alignments and fluorescence-based activity assays[60]. A DnaK/ClpB bi-chaperone system from the CPR (CPR-GrpE, CPR-DnaJ, CPR-DnaK, and CPR-ClpB) with clear biotechnology applications was also expressed and characterised, and preliminary cryo-electron microscopy studies performed with the CPR-ClpB disaggregase. Whilst these targets represent an incomprehensibly small fraction of the gene products concealed within uncultivated microorganisms, the successes described here exemplify and validate the metagenomic approach of the Virus-X project and highlight the need to further explore the uncharted sequence space of microbial dark matter.

Chapter 11. Experimental methods

11.1. General experimental information

All chemicals and materials were purchased from Merck or Thermo Fisher unless otherwise specified. Protein and DNA concentration estimates and OD_{600} measurements were made on a DeNovix D5-11+ Spectrophotometer. Protein molecular weights (MW) and estimated extinction coefficients (ε) at 280 nm were calculated using the ExPASy ProtParam tool[203].

11.2. Virus-X sampling, target selection, and annotation

The Virus-X biodiscovery pipeline has previously been described in detail[18]. CPR metagenome retrieval was conducted by members of the Virus-X consortium in Iceland. Briefly, hot spring samples were processed with consecutive steps of microfiltration, concentration, and DNA extraction to isolate genetic material for total metagenomic DNA analysis by NGS platforms (Illumina MiSeq, HiSeq; Oxford Nanopore). The output from sequencing was assembled and binned for downstream analysis, including quality control and assembly into longer contigs, followed by gene prediction, taxonomic and functional annotation of predicted genes, and binning into metagenome-assembled genomes. Gene products with high innovation potential were selected by Dr A. Ævarsson (Matís Iceland) for expression and characterisation, prioritising those showing extended conservation across genomes but with unknown function[60].

11.3. GX and CPR cell stock production

11.3.1. General protocol

The pJOE5751.1 vector was received from Virus-X partners in Stuttgart as a lyophilised plasmid and stored at −20 °C before delivery to GenScript on dry ice. Cloning of CPR and GX genes into the pJOE5751.1 vector was conducted commercially by GenScript: DNA fragments were inserted into pJOE5751.1 *via* BamHI and BsrGI restriction sites to create the pJOE5751.1-*CPR/GX* expression constructs. Genes were also cloned into pUC19 vectors for long-term storage.

Lyophilised pJOE5751.1-*CPR/GX* plasmids were centrifuged at 2000 x *g* for 1 min at 4 °C before adding 20 µl MilliQ water and vortexing for 1 min. 2 µl of a 10-fold dilution of the pJOE5751.1-*CPR/GX* plasmid stock in MilliQ water was used in the transformations of each plasmid into T7 Express Strain competent *Escherichia coli* cells (NEB) (see 11.5.). Transformed cells were plated on lysogeny broth (LB) agar (100 µg/ml ampicillin) for incubation overnight at 37 °C. Single colonies

were selected from transformation plates and used to inoculate 10 ml Lysogeny Broth (LB) overnight cultures (100 µg/ml ampicillin, 37 °C, 150 rpm), from which plasmid DNA was extracted and purified for sequencing (see 11.4. and 11.7.). Glycerol stocks were made with these overnight cultures using 500 µl culture + 500 µl 50% v/v glycerol. pJOE5751.1-*CPR/GX* plasmid DNA was also transformed into NEB® 5-alpha cells for long-term storage and glycerol stocks produced as above.

11.3.2. CPR-C4

The CPR-C4 gene was cloned separately into the pET28a(+)TEV vector by GenScript, with insertion *via* BamH1 and Xho1 restriction sites. The lyophilised plasmid was prepared as in 11.3.1. and similarly transformed into T7 Express Strain competent *E. coli* cells, with overnight cultures generated for production of glycerol stocks.

11.4. Plasmid recovery

Plasmid purifications were conducted at room temperature (RT) with 4 – 5 ml of overnight culture using the GeneJET plasmid miniprep kit (Thermo Scientific) according to instructions provided by the manufacturer. Purified plasmid DNA was stored at −20 °C.

11.5. Plasmid transformations

11.5.1. General protocol

Transformations of pJOE5751.1-*CPR/GX* plasmids into T7 Express and NEB® 5-alpha competent *E. coli* cells (NEB) were carried out using the following protocol, with variations in the length of heat shock and incubation temperatures as indicated. 2 µl (1 pg – 100 ng) purified plasmid DNA was added to thawed *E. coli* cells and the mixture placed on ice for 30 min. The cells were subjected to heat shock at 42 °C (10 s for T7 Express; 30 s for NEB® 5-alpha) and placed back on ice for a further 5 min. 950 µl RT SOC medium was added to the cells before incubation with shaking (1 h, 250 rpm, 37 °C). 100 µl of this mixture was plated on LB agar (100 µg /ml ampicillin) and incubated overnight at 37 °C. 100 µl of a 10-fold dilution in SOC was also plated. The success of cloning and subsequent transformations was confirmed by agarose gel electrophoresis and Sanger sequencing of purified plasmids from transformed cells (see 11.6. and 11.7.).

11.5.2. Transformation of CPR-B3 into SHuffle® cells

The pJOE5751.1-*CPRB3* plasmid was also transformed into SHuffle® T7 competent *E. coli* cells (NEB). 2 µl purified plasmid DNA was added to thawed cells and the mixture placed on ice for 30 min. The cells were subjected to heat shock at 42 °C for 30 s and placed back on ice for a further 5 min. 950 µl RT SOC medium was added to the cells before incubation with shaking (1 h, 250 rpm, 30 °C). 100 µl of this mixture, and 100 µl of a 10-fold dilution in SOC, were plated on LB agar (100 µg/ml ampicillin) and incubated overnight at 30 °C.

11.5.3. Transformation of pET28a(+)TEV-*CPRC4*

Transformation of pET28a(+)TEV-*CPRC4* into T7 Express and NEB® 5-alpha competent *E. coli* cells (NEB) was conducted using the protocol in 11.5.1., replacing ampicillin with 50 μg/ml kanamycin.

11.6. Agarose gel electrophoresis

All agarose gel electrophoresis experiments were performed using 1.5% agarose gels, run at 150 V for approximately 1 h. Samples were run against GeneRuler 100 bp and 1 kbp DNA ladders (Thermo Scientific). Gels were stained using ethidium bromide and imaged using a BioRad Molecular Imager® GelDoc™ XR+ with ImageLab™ Software.

11.7. DNA Sequencing

Purified plasmid (15 μl) was sequenced with T7 promoter and rrnB terminator primers (3.2 pmol/μl) by Sanger sequencing using Applied Biosystems 3730 capillary instrumentation by the DNA Sequencing service in the Department of Biosciences, Durham University. DNA sequences were translated using the ExPASy Translate tool[203].

11.8. Small-scale protein expression trials

11.8.1. General protocol

Single colonies were selected from transformation plates and used to generate 10 ml LB overnight starter cultures. Starter cultures were then used to inoculate two 10 ml LB cultures (250 μl culture, 100 μg/ml ampicillin; 37 °C, 150 rpm). These were grown to OD_{600} 0.4 - 0.6 before induction of one culture with L-rhamnose (Melford; 0.2% in culture). The second culture was left to grow without L-rhamnose. Induced and uninduced cultures were shaken for a further 3.5 h (37 °C, 150 rpm) and samples from each were then analysed by whole-cell SDS-PAGE (see 11.12.). This was repeated with four single colonies for each target.

11.8.2. Adjusted expression trials with CPR-B3

A T7 Express *E. coli* glycerol stock containing pJOE5751.1-*CPRB3* plasmid was used to generate four 20 ml LB cultures (100 μg/ml ampicillin; 37 °C, 150 rpm), which were grown to OD_{600} 0.4 - 0.6 before a 1 ml sample was removed for whole-cell SDS-PAGE (see 11.12.). Two cultures were then induced with L-rhamnose (Melford; 0.2% in culture) and the other two were left to grow without L-rhamnose. One pair of induced and non-induced cultures was transferred to a 30 °C incubator whilst the other pair was kept at 37 °C. All cultures were shaken for a further 4 h (150 rpm) before samples were taken for whole-cell SDS-PAGE analysis. Further samples were taken 16 h after induction. The optical densities of all samples were balanced by dilution in LB prior to SDS-PAGE analysis.

11.8.3. Expression trials of CPR-C4 with $ZnCl_2$

A T7 Express *E. coli* glycerol stock containing the pJOE5751.1-*CPRC4* plasmid was used to generate four 10 ml LB cultures (100 μg/ml ampicillin; 37 °C, 150 rpm), with varying concentrations of $ZnCl_{2(aq)}$ added to the media: 0 μM, 25 μM, 50 μM and 100 μM. All cultures were grown to OD_{600} 0.4 before induction with L-rhamnose (Melford; 0.2% in culture). After approx. 16 h, 1 ml samples were taken from each culture and the optical densities were balanced prior to whole-cell SDS-PAGE analysis (see 11.12.).

11.8.4. Expression trials of CPR-C4 from pET28a(+)TEV-*CPRC4*

Expression trials for CPR-C4 from the pET28a(+)TEV-*CPRC4* plasmid were conducted as in 11.8.1., with 50 μg/ml kanamycin rather than ampicillin, and induction with 1 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) in place of L-rhamnose.

11.9. Full-scale protein expression

11.9.1. General expression protocol

All GX and CPR targets were expressed from pJOE5751.1-*CPR/GX* plasmids as fusion proteins with an N-terminal His-tag (MTMITHHHHHHGS).

25 ml starter cultures (20 g/litre LB; 100 μg/ml ampicillin) were generated from transformed T7 Express *E. coli* cells and grown overnight (37 °C, 150 rpm). 6 x 1 litre LB was inoculated with the starter cultures (25 ml) and grown to OD_{600} 0.4 - 0.6 (~ 4 h, 37 °C, 150 rpm), at which point overexpression was induced with the addition of L-rhamnose (0.2% in culture) followed by further shaking overnight (~ 20 h, 25 °C, 150 rpm). The resulting cultures were centrifuged using a Beckmann Avanti Hi-Speed centrifuge (JLA-8.1000 rotor, 1300 x *g*, 25 min, 4 °C), the supernatant decanted, and the cell pellets frozen at −80 °C until required.

11.9.2. Adjusted expression of CPR-hel-2

The protocol in 11.9.1. was adjusted for CPR-hel-2 by reducing the growth time and temperature to 4 h at 18 °C after induction with L-rhamnose.

11.9.3. CPR-C4 expression with $ZnCl_2$

Optimised expression experiments for CPR-C4 from pJOE5751.1-*CPRC4* followed the procedure in 11.9.1. with the addition of $ZnCl_2$ (100 μM in culture) to LB media for 25 ml precultures and 1 litre cultures.

11.9.4. CPR-C4 expression from pET28a(+)TEV-*CPRC4*

CPR-C4 was also expressed as a second fusion protein with a cleavable N-terminal His-tag containing a TEV protease recognition site. The protocol outlined in 11.9.1. was followed, replacing ampicillin and L-rhamnose with 50 μg/ml kanamycin and 1 mM IPTG, respectively.

11.10. Immobilised metal ion affinity chromatography

11.10.1. General protocol

The following buffers were as standard used during purification:

Binding buffer:	40 mM imidazole, 300 mM NaCl, 20 mM HEPES pH 7.5
Elution buffer:	500 mM imidazole, 300 mM NaCl, 20 mM HEPES pH 7.5
SGC buffer:	10% v/v glycerol, 300 mM NaCl, 20 mM HEPES pH 7.5
Glycerol-free SGC:	300 mM NaCl, 20 mM HEPES pH 7.5

Cell pellets were thawed on ice and re-suspended in 25 ml cold binding buffer with cOmplete™ Mini EDTA-free Protease Inhibitor Cocktail (Roche). The suspension was sonicated on ice (4 min; 10 s on/off, 40% power) and centrifuged using a Beckmann Avanti Hi-Speed centrifuge (JA-25.50 rotor, 50,000 x *g*, 50 min, 4 °C). The lysate was filtered (0.22 μm) and loaded onto a 1 ml or 5 ml HisTrap™ HP affinity column (Cytiva) for capture and purification by immobilised metal ion affinity chromatography (IMAC). The column was washed with 5 column volumes (CV) of binding buffer on an ÄKTA Pure FPLC and target protein was eluted in 2 ml fractions using a 20 CV linear imidazole gradient (40 mM to 500 mM). Protein elution was detected by UV absorbance at 280 nm. Eluted fractions, unpurified lysate, and the cell debris pellet were analysed by SDS-PAGE (see 11.12.). Fractions containing target protein were combined for SEC analysis, or dialysed into SGC buffer for storage at −80 °C.

11.10.2. Optimised purification of CPR-C4

Thawed cell pellets were resuspended in 20 mM binding buffer (sodium phosphate pH 7.4, 500 mM NaCl, 40 mM imidazole) with cOmplete Mini EDTA-free Protease Inhibitor Cocktail (Roche), then sonicated and centrifuged at 50,000 x *g* for 50 min at 4 °C. The supernatant was filtered (0.22 μm) and loaded onto a 1 ml HisTrap™ HP affinity column (Cytiva), followed by column washing with 5 CVs binding buffer. CPR-C4 was eluted in 1 ml fractions by FPLC with a 10 CV imidazole gradient (40 mM to 1 M imidazole) using an ÄKTA Pure. Protein elution was detected by UV absorbance at 280 nm. Eluted fractions, unpurified lysate, and the cell debris pellet were analysed by SDS-PAGE (see 11.12.).

11.10.3. Purification of CPR-C4 from pET28a(+)TEV-*CPRC4*

Initial purification of CPR-C4 expressed from the pET28a(+)TEV-*CPRC4* plasmid followed the protocol described in 11.10.1. Fractions containing CPR-C4 from IMAC were then combined and incubated overnight at 4 °C with His-tagged TEV protease (Merck) for affinity tag removal (TEV to CPR-C4 ratio of 1:100 w/w). Tag-free CPR-C4 was reloaded onto a 1 ml HisTrap™ HP affinity column (Cytiva) and purified by imidazole gradient (0 - 500 mM imidazole) FPLC using an ÄKTA Pure.

11.10.4. Optimised purification of GX-hyp-16

Optimised purification of GX-hyp-16 was conducted using the protocol outlined in 11.10.1. with the following buffers:

Binding buffer: 40 mM imidazole, 500 mM NaCl, 20 mM sodium phosphate pH 7.5
Elution buffer: 500 mM imidazole, 500 mM NaCl, 20 mM sodium phosphate pH 7.5
SEC buffer: 500 mM NaCl, 20 mM sodium phosphate pH 7.5
Storage buffer: 10% v/v glycerol, 500 mM NaCl, 20 mM sodium phosphate pH 7.5

11.11. Size exclusion chromatography

11.11.1. General protocol

Size exclusion chromatography (SEC) experiments were performed using a 120 ml HiLoad 16/600 Superdex 75 pg column (Superdex 200 for CPR-DnaK and CPR-ClpB only) (Cytiva) on an ÄKTA Pure FPLC. Columns were equilibrated with 1.5 CV glycerol-free running buffer (SGC buffer, or SEC buffer for CPR-C4 and GX-hyp-16). Concentrated protein sample (1 ml) was then injected onto the column *via* a 2 ml injection loop. Protein was eluted in 1.5 ml fractions using two CVs running buffer at ~ 1 ml/min. Protein elution was detected by UV absorbance at 280 nm. The resulting fractions were analysed by SDS-PAGE (see 11.12.) and stored at 4 °C until required, or dialysed into SGC buffer with glycerol (storage buffer for CPR-C4 and GX-hyp-16) for long-term storage at −80 °C.

11.11.2. Column calibration

Size exclusion columns were calibrated using low molecular weight (LMW) and high molecular weight (HMW) kits (Cytiva) according to instructions provided by the manufacturer. Calibration standard proteins from the kits were suspended in glycerol-free SGC buffer, or MilliQ for carbonic anhydrase, to a concentration of 20 mg/ml. Calibration standards were then diluted as required to the concentrations shown in Table 11.1 to obtain UV absorbance peaks at 280 nm with similar amplitudes.

Table 11.1: MWs of protein standards and concentrations used for column calibration.

LMW calibration kit				**HMW calibration kit**		
Standard	**MW / kDa**	**Conc. / mg/ml**		**Standard**	**MW / kDa**	**Conc. / mg/ml**
Aprotinin	6.5	3		Ovalbumin	44	4
Ribonuclease A	13.7	3		Conalbumin	75	3
Carbonic anhydrase	29	3		Aldolase	158	4
Ovalbumin	44	4		Ferritin	440	0.3
Conalbumin	75	3		Thyroglobulin	669	5
Blue dextran	2000	1		Blue dextran	2000	1

A selection of calibration proteins for each column type (Table 11.2) were combined and 0.6 ml of the solution (0.5% of the geometrical column volume, V_c = 120 ml) was applied to the column.

Table 11.2: Protein standards selected for calibration of different column types.

	Column type	
	HiLoad 16/600 Superdex 75 pg	**HiLoad 16/600 Superdex 200 pg**
Calibration standards	Aprotinin	Aprotinin
	Ribonuclease A	Ribonuclease A
	Carbonic anhydrase	Carbonic anhydrase
	Conalbumin	Ovalbumin
		Conalbumin
		Aldolase
		Ferritin

From the UV absorbance curve, the elution volumes (V_e) for the calibration kit standards were determined by measuring the volume of eluent from the point of injection to the centre of the elution peak. A separate calibration experiment using blue dextran 2000 (1.0 mg/ml, 0.6 ml) was conducted to determine the void volume (V_o) of the column, which is equivalent to the V_e for blue dextran. Partition coefficients (K_{av}) for each standard were calculated using the following equation:

$$K_{av} = \frac{V_e - V_o}{V_c - V_o}$$

This allowed a logarithmic standard curve to be drawn for each column relating K_{av} to MW, from which the V_e of target proteins could be used to determine their MW and establish oligomeric state.

11.12. SDS-PAGE

11.12.1. General protocol

Polyacrylamide gels were prepared using 8%, 12%, or 16% resolving gel solution (~ 5 ml) and 5% stacking gel solution (~ 1 ml) and were run at 200 V for 45 - 50 min using a Bio-Rad Mini-PROTEAN® Tetra Cell system. Gels were subsequently stained with InstantBlue™ Protein Stain (Expedeon) and imaged using a BioRad Molecular Imager® GelDoc™ XR+ with ImageLab™ Software. Samples were run against 5 µl of a 10 - 180 kDa PageRuler™ Pre-stained Protein Ladder or 10 - 250 kDa PageRuler Plus™ Pre-stained Protein Ladder (Thermo Scientific). Gel figures were labelled using Microsoft PowerPoint.

11.12.2. Sample preparation

Whole cell sample preparation

1 ml of culture at OD_{600} ~ 1.0 was centrifuged for 5 min at 12000 x *g*. The supernatant was discarded, and the pellet re-suspended in 50 µl SDS loading buffer, before heating for 20 min at 96 °C and centrifuging for 20 min at 12000 x *g*. The top 10 µl of supernatant was loaded onto the gel.

Protein sample preparation

5 µl SDS loading buffer was added to 10 µl sample and the mixture was heated at 96 °C for 5 min. The samples were then briefly centrifuged before loading 10 µl onto the gel.

11.13. Mass spectrometry

11.13.1. ESI-TOF MS

Protein samples were buffer exchanged into MilliQ water and concentrated to ~ 1 mg/ml using a Vivaspin™ 500 MWCO 5000 spin column (Cytiva). Samples (150 µl) were analysed by electrospray ionisation (ESI) mass spectrometry using a Quad Time-of-Flight (QToF) Premier Spectrometer (Waters) by the mass spectrometry service in the Department of Chemistry, Durham University.

