

Integration of tail-anchored proteins by the twin-arginine translocase

José Jesús Gallego Parrilla

A Thesis Submitted for the qualification of Doctor of
Philosophy

Biosciences Institute

Newcastle University

Thesis Submitted May 2023



Abstract

The Tat pathway transports folded proteins across the cytoplasmic membranes of bacteria. Tat substrates have N-terminal signal peptides containing a conserved twin-arginine motif and are exported in a folded form. While most Tat substrates are soluble periplasmic proteins, a small fraction are anchored to the membrane by a single C-terminal transmembrane helix. In the model organism *Escherichia coli*, five Tat substrates are tail-anchored membrane proteins. However, the repertoire of tail-anchored Tat substrates across prokaryotes is currently unclear.

To identify new C-tail anchored Tat dependent proteins, a database search of bacterial proteins was carried out by Dr Govind Chandra, identifying many candidate proteins with both a predicted twin-arginine signal peptide and a hydrophobic stretch close to the C-terminus. A custom-written python code, utilising regular expression techniques was designed to classify the extracted proteins. From this around 80 distinct protein families were classified. To confirm the prediction that some of these represent new families of Tat-dependent C-tail proteins, the twin-arginine signal sequences and C-tail encoding regions of a small subset were fused to *E. coli* reporter proteins to assess for Tat-dependence and membrane integration. From this several new Tat-dependent tail-anchored substrates were verified.

The mechanism by which C-tails are integrated into the membrane is currently unknown. I designed genetic constructs with the aim to investigate this process. These constructs harboured full length SufI fused to the C-tail of FdnH, which was fused in frame to either maltose binding protein or β -lactamase. Experiments to determine the membrane localisation and Tat-dependence of these indicated that both fusion proteins were integrated into the membrane Tat-independently. Modification of the β -lactamase fusion through addition of linkers around the transmembrane domain improved stability and membrane integration. However, further modifications would be necessary before the fusion protein could be used as a reporter for Tat-dependent C-tail integration.

COVID-19 Impact Statement

The research in this thesis was significantly impacted due to the COVID-19 pandemic. I was unable to carry out laboratory research from March 2020 - July 2020 due to no access to the laboratory and minimal ability to do computer-based work for my project at that point in the project. For the remainder of July 2020 - October 2020, I was on shift work so lab work was also severely disrupted. I was granted a 12-week extension to my thesis submission deadline.

Acknowledgements

Firstly, I would like to express my profound appreciation to my PhD supervisor, Tracy Palmer. I am deeply grateful for your patience, support, and the opportunity you provided me to commence my PhD in your esteemed group. This opportunity allowed me to grow within a high-level academic research group, brimming with extraordinary individuals and opportunities. Your remarkable ability to assemble such a talented team is truly commendable, and I consider myself fortunate to have been a part of it. The composition of this group has evolved over the course of my PhD, and I would like to extend my gratitude to everyone who has been a part of this journey.

Starting with Jon Cherry, I extend my heartfelt thanks for your assistance with all those KLDs and your patience with a socially clumsy person like me. I was indeed fortunate to have someone with an Irish accent when I joined the group! I also want to express my gratitude to Grant Buchanan for introducing me to Scottish slang and teaching me how to cook Haggis!

Further, I would like to acknowledge two other marvelous Scots, Lisa Bowman and Fatima Ulhuq. Lisa, thank you for generously sharing your sacred Milli-Q water and for always being a cheerful presence in the lab - your laughter is sorely missed! Fatima, my hijabi girl, I am grateful for your unwavering support, comforting hugs, and the long conversations about the sunsets in Dundee.

Lastly, I wish to mention Chriselle and Guillermina, who have been instrumental in helping me in the lab on numerous occasions.

As part of my initial journey at CBCB, I would also like to acknowledge the Frank Sargent group, including Ciarán, Magali, Alice Banks, and the dynamic duo of Alex Finley and Johnny Allen, the latter of whom has a particularly lovely soul. Thank you all for the good times and the memorable nights out.

Finally, from the CBCB, I would like to express my gratitude to Sandra Laborda, Maria Puiu, Lizah van der Aart, Adam Azzi, Eleni, Jack, Tom, and Charles.

There are individuals who shared not only my time at the CBCB but also my experiences at the Medical School, those who have been part of my lab, my city, and my life over the past four years. Stephen, the other Irish soul who resides in my heart, is one of the most intelligent people and the best Pokémon trainer I know. Kieran, the perfectionist, is always concerned about those around him and has a heart of gold, never forget you are amazing. Yaping Yang, A.K.A YP, is always ready to play along with my jokes and is ever willing to lend a hand in the lab.

Andy, who has had to deal with my complaints and problems, has always been ready to listen and offer mature advice. Eunice, the one who inherited my bench and all my lab stuff, I hope you continue to enjoy everything and keep smiling. Curtis, a beautiful human being full of love and an incredible professional worker. Fliss, thank you for all the advice on looking for tails and signal peptides, and I apologize for not being able to swim with you!

Amy, who brings a lovely atmosphere to the lab and is always ready to help everyone, you have tremendous strength inside you, and I love how you take care of everyone. Merel, it's strange that I saw your talk back in 2018 when you were in the UK and I was with Yves. I was thrilled when I discovered that you were joining the group. The world is a small place, so I hope we keep in touch - bedankt.

My favorite German, what can I say! From not even saying hello to me on the first day to now being one of my favorite people, keep being you! I need to give special thanks to "my Italians" - in my mind, I always refer to you both like that. You saved me, and your support in my research was crucial. I will never forget it.

Jorge was never in Newcastle, in fact, it has been many years since we last saw each other, but I wanted to mention him here as he helped me with the code at times when I was unable to install Linux. He has been helping me since my university days, and I hope he continues to be a part of my life.

And last but certainly not least, we have Miss Eleanor Boardman. She was the second PhD to join the group when Tracy moved it to Newcastle. She made me feel better about my English when she told me that she, too, was unable to understand Grant. We have shared so much over these four years - pain, sadness, loneliness, happiness. We lost our grandmothers almost at the same time, and she became one of my partner's best friends. And of course, we cannot forget George! She has become something like a best friend, a sister, or I don't know, but I do know that I hope she stays in my life forever.

I would also like to acknowledge other individuals who have made my life easier at the Medical School, including David Saphira, David Lydall, and Vasilisa.

I am grateful to my panel members, David Bolam and Colin Harwood, for their care during those intimidating yearly reviews. They transformed them into a friendly environment, providing me with ideas and feedback.

It is of utmost importance for me to express my gratitude to Jennifer Stewart from the Disability Team. She has done a tremendous job supporting me. Being neurodivergent comes with its challenges, but she showed me that it also brings unique strengths. I am deeply thankful for her support during the difficult days and for her patience in providing me with all the tools I needed to complete my PhD. People with ADHD are underrepresented at high academic levels. Learning disabilities, pressure, and many other issues often lead us to quit and fade into oblivion. Therefore, I am truly grateful that the university supported me through Jen, Tracy, and others.

I also want to remember many friends. While I can't name you all, from Alcolea (Valeriano and Manuel), to Amsterdam (Duco, Jess, Pablo...) , passing through Seville (Carmen, Takeo, Sarai, Fuentes, Celia,Ale...) and many many others, you are all a part of me.

And to my "bichos", George, Lua, Gigante,Nami, Pipo, Wanda, and Cala. Thank you for taking care of me and always bringing a smile to my face.

To Charo, a person who has been with me not only through this journey but throughout my entire scientific career. Since 2012, when you came to see me while I was working on Gas Chromatography, until now, helping me not only with this thesis but also with life itself. My love, my friend, my partner... We've journeyed through three countries, more than fifteen houses, from the North Pole to Paris, under the Northern Lights and the fireworks. Thank you for each kiss and each caress, for each laugh and each word. Thank you for being there, for finding me, and for never letting me down.

The last bit of this Acknowledgements will be in Spanih, for my family.

Escribo esta parte en español, para vosotros, a quienes os debo todo. A ti, Antonio, por siempre estar ahí con esa alma infinita capaz de captar los más tímidos detalles, esa sensibilidad desmesurada, la inocencia con la que miras el mundo y a la vez la madurez con

la que enfrentas las cosas. Gracias por todo, por cuidar de la gente cuando yo estoy lejos, por no dejarme alejarme, por ser no solo mi hermano sino también mi mejor amigo.

A papá, por cuidar tan intensamente de mí, por tu amor que me llega aunque no te des cuenta, por haber heredado de ti esta capacidad de crecimiento y aprendizaje que me ha hecho imitarte y enfrentar problemas nuevos como la bioinformática. Parte de mi habilidad la generé aquellas tardes en Alcalá jugando con las herramientas. Gracias por la paciencia que has tenido junto a mamá conmigo cuando yo era un adolescente imposible.

Y a ti, mamá, gracias por jamás rendirte conmigo, por llenarme de amor, de atención, de ánimos, de optimismo. Si alguien se merece que la culpen de que yo haya conseguido esta tesis, eres tú, por arrastrarme cuando yo no quería andar, por valorarme cuando nadie más lo hacía. Por todas las noches que me hiciste los deberes y todos los horarios que me hiciste y plastificaste. No puedo, ni en mil dimensiones distintas, sentirme más orgulloso y afortunado de la familia que tengo. Os admiro, y os amo con locura.

Y no me olvido del resto de mi familia, mis titas y titos mis primas y primos, que son muchos y a los que quiero muchísimo.

It took me longer to reach this point than it does for many others, often with the wind against me. But I was fortunate to have a host of amazing people pushing me forward.

As Snoop Dogg says, "Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. I wanna thank me for, for never quitting."

Table of contents

Chapter 1. Introduction	1
1.1 <i>Escherichia coli</i> as a model organism	1
1.2 Protein transport and secretion systems	2
1.2.1 Cytoplasmic membrane protein transport systems: Sec and Tat	12
1.2.2 The Tat system	15
1.2.3 Tat signal peptides	18
1.2.4 Tat components	19
1.2.5 Tat mechanism	22
1.2.6 Tat-dependent membrane proteins	24
1.2.7 Hydrogenases.	26
1.2.8 Formate dehydrogenases.	27
1.2.9 Biogenesis of Tat-dependent tail-anchored proteins.	28
1.3 Computational approaches for membrane protein analysis.	29
1.3.1 Protein modelling using artificial intelligence.....	29
1.4. Aims of this thesis	30
Chapter 2. Materials and Methods.....	31
2.1 Bacterial strains.....	31
2.2 Buffers, solutions, and growth media	31
2.2.1 Growth media and additives	31
2.2.2 Antibiotics used in this study	32
2.2.3 Buffers and solutions	32
2.2.4 Biological and chemical reagents.....	33
2.3 Molecular biology techniques.	34
2.3.1 DNA manipulations: Plasmid, Gblocks and synthetic genes	34
2.3.2. Plasmid construction	37
2.3.3 Amplification of DNA by Polymerase Chain Reaction (PCR)	37
2.3.4 Q5® Site-Directed Mutagenesis	38
2.3.5 Agarose gel electrophoresis	38
2.3.6 DNA digestion and Ligation	39
2.3.7 DNA sequencing	39
2.3.8 Plasmid DNA preparation	39
2.4 Preparation of competent cells and transformation with plasmid DNA	39
2.4.1 Preparation of chemically competent cells.....	40
2.4.2 Commercial competent cells.....	40
2.5 Protein methods	40

2.5.1 SDS-PAGE	40
2.5.2 Semi-dry, wet Western Blotting and turbo-blotting	41
2.5.3 Preparation of soluble and membrane fractions	42
2.6 Growth assays	43
2.6.1 Minimum inhibitory concentration (MIC) assay	43
2.6.2 Spot assay for evaluation of Tat transport activity	43
2.6.3 MacConkey maltose agar.....	44
2.7 Computational tools.....	44
2.7.1 Python	44
2.7.2 bash	44
2.7.3 RegExr	44
2.7.4 Muscle	45
2.7.5 HMMER.....	45
2.8 Online resources	45
2.8.1 TMHMM	45
2.8.2 AnnoTree	45
2.8.3 DeepFRI	45
2.8.4 NCBI.....	46
2.8.5 SignalP.....	46
2.9 Protein structure prediction	46
2.9.1 AlphaFold 2	46
2.9.2 Robetta.....	46
Chapter 3. Bioinformatic analysis of Tat dependent tail-anchored proteins	47
3.1 Introduction	47
3.1.1 Python and Regex as bioinformatic tools for database analysis.	48
3.2 Results.....	58
3.3 Families with candidate Tat-dependent tail-anchored proteins.....	65
3.3.1 S1 family peptidase.....	67
3.3.2 Lytic polysaccharide monooxygenase	68
3.3.3 YcnI family proteins	70
3.3.5 LPXTG cell wall anchor domain-containing protein.....	74
3.3.6 Terpene cyclase-mutase family protein	79
3.3.7 Others	81
3.4 Discussion	82
Chapter 4. Experimental verification of novel Tat dependent tail-anchored proteins	85

4.1 Introduction	85
4.2 Results.....	86
4.2.1 WP_086565138.1 S1 family peptidase from <i>Streptomyces africanus</i> ...	87
4.2.2 WP_011931836.1 YcnI family protein from <i>Clavibacter michiganensis</i>	87
4.2.3 WP_049064233.1 HtaA domain-containing protein from <i>Corynebacterium striatum</i>	88
4.2.4 WP_031122887.1 LPXTG cell wall anchor domain-containing protein from <i>Streptomyces</i> sp. NRRL S-623.....	89
4.2.5 WP_019982084.1 MULTISPECIES: HtaA domain-containing protein from unclassified <i>Streptomyces</i>	89
4.2.6 WP_030568954.1 LPXTG cell wall anchor domain-containing protein from <i>Streptomyces cyaneofuscatus</i>	90
4.2.7 WP_046529179.1 LPXTG cell wall anchor domain-containing protein from <i>Cellulomonas</i> sp. FA1	91
4.2.8 WP_031517753.1 terpene cyclase/mutase family protein from <i>Streptomyces</i> sp. NRRL F-5123.....	92
4.2.9 WP_056088981.1 DUF4349 domain-containing protein from <i>Methylobacterium</i> sp. Leaf99.....	92
4.3 Investigating Tat dependence of the candidate twin-arginine signal peptides.	93
4.3.1 Liquid growth assays	99
4.5 Discussion	108
Chapter 5. Developing genetic reporters to assess the assembly of Tat-dependent tail-anchored proteins	110
5.1 Introduction	110
5.1.1 The use of fusion reporters in protein analysis.....	110
5.1.2 Modifications in fusion proteins: Linkers	112
5.2 Results.....	112
5.2.1 Fusion proteins used in the study.	112
5.2.2 Phenotypic analysis of the SufI::FdnH::MalE (SFM) fusion protein	113
5.2.3 Membrane localisation of the SFM fusion protein.	115
5.3 Use of the Suf::FdnH::Bla (SFB) fusion protein to assess Tat-dependence...	116
5.3.1 Structural prediction suggests that the SFB fusion may not fold as expected.	119
5.3.2 Design of linkers predicted to assist correct folding of the FdnH C-tail in the SFB fusion protein.....	122
5.3.3 Experimental validation of the linker design to improve the behaviour of the SFB fusion.	125
5.4 Discussion	127

Chapter 6. Conclusions and future outlook	130
Bibliography	135
Appendices.....	156

List of Figures

Figure 1.1. Schematic representation of all the currently identified Gram-negative secretion systems	3
Figure 1.2. Schematic representation of the Sec and Tat protein transport pathways and Sec and Tat signal peptides	13
Figure 1.3. Structures of the Tat components and Model for the resting state multimeric TatABC receptor complex	21
Figure 1.4. Proposed model for the Tat transport cycle.	23
Figure 1.5. Four classes of Tat-dependent membrane proteins	25
Figure 3.1. Database in HTML format displayed in web browser and displayed for data analysis.....	49
Figure 3.2. Name and counts for all proteins in the database.	52
Figure 3.3 CSV file with protein names and how many times they are present in the database	54
Figure 3.4. Grouped proteins displayed in console and Sorted groups of proteins in a spreadsheet document format	56
Figure 3.5. Possible outputs following alignments of protein sequences from each of the 84 ‘families’.	60
Figure 3.6. Work-flow for the analysis of candidate Tat-dependent C-tail anchored proteins.	62
Figure 3.7. Putative conserved domains in the HAD-IB family	64
Figure 3.8. Representation of the N-termini and C-tail of the S1 family peptidase group.....	66
Figure 3.9. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of the signal peptide region of the S1 family Peptidases using Jalview	
Figure 3.10. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of Lytic polysaccharide monooxygenase.	68
Figure 3.11. Representation of the N-termini and C-tail of the Lytic polysaccharide monooxygenase proteins sequences are coloured by percentage of identity and physicochemical properties.	69
Figure 3.12. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of YcnI family proteins.	70

Figure 3.13. Representation of the N-termini and C-tail of the YcnI family proteins	71
Figure 3.14. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of HtaA domain-containing proteins..	72
Figure 3.15. Representation of the N-termini and C-tails of the HtaA domain-containing proteins	73
Figure 3.16. Sortase-mediated cell wall ligation.....	75
Figure 3.17. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of LXPTG group 1	76
Figure 3.18. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of LXPTG group 2.	76
Figure 3.19. Representation of the N-termini and C-tail of the LPXTG family protein group 1	77
Figure 3.20. Representation of the N-termini and C-tail of the LPXTG family protein group 2	78
Figure 3.21. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of the terpene cyclase-mutase family.	79
Figure 3.22. Representation of the N-termini and C-tail of the terpene cyclase/mutase family protein.	80
Figure 3.23. Sequence of the protein with hypothesised Signal peptide and C-tail in bold	81
Figure 3.24. Sequence of the protein with hypothesised Signal peptide and C-tail in bold	81
Figure 3.25. Sequence of the protein with hypothesised Signal peptide and C-tail in bold	82
Figure 4.1. Structural predictions for WP_086565138.1 using RobeTTa fold.....	87
Figure 4.2. Structural predictions for WP_011931836.1 using RobeTTa fold.....	88
Figure 4.3. Structural predictions for WP_030238604.1 using RobeTTa fold.....	88
Figure 4.4. Structural predictions for WP_031122887.1 using RobeTTa fold.....	89
Figure 4.5. Structural predictions for WP_019982084.1 using RobeTTa fold.....	90
Figure 4.6. Structural predictions for WP_030568954.1 using RobeTTa fold.....	91
Figure 4.7. Structural predictions for WP_046529179.1 using RobeTTa fold.....	91
Figure 4.8. Structural predictions for WP_031517753.1 using RobeTTa fold.....	92

Figure 4.9. Structural predictions for WP_056088981.1 using RobeTTa fold.....	93
Figure 4.10. Schematic representation of the amidase reporter assay	95
Figure 4.11. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_086565138.1 S1 family peptidase from <i>Streptomyces africanus</i> fused to the mature part of AmiA	96
Figure 4.12. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_011931836.1 YcnI family protein from <i>Clavibacter michiganensis</i> fused to the mature part of AmiA	97
Figure 4.13. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_031122887.1 LPXTG cell wall anchor domain-containing protein from <i>Streptomyces</i> sp. NRRL S-623 fused to the mature part of AmiA.....	97
Figure 4.14. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_031517753.1 terpene cyclase/mutase family protein from <i>Streptomyces</i> sp. NRRL F-5123 fused to the mature part of AmiA.	98
Figure 4.15. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_056088981.1 DUF4349 domain-containing protein from <i>Methylobacterium</i> sp. Leaf99 fused to the mature part of AmiA.	98
Figure 4.16. Growth of strain MC4100 Δ amiA Δ amiC (tat ⁺) or MC4100 Δ amiA Δ amiC Δ tatABC (tat ⁻) harbouring either empty vector (pSUPROM; EV), or pSUPROM encoding the predicted signal peptide from the indicated candidate in the presence of 0.5% SDS	103
Figure 4.17. Western blot analysis to investigate membrane integration of Sufl fused to candidate C-tails	107
Figure 5.1. Fusion constructs to assess Tat-dependent C-tail insertion	113
Figure 5.2. MalE as a reporter for C-tail integration	114
Figure 5.3. The tat ⁺ and tat ⁻ strains producing the SFM fusion show red colouration on MacConkey maltose indicator plates	115
Figure 5.4. Western blot analysis of the SFM fusion protein in membrane fractions of the tat ⁺ and tat ⁻ strain.....	116

Figure 5.5. MIC assays for strains MC4100 and DADE and the same strains harbouring pSUPROM to determine basal levels of resistance to ampicillin	117
Figure 5.6. MIC assays for strains MC4100 and DADE harbouring pSUPROM SFB showing full resistance to ampicillin.	118
Figure 5.7. The sequence of the last 24 amino acids of the FdnH region in the SFB fusion (top) and the modified sequence in SFB-SECA.	119
Figure 5.8. Crystal structure of formate dehydrogenase with the C-tail in yellow and AlphaFold models.....	120
Figure 5.9. Simulation of the SufI::FdnH fusion protein	121
Figure 5.10 RoseTTAFold simulation of the SFB fusion protein (left panel), or the same protein but with inclusion of the proline residue that usually precedes the FdnH C-tail.....	122
Figure 5.11. RoseTTAFold Simulation of SFB folding after inclusion of different linker sequences flanking the FdnH region	123
Figure 5.12. Secondary structure prediction of the FdnH region within the SFB fusion protein when flanked by none, one, two, or three iterations of the [GGGGS] linker	124
Figure 5.13. Simulation of SFB with the different linkers	125
Figure 5.14. MIC assays for strains MC4100 and DADE harbouring pSUPROM SFB-DFL	126
Figure 5.15. Western Blot analysis of membranes, urea-washed membranes and the soluble cell fraction from strains MC4100 and DADE producing SFB-DFL using anti-Bla-antibody	127

List of Tables

Table 1.1. The table presents the known or probable E. coli Tat substrate proteins	17
Table 2.1. Bacterial strains used in this study.	31
Table 2.2. Growth media used in this study.	32
Table 2.3. Antibiotics used in this study with their stock and working conditions.	32
Table 2.4. General buffers and solutions used in this study.	32
Table 2.5. Antibodies used in this study.	33
Table 2.6. Plasmids used in this study.	34
Table 2.7. gBlocks® synthesised for use in this thesis.	36
Table 2.8. SDS-PAGE components.....	41
Table 3.1. Description of scripts used in this study. Full scripts are available in the annexes.....	51
Table 3.2. Script-classified protein ‘families’ from the original database of 34,605 items.....	57
Table 3.3. Identification of 38 candidate items of interest.	61
Table 3.4. Candidate families of Tat-dependent tail anchored proteins selected for study.	63
Table 3.5. A visual showing part of the HAD-IB family hydrolase WP list.	64
Table 4.1. Potential Tat-dependent tail anchored proteins selected for further study.	86
Table 4.2. Signal peptide and C-terminal regions for each protein candidate analysed in this chapter.	94
Table 4.3. Summary of the result for the experiments of candidate SP complement with Tat system.....	105

Abbreviation	Expansion
ABC	ATP binding cassette
ATP	adenosine triphosphate
AI	Artificial intelligence
APS	Ammonium persulfate
APH	amphipathic helix
Bam	Beta-barrel Assembly Machinery
bp	base pair(s)
CAMEO	continuously evaluate the accuracy and reliability of predictions
C-terminal	Carboxy terminal
C-terminal	carboxy terminal
C-terminus	carboxyl terminus
Da	Dalton
DMSO	Dimethyl sulphoxide
EDTA	ethylenediamine tetraacetate
ECL	enhanced chemiluminescence
EHEC	Enterohaemorrhagic Escherichia coli
EPEC	Enteropathogenic Escherichia coli
ESX	ESAT-6 secretion system
g	gram
GFP	green fluorescent protein
h	hour
HMM	Hidden Markov model
HMMER	Hidden Markov Model-based sequence comparison
HTML	Hypertext Markup Language
IM	Inner membrane
Kb	Kilobase pairs (1000 bp)
kDa	Kilodalton
Kan	Kanamycin
KLD	Kinase, ligase DpnI
l	litre
LB	Luria-Bertani
LPS	Lipopolysaccharide
μ	micro
M	molar
m	milli
M.I.C	Minimum Inhibitory Concentration
MGD	molybdopterin guanine dinucleotide
MDH	methylamine dehydrogenase
MPT	molybdopterin
MUSCLE	MULTiple Sequence Comparison by Log- Expectation
MUSCLE3	Multiple Sequence Comparison by Log-Expectation
N-terminal	amino terminal
N-terminus	Amino terminus
OM	outer membrane
PAGE	polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction

PMF	proton motive force
PMF	protonmotive force
RBS	Ribosome binding site
RNC	Ribosome nascent chain
rpm	revolutions per minute
RTX	Repeats in ToXin
SDS	Sodium dodecyl sulphate
SEM	Standard error of the mean
Sec	General secretory system
SFM	SufI FdnH Male
SFB	SufI FdnH Bla
SrtA	Sortase A
SRP	signal recognition particle
T1SS	Type I secretion system
T2SS	Type II secretion system
T3SS	Type III secretion system
T4SS	Type IV secretion system
T5SS	Type V secretion system
T6SS	Type VI secretion system
T7SS	Type VII secretion system
T8SS	Type VIII secretion system
T9SS	Type IX secretion system
T10SS	Type X secretion system
T11SS	Type XI secretion system
TBST	Tris-buffered saline with Tween 20
TBS	Tris-buffered saline
TMD	Transmembrane domain
TEMED	N,N,N',N'-tetramethylethane-1,2-diamine
TMH	transmembrane helix
TSB	Transformation and storage buffer
TTQ	tryptophan tryptophylquinone
V	Volt
v/v	Volume per volume
w/v	Weight per volume
WC	Whole cells
WT	Wild-type

Chapter 1. Introduction

1.1 *Escherichia coli* as a model organism

Escherichia coli is a Gram-negative, rod-shaped bacterium belonging to the family Enterobacteriaceae within the phylum γ -proteobacteria. It commonly resides in the mucus layer of the mammalian colon and is the most abundant facultative anaerobe of the human intestinal microflora (Martinson and Walk 2020). Although most strains are harmless, some pathogenic strains, such as enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC), can cause diarrheal diseases, urinary tract infections, and meningitis. Virulent strains often acquire virulence genes, such as those encoding the Shiga toxin, and can cause illness even in healthy hosts (Johnson and Nolan 2009).