11.13.2. Trypsin digest MS

Trypsin digest tandem MS experiments were performed by Dr A. Brown at the Proteomics Facility, Durham University. SDS-PAGE protein bands were provided for analysis.

11.14. Thermal shift analysis

Thermal shift analysis (TSA) was conducted using the Durham Screens®[210,211] (Molecular Dimensions). Protein samples were dialysed into 10 mM sodium phosphate pH 7.4, 100 mM NaCl. SYPRO orange dye (4 µl, 5000x in DMSO) was added to the protein sample (1 ml, 0.5 – 1.5 mg/ml) to give a protein-plus-dye solution. 10 µl of each Durham Screen® condition was added to 10 µl of the protein-plus-dye solution in a standard 96-well PCR plate, which was sealed with thermostable film before centrifugation (2 min, 160 x *g*). The temperature of the plate was held for 1 min at 1 °C intervals from 24 °C to 96 °C, and fluorescence data were collected using an Applied Biosystems 7500 Fast Real-Time PCR System with an excitation range of 540 – 550 nm. Data were analysed with in-house Microsoft Excel scripts and the graphical user interface-based python program NAMI[211] using the emission signal at 567 – 596 nm.

11.15. Circular dichroism

Circular dichroism (CD) spectra were collected using a Jasco J-1500 CD spectrophotometer by Dr B. Bromley in the Department of Physics, Durham University, with 1 mm pathlength and 3 nm bandwidth. Protein samples were provided in SGC buffer.

11.16. Protein crystallisation

11.16.1. Crystallisation screening

Protein samples were buffer exchanged into SGC buffer (10% v/v glycerol, 20 mM HEPES pH 7.5, 300 mM NaCl) and concentrated to between 4 and 6 mg/ml using a Vivaspin™ 6 MWCO 5000 spin column (Cytiva) for crystallisation unless otherwise specified.

Initial high-throughput (HT) crystallisation screening was conducted using a Mosquito® Xtal3 robot (SPT Labtech) with commercially available screens: JCSG-plus™ HT-96 ECO screen[137], Pact premier™ HT-96 ECO screen[137], Structure Screen 1+2 HT-96[135], and Morpheus® HT-96[431] (all Molecular Dimensions). 80 µl crystallisation reagent was added to reservoirs in MRC 96-well sitting drop plates (Jena Biosciences). Screen conditions and protein samples were combined in two ratios for each screen: 1:1 (100 nl : 100 nl) and 2:1 (200 nl : 100 nl) protein : crystallisation reagent. Plates were sealed with ClearVue Sheets (Molecular Dimensions) and drops were observed periodically for crystal formation using a Leica MZ16 Stereomicroscope.

11.16.2. Manual optimisation

Optimised screen conditions based on HT out-screening were prepared from 0.22 µm filtered buffer and precipitant stocks. Manual crystallisation experiments were conducted using 22 mm 24-well sitting drop plates (Hampton) with 500 µl crystallisation reagent in reservoirs, and ratios of 1:1 (1 µl : 1 µl) and 2:1 (2 µl : 1 µl) protein solution to crystallisation reagent. Plates were sealed with Crystal Clear Sealing Tape (Hampton) and drops were observed periodically for crystal formation using a Leica MZ16 Stereomicroscope.

11.16.3. Microseed matrix screening

Seed stock preparation

Seed stock preparation was conducted following an established protocol[275]. Crystals identified for seed stock production were gently crushed using a sterilised glass probe. 10 µl of crystallisation reagent from the reservoir was added to the drop containing crushed crystals and mixed by aspiration. The mixture was added to a microcentrifuge tube with a 3 mm PTFE seed bead cooled on ice. Further crystallisation reagent was added to the drop in 10 µl aliquots and transferred to the seed bead tube for a final volume of 50 µl. The tube was vortexed on ice for 3 min in 30 s bursts with 30 s cooling on ice between bursts. The resulting undiluted seed stock (Dropping Solution 1, DS1) was used to create a dilution series for seeding: 45 µl of crystallisation reagent with 5 µl of DS1 to form DS2. This dilution method was used to create a series of 6 crystal seed dropping solutions (DS1 – 6).

Crystallisation screening

Microseed Matrix Screening (MMS)[275,276] experiments were conducted by vapour diffusion in MRC 96-well sitting drop plates (Jena Biosciences) with a Mosquito® Xtal3 robot (SPT Labtech). The first round of seeding was conducted with DS1, moving to more dilute seed stocks in subsequent crystallisation trials, as necessary. Seed stock was combined with protein solution and crystallisation reagent in drops in the ratio 1:3:2 (100 nl : 300 nl : 200 nl), respectively, with 80 µl crystallisation

reagent added to reservoirs. Plates were sealed with ClearVue Sheets (Molecular Dimensions) and drops were observed periodically for crystal formation using a Leica MZ16 Stereomicroscope.

11.16.4. CPR-DprA crystallisation

CPR-DprA crystals were obtained at 20 °C from the Pact *premier*™ HT-96 ECO and JCSG-*plus*™ HT-96 ECO screens using a protein solution of 5 mg/ml in SGC buffer. Manual optimisation with two 24-condition screens (Tables C1 and C2) was performed as described in 11.16.2. Crystals leading to the structural solution of CPR-DprA were grown in 0.1 M MMT pH 7.5, 30% PEG 1500.

11.16.5. CPR-C4 crystallisation

Crystal form 1

Crystal form 1 (dodecahedrons) was obtained at 20 °C using a protein solution of 4 mg/ml in SGC buffer across 25% (24 of 96 conditions) of the Morpheus® HT-96 screen[431], predominantly in conditions containing a pH 6.5 MES monohydrate buffer, with mixtures of carboxylic acids, amino acids, or salt additives[431]. Crystals were optimised through manual crystallisation experiments with the Morpheus® screen using 22 mm 24-well hanging drop plates (Hampton) with 500 µl crystallisation reagent in reservoirs, and ratios of 1:1 (1 µl : 1 µl) and 2:1 (2 µl : 1 µl) protein solution to crystallisation reagent (Tables D1 and D2). Crystals leading to the structural solution of CPR-C4 were grown in 0.09 M NPS, 0.1 M buffer system 1 pH 6.5, 50% v/v precipitant mix 2 from the Morpheus® screen (Molecular Dimensions).

Crystal form 2

Crystal form 2 (cubic rods) were similarly grown at 20 °C with 4 mg/ml CPR-C4 in SGC buffer, in condition F9 from the JCSG-*plus*™ HT-96 screen[137] (2.4 M sodium malonate dibasic monohydrate pH 7.0). Out-screening was also performed as outlined in 11.16.2 (Table D3).

Crystal form 3

Crystal form 3 (plates) was generated using protein from the pET28a(+)TEV-*CPRC4* construct. Crystals were obtained from a 6 mg/ml protein solution in phosphate buffer (20 mM sodium phosphate pH 7.4, 500 mM NaCl) at 20 °C in 0.2 M lithium sulfate, 0.1 M Tris pH 8.5, 15% w/v PEG 4000.

11.17. Data collection, processing, and structural determination

11.17.1. Cryo-protection and harvesting of protein crystals

Crystals were cryo-protected with 50% v/v glycerol (1 µl glycerol : 1 µl crystallisation reagent) prior to looping with nylon Mounted CryoLoops™ (Hampton) and standard Uni-Puck pins, and flash cooling in liquid nitrogen[432] for storage and transport to Diamond Light Source (DLS).

11.17.2. Structure determination of CPR-DprA

Diffraction data were collected remotely from the I03 beamline at DLS, and manually processed using XDS[258]. The structure of CPR-DprA was solved by molecular replacement (MR) using

Phaser[261] in the CCP4i2 suite[278]. A suitable model (3uqz[254]) was selected using the PDB sequence alignment tool[117,433] and aligned to the CPR-DprA sequence using CLUSTALW[253]. The resulting alignment was used to create a homology model with CHAINSAW[13] for MR. Manual model building was carried out using *Coot*[434] and refinement was performed with REFMAC5[435] using jelly-body restraints for initial rounds of refinement. Ribbon diagrams were created using CCP4mg[372] and arranged using Microsoft Power Point. Dimer interface analysis was performed using PISA[436].

11.17.3. Structure determination of CPR-C4

Data were collected remotely using beamlines I03 (form 1), I24 (form 2), and I04 (form 3) at DLS. For crystal form 1, native diffraction data were initially processed using autoPROC[256,257] with STARANISO[257] *via* the ISPyB pipeline, and the auto-processed data were imported into CCP4i2[278]. Phasing was performed by multiple wavelength anomalous diffraction (MAD) at the Zn edge[142] (1.2824 Å) and native (0.9763 Å) wavelengths, using SHELXC/D/E[272–274] *via* the CRANK2 pipeline[277]. Peak data were subsequently reprocessed in XDS[258]. Rotational non-crystallographic symmetry (rNCS) was identified between two protein chains in the asymmetric unit; this was exploited through density modification with Parrot[279] applying local NCS restraints[280] and a solvent content of 69%, corresponding to two protein molecules in the asymmetric unit. Chain A of the model was built using iterative rounds of manual modifications in *Coot*[434] and REFMAC5[435] refinement cycles and used to fit chain B by molecular replacement with Phaser[261]. The CPR-C4 model was refined using *Coot*[434] and REFMAC5[435] with jelly-body restraints for initial rounds of refinement and using local NCS restraints[280].

Diffraction data for CPR-C4 crystal forms 2 and 3 were processed using the xia2 Dials DUI[281] and XDS[258], respectively. The structures were solved by MR with Phaser[261] using the structure of chain A from crystal form 1 as a homology model. The resulting structures were refined against the density using REFMAC5[435] with jelly-body restraints for initial rounds of refinement and local NCS restraints[280] for crystal form 3, with manual adjustments made in *Coot*[434].

All model building and evaluation was performed with *Coot*[434]. The final models were checked using MolProbity[437]. Coordinates and structure factors have been deposited in the PDB[117,433] with accession codes 7OB6 (form 1), 7OB7 (form 2), and 7PJO (form 3)[60]. Ribbon diagrams were generated using CCP4mg[372]. Protein-protein interfaces were analysed using PISA[436].

11.18. CPR-C4 structural analysis with 3DM

3DM[269] systems for the CPR-C4 protein superfamily were created by J. Lange at BioProdict, Netherlands, using the CPR-C4 amino acid sequence and a .pdb file of the refined crystal form 1 model. VASH1/VASH2 were identified as sharing structural homology with CPR-C4 through structure superpositions. A structure-based multiple sequence alignment (3D-MSA) was used to determine conserved ‘core’ regions between these structures, followed by BLAST[151] searches using the three protein sequences, resulting in a superfamily MSA containing 2441 sequences[60]. A

synchronised numbering system was used to assign all structurally equivalent residues in the 3DM system the same number (3D number) for direct comparison between sequences, linking sequence alignment data to the template structures. Yasara[438] was used to visualise structural alignment outputs from 3DM. Ribbon diagrams were created using CCP4mg[372].

11.19. CPR-C4 protease activity assays

Protease activity assays were conducted using the commercially available Molecular Probes' EnzChek® Protease Assay Kit (Invitrogen) with green fluorescence from BODIPY-FL casein[296]. Purified CPR-C4 was dialysed into 20 mM sodium phosphate pH 7.4, 500 mM NaCl, 10 mM TCEP. Assays were conducted in a 96-well plate format in Corning non-binding surface black fluorescence plates using a total reaction volume of 200 µl per well. 100 µl of BODIPY-FL casein substrate (10 µg/ml in 20 mM sodium phosphate pH 7.4, 500 mM NaCl) was added to 100 µl protein (concentrations ranging from 0 to 0.05 µM) before centrifugation (2 min, 160 x *g*). The plates were incubated at 30 °C and fluorescence was read at hourly intervals using a Synergy HTX plate reader with a fluorescein filter (ex/em = 485 ± 20 nm/528 ± 20 nm; gain 100). Experiments were carried out with controls for background fluorescence taken in triplicate at each protein concentration without the BODIPY-FL casein substrate to account for any intrinsic fluorescent effects of the CPR-C4 protein. 8 technical repeats were conducted at each concentration of CPR-C4. Experimental variability is reported as standard deviation[60].

11.20. CPR-C4 phylogenetics analysis

Separate BLAST[151] searches were run with the CPR-C4 and *Homo sapiens* VASH1/2 sequences using an E value threshold of 0.05. 92 representative protein sequences were selected from the results (Table D5), covering a wide range of taxa and E values, and were aligned using ClustalW[253]: 49 from the CPR-C4 search and 43 from VASH1/2 searches[60]. The alignment was used to generate a phylogenetic tree using the Maximum Likelihood method and Jones-Taylor-Thornton (JTT) matrix-based model in MEGAX[306–308]. 500 replicates were used to calculate bootstrap values and used to establish the strength of the consensus tree. Evolutionary analyses were conducted in MEGAX[306].

11.21. Electron microscopy studies of CPR-ClpB

11.21.1. Negative staining TEM

CPR-ClpB samples at 1 mg/ml were delivered on ice to the Electron Microscopy Research Service at Newcastle University for negative stain electron microscopy analysis. These were diluted 50-fold in MilliQ water for imaging.

11.21.2. Glutaraldehyde crosslinking

Purified CPR-ClpB (1 mg/ml) in SGC buffer with 5 mM $MgCl_2$ was incubated with 2 mM nucleotide (ATP, ADP, ATPγS, or AMP-PNP) for 15 min at RT. Freshly prepared 10% glutaraldehyde solution

was added to the protein-nucleotide sample for final concentration of 0.1% and incubated for 10 min at RT. The crosslinking reaction was quenched with addition of 1 M Tris-HCl pH 7.5 (10 µl for 1 ml protein solution). The sample was then run over a calibrated HiLoad 16/600 Superdex 200 pg column to separate the desired complex.

11.21.3. Single particle cryo-EM experiments

Purified CPR-ClpB samples in SGC buffer were delivered on ice to the Electron Microscopy facilities at the Astbury Centre for Structural Molecular Biology, Leeds, for single particle cryo-EM analysis as part of an Instruct-ERIC grant. Cryo-EM experiments were led by Dr D. Maskell. Two CPR-ClpB samples were provided: 1 mg/ml in SGC buffer, and 0.2 mg/ml with 2 mM ATPγS.

Grids were prepared in duplicate at varying concentrations of CPR-ClpB: 1 mg/ml, 0.5 mg/ml, and 0.25 mg/ml, as well as 0.2 mg/ml with ATPγS. 3 µl of CPR-ClpB sample was applied to Quantifoil™ R 1.2/1.3 on 300 copper mesh that had been plasma cleaned using a Tergeo-EM. Blotting was conducted using a Vitrobot Mark IV with 6/6 blot time/blot force, under controlled 100% humidity at 4 °C. The grids were plunge frozen in liquid ethane cooled by liquid nitrogen (~ −175 °C) and inserted into a Titan Krios transmission electron microscope (Thermo Fisher) operated at an accelerating voltage of 300 kV. Cryo-EM images were recorded with a Falcon 4 detector in counting mode at a nominal magnification of x96,000, with a pixel size of 0.85 Å and a defocus range of –3.3 µm to –2.1 µm, and an accumulated dose of 40.3 $e^{-}/Å^{2}$. EPU software (version 2.11.0.2368REL; Thermo Fisher) was used for automated data acquisition.

11.21.4. Cryo-EM data processing

EM data were processed using the cryoSPARC software[439]. Gain reference files were provided by the Astbury Centre in GAIN format and converted to MRC format using the e2proc2d.py script in EMAN2[440]. Movie files (EER format) were motion corrected followed by Contrast Transfer Function (CTF) estimation. Reference-free particle picking was performed using the cryoSPARC Blob Picker tool with minimum and maximum particle dimensions of 80 Å and 180 Å, respectively. Particles were extracted from CTF-correction micrographs and used to perform 2D class averaging across 200 classes with a circular mask diameter of 360 Å. Selected classes were used in template-based particle picking, followed by particle extraction and a second round of 2D class averaging across 200 classes with a circular mask diameter of 360 Å. Five *ab-initio* models were generated using heterogeneous *ab-initio* reconstruction in cryoSPARC[439]. Models were viewed in UCSF Chimera[379].

Appendices

Appendix A: TSA Durham Screens®

Table A1: Layout of the Durham Salt Screen® (Molecular Dimensions). Concentrations shown are final concentrations used in TSA assays; raw screen conditions are double the concentration shown.

	1	2	3	4	5	6	7	8	9	10	11	12
A	water	water	4 M urea	3.0 M Gu-HCl	1.0 M Gu-HCl	0.8 M Gu-HCl	0.6 M Gu-HCl	0.4 M Gu-HCl	0.2 M Gu-HCl	5 mM Gu-HCl	0.5 M Na_3 citrate	0.2 M Na_3 citrate
B	1.5 M Na_2 malonate	1.0 M Na_2 malonate	0.8 M Na_2 malonate	0.6 M Na_2 malonate	0.4 M Na_2 malonate	0.2 M Na_2 malonate	1.5 M $(NH_4)_2SO_4$	1.0 M $(NH_4)_2SO_4$	0.8 M $(NH_4)_2SO_4$	0.6 M $(NH_4)_2SO_4$	0.4 M $(NH_4)_2SO_4$	0.2 M $(NH_4)_2SO_4$
C	1.5 M NaCl	1.0 M NaCl	0.8 M NaCl	0.6 M NaCl	0.4 M NaCl	0.2 M NaCl	1.5 M NH_4Cl	1.0 M NH_4Cl	0.8 M NH_4Cl	0.6 M NH_4Cl	0.4 M NH_4Cl	0.2 M NH_4Cl
D	1.0 M $MgSO_4$	0.8 M $MgSO_4$	0.6 M $MgSO_4$	0.4 M $MgSO_4$	0.2 M $MgSO_4$	1.0 M Na_2SO_4	0.8 M Na_2SO_4	0.6 M Na_2SO_4	0.4 M Na_2SO_4	0.2 M Na_2SO_4	0.5 M KCl	0.2 M KCl
E	0.5 M LiCl	0.2 M LiCl	0.5 M RbCl	0.2 M RbCl	0.5 M CsCl	0.2 M CsCl	0.4 M NaF	0.1 M NaF	1.5 M NaBr	0.4 M NaBr	0.1 M NaBr	0.4 M NaI
F	0.1 M NaI	0.4 M $MgCl_2$	5 mM $MgCl_2$	5 mM $CaCl_2$	5 mM $SrCl_2$	1 mM $ZnCl_2$	0.1 mM $ZnCl_2$	1 mM $NiCl_2$	0.1 mM $NiCl_2$	5 mM $MnCl_2$	0.5 mM $MnCl_2$	1 mM $CoCl_2$
G	0.1 mM $CoCl_2$	1 mM $CuSO_4$	0.1 mM $CuSO_4$	1 mM $CdSO_4$	5 mM EDTA pH 7.5	5 mM EGTA pH 7.5	2 mM magic triangle pH 7.0	2 mM $La(NO_3)_3$	2 mM $PrCl_3$	2 mM $NdCl_3$	2 mM $SmCl_3$	2 mM $EuCl_3$
H	2 mM $GdCl_3$	2 mM $DyCl_3$	2 mM $HoCl_3$	2 mM $YbCl_3$	2 mM $LuCl_3$	5 mM Na_2HPO_4	5 mM Na_3VO_4	5 mM Na_2WO_4	5 mM Na_2MoO_4	5 mM DTT	5 mM TCEP pH 7.0	5 mM β-mercapto ethanol

EDTA: Ethylenediaminetetraacetic acid; **EGTA**: Ethylene glycol-bis(2-aminoethylether)-N,N,N′,N′-tetraacetic acid; **DTT**: 1,4-Dithiothreitol; **TCEP**: Tris(2-carboxyethyl)phosphine hydrochloride; **I3C**: 5-amino-2,4,6-triiodoisopthalic acid

Table A2: Layout of the Durham pH Screen® (Molecular Dimensions). Concentrations shown are final concentrations used in TSA assays; raw screen conditions are double the concentration shown. Full buffer names are given beneath*.