The genetic tractability, rapid reproduction time, and simple growth requirements of *E. coli* have made it a popular model organism and one of the most well-characterised prokaryotes. The sequencing of its entire genome has greatly enhanced understanding of its genetic and phenotypic diversity. The size of the genome varies among strains, with standard laboratory strains having genomes of approximately 4.5 million base pairs and around 4,000 genes, while pathogenic strains have over 5.9 million base pairs and at least 5,500 genes (Blattner et al. 1997; Lukjancenko, Wassenaar, and Ussery 2010). *E. coli* K-12 has emerged as a model bacterium for biochemical and cellular studies.

As a model organism, *E. coli* has been used to study a variety of biological processes, including DNA replication and repair, transcriptional regulation, protein synthesis, and metabolism. Its use in research has also facilitated the discovery of important cellular processes, such as the mechanism of bacterial conjugation, the discovery of CRISPR-Cas systems, and the identification of molecular chaperones (Idalia and Bernardo 2017; Taj et al. 2014).

Moreover, *E. coli* has been used to produce a variety of recombinant proteins, including medically important proteins such as insulin and human growth hormone. The ease of genetic manipulation, fast growth rate, and low cost of cultivation make

it a popular choice for large-scale production of proteins for both research and commercial applications (Blount 2015; Jørgensen et al. 1998; Chance and Frank 1993).

Overall, the versatility of *E. coli* as a model organism has made it an invaluable tool for the study of fundamental biological processes, the development of novel technologies, and the production of biopharmaceuticals.

1.2 Protein transport and secretion systems

While universal membrane transport systems, such as ATP-binding cassette (ABC) transporters, are present across all kingdoms of life, prokaryotes exhibit a higher level of specialisation due to the unique challenges they face. This is particularly true for Gram-negative bacteria which have two membrane barriers to overcome to acquire molecules from the environment, or to secrete intracellular products outside the cell. A particular challenge is the secretion of proteins, which are among the most complex macromolecules made by living organisms. Extracellular proteins play critical roles in nutrient acquisition and niche adaptation (Green and Mecsas 2016; Tseng, Tyler, and Setubal 2009), and protein secretion is ubiquitous across prokaryotes. Gram-negative bacteria have evolved numerous mechanisms to secrete proteins extracellularly and these are summarised below and in Fig. 1.1.

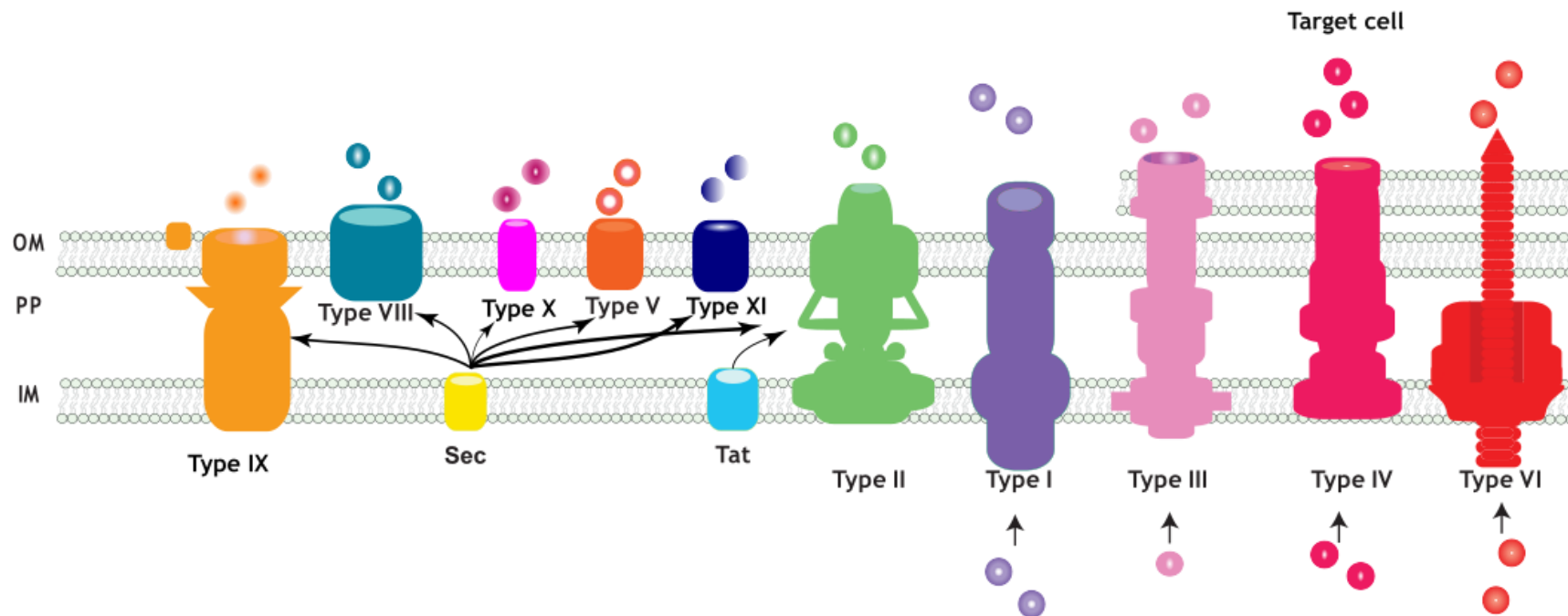


Figure 1.1. Schematic representation of all the currently identified Gram-negative secretion systems. The type VII secretion system is not included because it has not been shown to functionally occur in Gram-negatives. Note that the type III, type IV and type VI secretion systems additionally translocate proteins across a third membrane - that of the target cell. M - inner membrane, PP - periplasm, OM - outer membrane.

Type I Secretion System (T1SS)

Type I Secretion Systems (T1SS) are related to ABC transporters, but are specialised for the secretion of unfolded proteins, and primarily RTX (Repeats in ToXin)-type toxins (Costa et al. 2015; Linhartová et al. 2010; Felmlee and Welch 1988; Smith et al. 2018). The T1SS consists of an ABC transporter in the inner membrane, a periplasmic protein serving as an adaptor protein, and a porin-type protein in the outer membrane (Costa et al. 2015; Linhartová et al. 2010).

RTX proteins are typically defined by the presence of several blocks of nonapeptide-binding sequences with the consensus GGxGxDxUx (Linhartová et al. 2010). These motifs, also known as GG repeats, specifically bind Ca^{2+} and are implicated in post-translocation folding. The RTX repeats are located in the N-terminal half of the protein, while the secretion signal recognised by the T1SS is found at the extreme C-terminus (Griessl et al. 2013). The molecular identification of T1SS was first made in the 1980s and 1990s in the context of the pore-forming toxin HlyA from *E. coli* (Felmlee, Pellett, and Welch 1985).

T1 secretion is considered to be a one-step process, occurring directly from the cytosol to the extracellular space without any periplasmic intermediate (Spitz et al. 2019). This process is Sec-independent, with the C-terminal secretion signal being recognised by the cytoplasmic domains of the inner membrane ABC transporter (Benabdelhak et al. 2003). Transport is energised through ATP hydrolysis, but is also believed to be driven by Ca-dependent substrate folding as the RTX repeats emerge extracellularly (Bumba et al. 2016). Recent studies have proposed a two-step secretion process for certain large T1SS adhesin substrates identifying a "retention module" at the N-terminus that anchors the adhesin to the cell surface by stalling further translocation (Smith et al. 2018).

Type II Secretion System (T2SS)

Secretion of proteins by the T2SS is a two-step process. T2SS substrates are synthesised with N-terminal signal peptides and rely on the inner membrane Sec or Tat translocation systems for export into the periplasmic space (Abby et al. 2016). Once in the periplasm, these substrates are believed to be pushed through the secretin pore in the outer membrane by a pseudopilus in a piston-like action. This is energised by a cytoplasmic ATPase that provides the necessary power for pseudopilus extension and retraction, hence facilitating protein secretion (Naskar et al. 2021). The T2SS recognises folded substrates in the periplasm and mediates their transport across the outer membrane in a folded form (Pineau et al. 2014). The secretion system is a multicomponent machinery that spans the entire cell envelope, composed of a cytoplasmic ATPase, several inner membrane proteins, a periplasmic pseudopilus, and a secretin pore embedded in the outer membrane (McCallum, Burrows, and Howell 2019; Snively et al. 2014). This system is crucial for the survival of both extracellular and intracellular pathogens as well as environmental species of proteobacteria, as it secretes toxins and various hydrolytic enzymes, including proteases, lipases, and carbohydrate-active enzymes (Korotkov and Sandkvist 2019).

Many sequence and structural similarities exist between the T2SS and type IV pili, which suggests a common origin and argues for a pilus-mediated mechanism of secretion (Chernyatina and Low 2019).

Type III Secretion System (T3SS)

The T3SS is a highly conserved protein secretion system found in Gram-negative pathogens such as *Salmonella*, *Shigella*, *Yersinia*, and *Vibrio* (Abby and Rocha 2012; Yoshida, Frickel, and Mostowy 2017). It functions as an injectisome, transporting molecules directly from the bacterial cytoplasm into host cells, making it a potent virulence weapon (Abby and Rocha 2012). Attaching and effacing pathogens, such as EPEC, EHEC, and *Citrobacter rodentium*, rely on the T3SS for virulence, forming distinctive histological lesions in the intestinal epithelium (Gaytán et al. 2016).

The T3SS core architecture comprises a multi-ring basal body embedded in the bacterial membranes, a periplasmic inner rod, a transmembrane export apparatus, and cytosolic components. Two hollow appendages, a 23 nm needle and a filament extending up to 600 nm, create a channel for protein secretion and allowing the pathogens to penetrate the host cell's membrane barrier (Akedo and Galán 2005). Upon contact with target cells, a translocation pore assembles in the host membrane, allowing the passage of effector proteins (Büttner et al. 2006). Assembly of the T3SS is tightly regulated to ensure proper timing of substrate secretion, with hierarchical secretion determined by specialised chaperones, two molecular switches, and a sorting platform (Lara-Tejero et al. 2011).

Type IV Secretion System (T4SS)

Most T4SSs are involved in bacterial conjugation processes, which can be likened to bacterial sexual reproduction (Alvarez-Martinez and Christie 2009). The T4SS is capable of transporting both DNA and unfolded proteins and is responsible for introducing DNA into other organisms, as in the case of *Agrobacterium tumefaciens* (Alvarez-Martinez and Christie 2009). Human pathogens such as *Helicobacter pylori* and *Legionella pneumophila* also use the T4SS to secrete virulence factors (Llosa, Roy, and Dehio 2009).

T4SSs are divided into two large subfamilies: conjugation systems and effector translocators (Cascales and Christie 2003). Conjugation systems mediate interbacterial DNA transfer, leading to the rapid spread of antibiotic resistance genes in clinical settings (Juhas et al. 2009). Effector translocators are used by many Gram-negative bacterial pathogens to deliver virulence proteins to eukaryotic cells, modulating different physiological processes during infection (Karnholz et al. 2006).

Recently, considerable progress has been made in defining the structures of T4SS machine subunits and large machine subassemblies (Costa, Figueiredo, and Touati 2009). A DNA translocation route through the *Agrobacterium tumefaciens* VirB/VirD4 system was defined, and both intracellular (DNA ligand, ATP energy) and extracellular (phage binding) signals were shown to activate type IV-dependent translocation (Costa, Figueiredo, and Touati 2009).

Type V Secretion System (T5SS)

The T5SS, also known as the autotransporter system, operates as a two-step secretion system and exports adhesins, enzymes, toxins, and other virulence factors with varying sizes and structures (Salacha et al. 2010). Typically, T5SS substrates contain a signal peptide at the N-terminus mediating Sec-dependent transport across the inner membrane, a passenger domain exerting biological activity in the extracellular space, and a linker domain connecting the passenger and β -domain, which forms a β -barrel with a hydrophilic pore in the outer membrane (Salacha et al. 2010). In the periplasmic space of Gram-negative bacteria, energy is not available, and most secretion systems are energised through the presence of subunits in the inner membrane (Meuskens et al. 2019). The T5SS is unique because it lacks inner membrane components (Meuskens et al. 2019).

T5SS substrates reach the periplasm in an unfolded state, and they are protected by chaperones following their emergence from the Sec machinery (Dautin 2021). The Bam (Beta-barrel assembly machinery) complex catalyses their folding and insertion into the outer membrane (Dalbey and Kuhn 2012), facilitating exposure of the passenger domain at the extracellular side. Some Type V substrates remain anchored to the outer membrane through their β -domains, others have protease activity resulting in release of the passenger domain into the milieu (Van Ulsen et al. 2003).

The T5SS has been classified into several subtypes, with the main one being Type Va (Clantin et al. 2007). Some of the other subtypes, such as Type Vb (Two-Partner Secretion), have the β -domains and passenger domains as two separate polypeptides (Gawarzewski et al. 2013).

Type VI Secretion System (T6SS)

The T6SS was first discovered in 2006 in *Vibrio cholerae* and *Pseudomonas aeruginosa* (Filloux, Hachani, and Bleves 2008). The primary function of the T6SS is bacterial competition and defence against predation by unicellular eukaryotes. This system transports proteins from the bacterial cytoplasm directly inside a target cell

(often another Gram-negative bacterium) and is widely distributed, being found in plant and animal pathogens, marine organisms, and soil-dwelling bacteria (Filloux, Hachani, and Bleves 2008).

Protein secretion systems often evolve from pre-existing macromolecular complexes, and indeed two of the T6SS components (IcmF and DotU) are related to those found in the T4SS (Filloux, Hachani, and Bleves 2008). However, the bulk of the T6SS machinery evolved from contractile phage tails. It was recognised that the T6SS components, VgrG and Hcp, are homologous to the gp27/gp5 and gp19 bacteriophage proteins that form the contractile tail-tube structure (Mougous et al. 2006; Pukatzki et al. 2007; Pukatzki et al. 2006). The function of the phage gp27/gp5 is as a puncturing device, to create a hole in the outer membrane of the target cell and allow injection of the phage DNA through the gp19 tube (Leiman and Shneider 2012).

It is now clear that the T6SS functions as an inverted phage tail, forming a large contractile double-layered tube that is assembled in the cytoplasm and typically spans the width of the bacterial cell (Chang et al. 2017; Santin et al. 2019). The outer part of the tube, termed the sheath, is contractile and is made up of TssBC proteins. The inner part of the tube is rigid and formed from stacking rings of Hcp (Wang et al. 2017). The tip of the Hcp ring is topped with a VgrG protein which is capped by a PAAR protein (Shneider et al. 2013). Effector proteins are loaded into the Hcp ring or onto the VgrG/PAAR proteins and contraction of the sheath results in propulsion of the Hcp tube across the cell envelope and into a neighbouring cell, delivering its toxic cocktail of substrate proteins (Ho, Dong, and Mekalanos 2014).

Type VII Secretion System (T7SS)

The T7SS is present in the diderm bacteria including the deadly pathogens *Mycobacterium tuberculosis* and *Mycobacterium bovis* but is also abundantly found in Gram-positive bacteria including *Streptomyces coelicolor* and *Staphylococcus aureus*, among others (Bowman and Palmer 2021). Homologues of some components of the T7SS are encoded in the genomes of a few Gram-negative bacteria including

some strains of *Helicobacter pylori*, although to date it is not clear whether these are functional (Unnikrishnan et al. 2017).

This T7SS was first discovered during the study of *M. tuberculosis*, the main causative agent of tuberculosis, a significant infectious disease throughout history and still one of the top ten causes of death worldwide (Paulson 2013). The 6-kDa early secretory antigenic target (ESAT-6) and its co-secreted partner 10-kDa culture filtrate protein (CFP-10) are key antigens secreted by pathogenic mycobacteria. These have subsequently been renamed EsxA and EsxB, respectively (Cole et al. 1998). The *esxA-esxB* genes were found to localise within a gene cluster encoding a membrane-associated ATPase and other putative components of secretion systems (Rivera-Calzada et al. 2021). It was later demonstrated that the genes surrounding *esxA* and *esxB* encode an alternative secretion system, which has been named ESX (ESAT-6 secretion) (Brodin et al. 2004). Subsequently the more generic term "Type VII secretion system" (T7SS) was coined because it was argued that since mycobacteria have an outer membrane (albeit very different from the Gram-negative lipopolysaccharide-containing outer membrane) then it should be assigned a secretion number analogous to the Gram-negative secretion systems (Abdallah et al. 2007).

The T7SS has two main roles: mediating interbacterial competition and secreting virulence factors that act against eukaryotic hosts. The *Staphylococcus aureus* T7SS secretes two primary toxins: EsaD, a nuclease that degrades DNA, and TspA, a protein that forms pores in the membranes of other bacteria, although more toxins are suspected (Cao et al. 2016; Ulhuq et al. 2020). The actual secretion process is poorly understood, but common properties suggest the existence of a conserved secretion mechanism for the diverse T7SS substrates (Renshaw et al. 2005). T7SS substrates carry signal sequences required for transport, and typically form dimers that share a characteristic helix-turn-helix structure (Renshaw et al. 2005).

In the case of *M. tuberculosis*, T7SS substrates play a key role as virulence factors, allowing pathogenic mycobacteria to adapt and survive in various environments encountered at multiple stages of infection (Paulson 2013). Therefore, T7SS expression and activity have to be tightly regulated. T7SS substrates also have key

roles in bacterial fitness and survival, including DNA transfer and bacterial competition, and these roles, together with the contribution of tuberculosis substrates to the infectious process, are essential areas for further study (Bowman and Palmer 2021).

Type VIII Secretion System (T8SS)

The T8SS is responsible for the secretion and assembly of pre-pili for the biogenesis of fimbriae or curli and is also known as the nucleation-precipitation extracellular pathway (Chapman et al. 2002). Curli are extracellular polymeric amyloid fibres that are crucial for adhesion, biofilm formation, and colonisation of the host cell surface (Depluvere, Devos, and Devreese 2016).

Curli structure and biogenesis has been most thoroughly studied in *E. coli*. The major structural component of curli fibre is CsgA, which is exported across the inner membrane by the Sec pathway. CsgC is a periplasmic protein that prevents the toxic premature polymerisation of CsgA in the periplasm (Evans et al. 2015). CsgG is an outer membrane protein that forms a secretory channel in the bacterial outer membrane for the transport of CsgA across the bacterial outer membrane in an unfolded form (Depluvere, Devos, and Devreese 2016). The subunits subsequently polymerise on the extracellular surface facilitated by the extracellular lipoprotein CsgF (Nenninger, Robinson, and Hultgren 2009).

Type IX Secretion System (T9SS)

The T9SS is a complex secretion mechanism utilised by many members of the Bacteroidetes phylum, which are commonly found in the digestive tract of animals and humans and are widespread in the environment (Lasica et al. 2017; Veith et al. 2017; Gao et al. 2020; Kharade and McBride 2014; Narita et al. 2014). The T9SS secretes numerous effector proteins, such as proteases, adhesins, cellulases and surface layer proteins. It has a significant role in the periodontal pathogen *Porphyromonas gingivalis*, where it is responsible for the secretion of at least 30 proteins, including major virulence factors called gingipains (Sato et al. 2010; Sato et al. 2013; Grenier and Dang La 2011). Interestingly, the T9SS is also required for

gliding motility, with *Flavobacterium johnsoniae* being used as a model organism to study this type of motility (McBride 2019). The major motility adhesins, SprB and RemA, which are essential for gliding motility in *F. johnsoniae*, are delivered to the cell surface by the T9SS (McBride and Nakane 2015).

Protein substrates of the T9SS have N-terminal signal peptides for transport across the inner membrane by the Sec system and are targeted to the outer membrane translocon via their conserved C-terminal signal of approximately 80 amino acid residues (Shoji et al. 2018; Seers et al. 2006; Veith et al. 2013). The T9SS is composed of at least 19 protein components, which localise to the inner and outer membranes (Gorasia, Veith, and Reynolds 2020).

Type X Secretion System (T10SS)

The T10SS is a recently described secretion system that is related to bacteriophage lysis cassettes (Palmer et al. 2021; Mekasha and Linke 2021). This system is characterised by its utilisation of an inner membrane holin protein in conjunction with a cell wall-editing enzyme. These components work together to facilitate the transport of substrate proteins from the periplasm to the extracellular environment.

Specific elements of the TXSS can vary based on the biological system in which they are found. For instance, *Serratia marcescens*, uses a peptidoglycan endopeptidase enzyme in its chitinase secretion pathway. On the other hand, the secretion of Typhoid toxin in *Salmonella enterica* serovar Typhi is mediated by a muramidase (Palmer et al. 2021). Interestingly, the pairing of different families of holins with various types of peptidoglycan hydrolases suggests that this secretion pathway has evolved multiple times, demonstrating the adaptive nature of bacterial secretion systems (Palmer et al. 2021).

Type XI Secretion System (T11SS)

The T11SS is a novel secretion system found in several Gram-negative bacteria including various human pathogens such as *Neisseria meningitidis*, *Acinetobacter*

baumanii, *Haemophilus haemolyticus*, and *Proteus vulgaris*. It is involved in the transport of diverse proteins, including the hemophilin haemophore, which is essential for survival during haem starvation and facilitates nematode fitness in *Xenorhabdus nematophila* (Martens, Heungens, and Goodrich-Blair 2003). The lipidated symbiosis factor NilC, is transported across the inner membrane by the Sec pathway where it is lipidated and transported to the inner face of the outer membrane by the lipoprotein biogenesis pathway. NilC is subsequently surface exposed by the T11SS translocon NilB, an outer membrane β -barrel. Bioinformatic analysis has predicted 141 T11SS-dependent cargo proteins falling into 10 distinct architectures, including novel T11SS-dependent adhesins and glycoproteins (Grossman et al. 2021).

1.2.1 Cytoplasmic membrane protein transport systems: Sec and Tat

The general secretory (Sec) and twin arginine translocase (Tat) pathways are common protein transport pathways found in the cytoplasmic membranes of prokaryotes (Fig. 1.2). They translocate proteins across this membrane to the periplasm of Gram-negative bacteria, or to the extracellular space of Gram-positive bacteria and archaea. As they do not directly secrete proteins in Gram-negative bacteria, they have not been assigned a secretion number. However, it should be noted that they are *bona fide* secretion systems in many prokaryotes. These transport pathways also co-occur in the thylakoid membranes of plant chloroplasts (Zhu et al. 2022).

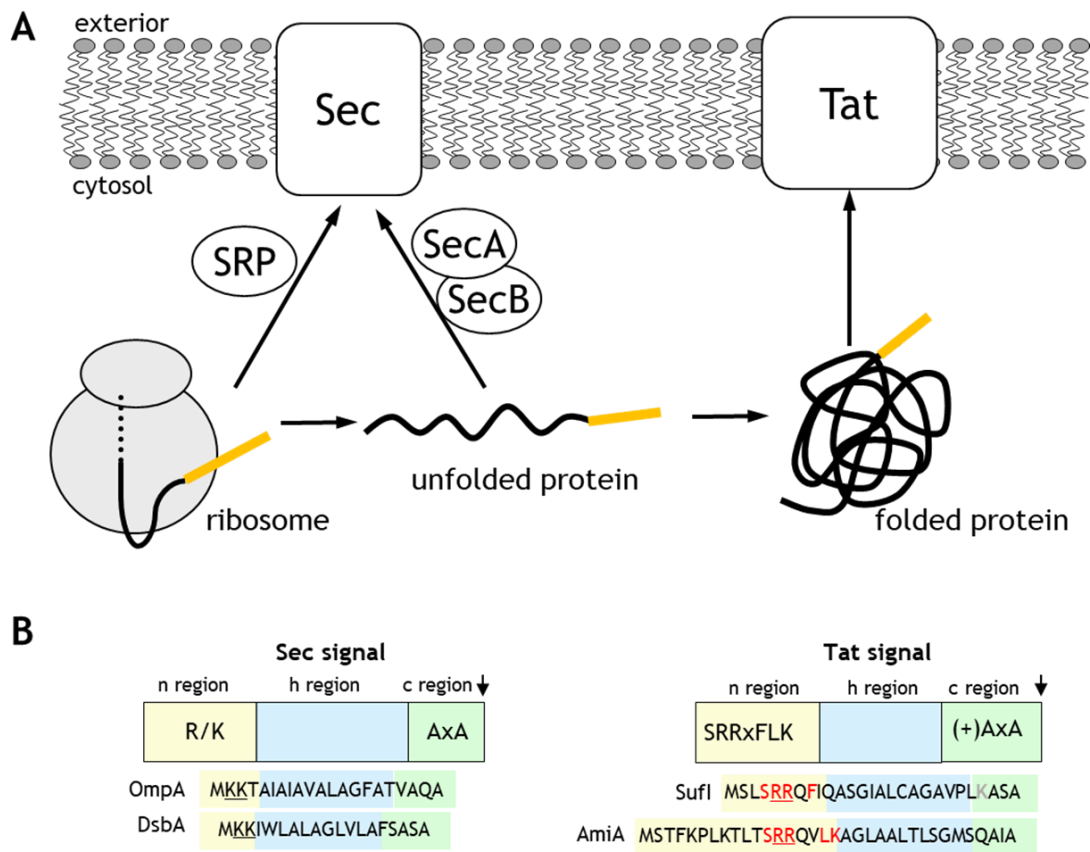


Figure 1.2. A. Schematic representation of the Sec and Tat protein transport pathways. B. Sec and Tat signal peptides. Signal peptide sequences of OmpA, that targets the post-translational Sec pathway, and DsbA, that targets Sec co-translationally, along with Tat-targeting SufI and AmiA signal sequences are shown. Positive charges in the signal peptide n-regions are shown in underline, and the amino acids matching the Tat consensus motif are shown in red. The arrow indicates the position of signal peptide cleavage. Adapted from Palmer and Stansfeld (Palmer and Stansfeld 2020).