	1	2	3	4	5	6	7	8	9	10	11	12
A	water	water	4 M urea	100 mM citric acid pH 4.1	100 mM citric acid pH 4.6	100 mM citric acid pH 5.1	100 mM acetic acid pH 4.2	100 mM acetic acid pH 4.7	100 mM acetic acid pH 5.2	100 mM succinic acid pH 4.4	100 mM succinic acid pH 4.9	100 mM succinic acid pH 5.4
B	100 mM malic acid pH 4.3	100 mM malic acid pH 4.8	100 mM malic acid pH 5.3	100 mM tartaric acid pH 4.3	100 mM tartaric acid pH 4.8	100 mM tartaric acid pH 5.3	100 mM propionic acid pH 4.3	100 mM propionic acid pH 4.8	100 mM propionic acid pH 5.3	100 mM malonic acid pH 5.2	100 mM malonic acid pH 5.7	100 mM malonic acid pH 6.2
C	100 mM citric acid pH 5.5	100 mM citric acid pH 6.0	100 mM citric acid pH 6.5	100 mM succinic acid pH 5.6	100 mM succinic acid pH 6.1	100 mM succinic acid pH 6.6	100 mM MES pH 5.6	100 mM MES pH 6.1	100 mM MES pH 6.6	100 mM maleic acid pH 5.7	100 mM maleic acid pH 6.2	100 mM maleic acid pH 6.7
D	100 mM sodium cacodylate pH 5.7	100 mM sodium cacodylate pH 6.2	100 mM sodium cacodylate pH 6.7	100 mM ADA pH 6.1	100 mM ADA pH 6.6	100 mM ADA pH 7.1	100 mM Bis-Tris pH 6.1	100 mM Bis-Tris pH 6.6	100 mM Bis-Tris pH 7.1	100 mM ACES pH 6.3	100 mM ACES pH 6.8	100 mM ACES pH 7.3
E	100 mM phosphate pH 6.3	100 mM phosphate pH 6.8	100 mM phosphate pH 7.3	100 mM PIPES pH 6.3	100 mM PIPES pH 6.8	100 mM PIPES pH 7.3	100 mM imidazole pH 6.6	100 mM imidazole pH 7.1	100 mM imidazole pH 7.6	100 mM MOPS pH 6.6	100 mM MOPS pH 7.1	100 mM MOPS pH 7.6
F	100 mM Bis-Tris propane pH 6.6	100 mM Bis-Tris propane pH 7.1	100 mM Bis-Tris propane pH 7.6	100 mM HEPES pH 7.0	100 mM HEPES pH 7.5	100 mM HEPES pH 8.0	100 mM tricine pH 7.5	100 mM tricine pH 8.0	100 mM tricine pH 8.5	100 mM EPPS pH 7.5	100 mM EPPS pH 8.0	100 mM EPPS pH 8.5
G	100 mM Tris pH 7.7	100 mM Tris pH 8.2	100 mM Tris pH 8.7	100 mM bicine pH 7.7	100 mM bicine pH 8.2	100 mM bicine pH 8.7	100 mM TAPS pH 7.9	100 mM TAPS pH 8.4	100 mM TAPS pH 8.9	100 mM Bis-Tris propane pH 8.5	100 mM Bis-Tris propane pH 9.0	100 mM Bis-Tris propane pH 9.5
H	100 mM boric acid pH 8.6	100 mM boric acid pH 9.1	100 mM boric acid pH 9.6	100 mM CHES pH 8.8	100 mM CHES pH 9.3	100 mM CHES pH 9.8	100 mM glycine pH 9.2	100 mM glycine pH 9.7	100 mM glycine pH 10.2	100 mM CAPS pH 9.9	100 mM CAPS pH 10.4	100 mM CAPS pH 10.9

***MES**: 2-(N-morpholino)ethanesulfonic acid; **ADA**: N-(2-Acetamido)iminodiacetic acid; **PIPES**: 1,4- Piperazinediethanesulfonic acid; **MOPS**: 3-(N-Morpholino)propanesulfonic acid; **TAPS**: N- [Tris(hydroxymethyl)methyl]-3-aminopropanesulfonic acid; **Tris**: Trizma base; **CHES:** 2- (Cyclohexylamino)ethanesulfonic acid; **CAPS**: 3-(Cyclohexylamino)-1-propanesulfonic acid; **EPPS**: 4-(2-Hydroxyethyl)-1-piperazinepropanesulfonic acid; **HEPES**: 4-(2-Hydroxyethyl)piperazine-1- ethanesulfonic acid; **ACES**: N-(2- Acetamido)-2-aminoethanesulfonic acid, N-(Carbamoylmethyl)taurine; **BICINE**: 2-(Bis(2- hydroxyethyl)amino)acetic acid

Table A3: Layout of the Durham Osmolyte Screen® (Molecular Dimensions). Concentrations shown are final concentrations used in TSA assays; raw screen conditions are double the concentration shown.

	1	2	3	4	5	6	7	8	9	10	11	12
A	water	water	4 M urea	30% glycerol	20% glycerol	10% glycerol	160 mM xylitol	20 mM xylitol	1 M D-sorbitol	200 mM D-sorbitol	250 mM D-maltose	25 mM D-maltose
B	500 mM D-sucrose	50 mM D-sucrose	250 mM D-galactose	25 mM D-galactose	500 mM D-glucose	50 mM D-glucose	65 mM D-trehalose	10 mM D-trehalose	1 M L-arabinose	200 mM L-arabinose	135 mM myo-inositol	15 mM myo-inositol
C	70 mM maltitol	10 mM maltitol	250 mM D-mannitol	25 mM D-mannitol	200 mM NDSB-195	200 mM NDSB-201	200 mM NDSB-211	200 mM NDSB-221	100 mM NDSB-256	50 mM pyridine	50 mM pyrimidine	2 mM adenine
D	10 mM dipicolinic acid	2 mM dipicolinic acid	1 mM spermidine	1 mM spermine HCl	250 mM methyl butyrate	20 mM methyl butyrate	250 mM sodium butyrate	20 mM sodium butyrate	100 mM ectoine	10 mM ectoine	100 mM hydroxy ectoine	10 mM hydroxy ectoine
E	750 mM TMAO	100 mM TMAO	100 mM sarcosine	10 mM sarcosine	100 mM carnitine	10 mM carnitine	100 mM hypo-taurine	10 mM hypo-taurine	100 mM taurine	10 mM taurine	100 mM acetyl-choline	10 mM acetyl-choline
F	100 mM choline	10 mM choline	25 mM glyphosate	3 mM glyphosate	100 mM Ala-Ala	10 mM Ala-Ala	100 mM Ala-Gly	10 mM Ala-Gly	100 mM Ala-Leu	10 mM Ala-Leu	100 mM Gly-Gly	10 mM Gly-Gly
G	120 mM L-glutamic acid	15 mM L-glutamic acid	200 mM L-glutamine	50 mM L-glutamine	200 mM L-proline	50 mM L-proline	200 mM 5-oxoproline	50 mM 5-oxoproline	125 mM L-arginine	25 mM L-arginine	200 mM L-lysine HCl	50 mM L-lysine HCl
H	140 mM L-citrulline	30 mM L-citrulline	200 mM glycine	50 mM glycine	200 mM trimethyl glycine	50 mM trimethyl glycine	200 mM L-alanine	50 mM L-alanine	200 mM β-alanine	50 mM β-alanine	60 mM L-histidine	15 mM L-histidine

NDSB: Non-detergent sulfobetains; **TMAO**: Trimethylamine N-oxide

Appendix B: CPR-exo-1

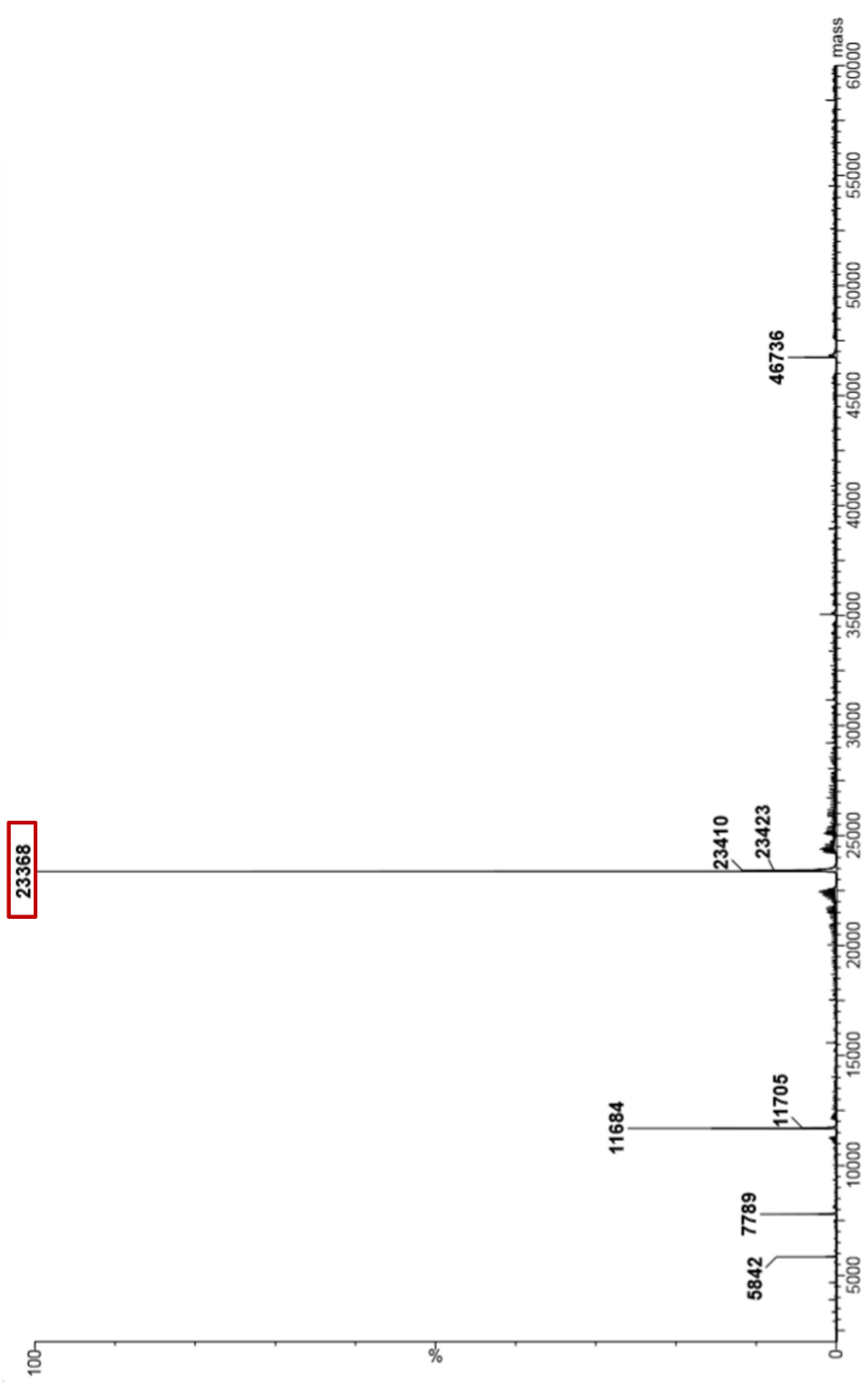

Fig. B1: Service mass spectrum (ESI-TOF) of CPR-exo-1. The molecular ion peak at 23368 Da is boxed in red and is within 3 Da of the expected mass (23371 Da), accounting for N-terminal methionine excision.

Appendix C: CPR-DprA

Table C1: Optimised crystallisation screen (OPT1) for CPR-DprA. The four original conditions yielding crystalline hits in high-throughput trials with the Pact premier™ HT-96 ECO screen are highlighted in red.

	1	**2**	**3**	**4**	**5**	**6**
A	0.1 M PMTP pH 7.5 20% PEG 1500	0.1 M PMTP pH 7.5 25% PEG 1500	0.1 M PMTP pH 7.5 30% PEG 1500	0.1 M MMT pH 7.5 20% PEG 1500	0.1 M MMT pH 7.5 25% PEG 1500	0.1 M MMT pH 7.5 30% PEG 1500
B	0.1 M PMTP pH 8.0 20% PEG 1500	0.1 M PMTP pH 8.0 25% PEG 1500	0.1 M PMTP pH 8.0 30% PEG 1500	0.1 M MMT pH 8.0 20% PEG 1500	0.1 M MMT pH 8.0 25% PEG 1500	0.1 M MMT pH 8.0 30% PEG 1500
C	0.1 M PMTP pH 8.5 20% PEG 1500	0.1 M PMTP pH 8.5 25% PEG 1500	0.1 M PMTP pH 8.5 30% PEG 1500	0.1 M MMT pH 8.5 20% PEG 1500	0.1 M MMT pH 8.5 25% PEG 1500	0.1 M MMT pH 8.5 30% PEG 1500
D	0.1 M PMTP pH 9.0 20% PEG 1500	0.1 M PMTP pH 9.0 25% PEG 1500	0.1 M PMTP pH 9.0 30% PEG 1500	0.1 M MMT pH 9.0 20% PEG 1500	0.1 M MMT pH 9.0 25% PEG 1500	0.1 M MMT pH 9.0 30% PEG 1500

Table C2: Optimised crystallisation screen (OPT2) for CPR-DprA. The original condition yielding crystalline hits during high-throughput trials with the JCSG-plus™ HT-96 ECO screen is highlighted in red.

	1	**2**	**3**	**4**	**5**	**6**
A	0.1 M HEPES pH 7.0 8% PEG 6000	0.1 M HEPES pH 7.0 10% PEG 6000	0.1 M HEPES pH 7.0 12% PEG 6000	0.1 M HEPES pH 7.0 20% PEG 1500	0.1 M HEPES pH 7.0 25% PEG 1500	0.1 M HEPES pH 7.0 30% PEG 1500
B	0.1 M HEPES pH 7.4 8% PEG 6000	0.1 M HEPES pH 7.4 10% PEG 6000	0.1 M HEPES pH 7.4 12% PEG 6000	0.1 M HEPES pH 7.4 20% PEG 1500	0.1 M HEPES pH 7.4 25% PEG 1500	0.1 M HEPES pH 7.4 30% PEG 1500
C	0.1 M HEPES pH 7.8 8% PEG 6000	0.1 M HEPES pH 7.8 10% PEG 6000	0.1 M HEPES pH 7.8 12% PEG 6000	0.1 M HEPES pH 7.8 20% PEG 1500	0.1 M HEPES pH 7.8 25% PEG 1500	0.1 M HEPES pH 7.8 30% PEG 1500
D	0.1 M HEPES pH 8.2 8% PEG 6000	0.1 M HEPES pH 8.2 10% PEG 6000	0.1 M HEPES pH 8.2 12% PEG 6000	0.1 M HEPES pH 8.2 20% PEG 1500	0.1 M HEPES pH 8.2 25% PEG 1500	0.1 M HEPES pH 8.2 30% PEG 1500

Table C3: Data collection, processing, and refinement statistics for the CPR-DprA crystal structure. Values in parentheses are for the outer shell.

Beamline	DLS I03
Wavelength / Å	0.9762
Resolution range / Å	49.56 – 2.10 (2.16 – 2.10)
Space group	$P6_522$
***a, b, c* / Å**	65.07, 65.07, 206.44
α, β, γ / °	90.00, 90.00, 120.00
Number of reflections	577643 (30497)
Unique reflections	16068 (1239)
Multiplicity	35.9 (24.6)
Completeness / %	99.8 (98.0)
$CC_{1/2}$	1.000 (0.879)
⟨I/σ(I)⟩	18.2 (1.4)
R_{meas}	0.124 (2.914)
R_{pim}	0.021 (0.574)
R_{work}	0.199
R_{free}	0.233
No. of atoms	3360
No. of protein residues	220
RMSD bond lengths / Å	0.0084
RMSD bond angles / °	1.535
Ramachandran favoured / allowed / outliers	213 / 4 / 1
Average B factor, overall / $Å^2$	69.6

Table C4: Statistics for the CPR-DprA structural solution from MR with Phaser.

Rotation Function Z-score (RFZ)	5.1
Translation Function Z-score (TFZ)	18.7
Refined TFZ equivalent	20.7
Packing clashes	5
Log Likelihood Gain (LLG)	276
Overall LLG	278

Appendix D: CPR-C4

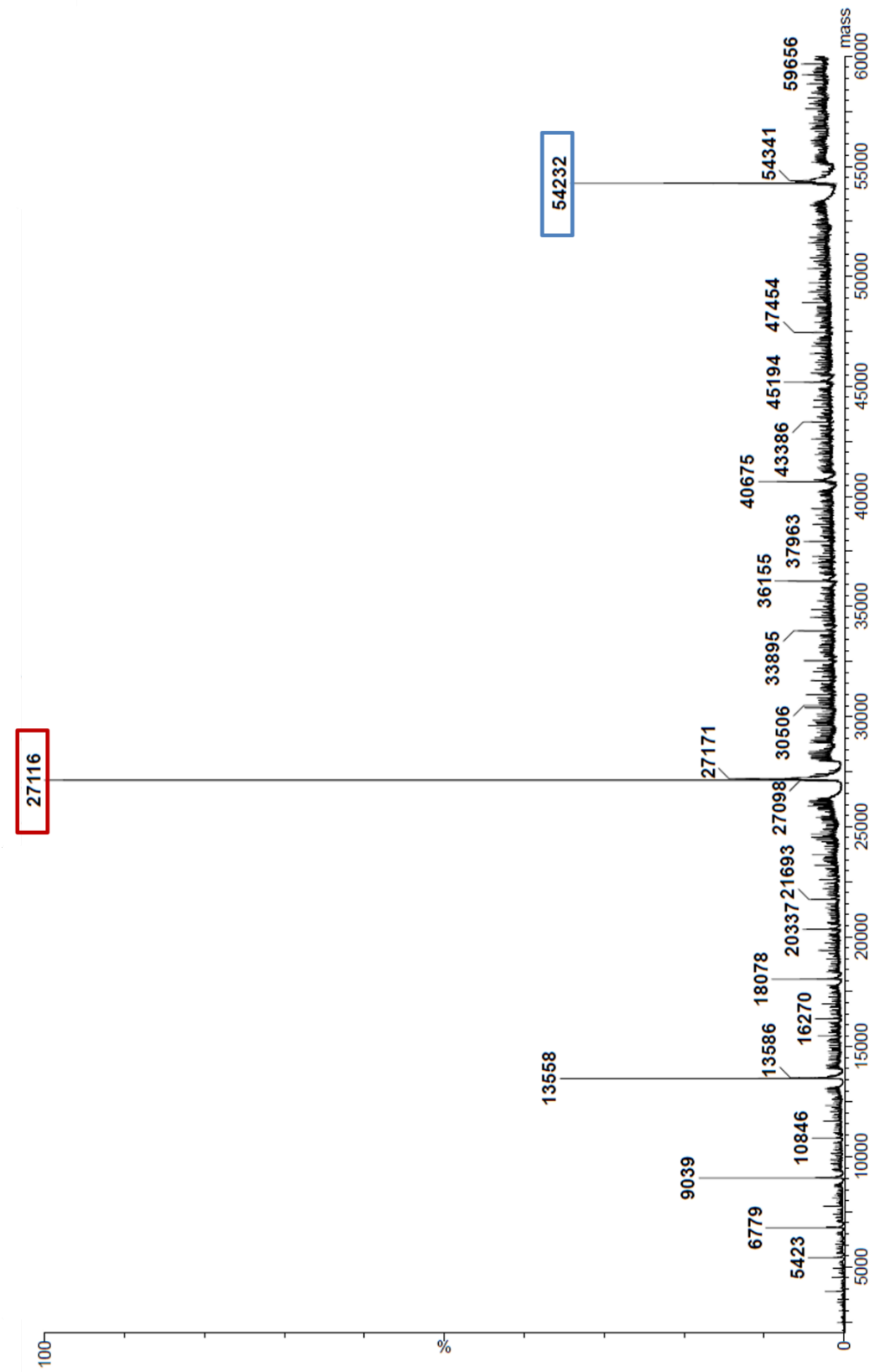

Fig. D1: Service mass spectrum (ESI-TOF) of CPR-C4 from the pJOE5751.1-*CPRC4* construct. The molecular ion peak at 27116 Da is boxed in red and is within 3 Da of the expected mass (27121 Da), accounting for N-terminal methionine excision. The peak at 54232 Da boxed in blue is the CPR-C4 dimer.