The Sec pathway

The Sec system is ubiquitous in biology, being found in all prokaryotes and in the endoplasmic reticulum of eukaryotes, and it is essential for viability in all organisms tested. In bacteria the Sec pathway translocates proteins across the cytoplasmic membrane and it also mediates the biogenesis of most of the cytoplasmic membrane proteins.

Targeting of proteins to the Sec pathway occurs either co-translationally or post-translationally. Proteins destined for the post-translational translocation pathway are synthesised as precursor proteins with N-terminal signal peptides. These signal peptides do not share absolute sequence identity but are generally 18-26 residues in length and have a conserved tripartite structure (Fig. 1.2B). The n-region contains

at least one positive charge and is followed by a core hydrophobic (h-) region and a polar c-region. The c-region contains a recognition sequence for cleavage by signal peptidase. This is usually an AxA motif, where A is alanine and x is any amino acid (Chen, Shanmugam, and Dalbey 2019).

Proteins that use the co-translational pathway also have signal sequences at their N-termini. These have the same tripartite arrangement as signal peptides that target the post-translational pathway, but the h-region is more hydrophobic (De Gier et al. 1998). Co-translational signals may contain signal peptidase cleavage sites, for example the DsbA signal sequence which engages this route. However, more commonly they are uncleaved and serve as the first transmembrane segment of polytopic membrane proteins.

Post-translationally exported proteins are recognised by the molecular chaperone SecB during emergence from the ribosome. SecB binds the precursor protein in its mature region in order to prevent the protein from folding (Gannon, Li, and Kumamoto 1989; Bechtluft et al. 2007; Collier et al. 1988). SecB is able to interact with the motor component of the Sec pathway, SecA, which recognises SecB through conserved residues present in its extreme C-terminus (Fekkes, van der Does, and Driessen 1997). SecA subsequently binds to the precursor through both its mature region and its signal peptide, guiding the protein to the SecYEG complex. SecA uses the energy of ATP hydrolysis to drive translocation of the precursor through the Sec translocon (Chatzi et al. 2014). Once the signal sequence reaches the periplasm it can be cleaved by signal peptidase and released into the periplasm.

E. coli uses the Sec pathway for around 96% of its exportome (Tsirigotaki et al. 2017). Most lipoproteins are also exported by the post-translational Sec pathway and are targeted via tripartite signal peptides that are very similar to those cleaved by signal peptidase. However, they contain a conserved lipobox sequence in their c-region with an invariant cysteine that is recognised and cleaved by lipoprotein signal peptidase after the cysteine residue has been fatty acylated by prelipoprotein diacylglycerol transferase (El Rayes, Rodríguez-Alonso, and Collet 2021).

In contrast, the signal recognition particle (SRP) pathway operates co-translationally, recognising and targeting the nascent protein as it is being synthesised by the ribosome, and facilitating its translocation across or integration into the membrane (Green and Meccas 2016). Here SRP binds the hydrophobic signal sequence as it emerges from the ribosome, leading to the formation of the Ribosome Nascent chain (RNC) complex. This complex is targeted to the membrane through the interaction of SRP with the SRP receptor, FtsY, a bacterial peripheral membrane protein. FtsY interacts with the SecY component of the Sec translocon, thus targeting the RNC complex to the Sec machinery (Kuhn et al. 2011; Angelini, Deitermann, and Koch 2005). The ribosome exit tunnel lines up with the SecY channel allowing translocation of the emerging polypeptide directly into the Sec channel (Menetret et al. 2007). Nascent polypeptide insertion triggers opening of a lateral gate in SecY, allowing the hydrophobic segment to escape into the membrane (Egea and Stroud 2010).

YidC is a membrane protein insertase that is also found throughout prokaryotes. It has been best characterised in *E. coli*, where it is essential for viability (Samuelson et al. 2000). YidC alone catalyses the insertion of small membrane proteins, usually comprising just one or two hydrophobic helices. However, it also forms part of a supercomplex with SecYEG and SecDF-YajC, and crosslinks to hydrophobic segments of polytopic membrane proteins during their insertion into the membrane (Scotti et al. 2000; Schulze et al. 2014). It has been proposed that YidC promotes the removal of hydrophobic segments from the SecY channel, facilitating their integration into the lipid bilayer and acting as an assembly site for multi-spanning TMD proteins (Zhu et al. 2012; Urbanus et al. 2001; Beck et al. 2001). The essentiality of YidC arises from its essential role in assembling some protein complexes that are functionally involved in respiration (van der Laan, Nouwen, and Driessen 2005).

1.2.2 The Tat system

The Tat system is a protein export pathway found in prokaryotes, including bacteria and archaea, as well as in chloroplasts, mitochondria of plants and some unicellular eukaryotic organisms, and in homoscleromorph sponges (Pett and Lavrov 2013). The Tat system is distinct from the Sec system in that it can transport fully folded

proteins. In *E. coli*, over 450 exported proteins utilise the Sec pathway, while only 41 use the Tat pathway (Palmer, Sargent, and Berks 2010; Huang and Palmer 2017). Of these 41 proteins, 31 have signal peptides, and 10 are co-exported with partners, as shown in Table 1.1. Despite the smaller number of proteins using the Tat pathway, these proteins are of different folded size and are exported fully folded without compromising the integrity of the cytoplasmic membrane.

The Tat system plays essential roles in many bacterial cellular processes, such as respiratory and photosynthetic energy metabolism, iron and phosphate acquisition, cell division, cell motility, quorum sensing, organophosphate metabolism, resistance to heavy metals and antimicrobial peptides, and symbiotic nitrogen fixation (Palmer and Berks 2012). It is crucial for virulence of many pathogenic bacteria, including *M. tuberculosis* in which is also essential for survival (Saint-Joanis et al. 2006). The system is also essential for survival under standard laboratory condition in a few other organisms, including some haloarchaea (Dilks, Gimenez, and Pohlschroder 2005; Lazarus et al. 2009; Thomas and Bolhuis 2006), the soil bacterium *Ensifer meliloti* (Kuzmanović et al. 2022; Pickering and Oresnik 2010), some pathogenic *Brucella* (Riquelme et al. 2023) and *Bdellovibrio bacteriovorus* (Chang et al. 2011a). It is dispensable for *E. coli* viability during aerobic respiration, facilitating study of the system (Weiner et al. 1998; Sargent, Bogsch, et al. 1998).

Table 1.1. The table presents the known or probable *E. coli* Tat substrate proteins. Proteins that bind iron-sulphur clusters (Fe-S), molybdopterin (MPT), molybdopterin guanine dinucleotide (MGD), molybdopterin cytosine dinucleotide (MCD), or MPT-tungstate cofactors are necessarily targeted via the Tat pathway because they fold in the cytoplasm. Among the remaining *E. coli* proteins, the export of CueO (formerly Yack), SufI, MdoD, AmiA, AmiC, FhuD, Pac, C3736, and EfeB (formerly YcdB) has been experimentally shown to require the Tat pathway. The signal peptides of YahJ, WcaM, and YcbK have been demonstrated to engage in the Tat pathway when fused to a Tat-specific reporter protein, and YaeI was similarly shown to engage in the Tat pathway when tested in a *Streptomyces* agarase assay. Note that PcoA is plasmid-encoded and related to CueO. Pac (penicillin acylase), is found in *E. coli* W strains. Interestingly, C3736 is encoded in the genomes of most strains of *E. coli*, but it appears to be inactive in the K12 strains due to two recent frame-shifts. Updated from (Palmer, Sargent, and Berks 2010).

Protein	Physiological role	Cofactors	Co-exported partner?	Tat-dependent Membrane Protein?
HyaA	Hydrogen oxidation	3 x Fe-S clusters	HyaB (Ni-Fe cofactor)	Yes
HybO	Hydrogen oxidation	3 x Fe-S clusters	HybC (Ni-Fe cofactor)	Yes
HybA	Hydrogen oxidation	4 x Fe-S clusters	Unknown	Yes
NapG	Nitrate reduction	4 x Fe-S clusters	Unknown	No
NrfC	Nitrite reduction	4 x Fe-S clusters	Unknown	No
YagT	Aldehyde oxidoreductase	2 x Fe-S clusters	YagR (MCD) YagS (FAD)	No
YdhX	Component of aldehyde ferredoxin oxidoreductase?	4 x Fe-S clusters	YdhV? (MPT-tungstate)	No
TorA	TMAO reduction	MGD	None	No
TorZ	TMAO reduction	MGD	None	No
NapA	Nitrate reduction	MGD, 1 x Fe-S cluster	None	No
DmsA	DMSO reduction	MGD, 1 x Fe-S cluster	DmsB (4 x Fe-S clusters)	No
YnfE	DMSO reduction	MGD, 1 x Fe-S cluster	YnfG (4 x Fe-S clusters)	No
YnfF	DMSO reduction	MGD, 1 x Fe-S cluster	YnfG (4 x Fe-S clusters)	No
FdnG	Formate oxidation	MGD, 1 x Fe-S cluster	FdnH (4 x Fe-S clusters)	Yes
FdoG	Formate oxidation	MGD, 1 x Fe-S cluster	FdoH (4 x Fe-S clusters)	Yes
YedY	TMAO/DMSO reduction?	MPT	None	No
CueO	Copper homeostasis	4 x Cu ions	None	No
PcoA	Copper resistance	4 x Cu ions	None	No
SufI	Cell division	None	None	No
YahJ	Putative deaminase	1 x Fe ion	None	No
WcaM	Colanic acid biosynthesis	Unknown	None	No

MdoD	Periplasmic glucans biosynthesis	Unknown	None	No
EfeB	Deferrochelataase	Haem <i>b</i>	None	No
YaeI	Possible phosphodiesterase	Unknown	None	No
AmiA	Cell wall amidase	None	None	No
AmiC	Cell wall amidase	None	None	No
FhuD	Ferrichrome binding	None	None	No
YcbK	Unknown - peptidase M15 superfamily	Unknown	None	No
Pac	Penicillin amidase	Ca ²⁺	None	No
C3736	Possible diene lactone hydrolase	Unknown	None	No
FecR	Iron di-citrate sensor	Unknown	None	Yes

1.2.3 Tat signal peptides

Tat substrates are targeted for export by the presence of N-terminal signal peptides. Tat signal peptides share the same tripartite structure as Sec signals, including the polar n-region, hydrophobic h-region, and polar c-region with the signal peptidase cleavage site (Fig. 1.2B). However, they are distinct due to the presence of two consecutive arginine residues in the n-region, forming part of the consensus motif SRRxFLK, where x can be any amino acid but it usually polar (Berks, Sargent, and Palmer 2000; Stanley, Palmer, and Berks 2000; Joshi et al. 2010). They also tend to be longer, usually due to a longer and more charged n-region (Tjalsma et al. 2000), and they have a less hydrophobic h-region (Cristobal et al. 1999) compared to Sec signal peptides. Tat signal peptides frequently contain a basic amino acid in their c-region or close to the N-terminus of the mature protein; known as a ‘Sec-avoidance’ signal it is not required for recognition by the Tat pathway but serves to prevent mis-targeting to Sec (Cristobal et al. 1999; Tooke et al. 2017; Tullman-Ercek et al. 2007). This is critical because the folded domains of Tat substrates would block the Sec pathway which would be highly deleterious to the cell.

The twin-arginine residues in Tat signals are conserved, and essential for Tat transport. Mutation studies show that substitutions of the arginines, even for conserved lysines drastically affects recognition by the Tat system. (Chaddock et al. 1995; DeLisa et al. 2002; Mendel et al. 2008; Stanley, Palmer, and Berks 2000). The

other motif residues are less highly conserved, and their substitution has less dramatic phenotypes, although the presence of a hydrophobic residue at the 'F' position is important for efficient transport (Huang and Palmer 2017; Stanley, Palmer, and Berks 2000). Interestingly, recent studies have demonstrated that the twin-arginine motif is not mechanistically essential for the operation of the Tat pathway. Inactivating substitutions in either the paired arginines or the binding site on the Tat translocon can be overcome by increasing the hydrophobicity of the signal peptide h-region (Huang et al. 2017; Ulfing et al. 2017; Palmer and Stansfeld 2020). It has been proposed that the lower hydrophobicity of Tat signal peptides is a key mechanism to avoid interaction with the Sec pathway (with which it always co-occurs in biological systems), and that the twin-arginine motif and its cognate recognition site on the Tat translocon, are a mechanism to increase its affinity for the weakly hydrophobic signal peptides (Huang and Palmer 2017; Palmer and Stansfeld 2020).

Several Tat signal peptide prediction software programmes exist, such as TatFind and TatP (Bagos et al. 2010; Bendtsen et al. 2005). However, taken in isolation they are not always reliable predictors of targeting pathway given the overlapping features found in Sec and Tat signals - this can lead to both false-positive and false-negative identification (Tjalsma et al. 2000; Jongbloed et al. 2002; Kouwen et al. 2009; Keller et al. 2012a). Therefore, while bioinformatics tools are invaluable for finding candidate substrates, experimental confirmation is necessary to verify potential Tat substrates. Moreover, it should be noted that some proteins without their own signal peptide bind to a Tat substrate with a Tat signal peptide and are co-exported as a complex (Wu et al. 2000) (Table 1.1). Examples of some 'hitchhiking' substrates will be discussed below.

1.2.4 Tat components

Tat transport is carried out by integral membrane proteins from the TatA and TatC families. In some organisms, just TatA and TatC components are sufficient for transport, while in others, an additional member of the TatA family, termed TatB, is required (Sargent, Bogsch, et al. 1998; Sargent et al. 1999; Palmer and Berks 2012).

TatA and TatA-like proteins are small membrane proteins characterised by their N-out C-in topology (Fig. 1.3). They possess a single very short transmembrane helix (TMH), and an amphipathic helix (APH) that, due to the short TMH, is partially embedded in the membrane on the cytoplasmic side. In various species, *tatA* genes have undergone multiple duplication events, giving rise to *tatA*-like genes found alongside *tatA* in the *tat* gene cluster or elsewhere in the genome (Wu et al. 2000; Yen et al. 2002). Some of these duplicates, such as *E. coli* TatE or *C. jejuni* TatA2, retain similar functions to TatA proteins (Sargent et al. 1999; Datta et al. 2001; Baglieri et al. 2011; Kikuchi et al. 2006; Goosens and van Dijl 2017).

However, other duplicate TatA-like proteins, referred to as TatB, have diverged in sequence and function (Sargent et al. 1999; Mori, Summer, and Cline 2001; Schaerlaekens et al. 2001; De Keersmaecker et al. 2005). For instance, in *E. coli*, *Streptomyces* species, and plant chloroplast thylakoids, TatA and TatB are both individually essential for translocation activity. TatB proteins are bigger, and have a longer APH, but also have a very short TMH, similar to TatA (Fig. 1.3).

Some Tat systems, particularly those in Gram-positive bacteria such as *Bacillus subtilis* and in Archaea require only a single TatA-like protein, which appears to be bifunctional, serving the role of both TatA and TatB (Hicks et al. 2003; Barnett et al. 2008). As a result, there is no clear definition that differentiates TatA from TatB, and some sequence annotations may be erroneous.

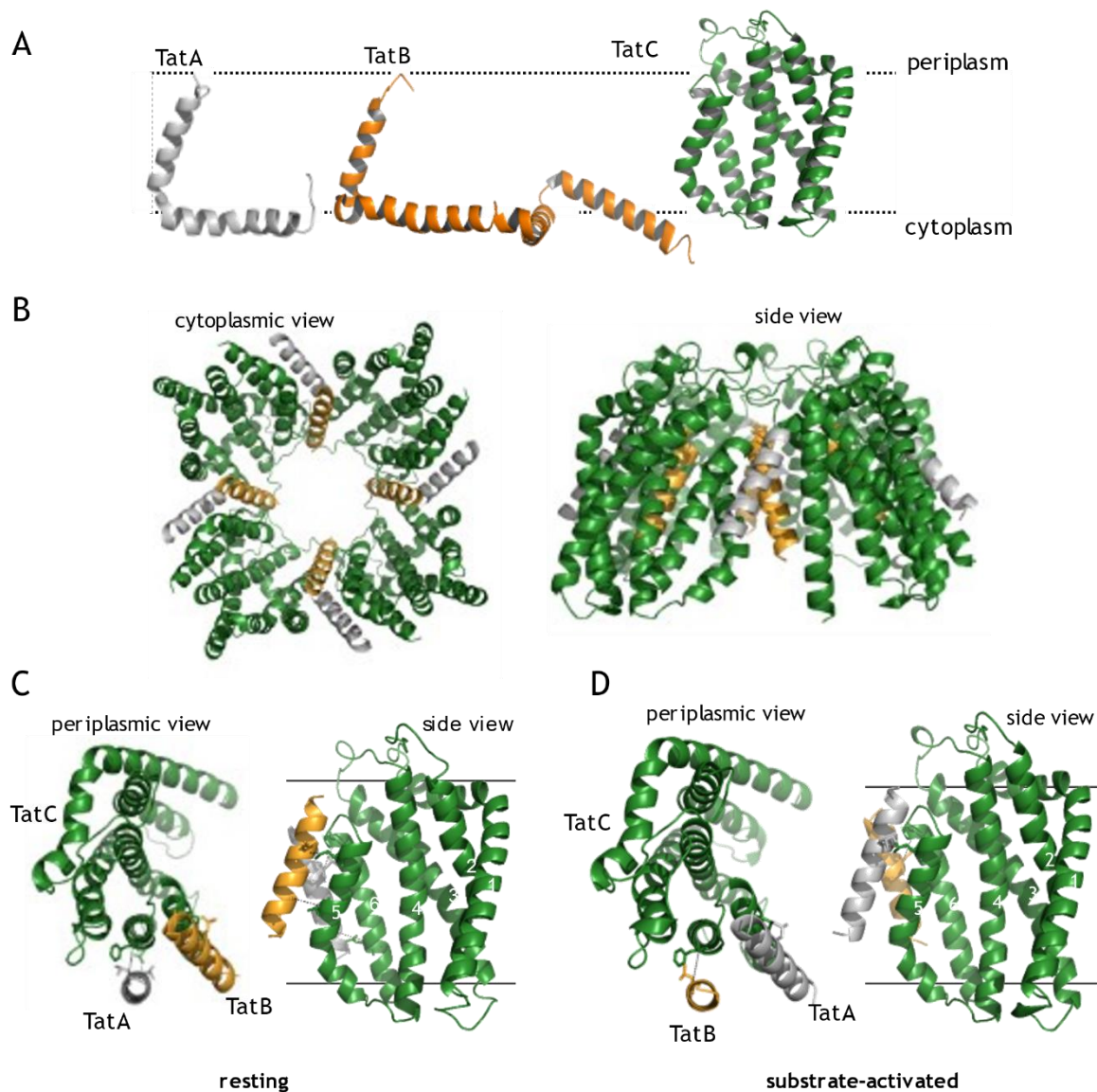


Figure 1.3. A. Structures of the Tat components (Ramasamy et al. 2013; Rollauer et al. 2012; Rodriguez et al. 2013; Zhang et al. 2014). Note that the unstructured C-terminal regions of TatA and TatB are not shown. B. Model for the resting state multimeric TatABC receptor complex. TatA is shown in silver, TatB in gold and TatC in green. The APHs of TatA and TatB are omitted. C and D. Periplasmic and side views of C. the resting TatABC heterotrimer and D. the substrate-activated TatABC heterotrimer. Adapted from (Habersetzer et al. 2017).

The other main components of the Tat system are the TatC proteins - polytopic integral membrane proteins that recognise substrate proteins and act as a scaffold for the assembly of the active Tat translocon. TatC proteins have six TMH, with TMH5 and TMH6 being notably short and barely spanning the membrane (Rollauer et al. 2012). Molecular dynamic simulations indicate that the short TMH of TatA, TatB and TatC will result in thinning of the membrane bilayer, which is likely to be crucial for the function of the Tat system (Rollauer et al. 2012; Rodriguez et al. 2013).

TatC proteins are the primary recognition elements for Tat substrates, specifically recognising the twin-arginine motif through a negatively-charged patch on the cytoplasmic surface (Rollauer et al. 2012). This recognition is essential for the translocation process. Although the extracytoplasmic loops of TatC proteins do not exhibit high sequence similarity, their conserved secondary structure is vital, as demonstrated by random mutagenesis studies (Strauch and Georgiou 2007; Kneuper et al. 2012). These loops form a periplasmic cap, which is crucial to stabilise the linear arrangement of the TatC transmembrane helices. Notably, TatC functions as a multimer, and the oligomeric arrangement of TatC is mediated through the periplasmic cap region, which forms extensive contacts with the same region of a neighbouring TatC (Cleon et al. 2015) (Fig 1.3B). In addition to forming homo-oligomers (Buchanan et al. 2002; Punginelli et al. 2007), TatC also interacts with TatA and TatB. This will be discussed further in the next section.

Over 50% of the Tat-encoding genomes sequenced to date specify only TatA and TatC components (Simone et al. 2013). This supports the idea of a core Tat system comprising a TatA-TatC pair. However, relatively little is known about the TatAC only systems and the remainder of this Chapter will focus on the TatABC systems.

1.2.5 Tat mechanism

The Tat pathway is unique in its ability to maintain the membrane's permeability barrier while allowing passage of large, fully folded protein molecules. Tat transport is driven solely by the protonmotive force and does not require ATP hydrolysis (Mould and Robinson 1991; Yahr and Wickner 2001). It has been estimated that each protein translocated by the Tat system requires the transport of around 10^5 protons from the proton gradient, equivalent to the energy stored in approximately 10^4 molecules of ATP (Driessen 1992; Palmer and Berks 2012).

In its resting state, the TatABC receptor complex, also known as the 'receptor complex', exists as a heterotrimer of TatA, TatB and TatC. The receptor complex is multimeric, comprising several copies of each component in a 1:1:1 ratio (Bolhuis et al. 2001; Alcock et al. 2016; Zoufaly et al. 2012). In the resting state receptor, TatB is bound to TMH5 of TatC, mediated by interaction between a cluster of polar

residues between the C-terminal end of TatC TMH5 and the start of TMH6 and a polar residue in the TMH of TatB (Habersetzer et al. 2017; Alcock et al. 2016). Interestingly, a second binding site on TatC at TMH6 has been identified, which TatA occupies under resting conditions (Fig. 1.3B,C). However, the functional relevance of this second site is currently less clear (Habersetzer et al. 2017; Severi et al. 2023).

Initiation of the Tat transport cycle involves the interaction of the Tat receptor with a signal peptide, where the conserved twin-arginine motif initially binds to the charged patch on TatC (Rollauer et al. 2012) (Fig. 1.4). The signal peptide then transitions to bind more deeply within the receptor, making extensive contacts with TatB TMH (Alami et al. 2003; Blummel et al. 2015). This is followed by a rearrangement at the TMH5 binding site, with TatA replacing TatB (Alcock et al. 2016; Habersetzer et al. 2017), forming the activated receptor complex (Fig. 1.3D). The switching of TatA into the TMH5 site allows further molecules of TatA to be recruited from the membrane to assemble onto the activated receptor (Fig. 1.4) (Mori and Cline 2002; Dabney-Smith, Mori, and Cline 2006; Alcock et al. 2013; Aldridge et al. 2014). The precise number of TatAs recruited to the receptor is not known but is estimated to be in the region of 20 - 30 (Leake et al. 2008). The assembled TatA oligomer facilitates transport of the folded substrate across the

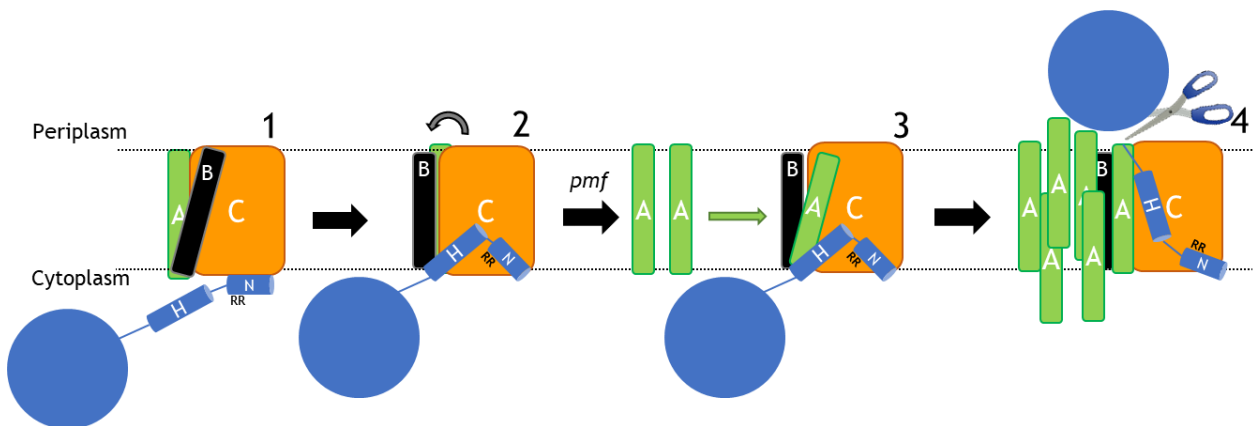


Figure 1.4. Proposed model for the Tat transport cycle. The resting state TatABC receptor complex is a heterotrimer, with TatB and TatA bound to TatC at distinct sites. The cycle initiates when a substrate protein bearing a Tat signal peptide binds to the cytoplasmic surface of TatC (step 1), triggering its deeper insertion into TatC (step 2). This displacement of the signal peptide forces TatB to relocate to the original TatA binding site. This change allows the recruitment of a TatA molecule, followed by the subsequent assembly of additional TatA molecules into a large oligomer (step 3). The substrate protein translocates across the membrane through this TatA oligomer (step 4). The signal peptide is then cleaved, releasing the mature protein domain into the periplasmic space, and the TatA oligomer disassembles, resetting the Tat receptor complex to its resting state. Adapted from (Huang et al. 2017).

membrane by an unknown mechanism, but potentially through localised weakening of the cytoplasmic membrane (Bruser and Sanders 2003; Rodriguez et al. 2013). Following translocation, the TatA oligomer disassembles, returning the system to its resting state, ready for another round of transport.

Despite our growing understanding of the Tat transport cycle, numerous questions remain. Some of these include the precise energy requirements, the process of membrane sealing in the absence of a substrate, the role of the PMF in driving protein transport, the exact mechanism of substrate translocation, and the functional significance of the second TatA and TatB binding site on TatC TMH6 (Habersetzer et al. 2017; Severi et al. 2023). Further studies, including extensive mutagenesis and biochemical analysis of variant Tat receptor complexes, are needed to address these remaining mysteries.

1.2.6 Tat-dependent membrane proteins

In addition to exporting fully folded proteins across the bacterial cytoplasmic membrane the Tat system is also capable of integrating some classes of membrane proteins into the bilayer. Tat-dependent membrane proteins can be assigned to four main categories: signal anchor proteins, bitopic membrane proteins, polytopic membrane proteins that use both Sec and Tat, and tail-anchored proteins (Fig. 1.5). Each of these classes has unique characteristics and functions, which are discussed below.

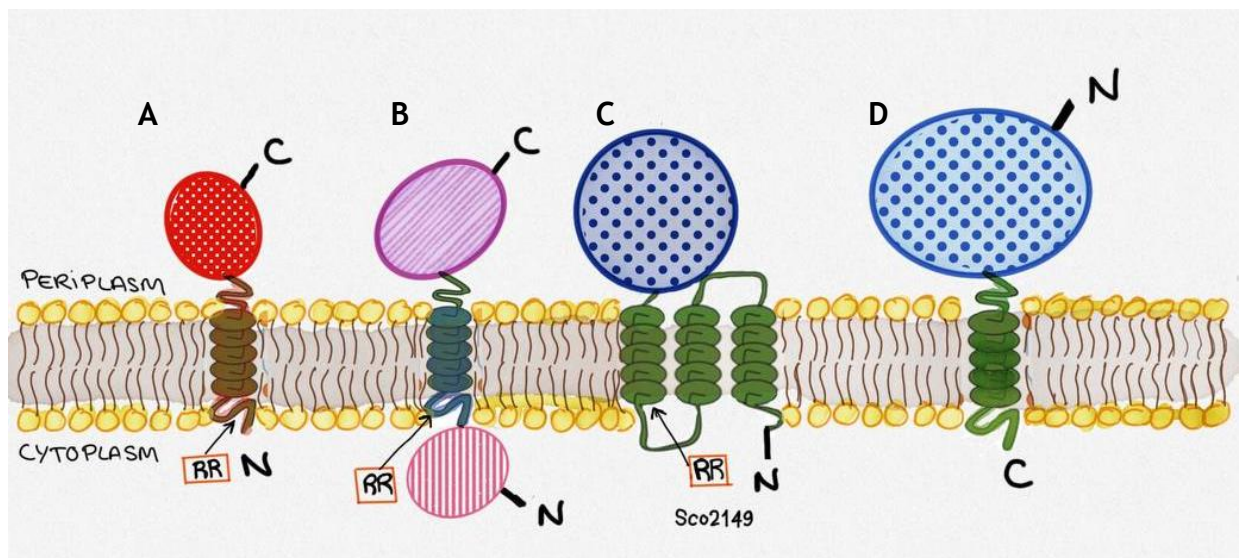


Figure 1.5. Four classes of Tat-dependent membrane proteins. Tat dependent membrane proteins with A. a non-cleaved N-terminal signal anchor B. a bitopic N-in arrangement with an internal signal peptide C. a polytopic protein where the final TMH is Tat-dependent and has a twin arginine motif and D. a C-terminal transmembrane helix (C-tail).

Signal anchor proteins: Signal anchor proteins, such as the single transmembrane Rieske iron-sulphur proteins from *Paracoccus* and *Bacillus*, are anchored to the membrane via their N-terminal twin arginine signal sequences (Bachmann et al. 2006; De Buck et al. 2007). As this sequence is not cleaved off during or after translocation, it is bifunctional, serving as both a signal for interaction with the Tat machinery and a membrane anchor. The Rieske proteins are the best studied example of this class and play a crucial role in electron transport, including the photosynthetic electron transport chain in plant chloroplasts (Molik et al. 2001).

Bitopic membrane proteins: Bitopic membrane proteins have a single transmembrane domain which acts as an internal twin arginine signal sequence. This is the most recently discovered class of Tat-dependent membrane proteins and only one example has been described to date - FecR, a protein involved in iron regulation in bacteria (Braun and Hantke 2020; Passmore et al. 2020).

Polytopic membrane proteins that use Sec and Tat: A very small fraction of polytopic membrane proteins utilise both the Sec and Tat pathways for their insertion into the membrane. Four examples of this class of protein have been discovered - the actinobacterial Rieske proteins, an actinobacterial molybdenum

cofactor-containing protein, a polyferredoxin and a polytopic YufE family protein (Keller et al. 2012; Tooke et al. 2017). The best characterised of these is the *Streptomyces coelicolor* Rieske protein, which has three TMHs at its N-terminus. The first two are inserted by the Sec pathway whereas the third resembles a Tat signal sequence and is recognised by the Tat pathway for translocation of the folded C-terminal iron-sulphur cluster-containing domain across the membrane (Keller et al. 2012). Intricate dissection of its biogenesis has shown that it is a combination of low hydrophobicity of the Tat-dependent TMH coupled with multiple basic amino acids close to the c-region that promote release of this TMH from the Sec machinery to allow recognition and insertion by the Tat pathway (Tooke et al. 2017).

Tail-anchored proteins: Tail-anchored proteins are a unique class of membrane proteins that are characterised by the presence of a single transmembrane domain near the C-terminus. Tail-anchored proteins are critical subunits of the periplasmic formate dehydrogenases and hydrogenases, are inserted into the membrane post-translationally by the Tat pathway (Table 1.1) (Hatzixanthis, Palmer, and Sargent 2003).

1.2.7 Hydrogenases.

First described by Stephenson and Stickland in 1931), hydrogenases are specialised in metabolising hydrogen, catalysing the reversible cleavage of molecular hydrogen into protons and electrons (Bowman, Palmer, and Sargent 2013; Bowman et al. 2014). These enzymes are characterised by metal-containing active sites essential for hydrogen activation. Depending on the metal cofactor present, hydrogenases are classified into three classes: [NiFe]-, [FeFe]-, and [Fe]-hydrogenases (Vignais and Billoud 2007).

All *E. coli* hydrogenases are [NiFe]-containing enzymes, and *E. coli* K12 encodes four of them (Benoit et al. 2020). Two of these, HYD-1 and HYD-2, have their active sites in the periplasm and are translocated across the membrane by the Tat pathway (Sargent, Bogsch, et al. 1998). HYD-1 and HYD-2 are arranged similarly - they comprise a large subunit containing the [NiFe] active site and a small subunit with 3 iron-sulphur clusters. A Tat signal peptide is found on the small subunit, whereas the

large subunit lacks a signal peptide and is transported across the membrane because it forms a complex with the small subunit in the cytoplasm before export - an example of the hitchhiker mechanism (Rodrigue et al. 1999). Cytoplasmic chaperones dedicated to each hydrogenase enzyme orchestrate insertion of cofactors and dimerisation, preventing interaction with the Tat pathway until these processes have been completed (Jack et al. 2004; Dubini and Sargent 2003; Sargent 2007).

The small subunits of HYD-1 and HYD-2 also harbour a C-terminal TMH that anchors the enzymes to the periplasmic face of the membrane. Apart from acting as membrane anchors, the C-tails also interact with the cytochrome *b* subunit of the hydrogenase enzymes (Volbeda et al. 2013). The HYD-2 enzyme contains an additional iron-sulphur protein component, HybA, that is not found in HYD-1 enzymes. HybA has its own Tat signal peptide and is transported separately from the large and small subunits. It also has a C-terminal TMH for membrane anchoring (Hatzixanthis, Palmer, and Sargent 2003; Sargent, Ballantine, et al. 1998).

1.2.8 Formate dehydrogenases.

Formate dehydrogenases (FDHs) catalyse the two-electron oxidation of formate to carbon dioxide, playing a vital role in cellular formate metabolism and the global carbon cycle. These enzymes are found across a broad spectrum of organisms, including bacteria, yeasts, and plants. Further subdivided into respiratory FDHs, which are coupled to the respiratory chain, and cytoplasmic FDHs, which function independently, formate dehydrogenases showcase functional diversity (Jormakka et al. 2002).

Two periplasmic FDHs are found in *E. coli*. The best-studied example is the nitrate-inducible formate dehydrogenase, FDH-N, a critical enzyme complex in *E. coli*. Under anaerobic conditions when nitrate is available, FDH-N, encoded by the *fdnGHI* operon, is induced. This complex comprises three subunits: FdnG, the multi-cofactor catalytic subunit which harbours a molybdenum cofactor and iron-sulphur clusters alongside a selenocysteine residue at the active site; FdnH, an iron-sulphur protein and FdnI, a cytochrome which participates in the electron transport process. The FDH-N complex significantly contributes to the metabolism and energy production of

E. coli, particularly under anaerobic conditions when nitrate serves as an electron acceptor (Jormakka et al. 2002; Sargent 2007; Berks, Palmer, and Sargent 2003).

The biogenesis of FDH-N is dependent on the Tat pathway (Sargent, Bogsch, et al. 1998). The FdnG subunit has a twin arginine signal peptide that directs transport of a heterodimer of FdnG with the iron-sulphur protein FdnH. FdnH lacks its own signal peptide, and its FdnG-dependent translocation is a further example of Tat 'hitchhiking' (Stanley et al. 2002). FdnH has a C-terminal TMH that anchors the heterodimer to the periplasmic side of the inner membrane (Jormakka et al. 2002). The third subunit of the complex, FdnI is a membrane-embedded cytochrome *b* that is inserted by the Sec pathway (Jormakka et al. 2002; Berks, Palmer, and Sargent 2003).

The second periplasmic formate dehydrogenase expressed by *E. coli* is the aerobic formate dehydrogenase (FDH-O) (Benoit, Abaibou, and Mandrand-Berthelot 1998). Although it has been designated an aerobic enzyme, it is synthesised constitutively and is present at low levels under all growth conditions (Benoit, Abaibou, and Mandrand-Berthelot 1998). It is highly related to FDH-N at the amino acid sequence level, but at present its biological role is still unclear.

1.2.9 Biogenesis of Tat-dependent tail-anchored proteins.

Hatzixanthis, Palmer, and Sargent (2003) undertook a study to investigate the integration of tail-anchored membrane proteins by the Tat pathway. The authors fused the C-terminal TMH regions from each of the five predicted Tat-dependent C-tail proteins, FdnG, FdoG, HyaA, HybA and HybO, to the soluble Tat substrate, Sufl, showing that each sequence was able to mediate Sufl anchoring to the inner membrane. As no periplasmic intermediate was detected for any of these constructs it was deduced that the TMH was directly integrated into the membrane by the Tat pathway, thus acting as a stop transfer sequence, rather than being fully exported across the membrane and integrating from the extracytoplasmic face (Hatzixanthis, Palmer, and Sargent 2003). Importantly it was shown that membrane integration was YidC-independent, making it distinct from the mechanism of membrane integration of other bacterial inner membrane proteins. This thesis will focus on the

identification of novel tail-anchored Tat substrates and developing tools to investigate their membrane integration.

1.3 Computational approaches for membrane protein analysis.

1.3.1 Protein modelling using artificial intelligence

Protein modelling using artificial intelligence (AI) has emerged as a powerful tool in membrane protein analysis. AI-based methods, such as deep learning algorithms, can predict the tertiary structure of proteins and help understand their functions. These techniques can significantly reduce the time and resources required for experimental methods, allowing researchers to investigate membrane proteins more efficiently and accurately. AI-driven protein modelling also aids in the discovery of novel drug targets and the development of new therapeutic strategies.

AlphaFold and RoseTTAFold (Jumper et al. 2021; Baek et al. 2021) represent significant advancements in the field of protein structure prediction, utilising deep learning approaches to outperform traditional methods. Rather than relying on physically-based models with their inherent assumptions about atomic interactions, these AI-based systems leverage vast numbers of parameters learned directly from training data, comprising tens of thousands of experimentally-determined protein structures. These algorithms are trained not just on individual amino acid sequences but on alignments of many homologous sequences, which allows them to derive rich structural information from evolutionary data. Interestingly, RoseTTAFold has demonstrated the ability to predict structures of de-novo-designed proteins using single amino acid sequences, suggesting a deep understanding of protein sequence-structure relationships embedded in the model itself (Baek and Baker 2022).

The combined use of RoseTTAFold and AlphaFold has proven especially powerful, enhancing the accuracy of protein-protein complex structure predictions. This integrated approach has enabled comprehensive proteome-scale modelling of protein-protein interactions, providing valuable insights into biological functions (Drake, Seffernick, and Lindert 2022). In terms of protein design, the capabilities of

these AI models have been leveraged to generate novel proteins. The 'inversion' of deep learning structure prediction networks has opened up new avenues for protein design, with the potential to increase the complexity of proteins that can be crafted (Baek and Baker 2022).

While these advancements are impressive, challenges remain. Notably, deep learning methods are data-hungry, requiring large and information-rich datasets for accurate model training. For areas with limited training data, the future may see a blend of deep learning with physically-based models, like Rosetta, to maximise predictive power.

1.4. Aims of this thesis

The overall goal of this thesis was to provide new information about the distribution and assembly of Tat-dependent tail-anchored membrane proteins. In Chapter 3, a bioinformatics approach will be employed to facilitate the identification and classification of novel Tat-dependent tail-anchored proteins. In Chapter 4 this bioinformatic study will be followed up by experimental approaches to verify that candidate proteins identified computationally are Tat-dependent and have a C-terminal membrane anchor. In Chapter 5, a synthetic approach will be used to design and test genetic constructs as tools to investigate the mechanism of Tat-dependent C-tail integration.

Chapter 2. Materials and Methods

2.1 Bacterial strains

The strains used in this study are derived from *Escherichia coli* K12 and are listed in Table 2.1. The strains XL10-gold and DH5 α were used for transformation of plasmid vectors and cloning procedures. NRS-3 is an MC4100 derivative that was used for experiments involving detection of the protein SufI. HS25 was used for all experiments involving maltose-binding protein (MalE) as this lacks the *malE* gene. For all the work related with β -lactamase-encoding constructs the strains used were MC4100 and DADE. The strain MC4100 Δ *amiA* Δ *amiC* was chosen for all experiments involving AmiA-encoding constructs.

Table 2.1. Bacterial strains used in this study.

Strains	Genotype	Resistance	Reference
MC4100	<i>F-ΔlacU169 araD139 rpsL150 relA1 ptsF rbs flbB5301</i>	None	(Casadaban and Cohen 1979)
DH5 α	<i>Δ80d Δ(lacZ)M15 recA1 endA1 gyrA96 thi-1 hsdR17 (<i>r_k⁻m_k⁺</i>) supE44 relA1 deoR Δ(lacZYA-argF) U169 (Promega)</i>	None	Thermofisher
ICB5	<i>F- araD139 ΔlacU169 rpsL thi MalTc-1 ΔtatABCD::<i>apra</i> ΔtatE ΔmalE444</i>	apramycin	(Caldelari, Palmer, and Sargent 2008)
MC4100 Δ <i>amiA</i> Δ <i>amiC</i> Δ <i>tatABC</i>	MC4100, Δ <i>amiA</i> , Δ <i>amiC</i> , Δ <i>tatABC</i> :: <i>Apra</i>	apramycin	(Keller et al. 2012)
NRS-3	MC4100 Δ <i>SufI</i>	None	(Stanley et al. 2001)
DADE	MC4100 Δ <i>tatABCD</i> Δ <i>tatE</i>	None	(Wexler et al. 2000)
XL10-gold Ultracompetent Cells	<i>Tet^rΔ (<i>mcrA</i>)183 Δ(<i>mcrCB-hsdSMR-mrr</i>)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte [F \square <i>proAB lacIqZ</i>ΔM15 Tn10 (Tet^r) Amy Cam^r].</i>	tetracycline and chloramphenicol	Agilent

2.2 Buffers, solutions, and growth media

2.2.1 Growth media and additives

Bacterial nutrient media are outlined in Table 2.2. Strains were commonly grown aerobically overnight at 37°C in Luria-Bertani (LB) broth with shaking. Growth on

solid media was also performed at 37°C. Long-term storage of strains was carried out at -80°C by the addition of a final concentration of 25% glycerol to a stationary phase culture, with this being flash frozen in liquid nitrogen before storage. Sodium dodecyl sulphate (SDS) was added to the media at a final concentration of 1-2% where indicated, in order to evaluate Tat transport activity.

Table 2.2. Growth media used in this study.

Media name	Component and final concentration
Luria-Bertani (LB) Liquid	Tryptone 10g/l Yeast extract 5g/l NaCl 10g/l
Luria-Bertani (LB) Solid	Tryptone 10g/l Yeast extract 5g/l NaCl 10g/l Agar 15g/l
MacConkey agar	MacConkey agar powder g/l 1% maltose 10mM MgCl ₂
Müller-Hilton Plus	Müller powder 20g/l MgSO ₄ 20g/l Agar 17g/l

2.2.2 Antibiotics used in this study

The antibiotics used in this study are listed in the Table 2.3. Stock solutions of ampicillin (Amp), kanamycin (Kan) and apramycin (Apra) were prepared in distilled water. All stock solutions were filter sterilised prior to use.

Table 2.3. Antibiotics used in this study with their stock and working conditions.

Antibiotics	Final concentration (µg/ml)	Solvent
Kanamycin (Kan)	50	Water
Ampicillin (Amp)	125	Water
Apramycin (Apra)	50	Water

2.2.3 Buffers and solutions

Buffers and solutions used in this study are listed in Table 2.4.

Table 2.4. General buffers and solutions used in this study.

Buffer/solution	Composition	Final concentration
APS	Ammonium persulphate	10 % (w/v)
50 x TAE	Tris-base	24.2% (w/v)

	Acetic acid EDTA	5.71 (v/v) 0.05 mM
DNA loading dye	Bromophenol blue Xylene cyanol blue Sucrose	0.25 % (w/v) 0.25 % (w/v) 40 % (w/v)
Laemmli sample buffer (2x)	Tris-HCl, pH 6.8 SDS β-mercaptoethanol Glycerol Bromophenol blue	65.8 mM 2.1 % (w/v) 355 mM 26.3 % (w/v) 0.01 % (w/v)
Resuspension buffer	Tris-HCl, pH 7.5 Glycerol Anti-protease cocktail	50 mM 10% (v/v) One tablet per 10mL
Transformation and storage buffer (TSB)	Tryptone Yeast extract NaCl PEG6000 MgCl ₂ MgSO ₄ DMSO	1% (w/v) 0.5% (w/v) 1% (w/v) 10% (w/v) 10 mM 10 mM 5% (v/v)
SDS running buffer (10x)	Tris-HCl, pH 8.3 Glycine SDS	250 mM 1.92 M 1.0 % (w/v)
Tris-buffered saline (TBS)	Tris-HCl, pH 7.5 NaCl	20 mM 137 mM
Tris-buffered saline with Tween 20 (TBST)	Tris-HCl, pH 7.5 NaCl Tween®20	20 mM 137 mM 0.1 % (v/v)
Tris-glycine transfer buffer	Tris/HCl, pH 8.8 Glycine Methanol	25 mM 192 mM 10 % (v/v)
Buffer 1 (fractionation buffer)	Sucrose Tris-HCl, pH 7.6 EDTA	20% (v/v) 20 mM 2 mM
Buffer 2 (membrane resuspension buffer)	Tris-HCl, pH 7.6 EDTA	50 mM 2 mM
Block buffer	TBS Skimmed milk	1X 3% (w/v)

2.2.4 Biological and chemical reagents

Antibodies used in this study are listed in Table 2.5.

Table 2.5. Antibodies used in this study.

Antibody	Dilution	Host	Supplier/Reference
Primary antibody			
Polyclonal Anti-Sufl	2:10 000	Rabbit	(Buchanan et al. 2002)
Monoclonal Anti MaIE	1:10 000	Mouse	New England Biolabs

Polyclonal Anti-BLA	2:10 000	Rabbit	Abcam® (Cat. #ab12251)
Secondary antibody			
Anti-mouse IgG HRP conjugate	1: 10 000	Goat	Bio-Rad (Cat. #: 170-6516)
Anti-rabbit IgG HRP conjugate	1: 10 000	Goat	Bio-Rad (Cat. #: 170-6515)

Other reagents used were restriction enzymes that were purchased from NEB or Roche. Precision Plus Protein™ All Blue Standards or Dual-Colour™ (Bio-rad) were used as protein markers to estimate protein masses during SDS-PAGE. 1 Kb Plus DNA Ladder (ThermoFisher Scientific) was used to estimate DNA sizes in agarose gel electrophoresis.

2.3 Molecular biology techniques.

2.3.1 DNA manipulations: Plasmid, Gblocks and synthetic genes

Table 2.6. Plasmids used in this study.

Plasmid name	Description	Resistance	Reference
pSUPROM	Vector for expression of genes under control of the <i>E. coli</i> <i>tatA</i> promoter (Kan ^R)	Kan	(Jack et al. 2004)
pSUPROM SufI-CT1	pSUPROM bearing <i>sufI</i> (full length minus stop codon) fused to the C-tail coding sequence of Candidate 1	Kan	This study
pSUPROM SufI-CT2	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 2	Kan	This study
pSUPROM SufI-CT3	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 3	Kan	E. Severi unpublished
pSUPROM SufI-CT4	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 4	Kan	E. Severi unpublished
pSUPROM SufI-CT5	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 5	Kan	E. Severi unpublished
pSUPROM SufI-CT6	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 6	Kan	This study
pSUPROM SufI-CT7	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 7	Kan	E. Severi unpublished
pSUPROM SufI-CT8	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 8	Kan	This study
pSUPROM SufI-CT9	As pSUPROM SufI-CT1 but with C-tail coding sequence of Candidate 9	Kan	E. Severi unpublished
pSUPROM SufI-CTFdnH	As pSUPROM SufI-CT1 but with C-tail coding sequence of FdnH	Amp	E. Severi unpublished
pSU40 UniAmiA	<i>amiA</i> (mature sequence) cloned into pSUPROM	Kan	This study

pSU40 UniAmiA SP1	Signal peptide coding sequence from Candidate 1 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	This study
pSU40 UniAmiA SP2	Signal peptide coding sequence from Candidate 2 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	This study
pSU40 UniAmiA SP3	Signal peptide coding sequence from Candidate 3 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	E. Severi unpublished
pSU40 UniAmiA SP4	Signal peptide coding sequence from Candidate 4 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	This study
pSU40 UniAmiA SP5	Signal peptide coding sequence from Candidate 5 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	E. Severi unpublished
pSU40 UniAmiA SP6	Signal peptide coding sequence from Candidate 6 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	E. Severi unpublished
pSU40 UniAmiA SP7	Signal peptide coding sequence from Candidate 7 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	E. Severi unpublished
pSU40 UniAmiA SP8	Signal peptide coding sequence from Candidate 8 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	This study
pSU40 UniAmiA SP9	Signal peptide coding sequence from Candidate 9 cloned <i>Bam</i> HI - <i>Xba</i> I into pSU40 UniAmiA	Kan	This study
pSUPROM-SFB	pSUPROM encoding full-length <i>SufI</i> lacking stop codon fused in frame with the <i>FdnH</i> transmembrane domain and the mature region of β -lactamase	Kan	This study
pSUPROM-SFB-SECA	As for pSUPROM-SFB, with removal of coding sequence for last nine amino acids of <i>FdnH</i> C-tail and addition of three Lys codons	Kan	This study
pSUPROM-SFB-DFL	As pSUPROM pSUPROM-SFB-SECA but with insertion of a proline codon at position 470 a double short flexible linker (GGGGS) ₂ flanking the <i>FdnH</i> C-tail	Kan	This study
pUNIPROM	Cloning vector for expression of genes under the control of the <i>tat</i> and T7 promoters; Amp ^R	Amp	(Jack et al. 2004)
pUNIPROM-SFM	pUNIPROM encoding full-length <i>SufI</i> lacking stop codon fused in frame with the <i>FdnH</i> transmembrane domain and the mature region of maltose binding protein (MalE)	Amp	This study

Gene fragments encoding the C-tails and signal peptides of Candidates 1,2,3,4,6,8 and 9, were synthesized as gBlocks® after performing codon optimization to improve gene expression (Table 2.7). The gBlocks® were ordered from Integrated DNA Technologies, Inc. (IDT®).

Table 2.7. gBlocks® synthesised for use in this thesis.