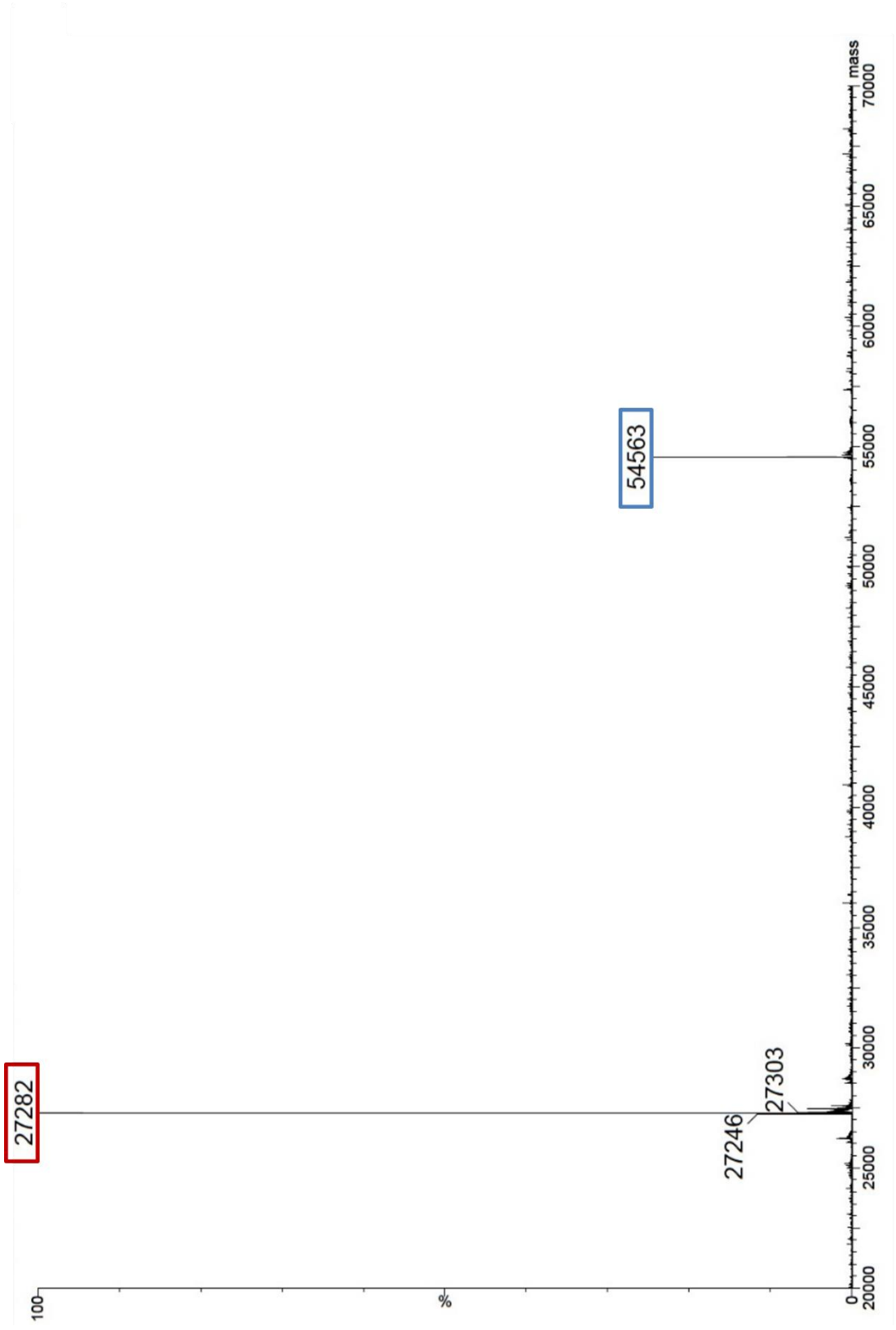

Fig. D2: Service mass spectrum (ESI-TOF) of CPR-C4 from the pET28a(+)TEV-*CPRC4* construct following tag removal with TEV protease. The molecular ion peak at 27282 Da is boxed in red and is identical to the expected mass, accounting for N-terminal methionine excision. The peak at 54563 Da boxed in blue is the CPR-C4 dimer.

Table D1: Conditions from Morpheus® HT-96 used in manual optimisations of CPR-C4 crystal form 1.

Condition	Reagent mixture
C2	0.09 M NPS, 0.1 M buffer system 1 pH 6.5, 50% v/v precipitant mix 2
D2	0.12 M alcohols, 0.1 M buffer system 1 pH 6.5, 50% v/v precipitant mix 2
E2	0.12 M ethylene glycols, 0.1 M buffer system 1 pH 6.5, 50% v/v precipitant mix 2
F2	0.12 M monosaccharides, 0.1 M buffer system 1 pH 6.5, 50% v/v precipitant mix 2
G2	0.1 M carboxylic acids, 0. 1M buffer system 1 pH 6.5, 50% v/v precipitant mix 2
H2	0.1 M amino acids, 0.1 M buffer system 4 pH 6.5, 50% v/v precipitant mix 2
E6	0.12 M ethylene glycols, 0.1 M buffer system 2 pH 7.5, 50% v/v precipitant mix 2
G6	0.1 M carboxylic acids, 0.1 M buffer system 2 pH 7.5, 50% v/v precipitant mix 2
G10	0.1 M carboxylic acids, 0.1 M buffer system 3 pH 8.5, 50% v/v precipitant mix 2
H6	0.1 M amino acids, 0.1 M buffer system 2 pH 7.5, 50% v/v precipitant mix 2
H7	0.1 M amino acids, 0.1 M buffer system 2 pH 7.5, 50% v/v precipitant mix 3
H8	0.1 M amino acids, 0.1 M buffer system 2 pH 7.5, 50% v/v precipitant mix 4

Table D2: Constituent components of the Morpheus® HT-96 screen reagent mixtures used during manual optimisation of CPR-C4 crystal form 1.

Mixture	Additives
Buffer system 1	1 M imidazole, MES monohydrate (acid), pH 6.5
Buffer system 2	1 M sodium HEPES, MOPS (acid), pH 7.5
Buffer system 3	1 M Tris base, BICINE, pH 8.5
Precipitant mix 2	40% ethylene glycol; 20% PEG 8000
Precipitant mix 3	40% glycerol; 20% PEG 4000
Precipitant mix 4	25% MPD; 25% PEG 1000; 25% PEG 3350
NPS	0.3 M sodium nitrate, 0.3 M sodium phosphate dibasic, 0.3 M ammonium sulphate
Alcohols	0.2 M 1,6-hexanediol, 0.2 M 1-butanol, 0.2 M 1,2-propanediol, 0.2 M propanol, 0.2 M 1,4-butanediol, 0.2 M propanediol
Ethylene glycols	0.3 M diethylene glycol, 0.3 M triethylene glycol, 0.3 M tetraethylene glycol, 0.3 M pentaethylene glycol
Monosaccharides	0.2 M D-glucose, 0.2 M D-mannose, 0.2 M D-galactose, 0.2 M L-fucose, 0.2 M D-xylose, 0.2 M N-acetyl-D-glucosamine
Carboxylic acids	0.2 M sodium formate, 0.2 M ammonium acetate, 0.2 M sodium citrate tribasic dihydrate, 0.2 M sodium potassium tartrate tetrahydrate, 0.2 M sodium oxamate
Amino acids	0.2 M L-Na-glutamate, 0.2 M alanine (racemic), 0.2 M glycine, 0.2 M lysine HCl (racemic), 0.2 M serine (racemic)

Table D3: Optimised crystallisation screen for CPR-C4 crystal form 2. The original condition yielding crystalline hits during high-throughput trials with the JCSG-plus™ HT-96 ECO screen is highlighted in red.

	1	**2**	**3**	**4**	**5**	**6**
A	2.0 M sodium malonate pH 6.6	2.0 M sodium malonate pH 6.8	2.0 M sodium malonate pH 7.0	2.0 M sodium malonate pH 7.2	2.0 M sodium malonate pH 7.4	2.0 M sodium malonate pH 7.6
B	2.2 M sodium malonate pH 6.6	2.2 M sodium malonate pH 6.8	2.2 M sodium malonate pH 7.0	2.2 M sodium malonate pH 7.2	2.2 M sodium malonate pH 7.4	2.2 M sodium malonate pH 7.6
C	2.4 M sodium malonate pH 6.6	2.4 M sodium malonate pH 6.8	2.4 M sodium malonate pH 7.0	2.4 M sodium malonate pH 7.2	2.4 M sodium malonate pH 7.4	2.4 M sodium malonate pH 7.6
D	2.6 M sodium malonate pH 6.6	2.6 M sodium malonate pH 6.8	2.6 M sodium malonate pH 7.0	2.6 M sodium malonate pH 7.2	2.6 M sodium malonate pH 7.4	2.6 M sodium malonate pH 7.6

Table D4: Crystallographic data collection and refinement statistics for CPR-C4. Values in parentheses are for the highest resolution shell.

Parameter	Form 1	Form 2	Form 3
Beamline	DLS I03	DLS I24	DLS I04
Wavelength / Å	1.2824	1.2819	0.979
Resolution range / Å	48.26 – 2.60 (2.71 – 2.60)	67.40 – 2.68 (2.81 – 2.68)	44.26 – 2.25 (2.32 – 2.25)
Space group	$P3_221$	I222	$P2_12_12_1$
***a, b, c* / Å**	123.39, 123.39, 96.53	56.31, 79.73, 126.20	44.78, 77.55, 161.72
α, β, γ / °	90, 90, 120	90, 90, 90	90, 90, 90
Number of reflections	224888 (26437)	79870 (8047)	337006 (16676)
Unique reflections	26494 (3143)	8223 (995)	27119 (2058)
Multiplicity	8.5 (8.4)	9.7 (8.1)	12.4 (8.1)
Completeness / %	99.8 (99.2)	98.8 (92.7)	98.1 (85.3)
$CC_{1/2}$	1.000 (0.368)	0.972 (0.055)	0.999 (0.360)
$\langle I/\sigma(I) \rangle$	14.4 (0.7)	7.4 (1.4)	17.1 (1.1)
R_{merge}	0.094 (4.086)	0.453 (6.072)	0.085 (1.770)
R_{pim}	0.050 (2.198)	0.153 (2.225)	0.025 (0.666)
R_{work}	0.234	0.238	0.179
R_{free}	0.285	0.259	0.225
No. of atoms	3412	1671	7223
No. of protein residues	435	211	433
RMSD bond lengths / Å	0.0060	0.0061	0.0076
RMSD bond angles / °	1.418	1.406	1.442
Ramachandran favoured / allowed / outliers	409 / 18 / 0	197 / 9 / 1	407 / 19 / 0
Average B factor, overall / $Å^2$	109.0	51.0	61.0
PDB accession codes	7OB6	7OB7	7PJO

Table D5: NCBI accession codes for the protein sequences used in CPR-C4 phylogenetic tree construction. Taxonomic classification is at the species level where known[60].

#	Accession code	Taxonomy	#	Accession code	Taxonomy	#	Accession code	Taxonomy
1	NP 796328.2	*Mus musculus*	34	XP 024530086.1	*Selaginella moellendorffii*	67	MBI2004736.1	*Patescibacteria group bacterium*
2	XP 005084063.1	*Mesocricetus auratus*	35	OAE33039.1	*Marchantia polymorpha*	68	MBP7770728.1	*Candidatus Pacebacteria bacterium*
3	XP 003808906.1	*Pan paniscus*	36	KAG6549164.1	*Marchantia paleacea*	69	MBC7836912.1	*Acetobacteraceae bacterium*
4	sp|Q7L8A9	*Homo sapiens*	37	XP 024370159.1	*Physcomitrium patens*	70	MBI4093848.1	*Candidatus Kaiserbacteria bacterium*
5	XP 036169856.1	*Myotis myotis*	38	KAG0629661.1	*Ceratodon purpureus*	71	OGG53456.1	*Candidatus Kaiserbacteria bacterium*
6	XP 003987908.1	*Felis catus*	39	PIK60116.1	*Apostichopus japonicus*	72	MBI3702436.1	*Hyphomicrobiales bacterium*
7	XP 006052929.1	*Bubalus bubalis*	40	KAG5702141.1	*Batillaria attramentaria*	73	MBS0533742.1	*Proteobacteria bacterium*
8	XP 004453621.1	*Dasypus novemcinctus*	41	TRY54892.1	*Danionella translucida*	74	MBL8158630.1	*Bacterium*
9	XP 013159882.1	*Falco peregrinus*	42	MBP1660180.1	*Candidatus Aminicenantes bacterium*	75	MBS4050093.1	*Methylomonas*
10	XP 006274820.2	*Alligator mississippiensis*	43	AJF62464.1	*Archaeon*	76	CAE8679133.1	*Polarella glacialis*
11	TKS86491.1	*Collichthys lucidus*	44	MBI4414812.1	*Candidatus Kerfeldbacteria bacterium*	77	CAE8638246.1	*Polarella glacialis*
12	XP 007910324.1	*Callorhinchus milii*	45	OGS51393.1	*Euryarchaeota archaeon*	78	GAQ84263.1	*Klebsormidium nitens*
13	XP 021467687.1	*Oncorhynchus mykiss*	46	TLZ64862.1	*Euryarchaeota archaeon*	79	OGU65384.1	*Ignavibacteria bacterium*
14	XP 005753174.2	*Pundamilia nyererei*	47	HEK51315.1	*Firmicutes bacterium*	80	HBH84536.1	*Bacteroidales bacterium*
15	XP 009277863.1	*Aptenodytes forsteri*	48	HEQ02367.1	*Firmicutes bacterium*	81	MBN1539844.1	*Candidatus Thermoplasmatota archaeon*
16	XP 014732398.1	*Sturnus vulgaris*	49	GDY17089.1	*Verrucomicrobia bacterium*	82	MBI1999858.1	*Candidate division NC10 bacterium*
17	XP 015283097.1	*Gekko japonicus*	50	NCZ71312.1	*Actinobacteria bacterium*	83	MBI2000069.1	*Candidate division NC10 bacterium*
18	KAG6928815.1	*Chelydra serpentina*	51	MBM3582227.1	*Alphaproteobacteria bacterium*	84	PYN57126.1	*Candidatus Rokubacteria bacterium*
19	sp|Q86V25	*Homo sapiens*	52	TMH00807.1	*Betaproteobacteria bacterium*	85	PYM30278.1	*Candidatus Rokubacteria bacterium*
20	XP 005540865.1	*Macaca fascicularis*	53	MBT6460655.1	*Planctomycetaceae bacterium*	86	MBN2498250.1	*Deltaproteobacteria bacterium*
21	XP 010851343.1	*Bison bison*	54	HIH23410.1	*Candidatus Woesearchaeota archaeon*	87	WP 116225417.1	*Pelolinea submarina*
22	NP 001073079.1	*Bos taurus*	55	OGY49875.1	*Candidatus Buchananbacteria bacterium*	88	MBM3152234.1	*Chloroflexi bacterium*
23	XP 027812095.1	*Ovis aries*	56	MBU0958765.1	*Nanoarchaeota archaeon*	89	MBK7703915.1	*Bacterium*
24	PVD25902.1	*Pomacea canaliculata*	57	MBM3303958.1	*Candidatus Aenigmarchaeota archaeon*	90	MBN1995914.1	*Candidate division KSB1 bacterium*
25	CAG2227822.1	*Mytilus edulis*	58	MBI4152238.1	*Candidatus Woesearchaeota archaeon*	91	MBC7257623.1	*Chloroflexi bacterium*
26	XP 002109112.1	*Trichoplax adhaerens*	59	KAA0206149.1	*Candidatus Uhrbacteria bacterium*	92	MBI2044722.1	*Candidatus Pacearchaeota archaeon*
27	XP 032237246.1	*Nematostella vectensis*	60	XP 005825204.1	*Guillardia theta*			
28	XP 023340823.1	*Eurytemora affinis*	61	HCY17607.1	*Candidatus Nomurabacteria bacterium*			
29	KNE57617.1	*Allomyces macrogynus*	62	OGN19854.1	*Candidatus Yanofskybacteria bacterium*			
30	ORY40458.1	*Rhizoclosmatium globosum*	63	PIP56008.1	*Candidatus Zambryskibacteria bacterium*			
31	OON07946.1	*Batrachochytrium salamandrivorans*	64	MBI2038209.1	*Candidatus Magasanikbacteria bacterium*			
32	EPZ36457.1	*Rozella allomycis*	65	(CPR-C4)	*CPR-C4 bacterium*			
33	KAG4090764.1	*Neocallimastix*	66	KKW35909.1	*Candidatus Adlerbacteria bacterium*			

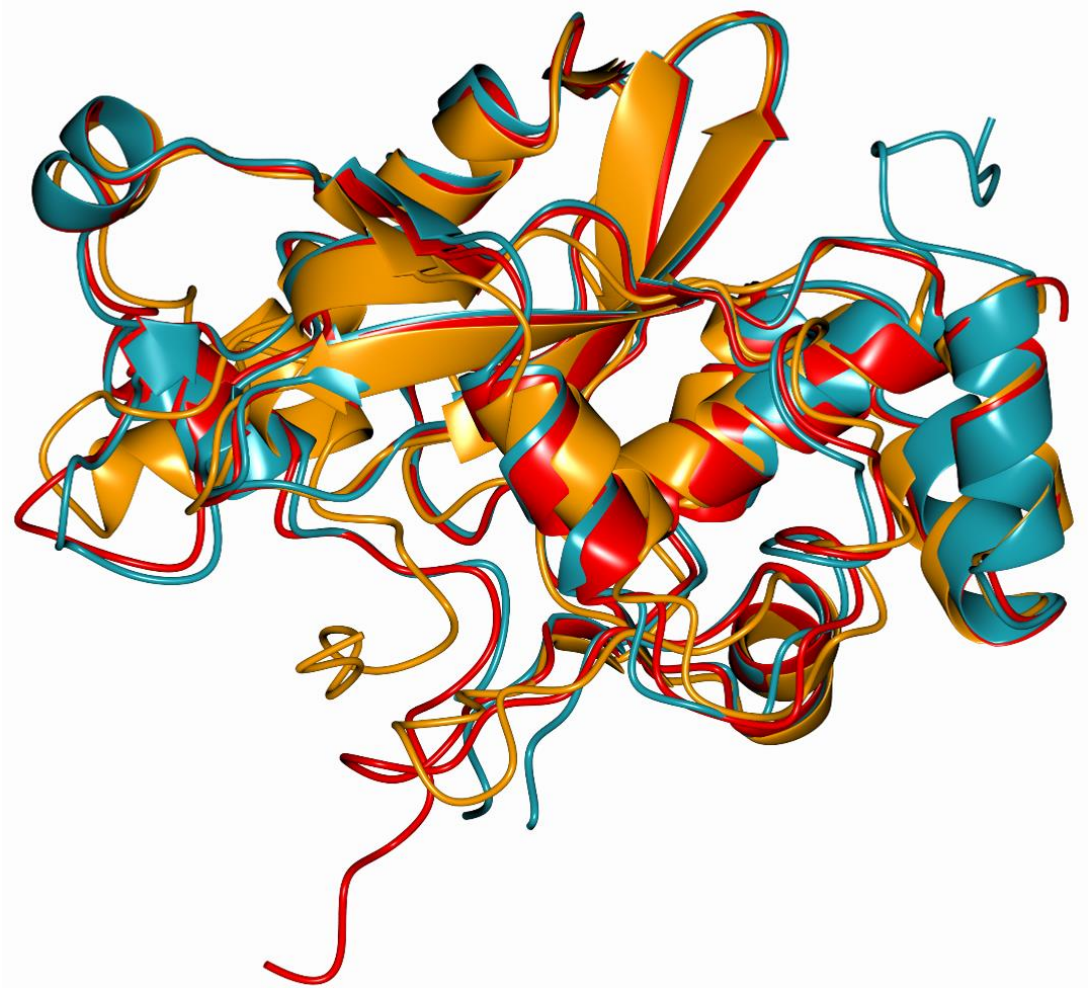

Fig. D3: Ribbon diagram showing the alignment of the experimentally determined CPR-C4 structure (form 1 chain A, teal) with models generated using AlphaFold2[118,263] (red, RMSD 0.84 Å) and RoseTTAFold[309] (orange, RMSD 1.95 Å) through least-squares superposition[60].