SeqJG-1
ATGCGTCACGCTCGTCTGTGGTTGTTTCGTCGTGTTACCCGCTCTGGCTGCTGTTGGTGGTCTGCTGCTG GGTGGTGCTATGGTTACCAACGCTGTTGCTCTGACCAGCCGCGTAACGTTGGTCCGGGTCTGCTGGTT ATCGCTGGTTCTCTGGTTGCTCTGGTTGCTACCCGTTGGATCCGTGCTGAACAGGACCGTAAAGCTTACC GTCAGCACTACTCTGCTACCTGGGGTTAA
SeqJG-2SP
ATGACCACCTCTTCTCCGTCTGCTCCGCGTCGTCTATCCTGCGTTCTGCTACCGCTCTGGTTGGTGGTG TTGCTCTGGCTGTTGCTGTTCCGCTGGCTGCTTCTGCTCACGTTCTGTTTTCTCCGGACCAGGCTGCTGC T
SeqJG-2CT
ACCGCTGCTCCGGACACCACCGTTACCACCGCTGCTTCTGACACCTCTGCTACCTCTTCTGCTGTTGCTG TTGGTCTGGGTGTTGGTGGTCTGGCTCTGGGTGCTGTTGCTCTGGTTGTTGCTGTTTTCTGCTCTGACCC GTGTTTCGTCGTGAAGGTGGTGGTCAGGCTTAA
SeqJG-3
ATGACCTCTCGTCGTGGTACCTTCCTGGCTGCTCTGGTTACCGCTTCTCTGATCCCCTGGCTCCGCCGG CTCTGGCTGCTGGTACCCCGCTGGCTGTTCTGCTGGGTCTGTTTCGCTGCTATCGCTGTTGCTGTTGGTG CTATCAAACCGCTGCACTCTTCTCTGCTGCAGGTTACGCGTACCCTGGGTCTGTAA
SeqJG-4
ATGGGTAACGCTGTTTTCTGGTCGTCTGCTACCCTGCTGTCTGGTACCGCTGTTCTGGCTGCTGTTGCTCTGA TCGCTCTGGGTCTGCTCCGGCTCAGGCTCAGACCGGTGGTGACCTGGCTGCTACCGGTTCTGACTCTA CCCTGCCGGTTGCTGGTGCTGCTGGTGCTGCTCTGCTGGCTGGTGGTGGTCTGTTCTACGCTATGCGTC GTCGTATGGCTGCTCGTAACGGTTAA
SeqJG-6
ATGGGTATCGCTGCTTCTGGTCGTCTGCTACCCTGCTGTCTGCTACCGCTGTTTCTGCTACCGCTGCTCTGA TCGCTCTGGGTGCTGCTCCGGCTCAGGCTGACGCTATCAAACCGGACCTGGGTGTTCTGCTCTGGCTT CTACCCTGCCGCTGGCTGGTGCTGCTGGTGCTGCTCTGCTGGCTGGTGGTCTATCGTTTGGGCTGTTT GTCGTCTGTTCTGCTGCTCGTGCTTCTTAA
SeqJG-8
ATGCTCTGCGTCGTCTGCTGCTCTGCTGGTTACCGCTGCTTCTACCCTGACCGCTCTGGCTGCTCCGG CTGCTCTGGCTGCTCTGGGTCCGGCTGCTACCAAACGTGCTACCGGTGCTGAACTGCGTTCTGACGACA AAAAAGACGACGGTCTGTCTTCTTCTGCTACCACCTGGATCATCGTTGGCGTTGTATTGTCGCTTCTGC TGGTTTCGGTCTGCTGCTGTCTGGTCGTAAACGTCGTCTGTCGTA
SeqJG-9
ATGCTGGTATGGCTGGTCGTCTGTTCTCTGGTTGCTGCTGCTCTGGGTCTGGCTGCTCTGCTGGGTGGT TGCTCTGACGCTCGTCCGCCGGCTCCGGAAGTTGCTGTTGCTTCTCTGCCGTGGCTGCCGGTTGCTGCTA TCGGTCTGCTGCTGGTTCGTGTTCTGTGGCGTCTGCGTCGTCTGCTGCTGCTATCGGTCGTCCGC TGACCGAAGCTCAGCCGTAA

Due to the more complex sequence of protein Candidates 5 and 7, GenScript Gene Synthesis service was used instead of gBlocks®. The signal peptide and C-tail encoding

regions of each protein was synthesized and cloned as a concatemer into the vector pUC57.

JG-5_7

```
ATGCCGGCGACCACCGGTGATCGTAGCCGTCGTCGTCCGCTGGCGTTTGC GGCGGGCGGTT
GCGACCGCGGGCGGCGATTGGCGCGGCGAGCTTGGCGGGCGGCCGACCACCGCGGGCGGC
GGGTAGCGACGCGCCGGTTGCGGCGCTGGGCACCGCGGGCGGCGCTGGCGGTGGCGGGCGG
GTGCGGGCGTGTTTTTCGCGGTTTCGTCTGTCGTCTGCTGGTGCGCGTGATGCGCAGGCGTAAA
TGAGCCGTCGTCTGTACCGCGCTGCTGACCGCGCCGGCTGCTGCGGGCGGGCGGGCGGCGCTGA
CCCTGGTGGGCGCGGCGACCAAGCGCGAGCGCGGACGATGCGACCCCGGCGGGCGCGTGCG
CAAGTGCGTGCGAGCGCGCAAGAGCGTAGCGAAGTTCTGGCGGGCGACCGGTGCGCGTACC
GGCGTGCTGCTGGCGGGCGGGTGCGCTGGCGCTGGGTCTGGGTGCGGGTCTGGTTACCTG
GCGTCGTCTGTCGTGCGGGCGGGTGCGTAA
```

2.3.2. Plasmid construction

Plasmids were constructed either using Gibson assembly (Gibson et al. 2009), or by PCR-amplification of DNA and restriction cloning. DNA substitutions were introduced using Q5[®] Site-Directed Mutagenesis and KLD reactions (see section 2.3.4 for a description of KLD). All constructed plasmids were fully sequenced after generation.

2.3.3 Amplification of DNA by Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is a technique that allows the amplification of specific regions of DNA by several orders of magnitude. This procedure utilises thermostable DNA polymerase enzymes and oligonucleotide primers, which are designed to be complementary to the 5' and 3' ends of a thermally denatured single stranded DNA template. DNA elongation occurs in a 5' to 3' direction, enabling the strands to extend towards one another.

In this study PCR was used for plasmid amplification, Q5® Site-Directed Mutagenesis and colony-PCR for sequencing. In the case of colony-PCR the reaction mix was composed of: 10µl template DNA (from a 50 µl boiled sample of bacterial biomass), 1 µl of forward/reverse primer (from a 100 µM stock), 12 µl Go Taq MIX and ultra-pure water up to 20µl of final volume. The PCR programme consists of three essential steps: a denaturing step at 96°C, an annealing step at an appropriate temperature for the primers and an elongation step at 72°C. These steps were repeated for 28 cycles and followed by a final elongation at 72°C. A similar procedure, but using 0.5 µl of target plasmid, was used for plasmid amplification. In the case of Q5® Site-Directed Mutagenesis, the mix contains 1 µl of the plasmid (50ng/µL), 1 µl of each primer (100µM), 12.5 µl of Q5 Master Mix and 14 µl of ultra-pure water. The PCR programme used was: 5 minutes denaturing at 98°C followed by 33 cycles of 30 seconds at 95°C, 1 minute at the appropriate temperature for the primer and seven minutes at 72°C. The final step was 10 minutes at 72°C for final elongation. In all cases PCR products were analysed by agarose gel electrophoresis and purified using a NEB Gel extraction kit or a NEB PCR Purification kit.

2.3.4 Q5® Site-Directed Mutagenesis

Q5® Site-Directed mutagenesis (New England BioLabs) allows the introduction of nucleotide changes in plasmids. The protocol consists of three steps: An exponential amplification PCR as described above, followed by a kinase, ligase and *DpnI* (KLD) treatment, and subsequent transformation into competent *E. coli* cells. The protocol was performed following the manufacturer's instructions. Eurogentec produced the custom primers for the directed mutagenesis.

2.3.5 Agarose gel electrophoresis

DNA samples were analysed by agarose gel electrophoresis using 1% (w/v) agarose gels prepared with 1xTAE buffer and containing 0.001% (v/v) Gel Red dye (Biotium). Loading dye was added to samples to create a visible running front and DNA size markers (1 kb Plus DNA Ladder, Invitrogen) were run on gels to allow identification of target bands.

2.3.6 DNA digestion and Ligation

DNA was digested using the appropriate restriction enzymes incubated in the corresponding restriction enzyme buffer. Samples were incubated for at least 2h at 37°C. Digested vectors were then treated with alkaline phosphatase to dephosphorylate the vector and avoid self-ligation.

DNA fragments were ligated using T4 DNA ligase and the appropriate buffer. Ligations were performed at room temperature for 120 min, and the entire reaction volume was used to transform competent cells.

2.3.7 DNA sequencing

After cloning, DNA sequencing was used to confirm sequences of plasmids. The Medical Research Council Protein Phosphorylation and Ubiquitylation Unit (MRC PPU) in Dundee, Scotland performed sequencing of DNA. Chromatogram files obtained from sequencing reactions were analysed using SnapGene software (from GSL Biotech; available at snapgene.com).

2.3.8 Plasmid DNA preparation

Plasmid extraction from *E. coli* strains was performed using the Monarch® Plasmid Miniprep Kit, as per the manufacturer's instructions. A single colony was used to inoculate a 5 ml culture which was grown overnight before centrifugation. Pelleted cells were subsequently resuspended in lysis buffer and then treated with neutralisation buffer before isolation of the plasmid DNA through adsorption onto a silica membrane under high salt conditions. After washing, the DNA was eluted from the membrane in 20 µl of ultra-pure water, repeating this process twice with 10 µl water followed by 5 minutes of centrifugation. The plasmids used in this study are shown in Table 2.6.

2.4 Preparation of competent cells and transformation with plasmid DNA

2.4.1 Preparation of chemically competent cells

5 ml of LB with the appropriate antibiotics was inoculated with 50 µl from an overnight culture in stationary phase and grown aerobically at 37°C with shaking until OD₆₀₀ of 0.3-4 was achieved. The cells were pelleted by centrifugation. The pellet was resuspended in 1 ml of 100mM ice-cold CaCl₂ and then either stored at -80°C or used immediately. Transformation was performed by incubation of 100 µl of competent cells with 1 µl of plasmid, and 5 - 10 µl of ligation mix for 15 min on ice. Cells were subjected to heat-shock at 42°C for 90 s followed by an additional 2 min incubation on ice. 1 ml of LB was added, and cells were grown, with shaking, at 37°C for at least 2 hours. Cells were then pelleted by centrifugation in a bench top centrifuge at max speed for 5 min. After that, the cells were plated onto LB agar plates containing required antibiotics and incubated overnight at 37°C.

2.4.2 Commercial competent cells

In some instances, commercial XL-10 ultracompetent cells were used. The transformation protocol involved pre-chilling of one Falcon polypropylene tube for each transformation plus one extra tube for the control on ice. In addition, a bottle of LB broth was pre-heated to 42°C. After the cells were thawed on ice, 22 µL of cells were added into each pre-chilled Falcon tube followed by 2µL of β-mercaptoethanol mix. The cells were incubated on ice for 10 minutes swirling every two minutes. For the transformation, 2µL of ligation mix was added and for the control another 2µL of a 1:10 dilution of pUC18. All tubes were incubated for 30 minutes on ice, and then heat-pulsed at 42°C for 30 seconds. After the heat-pulse, tubes were returned to ice for two minutes. For recovery, 0.9 mL of prewarmed LB was added to each tube and the tubes were placed in the orbital incubator at 37°C, with shaking at 225-250 rpm, for one hour. Cells were then pelleted and plated on LB agar plates containing the required antibiotics and incubated overnight at 37°C.

2.5 Protein methods

2.5.1 SDS-PAGE

SDS polyacrylamide gel electrophoresis (SDS-PAGE) is used to separate proteins under denaturing conditions according to molecular weight (Laemmli 1970). Tris-glycine SDS-PAGE gels of 0.75mm thickness were prepared for use with the Mini-PROTEAN II system (Bio-Rad) at a concentration of 10% polyacrylamide. The composition of resolving and stacking gels can be found in Table 2.8.

Table 2.8. SDS-PAGE components.

Resolving gel	Stacking gel
1.5 M Tris-HCl, pH 8.8: 4 mL	0.5 M Tris-HCl, pH 6.8: 2.5 mL
10% Sodium Dodecyl Sulfate (SDS): 160 μ L	10% SDS: 100 μ L
10% Ammonium Persulfate (APS): 160 μ L	10% APS: 100 μ L
N,N,N',N'-Tetramethylethylenediamine (TEMED): 16 μ L	TEMED: 10 μ L
Distilled Water: 6.3 mL	Distilled Water: 5.3 mL
30% Acrylamide: 5.33 mL	30% Acrylamide: 2 mL

Tris-glycine gels were submerged in a gel electrophoresis tank (Bio-Rad) filled with SDS running buffer. Samples were prepared by mixing 1:5 with 6x sample buffer or 1:1 with 2x Laemmli sample buffer both containing β -mercaptoethanol and loaded along with the marker Precision Plus Protein™ Standards Dual Colour (Bio-Rad). The gel was run at 100 V for 1 hour, after that the voltage was increased to 200 V until the blue dye reached the bottom of the gel.

2.5.2 Semi-dry, wet Western Blotting and turbo-blotting

Samples to be analysed were first separated using SDS-PAGE as described above, subsequently they were transferred to a nitrocellulose membrane using either the semi-dry or wet transfer systems.

For semi-dry transfer, gels were soaked in Tris-glycine transfer buffer before being placed on pre-soaked nitrocellulose membrane (Amersham Hybond-ECL, GE Healthcare). These were sandwiched between four pieces of pre-soaked 3MM Whatman paper before being placed inside a TransBlot SD SemiDry Transfer Cell (Bio-Rad) for protein transfer. The transfer was performed at 10 V for 60 min.

For turbo-transfer the protocol was the same as for semi-dry transfer, but the transfer was performed in a Bio-rad trans-Blot® Turbo™ Transfer system following the manufacturer's protocol.

After transfer, the membrane was blocked in TBS-Tween with 3% skimmed milk for 2 h at room temperature with shaking. This was followed by overnight incubation at 4°C with primary antibody suitably diluted in TBS-Tween. This was followed by three washes in TBS-Tween before the incubation of one hour-with the secondary antibody, also suitably diluted in TBS-Tween, see Table 2.4.

The detection of the bands was performed by the use of enhanced chemiluminescence (ECL) under UV light in the *G:Box Syngene*.

2.5.3 Preparation of soluble and membrane fractions

50 µl of a stationary phase culture was subcultured into 50 ml LB and grown at 37°C with shaking until OD₆₀₀ of 1 was reached. The cells were harvested by centrifugation at 3000 x g for 10 min at 4°C and the supernatant was discarded. The cell pellet was resuspended in 1 ml of Buffer 1 (Table 2.4). The cells were lysed by sonication at 30% amplitude for 1 min. The lysate was then centrifuged to remove cell debris at 17000 x g for 5 min followed by ultracentrifugation to pellet the membranes at 227000 x g for 30 min at 4°C. The supernatant from this step was retained as the soluble fraction, and the membrane pellet was resuspended in Buffer 2. To assess membrane integration of proteins, membranes were incubated for 20 min at room temperature in the presence of 4M urea, followed by re-pelleting of membranes by ultracentrifugation (this time at 20°C to avoid crystallisation of the urea).

2.6 Growth assays

2.6.1 Minimum inhibitory concentration (MIC) assay

In order to investigate whether the SFB fusion protein had been integrated into the inner membrane or exported to the periplasm, M.I.C.Evaluator™ (M.I.C.E.™) test strips containing ampicillin were used. These are used for the accurate determination of MIC of a test organism to an antimicrobial agent. The strips consist of a gradient of stabilised antimicrobial compounds covering 15 doubling dilutions. On the application of a strip to a pre-inoculated agar plate, the antimicrobial is released from the polymer strip, forming a defined concentration gradient in the area around it. After appropriate incubation, growth develops with a zone of inhibition around the strip.

In order to perform this technique, an overnight culture of the bacteria harbouring the plasmid of interest was grown in LB supplemented with kanamycin. A sample from the overnight culture was diluted to obtain a final OD₆₀₀ of 0.06. A pre-warmed petri dish with LB agar containing kanamycin was spotted with 80µL of the diluted bacterial culture and spread with sterile cotton swabs, and then the dish left to dry for 10 minutes at room temperature. After that, the strip were placed using sterilised tweezers and the zone of clearing assessed after 18 hours of growth at 37°C.

2.6.2 Spot assay for evaluation of Tat transport activity

The activity of the Tat system can be evaluated by detecting the presence of Tat substrates in the periplasm. One such protein is AmiA, that catalyses cleavage of the murein septum during cell division (Heidrich et al. 2001). The export of AmiA and AmiC is strictly dependent on the Tat system. Their mislocalisation results in defective cell division and an aberrant outer membrane which causes the cell to be sensitive to detergents, for example SDS (Ize et al. 2003a). Therefore, based on the transport of AmiA/AmiC, evaluating Tat transport activity can be achieved simply by testing growth of cells on SDS-containing media.

In order to address if signal peptide fusions to AmiA were transported in a Tat-dependent manner, a spot growth assay was used. Cultures were grown overnight in

LB plus the required antibiotic and subsequently normalised to OD₆₀₀ of 1. These were then serially diluted to 10⁻⁶ and either 10 or 20 µL of the series were spotted onto LB (antibiotic) plus/minus 2% SDS and grown at 37°C overnight.

2.6.3 MacConkey maltose agar

MacConkey agar plates (Sambrook and Russell 2001) supplemented with 1% maltose and 20 mM MgCl₂, were used to indirectly investigate the periplasmic or cytoplasmic localisation of MalE in SFM fusion proteins. Strains expressing the fusion were streaked onto MacConkey agar and incubated at 30°C for 48 hours. The colour of the colonies was recorded to score for maltose utilisation.

2.7 Computational tools

A range of computational tools has been used in this study in order to gather information from the protein databases and to predict protein structure and function

2.7.1 Python

Python is a high-level, general purpose programming language, and it has been the main language chosen for all the “in house scripts”. All scripts used in this study are available on GitHub under the project names “protein filter” and “protein tools” (Van Rossum and Drake Jr 1995) and appendix B

2.7.2 bash

Bash is a Unix shell and command language used in this study as a command line processor.

2.7.3 RegExr

Regex (regular expression), is a string of text that allows the user to create patterns that help match, locate, and manage text. In order to work with large

databases of proteins, the resource RegExr, which allows an easy use of regex, was used (RegExr v3.8.0, 2022 by Grant Skinner <https://regexr.com/>).

2.7.4 Muscle

MUltiple Sequence Comparison by Log- Expectation (MUSCLE) v5 was used to align protein sequences (Edgar 2022).

2.7.5 HMMER

The HMMER software package was used to identify homologous proteins using a profile-HMM (Hidden Markov model) encoded across bacterial genomes.

2.8 Online resources

2.8.1 TMHMM

TMHMM was used to predict the presence of transmembrane helices (Hallgren et al. 2022).

2.8.2 AnnoTree

AnnoTree was used to develop phylogenetic trees (Mendler et al. 2019).

2.8.3 DeepFRI

In order to predict the function of the candidate proteins a structure-based protein method called DeepFri was used (Gligorijević et al. 2021).

2.8.4 NCBI

The RefSeq database was used in this thesis as the main source of gene and protein sequences. RefSeq is a curated non-redundant collection of sequences representing genomes, transcripts, and proteins. (Pruitt, Tatusova, and Maglott 2007).

2.8.5 SignalP

SignalP 5.0 was used to predict signal peptide sequences and their likely export route (Almagro Armenteros et al. 2019).

2.9 Protein structure prediction

Two different methods were used for protein modelling.

2.9.1 AlphaFold 2

The first method used the Alphafold artificial intelligence network, which was designed as a deep learning system able to predict a protein's 3D structure from its amino acid sequence. The tool was run under a Google Colab (ColabFold) (Jumper et al. 2021).

2.9.2 Robetta

The second method followed the CAMEO (continuously evaluate the accuracy and reliability of predictions) system. The chosen tool was RobeTTa fold based on the algorithm of RoseTTAFold (Baek et al. 2021), which is distinct from the Rosetta software

Chapter 3. Bioinformatic analysis of Tat dependent tail-anchored proteins

3.1 Introduction

While most substrates of the bacterial Tat pathway are soluble periplasmic proteins, a small fraction are anchored to the membrane by a single C-terminal transmembrane helix. In the model organism *Escherichia coli*, five of the 41 Tat substrates are tail-anchored membrane proteins. As described in Chapter 1 these are the small subunits of periplasmically-facing hydrogenase and formate dehydrogenase enzymes, and the HybA iron sulphur protein (Hatzixanthis, Palmer, and Sargent 2003). However, it is currently unclear whether this represents the entire repertoire of tail-anchored Tat substrates across prokaryotes, or whether there are other Tat-dependent tail-anchored proteins that are yet to be found.

To partially address this issue, Professor Tracy Palmer worked together with collaborator Dr Govind Chandra (John Innes Centre, Norwich) to develop a search of the Refseq database of bacterial proteins. This search was carried out by Dr Govind Chandra using TATFind 1.4 (Rose et al. 2002; Dilks et al. 2003) and TMHMM-2.0c (Krogh et al. 2001) over 89,292 bacterial genomes labelled as "reference" or "representative" in GenBank, identifying 34,634 candidate proteins that met certain criteria, including being more than 150 amino acids long, having a Tat signal within the first 50 amino acids, having no more than two transmembrane helices (TMHs), and having one TMH within 50 amino acids of the C-terminus. These 34,634 proteins came from 20,558 distinct genomes belonging to 6,798 distinct organisms.

It should be noted at the outset that there are limitations to this analysis. The major limitation is that the search relies upon the twin arginine signal sequence and the C-tail anchor being present within the same polypeptide sequence. While this is true for the hydrogenase small subunits and HybA proteins, the small subunits of formate dehydrogenases (FdnH and FdoH) lack any targeting information. Instead, they are exported to the periplasmic side of the membrane because they form a complex with their partner proteins (FdnG or FdoG) which carry a twin arginine signal peptide (Stanley et al. 2002; Hatzixanthis, Palmer, and Sargent 2003). The nature of the search means any such 'piggybacking' proteins will not be identified.

This initial output identified 34,634 candidate proteins that met these criteria, and was generated in HTML. HTML (Hypertext Markup Language) is a standardised language used to create and structure web pages. It uses tags to define the content and layout of a web page, including headings, paragraphs, images, links, and other elements. However, HTML is not well suited for storing large amounts of data or complex relationships between data items, as it is primarily designed for presenting information in a readable format on the web. For this reason, it is generally considered a bad language for databases, as it does not provide the structure, security, and scalability necessary for managing large amounts of data efficiently. Database management systems such as SQL and NoSQL databases are more commonly used for this purpose.

3.1.1 Python and Regex as bioinformatic tools for database analysis.

As part of my PhD research, I chose to analyse these data carefully to identify additional families of Tat-dependent tail-anchored proteins. The initial database was created in HTML, the standard language for creating documents to be displayed in a web browser (as shown in Fig. 3.1). Its main advantage is that it makes it easier for people to view the documents, regardless of their level of computer literacy. However, this format makes it more challenging to search for large amounts of data within the main database as it is not compatible for data mining using computing language. So, the first step was to process the HTML database into a more computer-readable database.

To work with this large database, I chose Python 3 as the language and tool. Python is a high-level, general-purpose, interpreted programming language that was created by Guido van Rossum (Van Rossum and Drake Jr 1995). Its object-oriented approach is designed to help programmers write clear, logical code for both small and large-scale projects (Dave 2012). Python is dynamically typed and garbage-collected (automatic memory management) and supports multiple programming paradigms, including structured, object-oriented, and functional programming (<https://www.python.org>). Python was selected over others such as R, firstly, because Python is a well-established programming language that has a

TAT Signal and TMH near C-terminus

1	GCF_000688455.1_ASM68845v1_protein.faa.gz	Acidobacterium ailaui
Taxonomy	Acidobacteria; Acidobacteriia; Acidobacteriales; Acidobacteriaceae; Acidobacterium	
First 60 AAs	MSRRTFVSSATAGLALGALSSAAEGHAQLVWTSKNWKLAEFETLLREPARIRQVYDVTQ	
WP_026442391.1	hypothetical protein [Acidobacterium ailaui]	
TMHMM	WP_026442391.1	Length: 233
TMHMM	WP_026442391.1	Number of predicted TMHs: 1
TMHMM	WP_026442391.1	Exp number of AAs in TMHs: 21.25002
TMHMM	WP_026442391.1	Exp number, first 60 AAs: 1.35114
TMHMM	WP_026442391.1	Total prob of N-in: 0.67991
TMHMM	WP_026442391.1	WP_026442391.1 inside 1 201
TMHMM	WP_026442391.1	WP_026442391.1 TMhelix 202 224
TMHMM	WP_026442391.1	WP_026442391.1 outside 225 233
2	GCF_000022565.1_ASM2256v1_protein.faa.gz	Acidobacterium capsulatum ATCC 51196
Taxonomy	Acidobacteria; Acidobacteriia; Acidobacteriales; Acidobacteriaceae; Acidobacterium; Acidobacterium capsulatum	
First 60 AAs	MKSISRRSFVTTAAAGMAALGSLGPALPAAQGQAVEMASDWDISSFNQLAQSPARVKQLF	
WP_012680923.1	Tat pathway signal sequence domain-containing protein [Acidobacterium capsulatum]	
TMHMM	WP_012680923.1	Length: 237
TMHMM	WP_012680923.1	Number of predicted TMHs: 1
TMHMM	WP_012680923.1	Exp number of AAs in TMHs: 31.62059
TMHMM	WP_012680923.1	Exp number, first 60 AAs: 5.92535
TMHMM	WP_012680923.1	Total prob of N-in: 0.86701
TMHMM	WP_012680923.1	WP_012680923.1 inside 1 205
TMHMM	WP_012680923.1	WP_012680923.1 TMhelix 206 228
TMHMM	WP_012680923.1	WP_012680923.1 outside 229 237
3	GCF_000014005.1_ASM1400v1_protein.faa.gz	Candidatus Koribacter versatilis Ellin345

```

<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="utf-8" />
<style type="text/css">
body {
background: #d4d4d4;
font-family: sans-serif;
}
.tta (background: #ffff00)
.timestamp {font-size: x-small}
td, table, tr, th {
border-collapse: collapse;
border: solid black 1px;
padding: 3px;
}
.species {font-style: italic}
.genus {font-style: italic}
a:link, a:visited, a:active {color: blue; text-decoration: none}
a:hover {color: magenta; text-decoration: none}
</style>
<title>TAT-CterTMH</title>
</head>
<body>
<p class="timestamp">Fri 21 Jul 13:14:15 BST 2017</p>

<h3>TAT Signal and TMH near C-terminus</h3>
<table>
<tr style="background:#F7EBDB"><td>1</td><td>GCF_000688455.1_ASM68845v1_protein.faa.gz</td><td colspan="4">Acidobacterium ailaui</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">Taxonomy: Acidobacteria; Acidobacteriia; Acidobacteriales; Acidobacteriaceae; Acidobacterium; Acidobacterium capsulatum</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">First 60 AAs: MSRRTFVSSATAGLALGALSSAAEGHAQLVWTSKNWKLAEFETLLREPARIRQVYDVTQ</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">WP_026442391.1 hypothetical protein [Acidobacterium ailaui]</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 Length: 233</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 Number of predicted TMHs: 1</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 Exp number of AAs in TMHs: 21.25002</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 Exp number, first 60 AAs: 1.35114</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 Total prob of N-in: 0.67991</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 WP_026442391.1 inside 1 201</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 WP_026442391.1 TMhelix 202 224</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_026442391.1 WP_026442391.1 outside 225 233</td></tr>
</table>
<table>
<tr style="background:#F7EBDB"><td colspan="3">2</td><td colspan="2">GCF_000022565.1_ASM2256v1_protein.faa.gz</td><td colspan="2">Acidobacterium capsulatum ATCC 51196</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">Taxonomy: Acidobacteria; Acidobacteriia; Acidobacteriales; Acidobacteriaceae; Acidobacterium; Acidobacterium capsulatum</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">First 60 AAs: MKSISRRSFVTTAAAGMAALGSLGPALPAAQGQAVEMASDWDISSFNQLAQSPARVKQLF</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">WP_012680923.1 Tat pathway signal sequence domain-containing protein [Acidobacterium capsulatum]</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 Length: 237</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 Number of predicted TMHs: 1</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 Exp number of AAs in TMHs: 31.62059</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 Exp number, first 60 AAs: 5.92535</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 Total prob of N-in: 0.86701</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 WP_012680923.1 inside 1 205</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 WP_012680923.1 TMhelix 206 228</td></tr>
<tr style="background:#F7EBDB"><td colspan="5">TMHMM WP_012680923.1 WP_012680923.1 outside 229 237</td></tr>
</table>
<table>
<tr style="background:#F7EBDB"><td colspan="3">3</td><td colspan="2">GCF_000014005.1_ASM1400v1_protein.faa.gz</td><td colspan="2">Candidatus Koribacter versatilis Ellin345</td></tr>
</table>

```

Figure 3.1. This figure presents a protein database represented in HTML format, as viewed within a web browser and opened in an editor. The image serves to emphasize the intricacies and challenges associated with handling extensive HTML-based databases. In the web browser view, the accessibility of the database for online exploration is apparent, while the editor's display underscores the intricate nature of working with the underlying HTML code. This illustration highlights the importance of effective tools and strategies for navigating and extracting valuable data from large HTML databases, particularly within the context of bioinformatics and computational biology research.

large and active community of users. It is user-friendly and has a vast number of libraries and tools available that are suitable for scientific computing and data analysis. This makes Python an ideal choice for database analysis in biology. Additionally, Python has a relatively simple syntax and provides a comprehensive set of tools for data processing, making it a highly effective language for this purpose. Given the size and repetitive structure of the database, I chose to use regular expressions (regex or regexp) to perform searches. The concept of regex was formalised by mathematician Stephen Cole Kleene in 1950 (Leung 2010). A regex is a sequence of characters that defines a search pattern and is commonly used for string search algorithms, such as "find" or "find and replace" operations on strings, or for input validation (Mitkov, Le An, and Karamanis 2006; Lawson 2003).

Furthermore, Regex is a powerful tool that can be used to process and manipulate large amounts of text data, making it ideal for the analysis of protein databases. By using regular expressions, it is possible to search for specific patterns in the database, extract relevant information, and perform operations on this information in a streamlined and efficient manner. This makes Regex a valuable tool for the analysis of protein databases, particularly in the context of bioinformatics.

All the Python scripts that have been used in this study are outlined in Table 3.1, and also are available in GitHub under the project "protein_tools" (https://github.com/Ravenneo/proteins_tools) and "Scripts" (<https://github.com/Ravenneo/Scripts>). GitHub is a web-based platform that provides hosting for software development version control. It offers features such as bug tracking, feature requests, task management, and wikis for every project. In bioinformatics, GitHub is valuable for several reasons. Firstly, it provides a platform for bioinformaticians to share, collaborate, and contribute to large scale bioinformatics projects (Crystal-Ornelas et al. 2022). This enables bioinformaticians from different institutions and locations to work together on a project, thus promoting transparency and collaboration in the field. Secondly, GitHub makes it easier for researchers to share their code, data, and results with the wider scientific community.

Table 3.1. Description of scripts used in this study. Full scripts are available in the appendix B.

Scripts	Description
arcane.py (hmmseARCh aNd parsE)	Python based script that uses hmmsearch (from the HMMER suite (Potter et al. 2018)) to look for homologues of a protein of interest. The retrieved report will be parsed using Pandas in order to only select entries that are from the Refseq NCBI database
etna2.py (EfeTch aNd pArse)	Python based script that uses the Entrez utils Efetch to fetch the Identical Protein Groups (IPG) report for a list of proteins IDs. The retrieved report will be parsed using Pandas in order to only select entries that are from the Refseq NCBI database.
get_fasta	Python based script that uses the Entrez util Efetch to download the protein Fasta sequence for a list of proteins IDs
reptile.py	Python based helper script. It is meant to be used after arcane.py or etna.py. From the output table of arcane.py or etna.py, it drops all the lines where proteins are annotated as 'hypothetical proteins'
Main_Proteins.py	Python based script that is part of Cala.py
Protein_Sorter.py	Python based script that extracts information from an HTML file to return a dictionary of protein names and their locations within the HTML file. This exports a CSV file named "all_proteins.csv".
Incomplete_protein_counter.py	Python based script that groups families present in Dr Chandra's List.
Cala.py	Python script that allows the user to either display the protein information on the console or export it to a CSV file. The user is prompted to select one of the two options. If the first option is selected, the script creates an instance of the ProteinHtml class, reads the name of the file using getFileName() method, and then searches for a specific protein name using searchProtein() method. Finally, it prints all the protein names and their count using findAllProteinNames() method. If the second option is selected, the script creates a DataFrame from the dictionary returned by findAllProteinNames(), prompts the user for a file name, and exports the DataFrame to a CSV file with the specified name.
ProteinHtml.py	Python based script that can be used to parse an HTML file. The class has two methods: searchProtein(proteinName): searches for a given protein name in the HTML file and returns the number of occurrences of that protein in the file. findAllProteinNames(): returns a dictionary containing all the protein names in the HTML file and their count.

To extract a comprehensive list of proteins from the database, four Python scripts were employed. The first two scripts, Main_Proteins.py and ProteinHtml.py, work together to produce a complete list of all protein names, which is displayed in the console (Fig. 3.2).

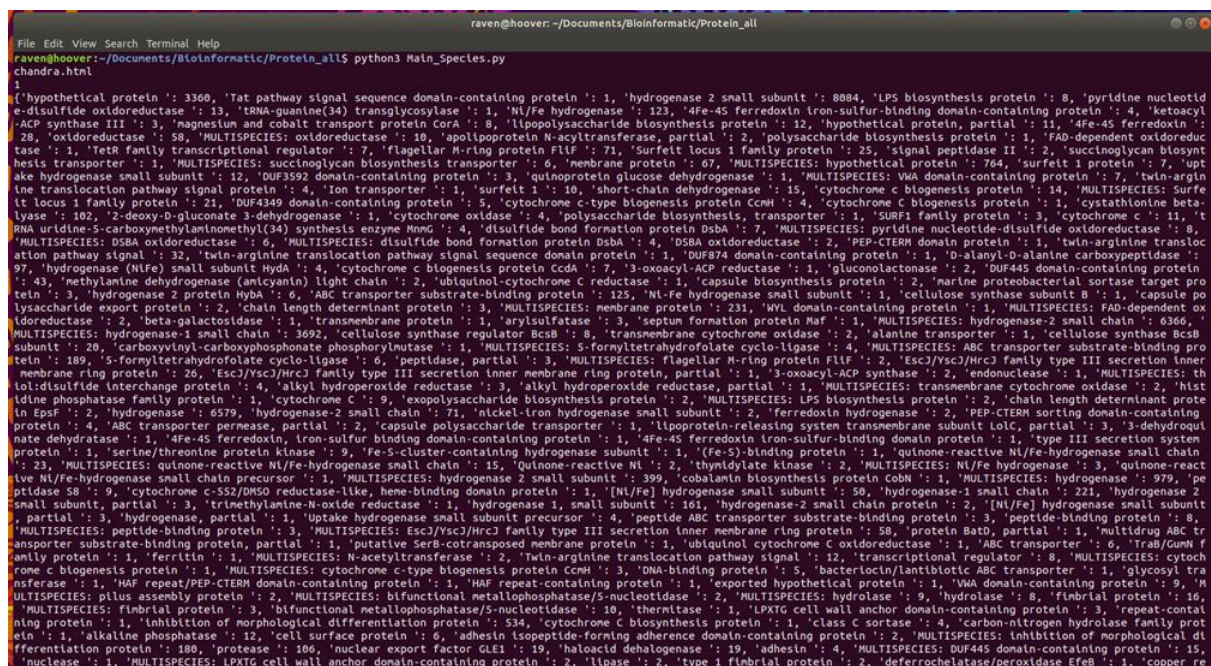
A terminal window screenshot with a dark background and light-colored text. The window title is 'raven@hoover: ~/Documents/Bioinformatic/Protein_all'. The prompt is 'raven@hoover:~/Documents/Bioinformatic/Protein_all\$ python3 Main_Species.py chandra.html'. The output is a long, single-line list of protein names and their counts, separated by semicolons. The list includes various proteins such as 'hypothetical protein', 'tetracycline resistance protein', 'hydrogenase 2 small subunit', 'LPS biosynthesis protein', 'pyridine nucleotide diphosphate kinase', 'transcriptional regulator', 'flagellar M-ring protein', 'surfactant protein', 'signal peptidase', 'succinoglycan biosynthesis transporter', 'succinoglycan biosynthesis transporter', 'membrane protein', 'MULTISPECIES: hypothetical protein', 'surfactant protein', 'uptake hydrogenase small subunit', 'DUF3592 domain-containing protein', 'quinoprotein glucose dehydrogenase', 'MULTISPECIES: VWA domain-containing protein', 'twin-arginine translocation pathway signal protein', 'ion transporter', 'surfactant protein', 'short-chain dehydrogenase', 'cytochrome c biogenesis protein', 'MULTISPECIES: Surfactant protein', 'DUF4349 domain-containing protein', 'cytochrome c-type biogenesis protein', 'cytochrome c biogenesis protein', 'cystathionine beta-lyase', '2-deoxy-D-glucanase', 'cytochrome oxidase', 'polysaccharide biosynthesis, transporter', 'SURF1 family protein', 'cytochrome c', 'RNA uridine-5-carboxymethylaminomethyl(34) synthetase enzyme MnmG', 'disulfide bond formation protein DsbA', 'MULTISPECIES: pyridine nucleotide-disulfide oxidoreductase', 'DSBA oxidoreductase', 'MULTISPECIES: disulfide bond formation protein DsbA', 'PEP-CTERM domain protein', 'twin-arginine translocation pathway signal protein', 'DUF374 domain-containing protein', 'D-alanyl-D-alanine carboxypeptidase', 'hydrogenase (NifE) small subunit HydA', 'cytochrome c biogenesis protein CcbA', '3-oxoacyl-ACP reductase', 'glucanase', 'DUF445 domain-containing protein', 'methylamine dehydrogenase (amicyanin) light chain', 'ubiquinol-cytochrome c reductase', 'capsule biosynthesis protein', 'marine proteobacterial sortase target protein', 'hydrogenase 2 protein HyaB', 'ABC transporter substrate-binding protein', 'Nif-Fe hydrogenase small subunit', 'cellulose synthase subunit B', 'capsule polysaccharide export protein', 'chain length determinant protein', 'MULTISPECIES: membrane protein', 'WYL domain-containing protein', 'MULTISPECIES: FAD-dependent oxidoreductase', 'beta-galactosidase', 'transmembrane protein', 'arylsulfatase', 'septum formation protein Maf', 'MULTISPECIES: hydrogenase-2 small chain', 'hydrogenase-1 small chain', 'cellulose synthase regulator BcsB', 'transmembrane cytochrome oxidase', 'alanine transporter', 'cellulose synthase BcsB subunit', 'carboxyvinyl-carboxyphosphonate phosphorylase', 'MULTISPECIES: 5-formyltetrahydrofolate cyclo-ligase', 'ABC transporter substrate-binding protein', '5-formyltetrahydrofolate cyclo-ligase', 'peptidase, partial', 'MULTISPECIES: flagellar M-ring protein FLIF', 'EscJ/YscJ/HrcJ family type III secretion inner membrane ring protein', 'EscJ/YscJ/HrcJ family type III secretion inner membrane ring protein, partial', '3-oxoacyl-ACP synthase', 'endonuclease', 'thiol:disulfide interchange protein', 'alkyl hydroperoxide reductase', 'alkyl hydroperoxide reductase, partial', 'MULTISPECIES: transmembrane cytochrome oxidase', 'histidine phosphatase family protein', 'cytochrome C', 'exopolysaccharide biosynthesis protein', 'MULTISPECIES: LPS biosynthesis protein', 'chain length determinant protein EpsF', 'hydrogenase', 'hydrogenase-2 small chain', 'nickel-iron hydrogenase small subunit', 'ferredoxin hydrogenase', 'PEP-CTERM sorting domain-containing protein', 'ABC transporter permease, partial', 'capsule polysaccharide transporter', 'lipoprotein-releasing system transmembrane subunit LolC, partial', '3-dehydroquinate dehydratase', '4Fe-4S ferredoxin, iron-sulfur binding domain-containing protein', '4Fe-4S ferredoxin iron-sulfur-binding domain protein', 'type III secretion system protein', 'serine/threonine protein kinase', 'Fe-S-cluster-containing hydrogenase subunit', 'Fe-S-binding protein', 'quinone-reactive Ni/Fe-hydrogenase small chain', 'MULTISPECIES: quinone-reactive Ni/Fe-hydrogenase small chain', 'thymidylate kinase', 'MULTISPECIES: Ni/Fe-hydrogenase', 'quinone-reactive Ni/Fe-hydrogenase small chain precursor', 'MULTISPECIES: hydrogenase 2 small subunit', 'cobaltin biosynthesis protein CobN', 'MULTISPECIES: hydrogenase', 'peptidase SB', 'cytochrome c-552/DMSO reductase-like, heme-binding domain protein', '[Ni/Fe] hydrogenase small subunit', 'hydrogenase-1 small chain', 'hydrogenase 2 small subunit, partial', 'trimethylamine-N-oxide reductase', 'hydrogenase 1, small subunit', 'hydrogenase-2 small chain protein', '[Ni/Fe] hydrogenase small subunit', 'hydrogenase, partial', 'uptake hydrogenase small subunit precursor', 'peptide ABC transporter substrate-binding protein', 'peptide-binding protein', 'MULTISPECIES: EscJ/YscJ/HrcJ family type III secretion inner membrane ring protein', 'protein BatD, partial', 'multidrug ABC transporter substrate-binding protein, partial', 'putative SerP-cotransposed membrane protein', 'ubiquinol cytochrome c oxidoreductase', 'ABC transporter', 'TraB/GumB family protein', 'ferritin', 'MULTISPECIES: N-acetyltransferase', 'twin-arginine translocation pathway signal', 'transcriptional regulator', 'MULTISPECIES: cytochrome c biogenesis protein', 'MULTISPECIES: cytochrome c-type biogenesis protein CcmH', 'DNA-binding protein', 'bacteriocin/lantibiotic ABC transporter', 'glycosyl transferase', 'HAF repeat/PEP-CTERM domain-containing protein', 'HAF repeat-containing protein', 'exported hypothetical protein', 'VWA domain-containing protein', 'MULTISPECIES: plus assembly protein', 'MULTISPECIES: bifunctional metallophosphatase/5-nucleotidase', 'hydrolase', 'hydrolase', 'fimbrial protein', 'MULTISPECIES: fimbrial protein', 'bifunctional metallophosphatase/5-nucleotidase', 'thermitase', 'LPXTG cell wall anchor domain-containing protein', 'repeat-containing protein', 'inhibition of morphological differentiation protein', 'S34, cytochrome c biogenesis protein', 'class C sortase', 'carbon-nitrogen hydrolase family protein', 'alkaline phosphatase', 'cell surface protein', 'adhesin isopeptide-forming adherence domain-containing protein', 'MULTISPECIES: inhibition of morphological differentiation protein', 'protease', 'nuclear export factor GLE1', 'haloacid dehalogenase', 'adhesin', 'MULTISPECIES: DUF445 domain-containing protein', 'nuclease', 'MULTISPECIES: LPXTG cell wall anchor domain-containing protein', 'lipase', 'type I fimbrial protein', 'deferochelatase/peroxidase EfeB', 'copper re

Figure 3.2. a terminal window screenshot reveals the results of a data extraction process conducted using Python scripts. To compile a comprehensive catalog of proteins from the database, a combination of four Python scripts was employed. The initial two scripts, Main_Proteins.py and ProteinHtml.py, collaborated to generate a comprehensive list of protein names. This list is presented here in the console, showcasing the successful extraction of protein names from the database.

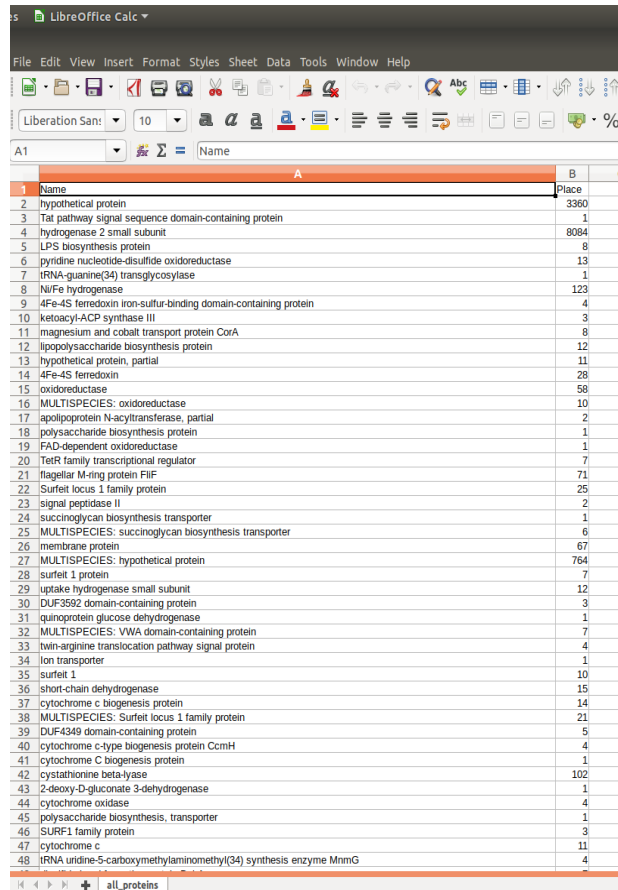
This helps to ensure reproducibility and transparency in bioinformatics research. Finally, GitHub provides an easy-to-use platform for researchers to access pre-existing bioinformatics tools and algorithms and allows them to adapt and extend these tools to meet their specific needs.

To analyse this information, the third script, Protein_Sorter.py, was used. This script not only extracts the information into a CSV file, compatible with spreadsheet programmes (Fig. 3.3A), but it also displays a small sample of names and hits along with the number of files and columns the CSV document will have in the console (Fig 3.3B).

However, the list of proteins contains all names present in the database, leading to an issue in which the programme cannot differentiate between proteins with similar names, such as "Multispecies hypothetical protein" and "hypothetical protein". To address this challenge, the fourth script, Incomplete_protein_counter.py, was employed. This script uses regular expressions to group similar protein names, for

instance, hydrogenases named "hydrogenase", "hydrogenase 2 small subunit", and "Multispecies hydrogenase 2 small subunit". The first set of scripts would categorise these hydrogenases differently, but this last script can group them into a single category and display the result in the console (Fig. 3.4A) and merge it in a spreadsheet (Fig. 3.4B).

A



LibreOffice Calc

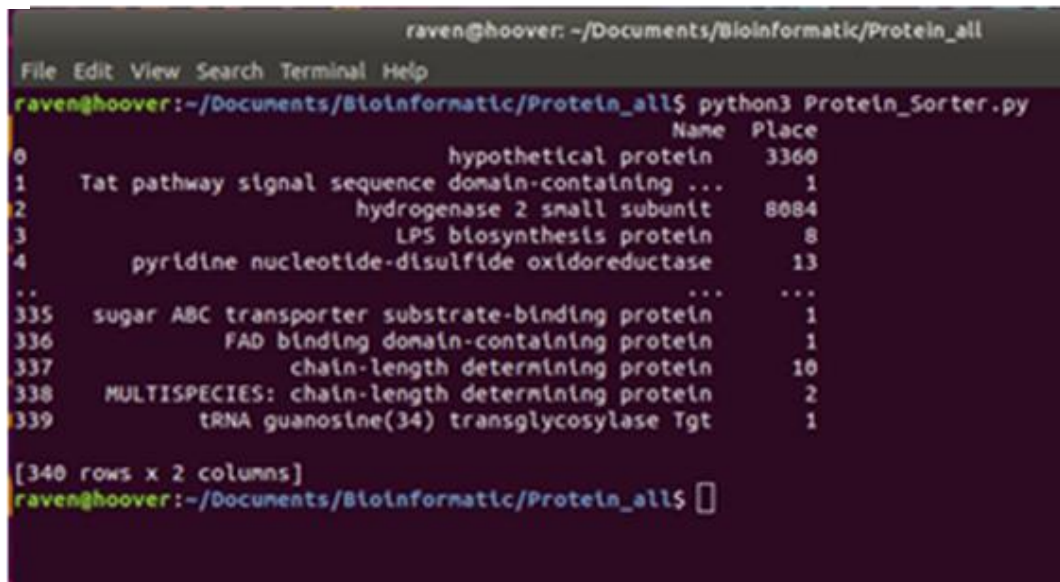
File Edit View Insert Format Styles Sheet Data Tools Window Help

Liberation Sans 10

Name	Place
hypothetical protein	3360
Tat pathway signal sequence domain-containing protein	1
hydrogenase 2 small subunit	8084
LPS biosynthesis protein	8
pyridine nucleotide-disulfide oxidoreductase	13
tRNA-guanine(34) transglycosylase	1
NiFe hydrogenase	123
4Fe-4S ferredoxin iron-sulfur-binding domain-containing protein	4
ketoacyl-ACP synthase III	3
magnesium and cobalt transport protein CorA	8
lipopolysaccharide biosynthesis protein	12
hypothetical protein, partial	11
4Fe-4S ferredoxin	28
oxidoreductase	58
MULTISPECIES: oxidoreductase	10
apolipoprotein N-acyltransferase, partial	2
polysaccharide biosynthesis protein	1
FAD-dependent oxidoreductase	1
TetR family transcriptional regulator	7
flagellar M-ring protein FlhF	71
Surfeit locus 1 family protein	25
signal peptidase II	2
succinoglycan biosynthesis transporter	1
MULTISPECIES: succinoglycan biosynthesis transporter	6
membrane protein	67
MULTISPECIES: hypothetical protein	764
surfeit 1 protein	7
uptake hydrogenase small subunit	12
DUF3592 domain-containing protein	3
quinoprotein glucose dehydrogenase	1
MULTISPECIES: VW domain-containing protein	7
twin-arginine translocation pathway signal protein	4
ion transporter	1
surfeit 1	10
short-chain dehydrogenase	15
cytochrome c biogenesis protein	14
MULTISPECIES: Surfeit locus 1 family protein	21
DUF4349 domain-containing protein	5
cytochrome c-type biogenesis protein CcmH	4
cytochrome C biogenesis protein	1
cystathionine beta-lyase	102
2-deoxy-D-glucate 3-dehydrogenase	1
cytochrome oxidase	4
polysaccharide biosynthesis, transporter	1
SURF1 family protein	3
cytochrome c	11
tRNA uridine-5-carboxymethylaminomethyl(34) synthesis enzyme MmmG	4

all_proteins

B



```

raven@hoover: ~/Documents/Bioinformatics/Protein_all
File Edit View Search Terminal Help
raven@hoover:~/Documents/Bioinformatics/Protein_all$ python3 Protein_Sorter.py
      Name  Place
0      hypothetical protein 3360
1  Tat pathway signal sequence domain-containing ... 1
2      hydrogenase 2 small subunit 8084
3      LPS biosynthesis protein 8
4  pyridine nucleotide-disulfide oxidoreductase 13
...
335 sugar ABC transporter substrate-binding protein 1
336 FAD binding domain-containing protein 1
337 chain-length determining protein 10
338 MULTISPECIES: chain-length determining protein 2
339 tRNA guanosine(34) transglycosylase Tgt 1

[340 rows x 2 columns]
raven@hoover:~/Documents/Bioinformatics/Protein_all$

```

Figure 3.3 A. CSV file that contains a list of protein names extracted from the database, along with the frequency of their occurrences. This CSV format is compatible with spreadsheet software, facilitating further analysis and manipulation. B. showcases grouped categories of proteins in the console, providing a snapshot of the data analysis process. This representation helps in understanding how the proteins have been categorized based on their names and occurrences.