Appendix E: CPR-GrpE

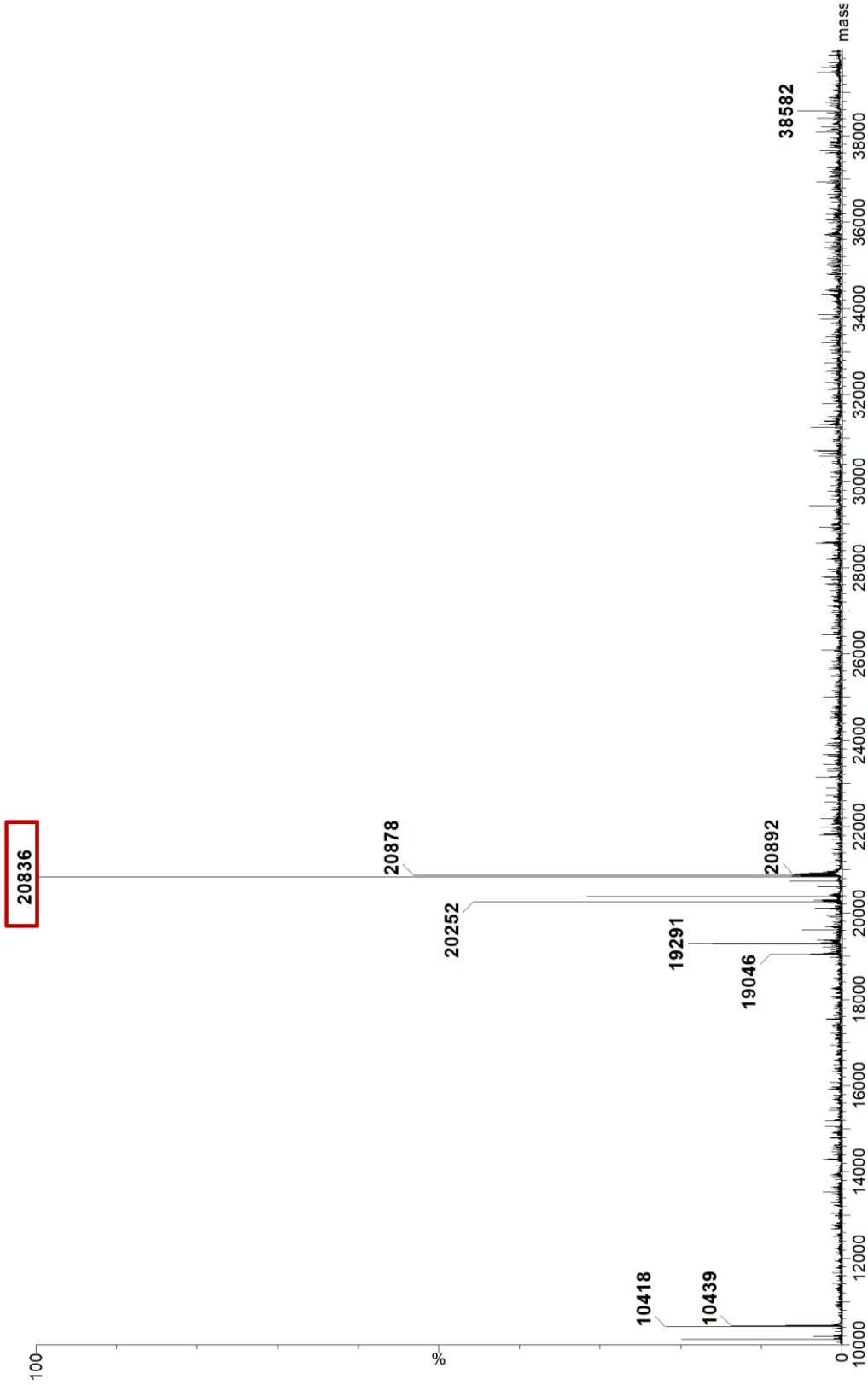

Fig. E1: Service mass spectrum (ESI-TOF) of CPR-GrpE. The molecular ion peak at 20836 Da is boxed in red and is within 3 Da of the expected mass (20839 Da), accounting for N-terminal methionine excision.

Table E1: Optimised crystallisation screen for CPR-GrpE. The original condition yielding crystalline hits during high-throughput trials with the Structure 1 + 2 screen is highlighted in red.

	1	**2**	**3**	**4**	**5**	**6**
A	1.4 M amm. phosphate* 0.1 M Tris pH 7.5	1.6 M amm. phosphate 0.1 M Tris pH 7.5	1.8 M amm. phosphate 0.1 M Tris pH 7.5	2.0 M amm. phosphate 0.1 M Tris pH 7.5	2.2 M amm. phosphate 0.1 M Tris pH 7.5	2.4 M amm. phosphate 0.1 M Tris pH 7.5
B	1.4 M amm. phosphate 0.1 M Tris pH 8.0	1.6 M amm. phosphate 0.1 M Tris pH 8.0	1.8 M amm. phosphate 0.1 M Tris pH 8.0	2.0 M amm. phosphate 0.1 M Tris pH 8.0	2.2 M amm. phosphate 0.1 M Tris pH 8.0	2.4 M amm. phosphate 0.1 M Tris pH 8.0
C	1.4 M amm. phosphate 0.1 M Tris pH 8.5	1.6 M amm. phosphate 0.1 M Tris pH 8.5	1.8 M amm. phosphate 0.1 M Tris pH 8.5	2.0 M amm. phosphate 0.1 M Tris pH 8.5	2.2 M amm. phosphate 0.1 M Tris pH 8.5	2.4 M amm. phosphate 0.1 M Tris pH 8.5
D	1.4 M amm. phosphate 0.1 M Tris pH 9.0	1.6 M amm. phosphate 0.1 M Tris pH 9.0	1.8 M amm. phosphate 0.1 M Tris pH 9.0	2.0 M amm. phosphate 0.1 M Tris pH 9.0	2.2 M amm. phosphate 0.1 M Tris pH 9.0	2.4 M amm. phosphate 0.1 M Tris pH 9.0

* ammonium phosphate monobasic

Appendix F: GX targets

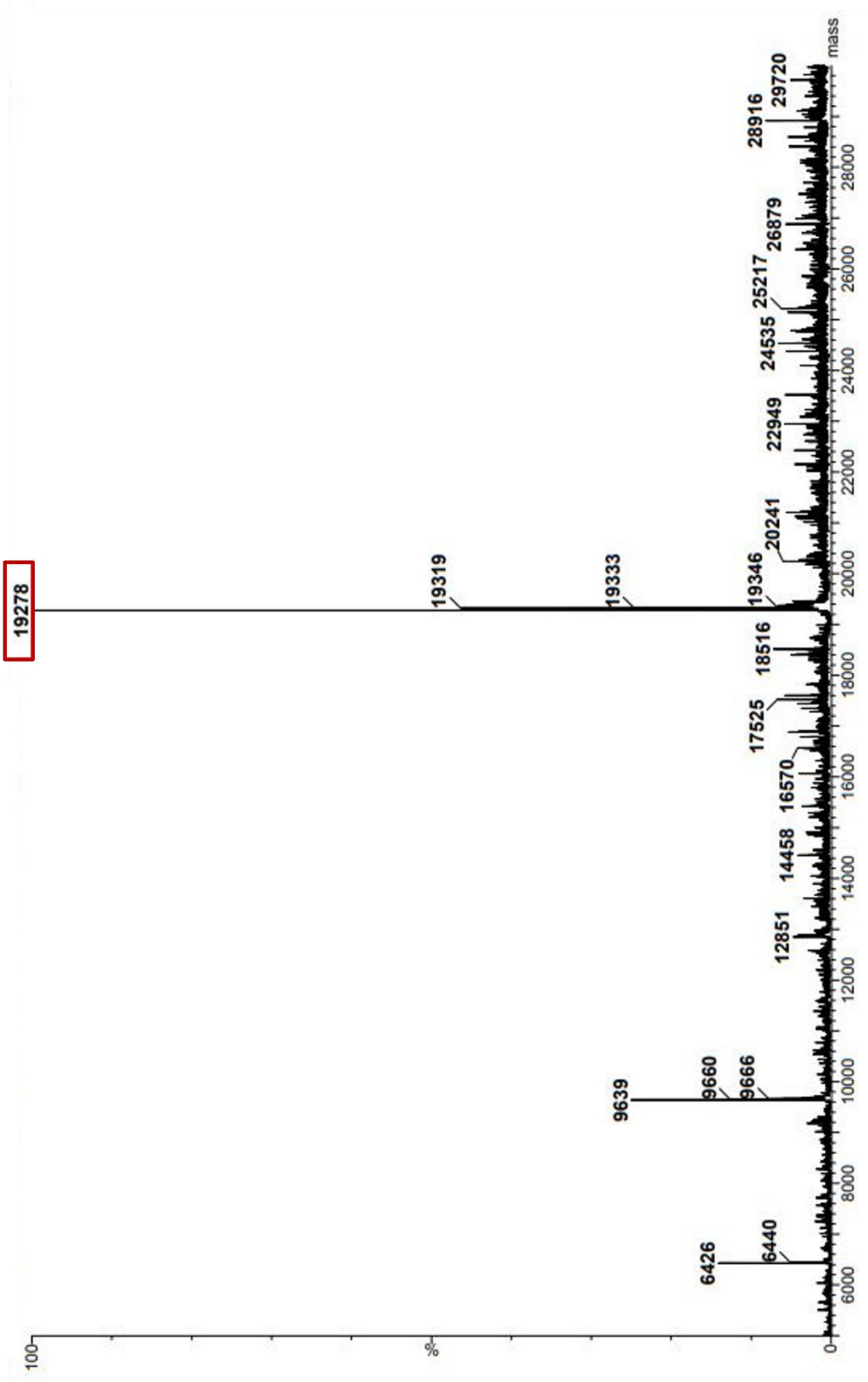

Fig. F1: Service mass spectrum (ESI-TOF) of GX-hyp-6. The molecular ion peak at 19278 Da is boxed in red and is within 1 Da of the expected mass (19277 Da), accounting for N-terminal methionine excision.

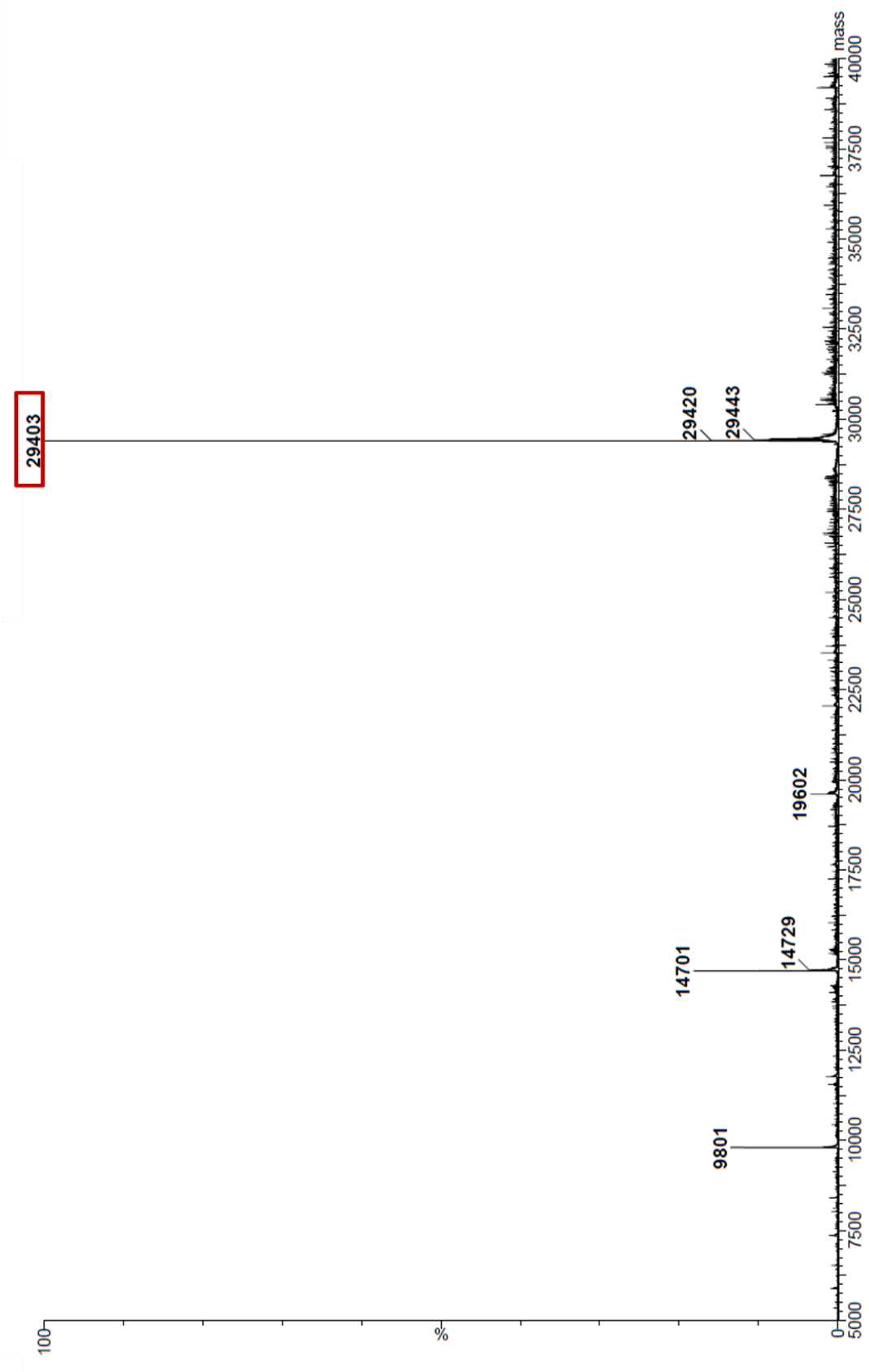

Fig. F2: Service mass spectrum (ESI-TOF) of GX-hyp-16. The molecular ion peak at 29403 Da is boxed in red and is within 10 Da of the expected mass (29413 Da), accounting for N-terminal methionine excision.

Bibliography

1. Sittenfeld, A. & Gamez, R. Biodiversity prospecting. *World Resources Institute* 69–97 (1993).

2. Krüger, A., Schäfers, C., Schröder, C. & Antranikian, G. Towards a sustainable biobased industry – Highlighting the impact of extremophiles. *New Biotechnology* **40**, 144–153 (2018).

3. Davids, T., Schmidt, M., Böttcher, D. & Bornscheuer, U. T. Strategies for the discovery and engineering of enzymes for biocatalysis. *Current Opinion in Chemical Biology* **17**, 215–220 (2013).

4. Adams, M. W. W., Perler, F. B. & Kelly, R. M. Extremozymes: Expanding the Limits of Biocatalysis. *Nature Biotechnology* **13**, 662–668 (1995).

5. Schiraldi, C. & de Rosa, M. The production of biocatalysts and biomolecules from extremophiles. *Trends in biotechnology* **20**, 515–21 (2002).

6. Sysoev, M. *et al.* Bioprospecting of novel extremozymes from Prokaryotes—The advent of culture-independent methods. *Frontiers in Microbiology* **12**, (2021).

7. Raddadi, N., Cherif, A., Daffonchio, D., Neifar, M. & Fava, F. Biotechnological applications of extremophiles, extremozymes and extremolytes. *Applied Microbiology and Biotechnology* **99**, 7907–7913 (2015).

8. Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & van Elsas, J. D. The great screen anomaly-a new frontier in product discovery through functional metagenomics. *Applied Microbiology and Biotechnology* **93**, 1005–1020 (2012).

9. Rosario, K. & Breitbart, M. Exploring the viral world through metagenomics. *Current opinion in virology* **1**, 289–297 (2011).

10. Steele, H. L., Jaeger, K. E., Daniel, R. & Streit, W. R. Advances in Recovery of Novel Biocatalysts from Metagenomes. *Microbial Physiology* **16**, 25–37 (2009).

11. Rohwer, F. Global phage diversity. *Cell* **113**, 141 (2003).

12. Delong, E. F. & Pace, N. R. Environmental Diversity of Bacteria and Archaea. *Syst. Biol* **50**, 470–478 (2001).

13. Kantor, R. S. *et al.* Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* **4**, (2013).

14. Warnecke, F. & Hess, M. A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts. *Journal of Biotechnology* **142**, 91–95 (2009).

15. Cowan, D. A., Ramond, J. B., Makhalanyane, T. P. & de Maayer, P. Metagenomics of extreme environments. *Current Opinion in Microbiology* **25**, 97–102 (2015).

16. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).

17. Wilmes, P., Heintz-Buschart, A. & Bond, P. L. A decade of metaproteomics: Where we stand and what the future holds. *Proteomics* **15**, 3409–3417 (2015).

18. Aevarsson, A. *et al.* Going to extremes – a metagenomic journey into the dark matter of life. *FEMS Microbiology Letters* **368**, (2021).

19. Aevarsson, A. *Viruses at the source*. www.esof.eu (2018).

20. Sandaa, R. A. *et al.* Seasonality drives microbial community structure, shaping both Eukaryotic and Prokaryotic host-viral relationships in an arctic marine ecosystem. *Viruses* **10**, (2018).

21. Liu, Y. *et al.* New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. *Environmental microbiology* **21**, 2002–2014 (2019).