The workflow had two limitations that had to be addressed in the process. Firstly, there was a limitation with the use of regular expressions, as some expressions could not be correctly grouped. For instance, a protein named "Type III fibronectin" was mistakenly counted among the Type III Secretion proteins due to similarities in their tags. To mitigate this issue, unique tags were utilised for each challenging group.

The second limitation was due to human error, as the categories and tags were manually written in the script based on our specific interests. This resulted in 211 out of 34,605 proteins not being classified. The main challenge behind this was that at least 114 of these proteins only appeared once in the database and some names were not correctly written in the database.

A

```

raven@hooover: ~/Documents/Bioinformatic/Protein_all
File Edit View Search Terminal Help
raven@hooover:~/Documents/Bioinformatic/Protein_all$ python3 Incomplete_protein_counter.py
Hydrogenases: 26829
Dehydrogenases: 29
Membrane proteins: 386
Cystathionine beta-lyase: 165
Hypothetical proteins: 4150
Exported hypothetical proteins: 1
Proteases: 366
Chaperones: 3
Tat pathway protein signal: 10
Capsule related proteins: 5
Exopolysaccharide biosynthesis protein: 3
Cellulose related proteins: 31
Lipoproteins: 6
Lipoproteins releasing systems: 3
Surfeit locus 1 family protein: 46
twin-arginine translocation pathway signal: 37
Tat related: 8
Apoproteins: 2
Carboxypeptidases: 115
Oxidoreductases: 103
tRNA: 6
Phospholipases: 33
Phosphatases: 27
Phosphodiesterase: 88
Metallophosphatase: 12
vinylcarboxyl carboxyphosphonate phosphorylmutase: 1
Dehalogenases: 67
Sortases: 12
Ferredoxins: 36
ABC transporter substrate-binding protein: 329
Inhibition of morphological differentiation proteins: 715
LPS biosynthesis proteins: 10
Phosphate acyltransferases: 12
Type I fimbrial protein: 2
Type II secretion: 1
Type III secretion: 86
Type VII secretion: 186
Fibronectins: 1
Cytochromes: 78
Zinc ribbon domain: 2
ketoacyl-ACP synthase III: 3
Hydrolases: 23
Dehydratases: 43
Penicillin-binding proteins: 40
Cell wall related proteins: 25
Magnesium and cobalt transport protein CorA: 8
HAF repeat: 2
PEP-CTERM: 6
chitin-binding protein: 69
VWA domain-containing: 16
Pilus related proteins: 2
Flagellar related proteins: 73
Deferochelatase/peroxidase EfeB: 1

```

B

Hydrogenases	26829
Hypothetical proteins	4150
Inhibition of morphological differentiation proteins	715
Membrane proteins	386
Proteases	366
ABC transporter substrate-binding protein	329
Type VII secretion	186
Cystathionine beta-lyase	165
Carboxypeptidases	115
Oxidoreductases	103
Phosphodiesterase	88
Type III secretion	86
cytochromes	78
Flagellar related proteins	73
chitin-binding protein	69
Dehalogenases	67
DUF445 domain-containing protein	58
Surfeit locus 1 family protein	46
Dehydratases	43
Penicillin-binding proteins	40
twin-arginine translocation pathway signal	37
Ferredoxins	36
Phospholipases	33
Cellulose related proteins	31
Dehydrogenases	29
Phosphatases	27
Cell wall related proteins	25
Hydrolases	23
Copper resistance protein	19
VWA domain-containing	16
Metallophosphatase	12
Sortases	12
Phosphate acyltransferases	12
Tat pathway protein signal	10
LPS biosynthesis proteins	10
Magnesium and cobalt transport protein CorA	8
Lipoproteins	6
tRNA	6
PEP-CTERM	6
Capsule related proteins	5
Balkyl	4
Chaperones	3
Exopolysaccharide biosynthesis protein	3
lipoproteins releasing systems	3
ketoacyl-ACP synthase III	3
Apoproteins	2

Figure 3.4. A. displays the results of a script that has effectively grouped proteins based on their names, addressing the challenge of differentiating between proteins with similar names. This grouped representation in the console provides insights into how proteins with related names have been categorized for analysis. B. shows sorted groups of proteins presented in a spreadsheet document format. This format allows for a more organized and structured view of the categorized proteins, making it easier to explore and analyze the data. Please note that these figures show only a partial view of the database, providing a glimpse of the data analysis outcomes.

Despite these limitations, the four python scripts (Main_Proteins.py, ProteinHtml.py, Protein_Sorter.py and Incomplete_protein_counter.py) allowed me to generate 249 groups of proteins from the initial list of 34,605. These groups included heterogeneous items, hypothetical proteins, duplications, and well-known Sec-dependent proteins. To clean up these groups and select one WP randomly from each bacterial species representing one protein per group, the Cala.py, reptile.py, and Etna2.py scripts were used. Etna2.py, originally written by Dr Giuseppina Mariano, was modified for this study to make it compatible with the required Python libraries. These scripts checked the IDs of the groups in the Refseq NCBI database,

removed hypothetical proteins, and fetched the identical protein groups. Finally, one WP was selected randomly to represent the proteins present in the database. This resulted in 84 protein ‘families’, as seen in Table 3.2.

Table 3.2. Script-classified protein ‘families’ from the original database of 34,605 items.

<ul style="list-style-type: none"> • alpha-lytic protease prodomain-containing protein • terpene cyclase-mutase family protein • YcnI family protein • M4 family metalloproteinase • alkaline phosphatase family protein • bifunctional metallophosphatase-5'-nucleotidase • carboxypeptidase regulatory-like domain-containing protein • choice-of-anchor A-G-M family protein • copper resistance protein CopC • YPDG domain-containing protein • YncE family protein • signal peptidase I • zinc metalloproteinase • LPXTG cell wall anchor domain-containing protein • PepSY-associated TM helix domain-containing protein • signal peptidase II • Cys-Gln thioester bond-forming surface protein • gluconate 2-dehydrogenase subunit 3 family protein • alginate lyase family protein • ABC transporter substrate-binding protein • 4Fe-4S dicluster domain-containing protein • D-alanyl-D-alanine carboxypeptidase • NAD(P)-FAD-dependent oxidoreductase • hydrogenase small subunit • DsbA family protein • hydrogenase 2 small subunit • SDR family oxidoreductase • Hydrogenase-2 small chain • hydrogenase 2 operon protein HybA • type III secretion inner membrane ring lipoprotein SctJ • SURF1 family protein • membrane protein • lipoprotein-releasing ABC transporter permease subunit 	<ul style="list-style-type: none"> • CHR domain-containing protein • flagellar M-ring protein FlhF • c-type cytochrome • cytochrome c1 • HAD family hydrolase • HAD-IB family hydrolase • S8 family serine peptidase • FIVAR domain-containing protein • GumC family protein • LamG domain-containing protein • glycoside hydrolase family 3 C-terminal domain-containing protein • prealbumin-like fold domain-containing protein • RICIN domain-containing protein • right-handed parallel beta-helix repeat-containing protein • tetratricopeptide repeat protein • low representation • type VII secretion-associated serine protease mycosin • pyrroloquinoline quinone-dependent dehydrogenase • aldehyde dehydrogenase family protein • S1 family peptidase • lytic polysaccharide monooxygenase • extracellular solute-binding protein • GldG family protein • ExeM-NucH family extracellular endonuclease • multifunctional 2',3'-cyclic-nucleotide 2'-phosphodiesterase-3'-nucleotidase-5'-nucleotidase • lamin tail domain-containing protein • peptidase • Tat pathway • Exceptions • FAD-binding protein • leucine-rich repeat domain-containing protein • VWA domain-containing protein • type II secretion system F family protein • TIGR family protein • tandem-95 repeat protein
---	---

<ul style="list-style-type: none"> • HtaA domain-containing protein • hydrogenase 1 small subunit • DUF domain-containing protein • exo-alpha-sialidase • endo-1,4-beta-xylanase • chaplin • metallophosphoesterase • WP_081857365.1 • discoidin domain-containing protein 	<ul style="list-style-type: none"> • superinfection immunity protein • M1 family metalloproteinase • fibronectin type III domain-containing protein • serine hydrolase • phosphatase PAP2 family protein • Ig-like domain repeat protein • carbohydrate binding domain-containing protein
---	--

3.2 Results

To further analyse the 84 ‘families’ in my extracted database, I employed two approaches. Firstly, I utilised MUSCLE3 and HMMER to ‘clean up’ the protein groups which are likely to contain some groupings of unrelated proteins. HMMER (Hidden Markov Model-based sequence comparison) and MUSCLE3 (Multiple Sequence Comparison by Log-Expectation) are widely used bioinformatics tools for protein database analysis. HMMER uses Hidden Markov Models (HMMs) to identify and analyse patterns in protein sequences, allowing for the detection of conserved domains and functions, making it a useful tool for comparing and analysing groups of similar proteins. On the other hand, MUSCLE3 is a tool for multiple sequence alignment, which aligns multiple protein sequences in a way that maximises their similarities. This tool is used to identify evolutionary relationships between sequences, making it useful for grouping similar proteins based on their sequences and structures (Edgar 2004; Eddy 2011).

By utilising these two tools, I analysed and compared members of each protein group to determine if they were similar and if the twin arginines present in their signal peptides were conserved across a range of similar proteins. This workflow allowed me to categorise each group into three categories: conserved, mixed, or not conserved (Fig. 3.5). The not conserved groups were composed of heterogenic proteins, while conserved groups showed conserved sequences. The mixed groups were further divided into conserved and not conserved groups (Fig. 3.6).

In the second step, I filtered out well-known Sec-dependent proteins and assessed the conservation of the twin arginine signal and the hydrophobic C-tail in similar proteins by using a BLAST search in NCBI. This was an additional step to verify the results obtained from HMMER, which identified patterns in protein sequences. This process resulted in a reduction of the number of items from 84 to 42, and finally, by considering both my own knowledge and the data from NCBI, I selected the 38 most suitable items, as shown in Table 3.3.

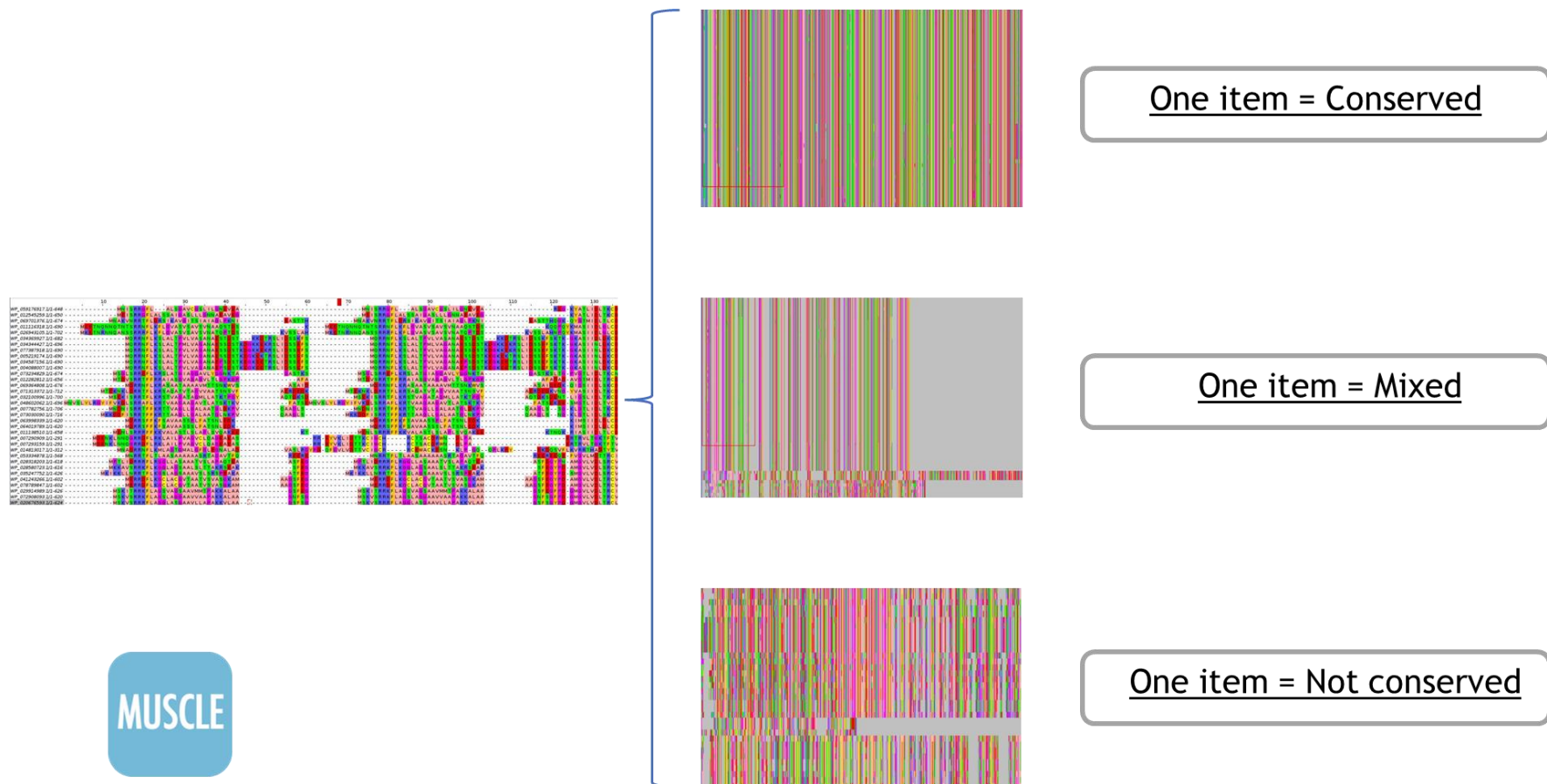


Figure 3.5. illustrates the potential outcomes of aligning protein sequences from each of the 84 distinct 'families' as part of a comprehensive analysis. The alignment process, performed using specific bioinformatics tools, aimed to assess the similarity and conservation of protein sequences within these families. The results of this analysis led to the categorization of each group into three distinct categories: conserved, mixed, or not conserved, as depicted in the figure.

Table 3.3. Identification of 38 candidate items of interest.

Manually curated items of interest	
<ul style="list-style-type: none"> • hydrogenase 1 small subunit • HAD-IB family hydrolase • hydrogenase 2 small subunit • type VII secretion-associated serine protease mycosin • HAD family hydrolase • DUF4349 • multifunctional 2',3'-cyclic-nucleotide 2'-phosphodiesterase-3 • ABC transporter substrate-binding protein • copper resistance protein CopC • 4Fe-4S dicluster domain-containing protein • D-alanyl-D-alanine carboxypeptidase • HtaA domain-containing protein • choice-of-anchor A-G-M family protein • DUF445 • LPXTG cell wall anchor domain-containing protein • S8 family serine peptidase • S1 family peptidase • lytic polysaccharide monooxygenase 	<ul style="list-style-type: none"> • carboxypeptidase regulatory-like domain-containing protein • YPDG domain-containing protein • terpene cyclase-mutase family protein • DUF1996 • NAD(P)-FAD-dependent oxidoreductase • DUF1134 • hydrogenase 2 operon protein HybA • signal peptidase I • type III secretion inner membrane ring lipoprotein SctJ • DUF1775 • TIGR family protein • flagellar M-ring protein FliF • Tat pathway • VWA domain-containing protein • DUF4185 • bifunctional metallophosphatase • alginate lyase family protein • GldG family protein • SURF1 family protein • YcnI family protein

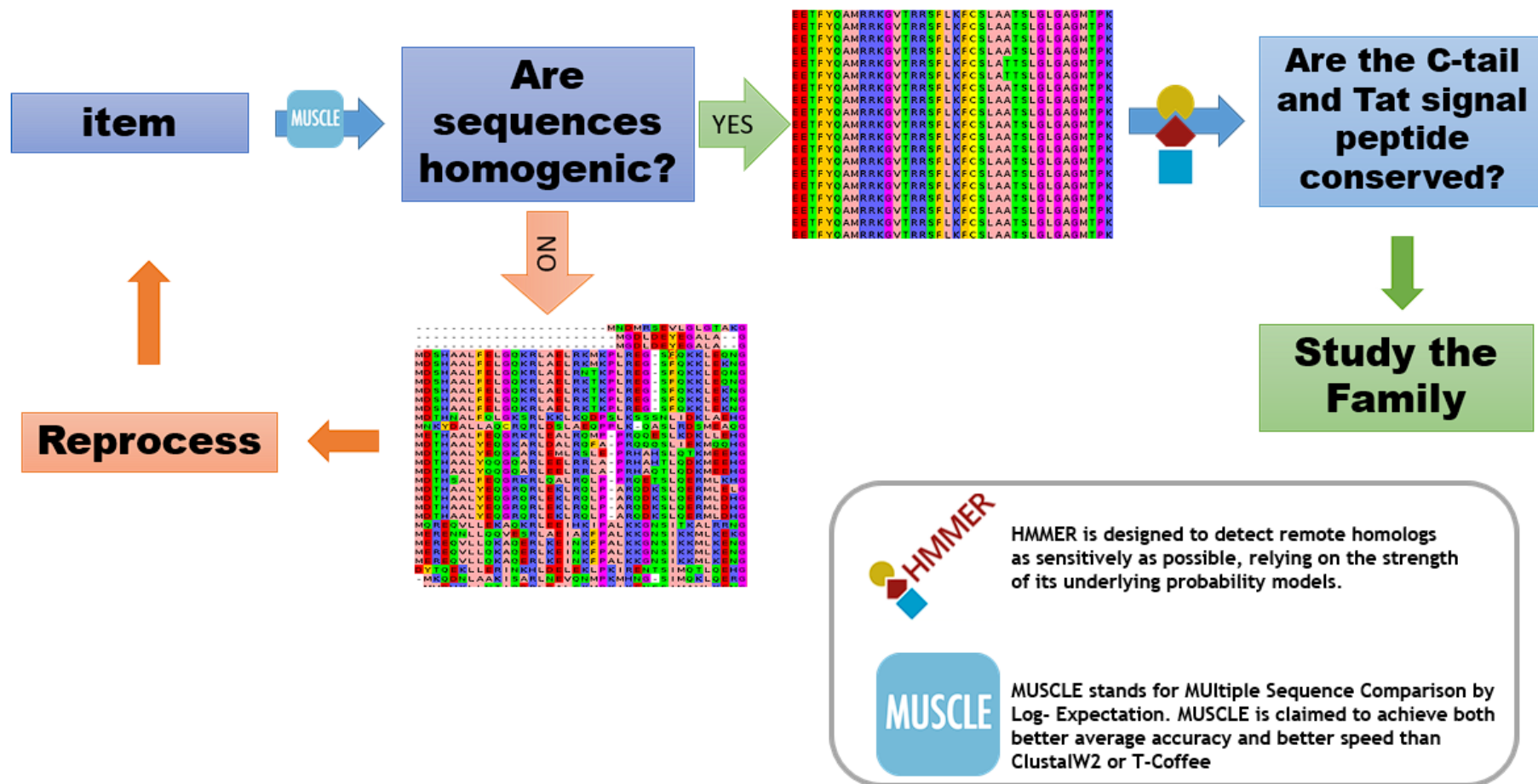


Figure 3.6. Work-flow for the analysis of candidate Tat-dependent C-tail anchored proteins.

In the final stage of the analysis, I focused on a deeper examination of the 38 selected protein families to assess the presence of other features that might indicate Tat dependence - for example the predicted presence of a (metal) cofactor, or whether a globular homologue was a known Tat substrate. This process resulted in a reduction of the list to 27 protein families, as shown in Table 3.4

Table 3.4. Candidate families of Tat-dependent tail anchored proteins selected for study.

Family proteins
<ul style="list-style-type: none"> • LPxTG cell wall anchor domain-containing protein • Lytic polysaccharide monooxygenase • S1 family peptidase • terpene cyclase/mutase family protein • YcnI family protein • HAD-IB hydrolase family • Type VII secretion-associated serine protease mycosin • SURF1 family protein • Copper resistance protein CopC • S8 family serine peptidase • TIGR family protein • Tat pathway signal sequence domain-containing protein • DUF11 • DUF1996 • DUF1134 • DUF1775 • VWA domain-containing protein • DUF4185 • alginate lyase family protein • GldG family protein • EipA • DUF4349 • ABC transporter substrate-binding protein • D-alanyl-D-alanine carboxypeptidase • HtaA domain-containing protein • choice-of-anchor A-G-M family protein • DUF445

A close manual analysis was next conducted on each group to examine the positioning of predicted domains. During this process, the HAD-IB hydrolase family (Table 3.5), consisting of 380 WPs, was filtered out because the active site motif overlapped with the twin arginines of the predicted signal peptide (Fig. 3.7).

Table 3.5. A visual showing part of the HAD-IB family hydrolase WP list.

ID	Source	Protein	Protein Name	Start	Stop	Strand	Organism
27997329	RefSeq	WP_0147387	HAD-IB family hydrolase	676578	677450	-	Modestobacter marinus
37082763	RefSeq	WP_0166427	HAD-IB family hydrolase	3168	3992	+	Streptomyces aurantiacus JA 4570
38557576	RefSeq	WP_0179478	HAD-IB family hydrolase	3704047	3704892	-	Streptomyces tirandamycinicus
39079098	RefSeq	WP_0184703	HAD-IB family hydrolase	31988	32860	+	Streptomyces albidoflavus
39098162	RefSeq	WP_0184894	HAD-IB family hydrolase	168622	169464	+	Streptomyces sp. SID8356
39169365	RefSeq	WP_0185479	HAD-IB family hydrolase	575607	576440	-	Streptomyces sp. LaPpAH-108
39183592	RefSeq	WP_0185621	HAD-IB family hydrolase	26477	27307	-	Streptomyces sp. SID8377
39572020	RefSeq	WP_0189588	HAD-IB family hydrolase	236694	237539	+	Streptomyces sp. CNB091
39676591	RefSeq	WP_0190637	HAD-IB family hydrolase	26256	27098	-	Streptomyces prunicolor NBRC 13075
39683108	RefSeq	WP_0190702	HAD-IB family hydrolase	33901	34743	+	Streptomyces hokutonensis
39814572	RefSeq	WP_0192018	HAD-IB family hydrolase	1651398	1652219	-	Tsukamurella sp. 1534
40138690	RefSeq	WP_0195270	HAD-IB family hydrolase	12776	13609	+	Streptomyces sp. FxanaD5
16854868	RefSeq	WP_0197466	HAD-IB family hydrolase	14699	15517	+	Rhodococcus rhodochrous ATCC 17895
40482431	RefSeq	WP_0198882	HAD-IB family hydrolase	3997330	3998166	-	Streptomyces purpureus KA281
40796018	RefSeq	WP_0202072	HAD-IB family hydrolase	517361	518200	-	Streptomyces sp. SID4923
41000356	RefSeq	WP_0203927	HAD-IB family hydrolase	20084	20914	+	Kribbella catacumbae DSM 19601
41182816	RefSeq	WP_0205549	HAD-IB family hydrolase	76522	77328	-	Embleya scabrispora DSM 41855
41259945	RefSeq	WP_0206321	HAD-IB family hydrolase	39346	40143	-	Amycolatopsis alba DSM 44262
41291999	RefSeq	WP_0206641	HAD-IB family hydrolase	7648123	7648944	-	Amycolatopsis benzoatilytica AK 16/65
43077390	RefSeq	WP_0213342	HAD-IB family hydrolase	73652	74470	-	Rhodococcus erythropolis DN1
45692500	RefSeq	WP_0235396	HAD-IB family hydrolase	3803475	3804311	+	Streptomyces niveus NCIMB 11891
45695302	RefSeq	WP_0235477	HAD-IB family hydrolase	4154625	4155455	+	Streptomyces roseochromogenus subsp. oscitans DS 12.976
45805237	RefSeq	WP_0235873	HAD-IB family hydrolase	2852881	2853735	+	Streptomyces thermolilacinus SPC6
45881435	RefSeq	WP_0236472	HAD-IB family hydrolase	2334773	2335597	+	actinobacterium LLX17
51999782	RefSeq	WP_0244938	HAD-IB family hydrolase	21921	22760	+	Streptomyces sp. AW19M42
52284249	RefSeq	WP_0247601	HAD-IB family hydrolase	107774	108640	-	Streptomyces exfoliatus DSM 41693
52398323	RefSeq	WP_0248754	HAD-IB family hydrolase	804646	805461	+	Saccharomonospora sp. CNQ490
47936181	RefSeq	WP_0253539	HAD-IB family hydrolase	328452	329258	-	Kutzneria albida DSM 43870
54082571	RefSeq	WP_0262495	HAD-IB family hydrolase	89256	90095	-	Streptomyces sp. ATexAB-D23
54359494	RefSeq	WP_0264532	HAD-IB family hydrolase	663	1481	+	[Actinopolyspora] iraqiensis IQ-H1
54821910	RefSeq	WP_0268743	HAD-IB family hydrolase	347841	348662	-	Jiangella gansuensis DSM 44835
55451296	RefSeq	WP_0275003	HAD-IB family hydrolase	94887	95705	-	Rhodococcus sp. UNC363MFTsu5.1
55457629	RefSeq	WP_0275067	HAD-IB family hydrolase	194930	195745	-	Rhodococcus sp. UNC23MFCrub1.1
55724152	RefSeq	WP_0277337	HAD-IB family hydrolase	157624	158469	+	Streptomyces sp. CNR698
55949101	RefSeq	WP_0279428	HAD-IB family hydrolase	125367	126188	-	Amycolatopsis taiwanensis DSM 45107
56419393	RefSeq	WP_0284160	HAD-IB family hydrolase	10388	11227	-	Streptomyces sp. ZEA171
56444485	RefSeq	WP_0284421	HAD-IB family hydrolase	71027	71866	+	Streptomyces sp. DpondAA-D4
56658171	RefSeq	WP_0286559	HAD-IB family hydrolase	65529	66338	-	Nocardioides sp. J9
56664847	RefSeq	WP_0286626	HAD-IB family hydrolase	105453	106271	-	Saccharomonospora paurometabolica YIM 90007
56801042	RefSeq	WP_0287991	HAD-IB family hydrolase	157585	158418	+	Streptomyces sp. 142MFC03.1
56850994	RefSeq	WP_0288485	HAD-IB family hydrolase	392179	393009	-	Thermocrispum agreste DSM 44070
56852223	RefSeq	WP_0288497	HAD-IB family hydrolase	309227	310036	+	Thermocrispum municipale DSM 44069
57007932	RefSeq	WP_0289594	HAD-IB family hydrolase	52469	53302	-	Streptomyces sp. SID4956
57405778	RefSeq	WP_0292897	HAD-IB family hydrolase	42251	43039	-	Cellulomonas sp. HZM
57529726	RefSeq	WP_0293825	HAD-IB family hydrolase	46169	47002	-	Streptomyces leeuwenhoekii

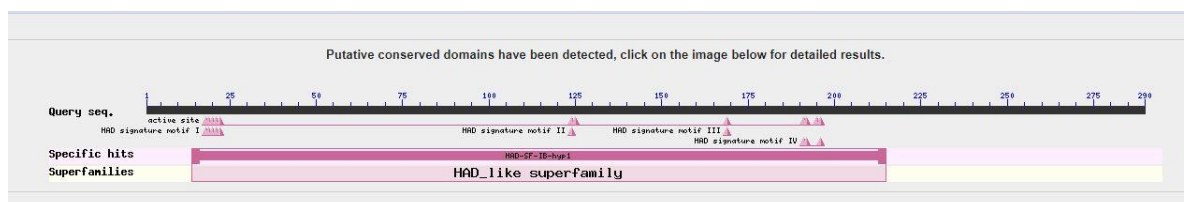


Figure 3.7. Putative conserved domains in the HAD-IB family. Output from NCBI.