22. Liu, Y. *et al.* A novel type of polyhedral viruses infecting hyperthermophilic Archaea. *Journal of Virology* **91**, (2017).

23. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* **16**, 603–606 (2019).

24. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).

25. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biology* **3**, 1–8 (2002).

26. Amann, R. I., Ludwig, W. & Schleifer, K.-H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143 (1995).

27. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science* **5**, 209 (2014).

28. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).

29. Hugenholtz, P. & Pace, N. R. Identifying microbial diversity in the natural environment: A molecular phylogenetic approach. *Trends in Biotechnology* **14**, 190–197 (1996).

30. Harris, J. K., Kelley, S. T. & Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. *Applied and environmental microbiology* **70**, 845–9 (2004).

31. Rosselli, R. *et al.* Direct 16S rRNA-seq from bacterial communities: a PCR-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. *Scientific Reports* **6**, 1–12 (2016).

32. Louca, S., Doebeli, M. & Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**, 1–12 (2018).

33. Bansal, A. K. & Meyer, T. E. Evolutionary analysis by whole-genome comparisons. *Journal of bacteriology* **184**, 2260–2272 (2002).

34. Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. The analysis of natural microbial populations by ribosomal RNA sequences. in *Advances in Microbial Ecology* 1–55 (Springer, Boston, MA, 1986). doi:10.1007/978-1-4757-0611-6_1.

35. de León, K. B., Gerlach, R., Peyton, B. M. & Fields, M. W. Archaeal and bacterial communities in three alkaline hot springs in heart lake geyser basin, Yellowstone National Park. *Frontiers in Microbiology* **4**, 330 (2013).

36. Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. Novel division level bacterial diversity in a Yellowstone Hot Spring. *Journal of Bacteriology* **180**, 366 (1998).

37. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences* **82**, 6955 (1985).

38. Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology* **173**, 4371 (1991).

39. Amann, R. I., Ludwig, W. & Schleifer, K.-H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143 (1995).

40. Murugkar, P. P., Collins, A. J., Chen, T. & Dewhirst, F. E. Isolation and cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in coculture with bacterial hosts. *Journal of Oral Microbiology* **12**, (2020).

41. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

42. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–1665 (2012).

43. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

44. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

45. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920 (2005).

46. DNA Sequencing Costs: Data. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

47. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533–1542 (2017).

48. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annual Review of Microbiology* **57**, 369–394 (2003).

49. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* **31**, 533–538 (2013).

50. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature 2016 536:7617* **536**, 425–430 (2016).

51. Wrighton, K. C. *et al.* Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *The ISME Journal* **8**, 1452–1463 (2014).

52. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3**, 1–12 (2015).

53. Raghoebarsing, A. A. *et al.* A microbial consortium couples anaerobic methane oxidation to denitrification. *Nature* **440**, 918–921 (2006).

54. Davis, N. M., Proctor, Di. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 1–14 (2018).

55. Garza, D. R. & Dutilh, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cellular and molecular life sciences* **72**, 4287–308 (2015).

56. Simmons, S. L. *et al.* Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLOS Biology* **6**, e177 (2008).

57. Zhang, L. *et al.* Advances in metagenomics and its application in environmental microorganisms. *Frontiers in Microbiology* **12**, 3847 (2021).

58. Turaev, D. & Rattei, T. High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. *Current Opinion in Biotechnology* **39**, 174–181 (2016).

59. Sczyrba, A. *et al.* Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods* **14**, 1063–1071 (2017).

60. Cornish, K. A. S., Lange, J., Aevarsson, A. & Pohl, E. CPR-C4 is a highly conserved novel protease from the Candidate Phyla Radiation with remote structural homology to human vasohibins. *Journal of Biological Chemistry* **298**, 101919 (2022).

61. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences* **104**, 11889 (2007).

62. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

63. Rheims, H., Rainey, F. A. & Stackebrandt, E. A molecular approach to search for diversity among bacteria in the environment. *Journal of Industrial Microbiology* **17**, 159–169 (1996).

64. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 16048 (2016).

65. Attar, N. CPR breathes new air into the tree of life. *Nature Reviews Microbiology* **14**, 332–332 (2016).

66. Schulz, F. *et al.* Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017).

67. Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S. & Banfield, J. F. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology* **18**, 69 (2020).

68. Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews Microbiology* **16**, 629–645 (2018).

69. Hanke, A. *et al.* Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Frontiers in Microbiology* **5**, (2014).

70. Yutin, N., Puigbò, P., Koonin, E. v. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLOS ONE* **7**, e36972 (2012).

71. Lecompte, O., Ripp, R., Thierry, J., Moras, D. & Poch, O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Research* **30**, 5382–5390 (2002).

72. Gong, J., Qing, Y., Guo, X. & Warren, A. "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). *Systematic and Applied Microbiology* **37**, 35–41 (2014).

73. Batinovic, S., Rose, J. J. A., Ratcliffe, J., Seviour, R. J. & Petrovski, S. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. *Nature Microbiology* **6**, 703–711 (2021).

74. He, X. *et al.* Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proceedings of the National Academy of Sciences* **112**, 244–249 (2015).

75. Cross, K. L. *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nature biotechnology* **37**, 1314 (2019).

76. Baker, B. J. *et al.* Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences* **107**, 8806–8811 (2010).

77. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature Communications* **10**, 4173 (2019).

78. Bor, B. *et al.* Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proceedings of the National Academy of Sciences* **115**, 12277–12282 (2018).

79. Utter, D. R., He, X., Cavanaugh, C. M., McLean, J. S. & Bor, B. The saccharibacterium TM7x elicits differential responses across its host range. *The ISME Journal* **14**, 3054 (2020).

80. Hugenholtz, P., Tyson, G. W., Webb, R. I., Wagner, A. M. & Blackall, L. L. Investigation of Candidate Division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. *Applied and Environmental Microbiology* **67**, 411 (2001).

81. Borneman, J. & Triplett, E. W. Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. *Applied and Environmental Microbiology* **63**, 2647 (1997).

82. Podar, M. *et al.* Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology* **73**, 3205–3214 (2007).

83. Chen, L. X. *et al.* Candidate phyla radiation roizmanbacteria from hot springs have novel and unexpectedly abundant CRISPR-cas systems. *Frontiers in Microbiology* **10**, (2019).

84. Dewhirst, F. E. *et al.* The human oral microbiome. *Journal of bacteriology* **192**, 5002–17 (2010).

85. Merino, N. *et al.* Living at the extremes: Extremophiles and the limits of life in a planetary context. *Frontiers in Microbiology* **10**, 780 (2019).

86. van den Burg, B. Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology* **6**, 213–218 (2003).

87. Niehaus, F., Bertoldo, C., Kähler, M. & Antranikian, G. Extremophiles as a source of novel enzymes for industrial application. *Applied Microbiology and Biotechnology* **51**, 711–729 (1999).

88. Dalmaso, G. Z. L., Ferreira, D. & Vermelho, A. B. Marine extremophiles: a source of hydrolases for biotechnological applications. *Marine Drugs* **13**, 1925 (2015).

89. Amaral-Zettler, L. A. Eukaryotic diversity at pH extremes. *Frontiers in Microbiology* **3**, 441 (2012).

90. Deming, J. W. Life in ice formations at very cold temperatures. in *Physiology and Biochemistry of Extremophiles* 133–144 (John Wiley & Sons, Ltd, 2014). doi:10.1128/9781555815813.CH10.

91. Takai, K. *et al.* Cell proliferation at 122°C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proceedings of the National Academy of Sciences* **105**, 10949 (2008).

92. Kashefi, K. & Lovley, D. R. Extending the upper temperature limit for life. *Science* **301**, 934 (2003).

93. Zaparty, M. & Siebers, B. Physiology, metabolism, and enzymology of thermoacidophiles. in *Extremophiles Handbook* 601–639 (Springer Japan, 2011). doi:10.1007/978-4-431-53898-1_28.

94. Schleper, C., Pühler, G., Klenk, H. P. & Zillig, W. Picrophilus oshimae and Picrophilus torridus, two species of hyperacidophilic, thermophilic, heterotrophic, aerobic archaea. *International Journal of Systematic Bacteriology* **46**, 814–816 (1996).

95. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems* **3**, (2018).

96. Suzuki, S. *et al.* Physiological and genomic features of highly alkaliphilic hydrogen-utilizing Betaproteobacteria from a continental serpentinizing site. *Nature Communications* **5**, (2014).

97. Krulwich, T. A., Sachs, G. & Padan, E. Molecular aspects of bacterial pH sensing and homeostasis. *Nature Reviews Microbiology* **9**, 330–343 (2011).

98. Jin, Q. & Kirk, M. F. pH as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science* **6**, 21 (2018).

99. Lane, N. & Martin, W. F. The origin of membrane bioenergetics. *Cell* **151**, 1406–1416 (2012).

100. Zhang, L. *et al.* Ratiometric fluorescent pH-sensitive polymers for high-throughput monitoring of extracellular pH. *RSC Advances* **6**, 46134–46142 (2016).

101. Littlechild, J. A. Enzymes from Extreme Environments and Their Industrial Applications. *Frontiers in Bioengineering and Biotechnology* **3**, (2015).

102. Frösler, J., Panitz, C., Wingender, J., Flemming, H. C. & Rettberg, P. Survival of Deinococcus geothermalis in biofilms under desiccation and simulated space and martian conditions. *Astrobiology* **17**, 431–447 (2017).

103. Kashefi, K. & Lovley, D. R. Extending the upper temperature limit for life. *Science* **301**, 934 (2003).

104. Rivkina, E. M., Friedmann, E. I., McKay, C. P. & Gilichinsky, D. A. Metabolic activity of Permafrost Bacteria below the freezing point. *Applied and Environmental Microbiology* **66**, 3230–3233 (2000).

105. Clarke, A. The thermal limits to life on Earth. *International Journal of Astrobiology* **14**, 141–154 (2014).

106. Elleuche, S., Schrö, C., Sahm, K. & Antranikian, G. Extremozymes-biocatalysts with unique properties from extremophilic microorganisms. *Current Opinion in Biotechnology* **29**, 116–123 (2014).

107. Kashefi, K., Holmes, D. E., Reysenbach, A. L. & Lovley, D. R. Use of Fe(III) as an electron acceptor to recover previously uncultured hyperthermophiles: Isolation and characterization of Geothermobacterium ferrireducens. *Applied and Environmental Microbiology* **68**, 1735–1742 (2002).

108. Gerday, C. *et al.* Cold-adapted enzymes: from fundamentals to biotechnology. *Trends in Biotechnology* **18**, 103–107 (2000).

109. Vieille, C. & Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and Molecular Biology Reviews* **65**, 1–43 (2001).

110. Farrell, J. & Campbell, L. L. Thermophilic bacteria and bacteriophages. *Advances in Microbial Physiology* **3**, 83–109 (1969).

111. Feller, G. & Gerday, C. Psychrophilic enzymes: hot topics in cold adaptation. *Nature Reviews Microbiology* **1**, 200–208 (2003).

112. Cavicchioli, R., Siddiqui, K. S., Andrews, D. & Sowers, K. R. Low-temperature extremophiles and their applications. *Current Opinion in Biotechnology* **13**, 253–261 (2002).

113. Woodley, J. M. Protein engineering of enzymes for process applications. *Current Opinion in Chemical Biology* **17**, 310–316 (2013).

114. Demirjian, D. C., Morís-Varas, F. & Cassidy, C. S. Enzymes from extremophiles. *Current opinion in chemical biology* **5**, 144–51 (2001).

115. Toogood, H. S. *et al.* A thermostable L-aminoacylase from Thermococcus litoralis: cloning, overexpression, characterization, and applications in biotransformations. *Extremophiles* **6**, 111–122 (2002).

116. Orengo, C. A., Todd, A. E. & Thornton, J. M. From protein structure to function. *Current Opinion in Structural Biology* **9**, 374–382 (1999).

117. Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47**, D520–D528 (2019).

118. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

119. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology 2007 8:12* **8**, 995–1005 (2007).

120. Sivashankari, S. & Shanmughavel, P. Functional annotation of hypothetical proteins - A review. *Bioinformation* **1**, 335–8 (2006).

121. Bork, P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome research* **10**, 398–400 (2000).

122. Zarembinski, T. I. *et al.* Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proceedings of the National Academy of Sciences* **95**, 15189 (1998).

123. Terwilliger, T. C., Stuart, D. & Yokoyama, S. Lessons from structural genomics. *Annual review of biophysics* **38**, 371–83 (2009).

124. Berman, H. M. *et al.* The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Research* **37**, D365 (2009).

125. Jenney, F. E. & Adams, M. W. W. The impact of extremophiles on structural genomics (and vice versa). *Extremophiles* **12**, 39–50 (2008).

126. Joachimiak, A. High-throughput crystallography for Structural Genomics. *Current opinion in Structural Biology* **19**, 573 (2009).

127. Montelione, G. T. & Anderson, S. Structural genomics: Keystone for a human proteome project. *Nature Structural Biology* **6**, 11–12 (1999).

128. Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomic resolution macromolecular structure determination. *F1000Research* **9**, (2020).

129. Dauter, Z. & Wlodawer, A. Progress in protein crystallography. *Protein and peptide letters* **23**, 201 (2016).

130. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).

131. Baldwin, P. R. *et al.* Big data in cryoEM: automated collection, processing and accessibility of EM data. *Current Opinion in Microbiology* **43**, 1–8 (2018).

132. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).

133. Chayen, N. E. & Saridakis, E. Protein crystallization: from purified protein to diffraction-quality crystal. *Nature Methods* **5**, 147–153 (2008).

134. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallographica. Section F, Structural Biology Communications* **70**, 2 (2014).

135. Jancarik, J. & Kim, S. H. Sparse matrix sampling. A screening method for crystallization of proteins. *Journal of Applied Crystallography* **24**, 409–411 (1991).

136. Brzozowski, A. M. & Walton, J. Clear strategy screens for macromolecular crystallization. *Journal of Applied Crystallography* **34**, 97–101 (2001).

137. Newman, J. *et al.* Towards rationalization of crystallization screening for small- To medium-sized academic laboratories: The PACT/JCSG+ strategy. *Acta Crystallographica Section D: Biological Crystallography* **61**, 1426–1431 (2005).

138. Bard, J., Ercolani, K., Svenson, K., Olland, A. & Somers, W. Automated systems for protein crystallization. *Methods* **34**, 329–347 (2004).

139. Taylor, G. L. Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography* **66**, 325 (2010).

140. Evans, P. & McCoy, A. An introduction to molecular replacement. *Acta Crystallographica Section D: Biological Crystallography* **64**, 1 (2008).

141. Rossmann, M. G. The molecular replacement method. *Acta Crystallographica Section A* **46**, 73–82 (1990).

142. Cha, S. S. *et al.* Experimental phasing using zinc anomalous scattering. *Acta Crystallographica Section D: Biological Crystallography* **68**, 1253–1258 (2012).

143. Pike, A. C. W., Garman, E. F., Krojer, T., von Delft, F. & Carpenter, E. P. An overview of heavy-atom derivatization of protein crystals. *Acta Crystallographica. Section D, Structural Biology* **72**, 303 (2016).

144. Hendrickson, W. A., Horton, J. R. & Lemaster, D. M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *The EMBO Journal* **9**, 1–665 (1990).

145. Christendat, D. *et al.* Structural proteomics of an archaeon. *Nature Structural Biology* **7**, 903–909 (2000).

146. Adams, M. W. W. Enzymes and proteins from organisms that grow near and above 100 degrees C. *Annual review of microbiology* **47**, 627–658 (1993).

147. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

148. Luef, B. *et al.* Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature Communications* **6**, 6372 (2015).

149. Vigneron, A. *et al.* Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnology and Oceanography Letters* **5**, 212–220 (2020).

150. Kantor, R. S. *et al.* Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio* **4**, e00708-13 (2013).

151. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. BLAST: Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

152. Kesici, M.-Z., Tinnefeld, P., Andrés, A. A. & Vera, M. A simple and general approach to generate photoactivatable DNA processing enzymes. *Nucleic Acids Research* **50**, (2021).

153. Rittié, L. & Perbal, B. Enzymes used in molecular biology: a useful guide. *Journal of cell communication and signaling* **2**, 25–45 (2008).

154. Boshoff, A. *et al.* Molecular chaperones in biology, medicine and protein biotechnology. *South African Journal of Science* **100**, 665–677 (2004).

155. Schlieker, C., Bukau, B. & Mogk, A. Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology. *Journal of Biotechnology* **96**, 13–21 (2002).

156. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. & Ginalski, K. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Research* **40**, 7016–7045 (2012).

157. Slupska, M. M. *et al.* Genes involved in the determination of the rate of inversions at short inverted repeats. *Genes to cells* **5**, 425–437 (2000).

158. Bork, P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Research* **10**, 398–400 (2000).

159. Demirjian, D. C., Morís-Varas, F. & Cassidy, C. S. Enzymes from extremophiles. *Current Opinion in Chemical Biology* **5**, 144–151 (2001).

160. Turner, P., Mamo, G. & Karlsson, E. N. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microbial Cell Factories* **6**, 1–23 (2007).

161. Haki, G. D. & Rakshit, S. K. Developments in industrially important thermostable enzymes: a review. *Bioresource Technology* **89**, 17–34 (2003).

162. Baneyx, F. Recombinant protein expression in Escherichia coli. *Current Opinion in Biotechnology* **10**, 411–421 (1999).

163. Cherbas, L. & Cherbas, P. Drosophila cell culture and transformation. *CSH protocols* (2007) doi:10.1101/PDB.TOP6.

164. van Oers, M. M., Pijlman, G. P. & Vlak, J. M. Thirty years of baculovirus-insect cell protein expression: From dark horse to mainstream technology. *Journal of General Virology* **96**, 6–23 (2015).

165. Zhu, J. Mammalian cell protein expression for biopharmaceutical production. *Biotechnology advances* **30**, 1158–1170 (2012).

166. Silverman, A. D., Karim, A. S. & Jewett, M. C. Cell-free gene expression: an expanded repertoire of applications. *Nature Reviews Genetics* **21**, 151–170 (2020).

167. Makrides, S. C. Strategies for achieving high-level expression of genes in Escherichia coli. *Microbiological Reviews* **60**, 512–538 (1996).

168. de Marco, A., Deuerling, E., Mogk, A., Tomoyasu, T. & Bukau, B. Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. *BMC Biotechnology* **7**, 32 (2007).

169. Giacalone, M. J. *et al.* Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system. *BioTechniques* **40**, 355–364 (2006).

170. Lobstein, J. *et al.* SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial Cell Factories* **11**, 1–16 (2012).

171. di Guana, C., Lib, P., Riggsa, P. D. & Inouyeb, H. Vectors that facilitate the expression and purification of foreign peptides in Escherichia coli by fusion to maltose-binding protein. *Gene* **67**, 21–30 (1988).

172. Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology* **189**, 113–130 (1986).

173. William Studier, F., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods in Enzymology* **185**, 60–89 (1990).

174. Jeong, H. *et al.* Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *Journal of Molecular Biology* **394**, 644–652 (2009).

175. Egan, S. M. & Schleif, R. F. A regulatory cascade in the induction of rhaBAD. *Journal of Molecular Biology* **234**, 87–98 (1993).

176. Haldimann, A., Daniels, L. L. & Wanner, B. L. Use of new methods for construction of tightly regulated arabinose and rhamnose promoter fusions in studies of the Escherichia coli phosphate regulon. *Journal of Bacteriology* **180**, 1277 (1998).

177. Ciarán, C. *et al.* A rhamnose-inducible system for precise and temporal control of gene expression in cyanobacteria. *ACS Synthetic Biology* **7**, 1056–1066 (2018).

178. Kelly, C. L. *et al.* Synthetic Chemical Inducers and Genetic Decoupling Enable Orthogonal Control of the rhaBAD Promoter. *ACS Synthetic Biology* **5**, 1136–1145 (2016).

179. Holcroft, C. C. & Egan, S. M. Interdependence of activation at rhaSR by cyclic AMP receptor protein, the RNA polymerase alpha subunit C-terminal domain, and RhaR. *Journal of Bacteriology* **182**, 6774–6782 (2000).

180. Uhlén, M., Forsberg, G., Moks, T., Hartmanis, M. & Nilsson, B. Fusion proteins in biotechnology. *Current Opinion in Biotechnology* **3**, 363–369 (1992).

181. Harper, S. & Speicher, D. W. Purification of proteins fused to glutathione S-transferase. *Methods in Molecular Biology* **681**, 259–280 (2011).

182. Smith, D. B. & Johnson, K. S. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* **67**, 31–40 (1988).

183. Bornhorst, J. A. & Falke, J. J. Purification of proteins using polyhistidine affinity tags. *Methods in Enzymology* **326**, 245–254 (2000).

184. Porath, J. Immobilized metal ion affinity chromatography. *Protein Expression and Purification* **3**, 263–281 (1992).

185. Porath, J., Carlsson, J., Olsson, I. & Belfrage, G. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* **258**, 598–599 (1975).

186. LaVallie, E. R. *et al.* A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. *Nature Biotechnology* **11**, 187–193 (1993).

187. Malakhov, M. P. *et al.* SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *Journal of structural and functional genomics* **5**, 75–86 (2004).

188. Marblestone, J. *et al.* Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Science* **15**, 182–189 (2006).

189. Niedzialkowska, E. *et al.* Protein purification and crystallization artifacts: The tale usually not told. *Protein Science* **25**, 720–733 (2016).

190. Waugh, D. S. An overview of enzymatic reagents for the removal of affinity tags. *Protein Expression and Purification* **80**, 283 (2011).

191. Kapust, R. B. & Waugh, D. S. Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expression and Purification* **19**, 312–318 (2000).

192. Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. His-tag impact on structure. *Acta crystallographica. Section D, Biological crystallography* **63**, 295–301 (2007).

193. Townsend, P. D. *et al.* The crystal structures of apo and cAMP-bound GlxR from Corynebacterium glutamicum reveal structural and dynamic changes upon cAMP binding in CRP/FNR family transcription factors. *PLOS ONE* **9**, e113265 (2014).

194. Smyth, D. R., Mrozkiewicz, M. K., McGrath, W. J., Listwan, P. & Kobe, B. Crystal structures of fusion proteins with large-affinity tags. *Protein Science* **12**, 1313 (2003).

195. Kopaciewicz, W., Rounds, M. A., Fausnaugh, J. & Regnier, F. E. Retention model for high-performance ion-exchange chromatography. *Journal of Chromatography A* **266**, 3–21 (1983).

196. Jungbauer, A. & Hahn, R. *Ion-Exchange Chromatography*. *Methods in Enzymology* vol. 463 (Academic Press, 2009).

197. Porath, J. & Flodin, P. Gel filtration: a method for desalting and group separation. *Nature* **183**, 1657–1659 (1959).

198. Hellberg, U., Ivarsson, J. P. & Johansson, B. L. Characteristics of Superdex® prep grade media for gel filtration chromatography of proteins and peptides. *Process Biochemistry* **31**, 163–172 (1996).

199. Gräslund, S. *et al.* Protein production and purification. *Nature methods* **5**, 135 (2008).

200. Laemmli, U. K. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature 1970 227:5259* **227**, 680–685 (1970).

201. Weber, K. & Osborn, M. The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry* **244**, 4406–4412 (1969).

202. MM, B. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry* **72**, 248–254 (1976).

203. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research* **31**, 3784–8 (2003).

204. Hong, P., Koza, S. & Bouvier, E. S. P. Size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *Journal of liquid chromatography & related technologies* **35**, 2923–2950 (2012).

205. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).

206. Singhal, N., Kumar, M., Kanaujia, P. K. & Virdi, J. S. MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology* **6**, 791 (2015).

207. Maria Nadler, W. *et al.* MALDI versus ESI: the impact of the ion source on peptide identification. *Journal of Proteome Research* **16**, 1207–1215 (2017).

208. Chait, B. T. Mass spectrometry in the postgenomic era. *Annual Review of Biochemistry* **80**, 239–246 (2011).

209. Henzel, W. J., Watanabe, C. & Stults, J. T. Protein identification: The origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry* **14**, 931–942 (2003).

210. Bruce, D., Cardew, E., Freitag-Pohl, S. & Pohl, E. How to Stabilize Protein: Stability Screens for Thermal Shift Assays and Nano Differential Scanning Fluorimetry in the Virus-X Project. *Journal of Visualized Experiments* (2019).

211. Grøftehauge, M. K., Hajizadeh, N. R., Swann, M. J. & Pohl, E. Protein-ligand interactions investigated by thermal shift assays (TSA) and dual polarization interferometry (DPI). *Acta crystallographica. Section D, Biological crystallography* **71**, 36–44 (2015).

212. Reinhard, L., Mayerhofer, H., Geerlof, A., Mueller-Dieckmann, J. & Weiss, M. S. Optimization of protein buffer cocktails using Thermofluor. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **69**, 209 (2013).

213. Freitag-Pohl, S. *et al.* Crystal structures of the Bacillus subtilis prophage lytic cassette proteins XepA and YomS. *Acta Crystallographica Section D: Structural Biology* **75**, 1028–1039 (2019).

214. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical biochemistry* **357**, 289–298 (2006).

215. Reinhard, L., Mayerhofer, H., Geerlof, A., Mueller-Dieckmann, J. & Weiss, M. S. Optimization of protein buffer cocktails using Thermofluor. *Acta crystallographica. Section F, Structural biology and crystallization communications* **69**, 209–214 (2013).

216. Boivin, S., Kozak, S. & Meijers, R. Optimization of protein purification and characterization using Thermofluor screens. *Protein expression and purification* **91**, 192–206 (2013).

217. Zamost, B. L., Nielsen, H. K. & Starnes, R. L. Thermostable enzymes for industrial applications. *Journal of Industrial Microbiology* **8**, 71–81 (1991).

218. Rigoldi, F., Donini, S., Redaelli, A., Parisini, E. & Gautieri, A. Review: Engineering of thermostable enzymes for industrial applications. *APL bioengineering* **2**, 011501 (2018).

219. Lilie, H., Schwarz, E. & Rudolph, R. Advances in refolding of proteins produced in E. coli. *Current opinion in biotechnology* **9**, 497–501 (1998).

220. Singh, S. M. & Panda, A. K. Solubilization and refolding of bacterial inclusion body proteins. *Journal of Bioscience and Bioengineering* **99**, 303–310 (2005).

221. Booth, W. T. *et al.* Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* **3**, 760–768 (2018).

222. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology* **112**, 531–552 (1999).

223. Sherman, F., Stewart, J. W. & Tsunasawa, S. Methionine or not methionine at the beginning of a protein. *BioEssays : news and reviews in molecular, cellular and developmental biology* **3**, 27–31 (1985).

224. Giglione, C., Boularot, A. & Meinnel, T. Protein N-terminal methionine excision. *Cellular and molecular life sciences* **61**, 1455–1474 (2004).

225. Ladenstein, R. & Antranikian, G. Proteins from hyperthermophiles: Stability and enzymatic catalysis close to the boiling point of water. *Advances in biochemical engineering/biotechnology* **61**, 37–85 (1998).

226. Deutscher, M. P., Marlor, C. W. & Zaniewski, R. Ribonuclease T: new exoribonuclease possibly involved in end-turnover of tRNA. *Proceedings of the National Academy of Sciences* **81**, 4290 (1984).

227. Viswanathan, M., Dower, K. W. & Lovett, S. T. Identification of a potent DNase activity associated with RNase T of Escherichia coli. *Journal of Biological Chemistry* **273**, 35126–35131 (1998).

228. Bošnjak, I., Bojović, V., Šegvić-Bubić, T. S. & Bielen, A. Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank. *Protein Engineering, Design and Selection* **27**, 65–72 (2014).

229. Chen, J. *et al.* Chaperone activity of DsbC. *Journal of Biological Chemistry* **274**, 19601–19605 (1999).

230. Aschenbrenner, J. & Marx, A. DNA polymerases and biotechnological applications. *Current Opinion in Biotechnology* **48**, 187–195 (2017).

231. Kur, J., Olszewski, M., Długołęcka, A. & Filipkowski, P. Single-stranded DNA-binding proteins (SSBs) -- sources and applications in molecular biology. *Acta Biochimica Polonica* **52**, 569–574 (2005).

232. Wang, Y., Yau, Y. Y., Perkins-Balding, D. & Thomson, J. G. Recombinase technology: applications and possibilities. *Plant Cell Reports* **30**, 267 (2011).

233. Aschenbrenner, J., Drum, M., Topal, H., Wieland, M. & Marx, A. Direct Sensing of 5-Methylcytosine by Polymerase Chain Reaction. *Angewandte Chemie International Edition* **53**, 8154–8158 (2014).

234. Aschenbrenner, J. & Marx, A. Direct and site-specific quantification of RNA 2′-O-methylation by PCR with an engineered DNA polymerase. *Nucleic Acids Research* **44**, 3495–3502 (2016).

235. Huber, C., von Watzdorf, J. & Marx, A. 5-methylcytosine-sensitive variants of Thermococcus kodakaraensis DNA polymerase. *Nucleic Acids Research* **44**, 9881–9890 (2016).

236. Werbowy, O. *et al.* The characteristics of new SSB proteins from metagenomic libraries and their use in biotech applications. *Proceedings* **50**, 135 (2020).

237. Adrio, J. L. & Demain, A. L. Microbial enzymes: tools for biotechnological processes. *Biomolecules* **4**, 117 (2014).

238. de Carvalho, C. Enzymatic and whole cell catalysis: Finding new strategies for old processes. *Biotechnology Advances* **29**, 75–83 (2011).

239. Ando, T., Israel, D. A., Kusugami, K. & Blaser, M. J. HP0333, a member of the dprA family, as involved in natural transformation in Helicobacter pylori. *Journal of Bacteriology* **181**, 5572 (1999).

240. Karudapuram, S., Zhao, X. & Barcak, G. J. DNA sequence and characterization of Haemophilus influenzae dprA+, a gene required for chromosomal but not plasmid DNA transformation. *Journal of Bacteriology* **177**, 3235 (1995).

241. Mortier-Barrière, I. *et al.* A Key Presynaptic Role in Transformation for a Widespread Bacterial Protein: DprA Conveys Incoming ssDNA to RecA. *Cell* **130**, 824–836 (2007).

242. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* **3**, 679–687 (2005).

243. Smith, G. R. Homologous recombination in procaryotes. *Microbiological Reviews* **52**, 1–28 (1988).

244. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *Journal of Bacteriology* **182**, 1016–1023 (2000).

245. Lisboa, J. *et al.* The C-terminal domain of HpDprA is a DNA-binding winged helix domain that does not bind double-stranded DNA. *The FEBS Journal* **286**, 1941–1958 (2019).

246. Lisboa, J. *et al.* Molecular determinants of the DprA−RecA interaction for nucleation on ssDNA. *Nucleic Acids Research* **42**, 7395 (2014).

247. Yadav, T. *et al.* Bacillus subtilis DprA Recruits RecA onto Single-stranded DNA and Mediates Annealing of Complementary Strands Coated by SsbB and SsbA. *The Journal of Biological Chemistry* **288**, 22437 (2013).

248. Bergé, M., Mortier-Barrière, I., Martin, B. & Claverys, J. P. Transformation of Streptococcus pneumoniae relies on DprA- and RecA-dependent protection of incoming DNA single strands. *Molecular Microbiology* **50**, 527–536 (2003).

249. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354 (2021).

250. Wang, W., Ding, J., Zhang, Y., Hu, Y. & Wang, D. C. Structural insights into the unique single-stranded DNA-binding mode of Helicobacter pylori DprA. *Nucleic Acids Research* **42**, 3478 (2014).

251. Kim, C. A., Gingery, M., Pilpa, R. M. & Bowie, J. U. The SAM domain of polyhomeotic forms a helical polymer. *Nature Structural Biology* **9**, 453–457 (2002).

252. Schwartz, T., Behlke, J., Lowenhaupt, K., Heinemann, U. & Rich, A. Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nature Structural Biology* **8**, 761–765 (2001).

253. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011).

254. Quevillon-Cheruel, S. *et al.* Structure-function analysis of pneumococcal DprA protein reveals that dimerization is crucial for loading RecA recombinase onto DNA during transformation. *Proceedings of the National Academy of Sciences* **109**, E2466 (2012).

255. Grøftehauge, M. K., Hajizadeh, N. R., Swann, M. J. & Pohl, E. Protein-ligand interactions investigated by thermal shift assays (TSA) and dual polarization interferometry (DPI). *Acta crystallographica. Section D, Biological crystallography* **71**, 36–44 (2015).

256. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallographica Section D: Biological Crystallography* **67**, 293–302 (2011).

257. Vonrhein, C. *et al.* Advances in automated data analysis and processing within autoPROC, combined with improved characterisation, mitigation and visualisation of the anisotropy of diffraction limits using STARANISO. *Acta Crystallographica Section A Foundations and Advances* **74**, a360–a360 (2018).

258. Kabsch, W. XDS. *Acta crystallographica. Section D, Biological crystallography* **66**, 125–32 (2010).

259. Grüne, T. & Scherrer, P. Data Processing with XDS. (2017).

260. Stein, N. CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. *Journal of Applied Crystallography* **41**, 641–643 (2008).

261. McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658–674 (2007).

262. Matthews, B. W. Solvent content of protein crystals. *Journal of Molecular Biology* **33**, 491–497 (1968).

263. Cramer, P. AlphaFold2 and the future of structural biology. *Nature Structural & Molecular Biology* **28**, 704–705 (2021).

264. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

265. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**, 774–97 (2007).

266. Krissinel, E. Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Research* **43**, W314–W319 (2015).

267. Galperin, M. Y. & Koonin, E. v. "Conserved hypothetical" proteins: prioritization of targets for experimental study. *Nucleic Acids Research* **32**, 5452–5463 (2004).

268. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology* **9**, 467–477 (2011).

269. Kuipers, R. K. *et al.* 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins* **78**, 2101–13 (2010).

270. McPherson, A. & Cudney, B. Optimization of crystallization conditions for biological macromolecules. *Acta Crystallographica. Section F, Structural Biology Communications* **70**, 1445 (2014).

271. Rodríguez, D. D. *et al.* Crystallographic ab initio protein structure solution below atomic resolution. *Nature Methods* **6**, 651–653 (2009).

272. Sheldrick, G. M. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography* **64**, 112–122 (2008).

273. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta crystallographica. Section D, Biological crystallography* **58**, 1772–1779 (2002).

274. Thorn, A. & Sheldrick, G. M. Extending molecular-replacement solutions with SHELXE. *Acta Crystallographica Section D: Biological Crystallography* **69**, 2251 (2013).

275. D'Arcy, A., Bergfors, T., Cowan-Jacob, S. W. & Marsh, M. Microseed matrix screening for optimization in protein crystallization: What have we learned? *Acta Crystallographica Section:F Structural Biology Communications* **70**, 1117–1126 (2014).

276. Shaw Stewart, P. D., Kolek, S. A., Briggs, R. A., Chayen, N. E. & Baldock, P. F. M. Random microseeding: a theoretical and practical exploration of seed stability and seeding techniques for successful protein crystallization. *Crystal Growth & Design* **11**, 3432–3441 (2011).

277. Pannu, N. S. *et al.* Recent advances in the *CRANK* software suite for experimental phasing. *Acta Crystallographica Section D Biological Crystallography* **67**, 331–337 (2011).

278. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* **67**, 235–242 (2011).

279. Cowtan, K. & IUCr. Recent developments in classical density modification. *Acta Crystallographica Section D Biological Crystallography* **66**, 470–478 (2010).

280. Usón, I. *et al.* 1.7 Å structure of the stabilized REI $_{v}$ mutant T39K. Application of local NCS restraints. *Acta Crystallographica Section D Biological Crystallography* **55**, 1158–1167 (1999).

281. Winter, G. *et al.* DIALS: implementation and evaluation of a new integration package. *Acta Crystallographica. Section D, Structural Biology* **74**, 85–97 (2018).

282. Weiss, M. S. & IUCr. Global indicators of X-ray data quality. *Journal of Applied Crystallography* **34**, 130–135 (2001).

283. Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. His-tag impact on structure. *Acta crystallographica. Section D, Biological crystallography* **63**, 295–301 (2007).

284. Chant, A., Kraemer-Pecore, C. M., Watkin, R. & Kneale, G. G. Attachment of a histidine tag to the minimal zinc finger protein of the Aspergillus nidulans gene regulatory protein AreA causes a conformational change at the DNA-binding site. *Protein Expression and Purification* **39**, 152–159 (2005).

285. Dauter, Z. Collection of X-ray diffraction data from macromolecular crystals. *Methods in Molecular Biology* **1607**, 165 (2017).

286. Mccall, K. A., Huang, C.-C. & Fierke, C. A. Function and mechanism of zinc metalloenzymes. *Journal of Nutrition* **130**, 1437–1446 (2000).

287. Auld, D. S. Zinc coordination sphere in biochemical zinc sites. *Biometals* **14**, 271–313 (2001).

288. Harding, M. M. Geometry of metal-ligand interactions in proteins. *Acta Crystallographica. Section D, Biological Crystallography* **57**, 401–411 (2001).

289. Harding, M. M. & IUCr. Metal–ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallographica Section D Biological Crystallography* **58**, 872–874 (2002).

290. Wang, N. *et al.* Structural basis of tubulin detyrosination by the vasohibin–SVBP enzyme complex. *Nature Structural & Molecular Biology* **26**, 571–582 (2019).

291. Li, F., Hu, Y., Qi, S., Luo, X. & Yu, H. Structural basis of tubulin detyrosination by vasohibins. *Nature Structural & Molecular Biology* **26**, 583–591 (2019).

292. Nieuwenhuis, J. *et al.* Vasohibins encode tubulin detyrosinating activity. *Science* **358**, 1453–1456 (2017).

293. Slep, K. C. Cytoskeletal cryptography: structure and mechanism of an eraser. *Nature Structural & Molecular Biology* **26**, 532–534 (2019).

294. Sanchez-Pulido, L. & Ponting, C. P. Vasohibins: new transglutaminase-like cysteine proteases possessing a non-canonical Cys-His-Ser catalytic triad. *Bioinformatics* **32**, 1441–1445 (2016).

295. Pilhofer, M., Ladinsky, M. S., McDowall, A. W., Petroni, G. & Jensen, G. J. Microtubules in bacteria: ancient tubulins build a five-protofilament homolog of the eukaryotic cytoskeleton. *PLOS Biology* **9**, e1001213 (2011).

296. Thompson, V. F., Saldaña, S., Cong, J. & Goll, D. E. A BODIPY fluorescent microplate assay for measuring activity of calpains and other proteases. *Analytical Biochemistry* **279**, 170–8 (2000).

297. Miseta, A. & Csutora, P. Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Molecular Biology and Evolution* **17**, 1232–1239 (2000).

298. Paulsen, C. E. & Carroll, K. S. Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chemical Reviews* **113**, 4633–4679 (2013).

299. Sechi, S. & Chait, B. T. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Analytical Chemistry* **70**, 5150–5158 (1998).

300. Alcock, L. J., Perkins, M. v & Chalker, J. M. Chemical methods for mapping cysteine oxidation. *Chemical Society Reviews* **47**, 231 (2018).

301. Coulombe, R. *et al.* Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *The EMBO Journal* **15**, 5492–5503 (1996).

302. Verma, S., Dixit, R. & Pandey, K. C. Cysteine proteases: Modes of activation and future prospects as pharmacological targets. *Frontiers in Pharmacology* **7**, 107 (2016).

303. Turk, B. *et al.* Acidic pH as a physiological regulator of human cathepsin L activity. *European Journal of Biochemistry* **259**, 926–932 (1999).

304. Sundararaj, S. *et al.* The ionic and hydrophobic interactions are required for the auto activation of cysteine proteases of Plasmodium falciparum. *PLOS One* **7**, e47227 (2012).

305. Mason, R. W. & Massey, S. D. Surface activation of pro-cathepsin L. *Biochemical and Biophysical Research Communications* **189**, 1659–1666 (1992).

306. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution* **35**, 1547–1549 (2018).

307. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–82 (1992).

308. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).

309. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

310. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

311. Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).

312. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proceedings of the National Academy of Sciences* **111**, 15873 (2014).

313. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 324–332 (2011).

314. Saibil, H. Chaperone machines for protein folding, unfolding and disaggregation. *Nature Reviews. Molecular Cell Biology* **14**, 630 (2013).

315. Mogk, A., Mayer, M. P. & Deuerling, E. Mechanisms of protein folding: molecular chaperones and their application in biotechnology. *ChemBioChem* **3**, 807–814 (2002).

316. Schroder, H., Langer, T., Hartl, F. U. & Bukau, B. DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. *The EMBO Journal* **12**, 4137 (1993).

317. Schonfeld, H. J., Schmidt, D., Schroder, H. & Bukau, B. The DnaK Chaperone System of Escherichia coli: Quaternary Structures and Interactions of the DnaK and GrpE Components. *Journal of Biological Chemistry* **270**, 2183–2189 (1995).

318. Buchberger, A., Schröder, H., Hesterkamp, T., Schönfeld, H. J. & Bukau, B. Substrate shuttling between the DnaK and GroEL systems indicates a chaperone network promoting protein folding. *Journal of Molecular Biology* **261**, 328–333 (1996).

319. Rebeaud, M. E., Mallik, S., Goloubinoff, P. & Tawfik, D. S. On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. *Proceedings of the National Academy of Sciences* **118**, (2021).

320. Harrison, C. GrpE, a nucleotide exchange factor for DnaK. *Cell Stress & Chaperones* **8**, 218 (2003).

321. Liberek, K., Marszalek, J., Ang, D., Georgopoulos, C. & Zylicz, M. Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. *Proceedings of the National Academy of Sciences* **88**, 2874–2878 (1991).

322. Glover, J. R. & Lindquist, S. Hsp104, Hsp70, and Hsp40: a novel chaperone system that rescues previously aggregated proteins. *Cell* **94**, 73–82 (1998).

323. Doyle, S. M. & Wickner, S. Hsp104 and ClpB: protein disaggregating machines. *Trends in Biochemical Sciences* **34**, 40–48 (2009).

324. Yu, H. *et al.* ATP hydrolysis-coupled peptide translocation mechanism of Mycobacterium tuberculosis ClpB. *Proceedings of the National Academy of Sciences* **115**, E9560–E9569 (2018).

325. Zeymer, C., Barends, T. R. M., Werbeck, N. D., Schlichting, I. & Reinstein, J. Elements in nucleotide sensing and hydrolysis of the AAA+ disaggregation machine ClpB: a structure-based mechanistic dissection of a molecular motor. *Acta Crystallographica Section D: Biological Crystallography* **70**, 582 (2014).

326. Barnett, M. E., Zolkiewska, A. & Zolkiewski, M. Structure and activity of ClpB from Escherichia coli. Role of the amino-and -carboxyl-terminal domains. *Journal of Biological Chemistry* **275**, 37565–71 (2000).

327. Mogk, A. *et al.* Identification of thermolabile Escherichia coli proteins: prevention and reversion of aggregation by DnaK and ClpB. *The EMBO journal* **18**, 6934–6949 (1999).

328. Carroni, M. *et al.* Head-to-tail interactions of the coiled-coil domains regulate ClpB activity and cooperation with Hsp70 in protein disaggregation. *Elife* **3**, e02481–e02481 (2014).

329. de Marco, A. Protocol for preparing proteins with improved solubility by co-expressing with molecular chaperones in Escherichia coli. *Nature Protocols 2007 2:10* **2**, 2632–2639 (2007).

330. de Marco, A. Molecular and chemical chaperones for improving the yields of soluble recombinant proteins. *Methods in Molecular Biology* **705**, 31–51 (2011).

331. Gelinas, A. D. *et al.* A Structure-based interpretation of E. coli GrpE thermodynamic properties. *Journal of Molecular Biology* **323**, 131–142 (2002).

332. Groemping, Y. & Reinstein, J. Folding properties of the nucleotide exchange factor GrpE from Thermus thermophilus: GrpE is a thermosensor that mediates heat shock response. *Journal of Molecular Biology* **314**, 167–178 (2001).

333. Holzwarth, G. & Doty, P. The ultraviolet circular dichroism of polypeptides. *Journal of the American Chemical Society* **87**, 218–228 (1965).

334. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols* **1**, 2876 (2006).

335. Grimshaw, J. P. A., Jelesarov, I., Schönfeld, H. J. & Christen, P. Reversible thermal transition in GrpE, the nucleotide exchange factor of the DnaK heat-shock system. *Journal of Biological Chemistry* **276**, 6098–6104 (2001).

336. Arsène, F., Tomoyasu, T. & Bukau, B. The heat shock response of Escherichia coli. *International Journal of Food Microbiology* **55**, 3–9 (2000).

337. Bolanos-Garcia, V. M. & Davies, O. R. Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochimica et Biophysica Acta* **1760**, 1304–1313 (2006).

338. Robichon, C., Luo, J., Causey, T. B., Benner, J. S. & Samuelson, J. C. Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli protein contamination after purification by immobilized metal affinity chromatography. *Applied and Environmental Microbiology* **77**, 4634 (2011).

339. Liao, Y. T., Manson, A. C., DeLyser, M. R., Noid, W. G. & Cremer, P. S. Trimethylamine N-oxide stabilizes proteins via a distinct mechanism compared with betaine and glycine. *Proceedings of the National Academy of Sciences* **114**, 2479–2484 (2017).

340. Cytiva. Selection guide: size exclusion chromatography columns and resins. https://cytiva-delivery.sitecorecontenthub.cloud/api/public/content/digi-13947-pdf.

341. Deller, M. C. & Rupp, B. Crystallisation of proteins and macromolecular complexes: past, present and future. in *eLS, John Wiley & Sons, Ltd (Ed.)*. (2014).

342. Deville, C. *et al.* Structural pathway of regulated substrate transfer and threading through an Hsp100 disaggregase. *Science Advances* **3**, e1701726–e1701726 (2017).

343. Biter, A. B., Lee, S., Sung, N. & Tsai, F. T. F. Structural basis for intersubunit signaling in a protein disaggregating machine. *Proceedings of the National Academy of Sciences* **109**, 12515–12520 (2012).

344. Lee, S. *et al.* The structure of ClpB: a molecular chaperone that rescues proteins from an aggregated state. *Cell* **115**, 229–240 (2003).

345. Li, J. & Sha, B. Crystal Structure of the E. coli Hsp100 ClpB N-Terminal Domain. *Structure* **11**, 323–328 (2003).

346. Cheng, Y. Single-particle cryo-EM at crystallographic resolution. *Cell* **161**, 450 (2015).

347. Callaway, E. Revolutionary cryo-EM is taking over structural biology. *Nature* **578**, 201–202 (2020).

348. Callaway, E. The revolution will not be crystallized: A new method sweeps through structural biology. *Nature* **525**, 172–174 (2015).

349. EMDB < Statistics. https://www.ebi.ac.uk/emdb/statistics/emdb_entries_year.

350. Weissenberger, G., Henderikx, R. J. M. & Peters, P. J. Understanding the invisible hands of sample preparation for cryo-EM. *Nature Methods* **18**, 463–471 (2021).

351. Smith, M. T. J. & Rubinstein, J. L. Beyond blob-ology. *Science* **345**, 617–619 (2014).

352. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444 (2014).

353. Maruthi, K., Kopylov, M. & Carragher, B. Automating decision making in the Cryo-EM pre-processing pipeline. *Structure* **28**, 727–729 (2020).

354. Faruqi, A. R. & Henderson, R. Electronic detectors for electron microscopy. *Current opinion in structural biology* **17**, 549–555 (2007).

355. Brilot, A. F. *et al.* Beam-induced motion of vitrified specimen on holey carbon film. *Journal of structural biology* **177**, 630–637 (2012).

356. Li, Y., Cash, J. N., Tesmer, J. J. G. & Cianfrocco, M. A. High-throughput Cryo-EM enabled by user-free preprocessing routines. *Structure* **28**, 858-869.e3 (2020).

357. Frauenfeld, J. *et al.* A saposin-lipoprotein nanoparticle system for membrane proteins. *Nature Methods* **13**, 345–351 (2016).

358. Padmanabha Das, K. M., Shih, W. M., Wagner, G. & Nasr, M. L. Large nanodiscs: a potential game changer in structural biology of membrane protein complexes and virus entry. *Frontiers in Bioengineering and Biotechnology* **8**, 539 (2020).

359. Gao, Y., Cao, E., Julius, D. & Cheng, Y. TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature* **534**, 347–351 (2016).

360. Ritchie, T. K. *et al.* Reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods in Enzymology* **464**, 211–231 (2009).

361. Carragher, B. *et al.* Current outcomes when optimizing 'standard' sample preparation for single-particle cryo-EM. *Journal of Microscopy* **276**, 39–45 (2019).

362. Drulyte, I. *et al.* Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta crystallographica. Section D, Structural biology* **74**, 560–571 (2018).

363. Gewering, T., Januliene, D., Ries, A. B. & Moeller, A. Know your detergents: A case study on detergent background in negative stain electron microscopy. *Journal of structural biology* **203**, 242–246 (2018).

364. de Carlo, S. & Harris, J. R. Negative staining and Cryo-negative Staining of Macromolecules and Viruses for TEM. *Micron (Oxford, England : 1993)* **42**, 117 (2011).

365. Scarff, C. A., Fuller, M. J. G., Thompson, R. F. & Iadaza, M. G. Variations on negative stain electron microscopy methods: tools for tackling challenging systems. *Journal of Visualized Experiments* **2018**, 57199 (2018).

366. Booth, D. S., Avila-Sakar, A. & Cheng, Y. Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *Journal of Visualized Experiments* **58**, e3227 (2011).

367. Ohi, M., Li, Y., Cheng, Y. & Walz, T. Negative staining and image classification – powerful tools in modern electron microscopy. *Biological Procedures Online* **6**, 23 (2004).

368. Houry, W. A. *et al.* Cooperation of Hsp70 and Hsp100 chaperone machines in protein disaggregation. *Frontiers in Molecular Biosciences* **2**, 22–22 (2015).

369. Doyle, S. M., Genest, O. & Wickner, S. Protein rescue from aggregates by powerful molecular chaperone machines. *Nature Reviews Molecular Cell Biology* **14**, 617–629 (2013).

370. Rizo, A. N. *et al.* Structural basis for substrate gripping and translocation by the ClpB AAA+ disaggregase. *Nature Communications* **10**, 2393–2393 (2019).

371. Inoue, Y. *et al.* Split conformation of Chaetomium thermophilum Hsp104 disaggregase. *Structure* **29**, 721-730.e6 (2021).

372. McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica Section D Biological Crystallography* **67**, 386–394 (2011).

373. Lee, S. *et al.* Cryo-EM structures of the Hsp104 protein disaggregase captured in the ATP conformation. *Cell Reports* **26**, 29–36 (2019).

374. Wu, S., Armache, J. P. & Cheng, Y. Single-particle cryo-EM data acquisition by using direct electron detection camera. *Microscopy* **65**, 35 (2016).

375. Rice, W. J. *et al.* Routine determination of ice thickness for Cryo-EM grids. *Journal of Structural Biology* **204**, 38 (2018).

376. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods 2017 14:3* **14**, 290–296 (2017).

377. Roseman, A. M. Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy* **94**, 225–236 (2003).

378. Cryo-EM Data Processing in cryoSPARC: Introductory Tutorial - CryoSPARC Guide. https://guide.cryosparc.com/processing-data/cryo-em-data-processing-in-cryosparc-introductory-tutorial.

379. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

380. del Castillo, U. *et al.* A quantitative analysis of the effect of nucleotides and the M Domain on the Association Equilibrium of ClpB. *Biochemistry* **50**, 1991–2003 (2011).

381. Akoev, V., Gogol, E. P., Barnett, M. E. & Zolkiewski, M. Nucleotide-induced switch in oligomerization of the AAA+ ATPase ClpB. *Protein Science* **13**, 567 (2004).

382. Glaeser, R. M. & Han, B.-G. Opinion: hazards faced by macromolecules when confined to thin aqueous films. *Biophysics reports* **3**, 1–7 (2017).

383. Glaeser, R. M. Proteins, interfaces, and cryo-EM grids. *Current Opinion in Colloid & Interface Science* **34**, 1 (2018).

384. Mei, K. *et al.* Cryo-EM structure of the exocyst complex. *Nature structural & molecular biology* **25**, 139–146 (2018).

385. Ofir, G. & Sorek, R. Contemporary phage biology: from classic models to new insights. *Cell* **172**, 1260–1270 (2018).

386. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801–812 (2007).

387. Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences* **96**, 2192–2197 (1999).

388. Mushegian, A. R. Are there 10^31 virus particles on Earth, or more, or fewer? *Journal of Bacteriology* **202**, 2192–2197 (2020).

389. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).

390. Harrison, J. P., Gheeraert, N., Tsigelnitskiy, D. & Cockell, C. S. The limits for life under multiple extremes. *Trends in Microbiology* **21**, 204–212 (2013).

391. Gil, J. F. *et al.* Viruses in extreme environments, current overview, and biotechnological potential. *Viruses* **13**, (2021).

392. Kuhn, J. H. *et al.* Classify viruses — the gain is worth the pain. *Nature* **566**, 318–320 (2019).

393. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).

394. Krupovic, M., Dolja, V. v. & Koonin, E. v. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology* 1 (2019) doi:10.1038/s41579-019-0205-6.

395. Angly, F. E. *et al.* The Marine Viromes of Four Oceanic Regions. *PLOS Biology* **4**, e368 (2006).

396. Kristensen, D. M., Mushegian, A. R., Dolja, V. v. & Koonin, E. v. New dimensions of the virus world discovered through metagenomics. *Trends in Microbiology* **18**, 11 (2010).

397. Schoenfeld, T. *et al.* Functional viral metagenomics and the next generation of molecular tools. *Trends in Microbiology* **18**, 20 (2010).

398. Beerenwinkel, N., Günthard, H. F., Roth, V. & Metzner, K. J. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology* **3**, (2012).

399. Ofir, G. & Sorek, R. Contemporary phage biology: from classic models to new insights. *Cell* **172**, 1260–1270 (2018).

400. Murray, N. E. & Gann, A. What has phage lambda ever done for us? *Current Biology* **17**, R305–R312 (2007).

401. van den Burg, B. Extremophiles as a source for novel enzymes. *Current opinion in microbiology* **6**, 213–218 (2003).

402. Jin, M., Gai, Y., Guo, X., Hou, Y. & Zeng, R. Properties and Applications of Extremozymes from Deep-Sea Extremophilic Microorganisms: A Mini Review. *Marine drugs* **17**, (2019).

403. Castelán-Sánchez, H. G. *et al.* Extremophile deep-sea viral communities from hydrothermal vents: Structural and functional analysis. *Marine Genomics* **46**, 16–28 (2019).

404. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

405. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology* **7**, 828–836 (2009).

406. Clokie, M. R. J., Millard, A. D., Letarov, A. v. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31 (2011).

407. Sandaa, R. A. *et al.* Seasonality drives microbial community structure, shaping both Eukaryotic and Prokaryotic host-viral relationships in an arctic marine ecosystem. *Viruses* **10**, (2018).

408. Middelboe, M., Jørgensen, N. O. G. & Kroer, N. Effects of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton. *Applied and environmental microbiology* **62**, 1991–1997 (1996).

409. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* **45**, 1320–1328 (2000).

410. Koonin, E. v & Dolja, V. v. A virocentric perspective on the evolution of life. *Current opinion in virology* **3**, 546–57 (2013).

411. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiology Reviews* **39**, 779–796 (2015).

412. Lwoff, A. The Concept of Virus. *Microbiology* **17**, 239–253 (1957).

413. Lwoff, A. & Tournier, P. The classification of viruses. *Annual review of microbiology* **20**, 45–74 (1966).

414. la Scola, B. *et al.* A giant virus in amoebae. *Science* **299**, 2033 (2003).

415. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).

416. Claverie, J. M. *et al.* Mimivirus and the emerging concept of "giant" virus. *Advances in Virus Research* **117**, 133–144 (2006).

417. Colson, P., la Scola, B., Levasseur, A., Caetano-Anollés, G. & Raoult, D. Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nature Reviews Microbiology* **15**, 243–254 (2017).

418. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).

419. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nature Communications* **10**, 1–11 (2019).

420. Abrahão, J. S., Araújo, R., Colson, P. & la Scola, B. The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLOS Genetics* **13**, e1006532 (2017).

421. Forterre, P. Giant viruses: conflicts in revisiting the virus concept. *Intervirology* **53**, 362–378 (2010).

422. Colson, P., de Lamballerie, X., Fournous, G. & Raoult, D. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* **55**, 321–332 (2012).

423. Yuan, Y. & Gao, M. Jumbo bacteriophages: An overview. *Frontiers in Microbiology* **8**, 403 (2017).

424. Hendrix, R. W. Jumbo Bacteriophages. *Current Topics in Microbiology and Immunology* **328**, 229–240 (2009).

425. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *The ISME Journal* **16**, 1657–1667 (2022).

426. Chayen, N. E. Turning protein crystallisation from an art into a science. *Current Opinion in Structural Biology* **14**, 577–583 (2004).

427. Grabowski, M., Niedzialkowska, E., Zimmerman, M. D. & Minor, W. The Impact of Structural Genomics: the First Quindecennial. *Journal of structural and functional genomics* **17**, 1 (2016).

428. Michalska, K. & Joachimiak, A. Structural genomics and the Protein Data Bank. *Journal of Biological Chemistry* **296**, 100747 (2021).

429. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature Communications* **10**, 1–12 (2019).

430. Wrighton, K. C. *et al.* RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *The ISME Journal* **10**, 2702–2714 (2016).

431. Gorrec, F. The MORPHEUS protein crystallization screen. *Journal of Applied Crystallography* **42**, 1035–1042 (2009).

432. Teng, T.-Y. & IUCr. Mounting of crystals for macromolecular crystallography in a free-standing thin film. *Journal of Applied Crystallography* **23**, 387–391 (1990).

433. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).

434. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst* **66**, 486–501 (2010).

435. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta crystallographica. Section D, Biological crystallography* **67**, 355–67 (2011).

436. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* **372**, 774–797 (2007).

437. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography* **66**, 12–21 (2010).

438. Krieger, E. & Vriend, G. YASARA View - molecular graphics for all devices - from smartphones to workstations. *Bioinformatics* **30**, 2981–2 (2014).

439. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290–296 (2017).

440. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *Journal of Structural Biology* **157**, 38–46 (2006).