Another candidate, the Tat pathway signal sequence domain-containing protein, showed the presence of a conserved cofactor binding domain (pyridoxal phosphate-dependent aminotransferase). The remaining groups were evaluated similarly, and based on the results and limitations of time, a final selection of seven groups was made. The selected groups included the S1 peptidase family, lytic polysaccharide monooxygenase, YcnI family protein, HtaA domain-containing protein, LPXTG cell wall anchor domain-containing protein, terpene cyclase-mutase family protein.

3.3 Families with candidate Tat-dependent tail-anchored proteins

The final selection of candidate proteins was made from the seven chosen groups. Upon further analysis it was clear that the S1 family peptidase was the most homogeneous group while the other five groups contained several unrelated protein families, each containing multiple proteins from different organisms, that had been grouped together due to common naming. The seventh group, referred to as the "others" group, also consisted of three individual proteins from various families. The characteristics and sequence conservation of the candidate proteins within each group are discussed in the subsequent sections.

Figs. 3.8, 3.11, 3.13, 3.15, 3.19, 3.20 and 3.22 showcase a representation of the selected candidates using Jalview software (Procter et al. 2021). The representation was designed to show the N-terminal signal peptide region and C-terminal hydrophobic stretch with the middle part of the sequence omitted. The figures use a combination of two colour schemes, Zappo and PID, to provide an in-depth analysis of the proteins. In the Zappo representation, residues are coloured according to their physico-chemical properties, while in PID, residues are coloured in shades of blue, with the intensity reflecting the percentage of residues in each column that match the consensus sequence. Only the residues that agree with the consensus residue for each column are coloured, providing a clear visual representation of the protein's properties.

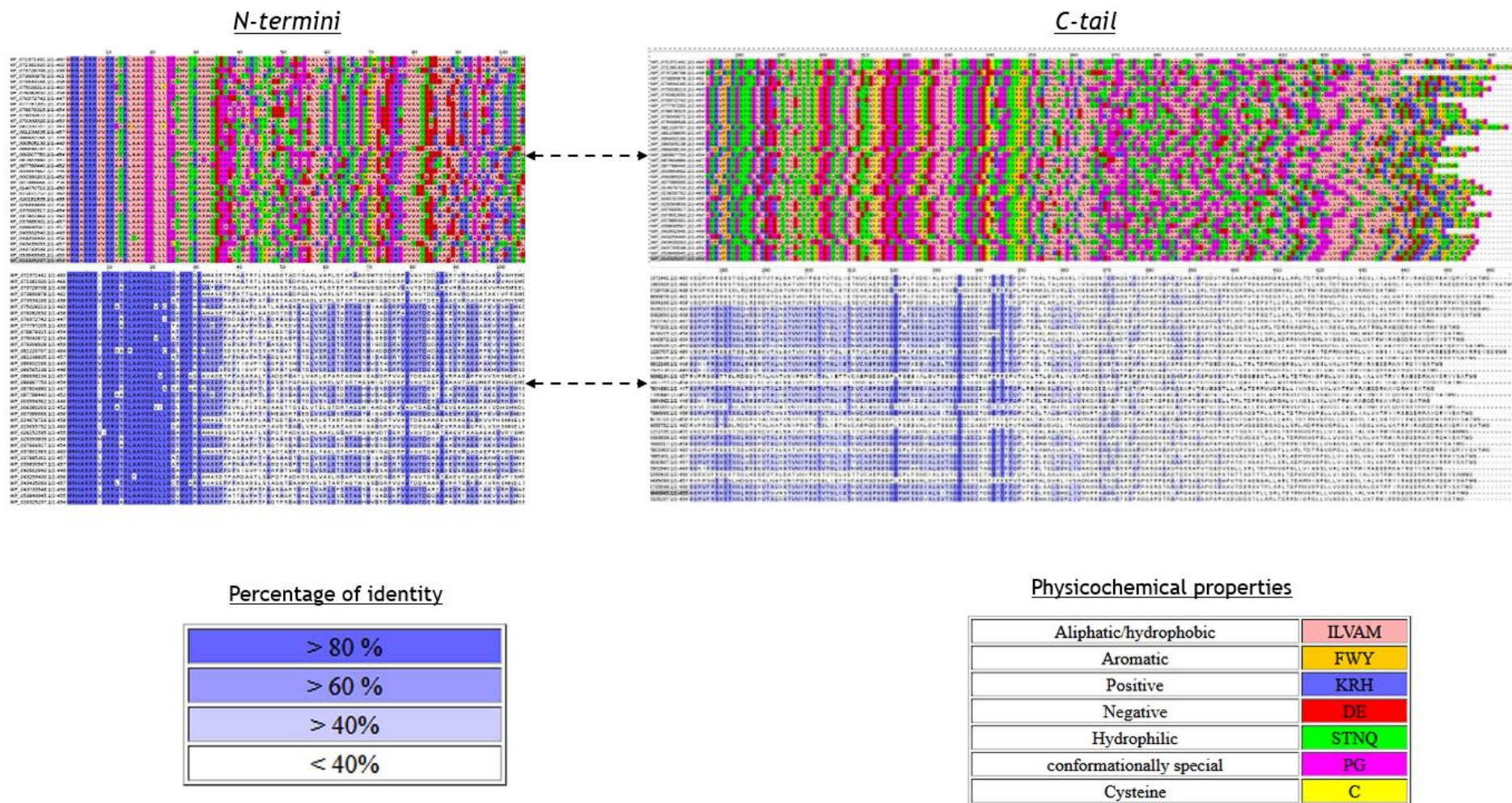


Figure 3.8. Representation of the N-termini and C-tail of the S1 family peptidase group. Sequences are coloured by percentage identity and physicochemical properties.

3.3.1 S1 family peptidase

The S1 family of peptidases are a group of hydrolases that cleave peptide bonds within proteins, and they play a key role in the degradation and recycling of proteins in cells. This peptidase family is frequently found in eukaryotes, where family members are often transmembrane proteins (Rawlings 2020). In the context of transmembrane transport, the S1 family peptidases may help in the cleavage and release of proteins from the membrane, as well as the processing of newly synthesised proteins to remove the signal peptides that target them for translocation across the membrane. The bacterial family members, all belonging to the *Streptomyces* genus, which were identified in this analysis show 100% conservation of the twin arginines in the signal peptide (Fig. 3.8, Fig. 3.9). The sequence conservation in the C-tail is lower across the members, but all shows high levels of hydrophobicity (Fig. 3.8).

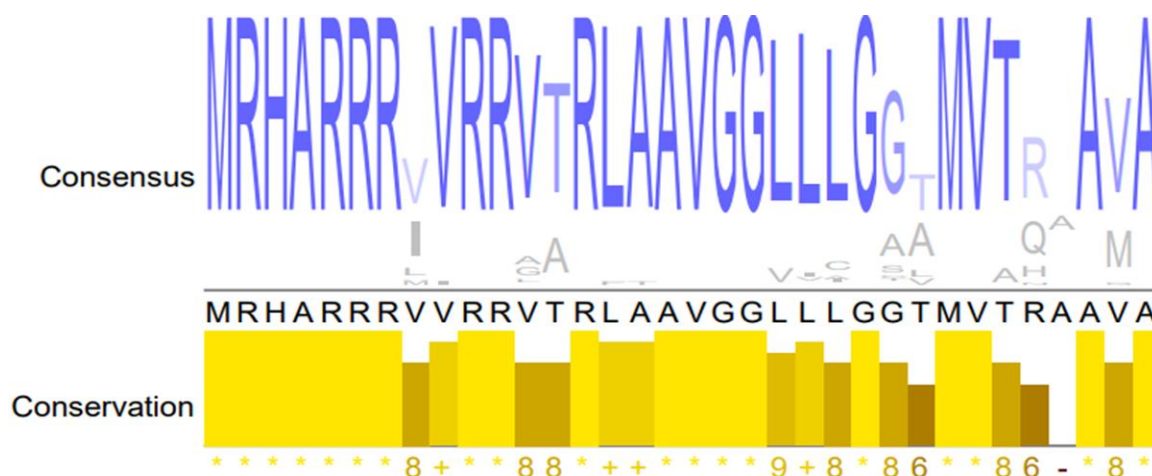


Figure 3.9. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of the signal peptide region of the S1 family Peptidases using Jalview. The consensus logo is shown at the top, where larger letters indicate the most conserved residues in the sequence. Below, the alignment conservation annotation is shown in a numerical index that quantifies the conservation, with the more conserved residue the bigger the number. 11 represents total conservation and is marked with a '*' symbol, while columns with a score of 10 have some diversity but maintain conserved properties and are marked with a '+' symbol.

3.3.2 Lytic polysaccharide monooxygenase

The lytic polysaccharide monooxygenases (LPMOs) are a group of copper-dependent enzymes that cleave polysaccharides such as cellulose, hemicellulose, and chitin. These enzymes play an important role in the degradation of plant cell walls and are often associated with plant pathogenicity. They have been described as biomass deconstruction boosters as they are able to mediate hydrolysis of recalcitrant cellulose (Singhania et al. 2021). Some LPMOs are predicted to have a transmembrane anchoring sequence at the C-terminus (Batth et al. 2022). The bacterial family members identified in this search all belong to the *Streptomyces* phylum. The amino acid conservation of the predicted signal peptide is shown in Fig. 3.10 and Fig. 3.11. The twin arginines are highly conserved, with the first arginine showing 98% conservation and the second being 100% conserved. The AxA signal peptidase cleavage site also conserved. The conservation in the C-tail is lower in comparison but shows high levels of hydrophobicity (Fig. 3.11).

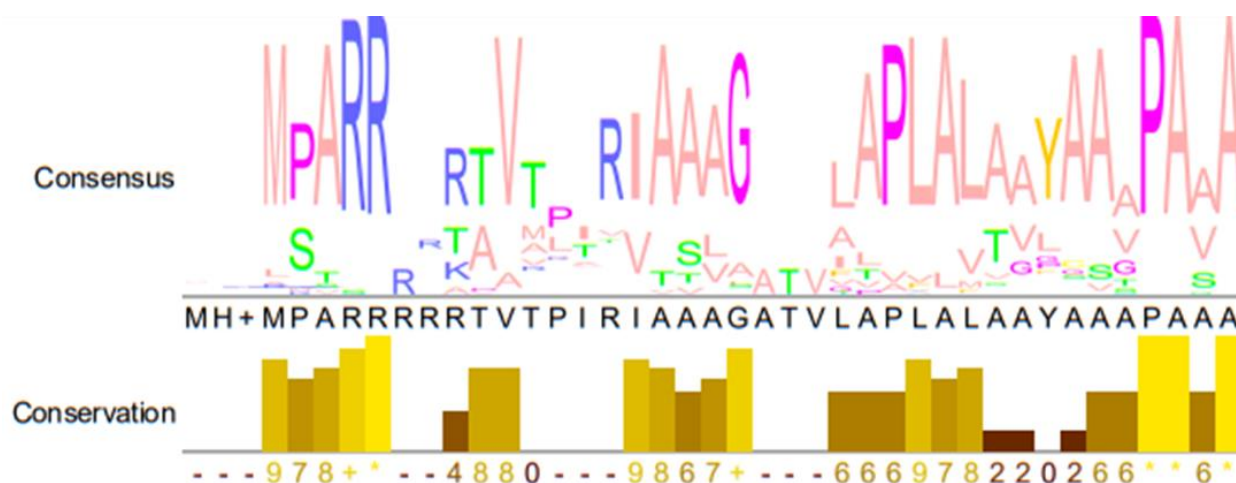


Figure 3.10. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of Lytic polysaccharide monooxygenase.

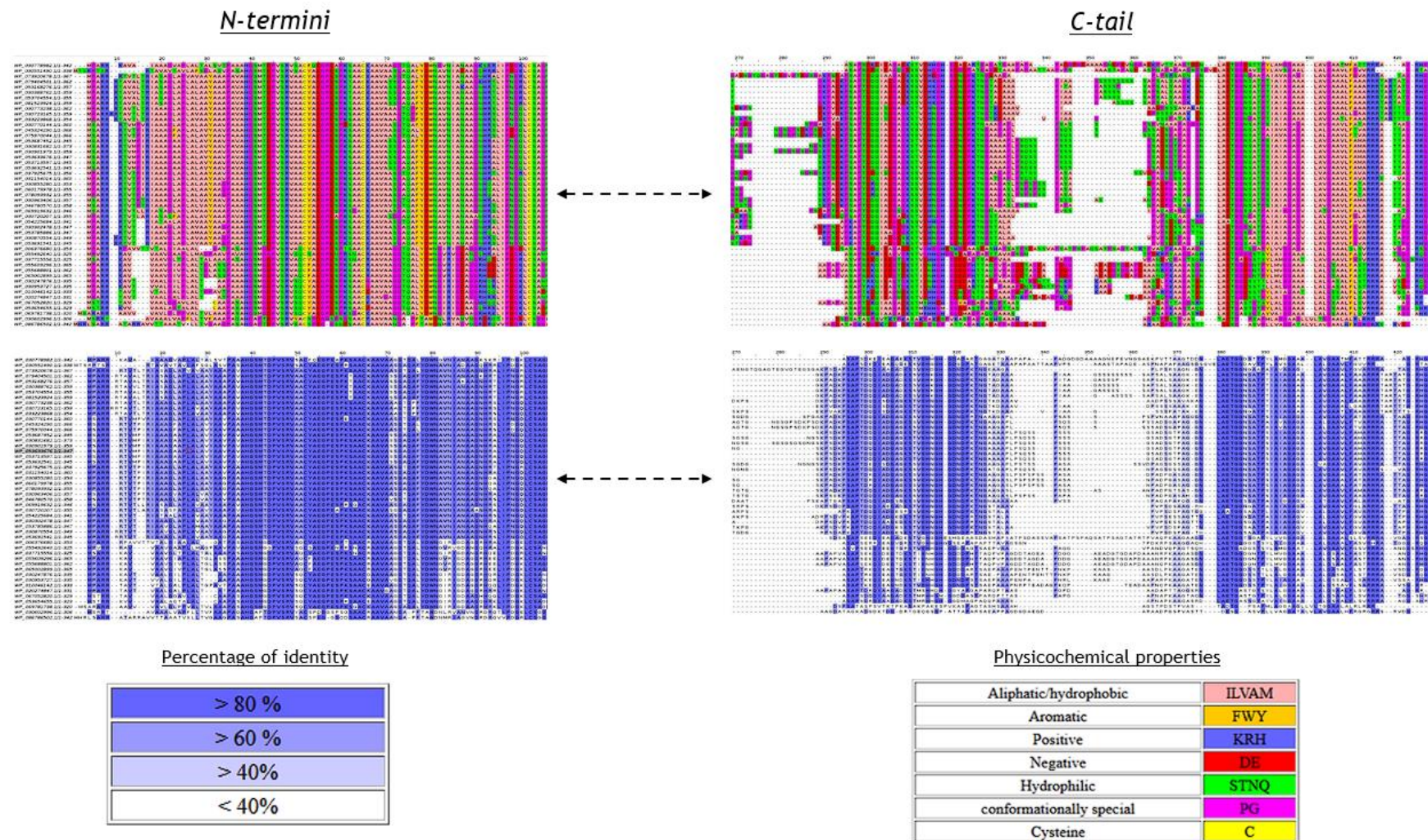


Figure 3.11. Representation of the N-termini and C-tail of the Lytic polysaccharide monooxygenase proteins sequences are coloured by percentage of identity and physicochemical properties.

3.3.3 YcnI family proteins

The *ycn* operon has been characterised in *Bacillus subtilis* and comprises genes encoding three proteins: the putative copper importer YcnJ, the copper-dependent transcriptional repressor YcnK, and YcnI, which harbours the uncharacterised Domain of Unknown Function 1775 (DUF1775) (Hirooka et al. 2012). DUF1775 domains are found across bacterial phylogeny, and bioinformatics analyses indicate that they are frequently encoded next to genes implicated in copper homeostasis and transport. YcnI from *B. subtilis* has been characterised biochemically and structurally, and shown to bind Cu(II) in a 1:1 stoichiometry, and has been proposed to be a copper chaperone (Damle et al. 2021). Fig. 3.12 and Fig. 3.13 shows that the twin arginines are poorly conserved across YcnI proteins, with the first arginine having a conservation of 81% and the second one only 51%. Indeed the *B. subtilis* YcnI has two lysines in the n-region of its signal peptide and has not been identified as a Tat substrate in that organism. The absolute conservation in the C-tail is also low but the proteins all show high levels of hydrophobicity (Fig. 3.13).

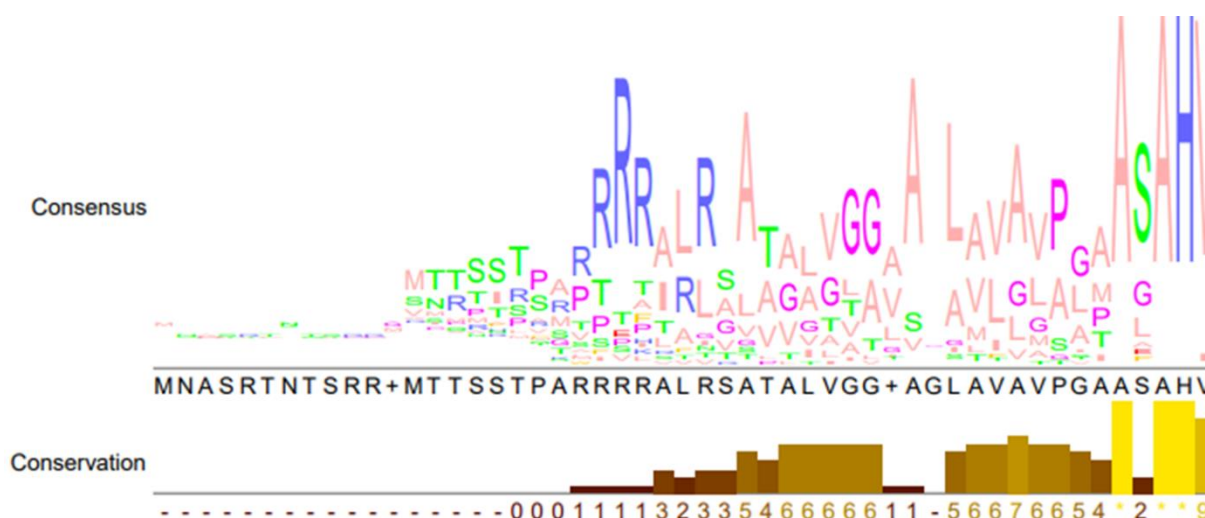


Figure 3.12. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of YcnI family proteins.

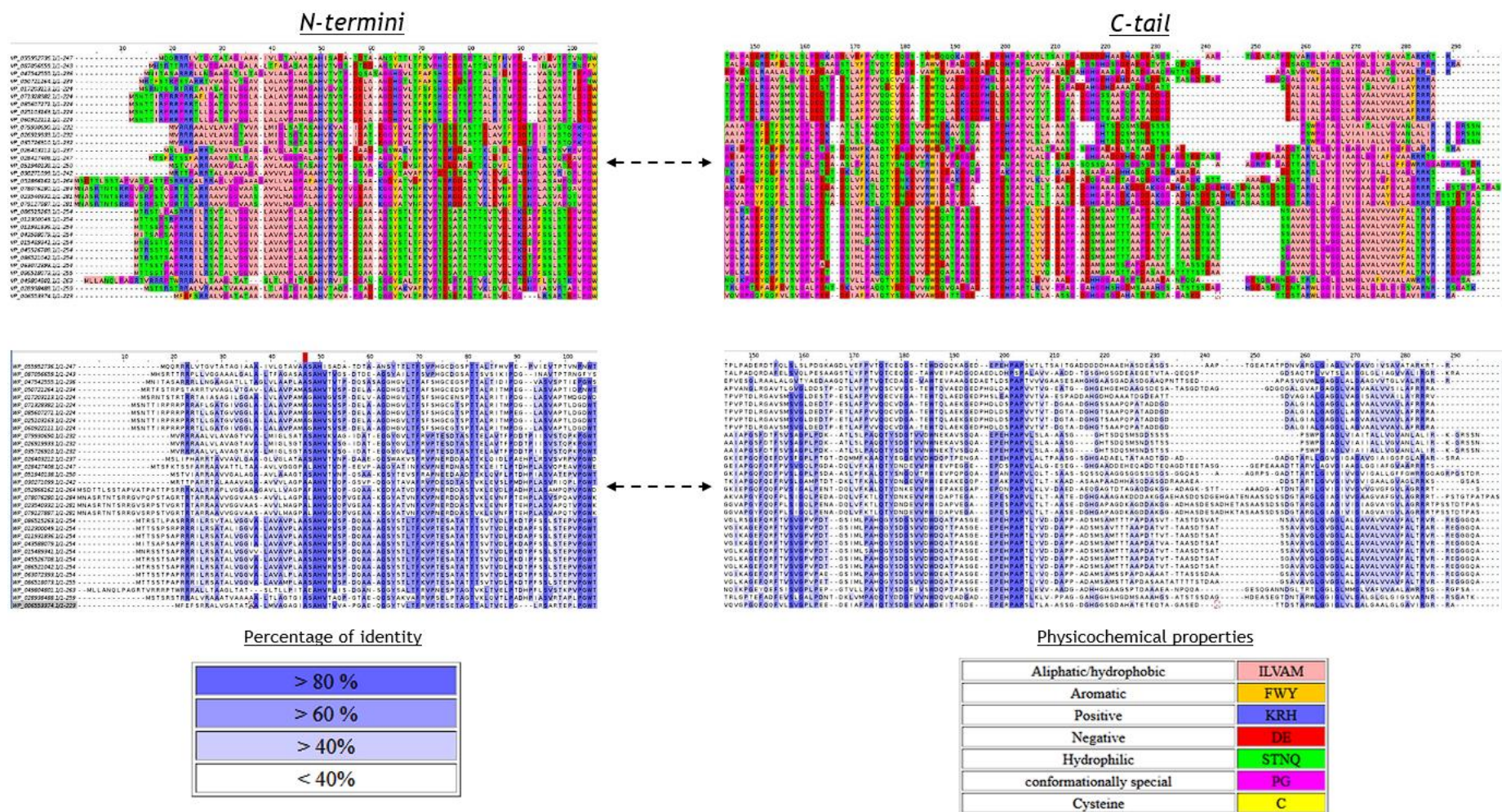


Figure 3.13. Representation of the N-termini and C-tail of the YcnI family proteins. Sequences are coloured by percentage of identity and physicochemical properties.

3.3.4 HtaA domain-containing protein

The HtaA domain-containing proteins are proteins that bind haem, a component of haemoglobin and other iron-containing proteins (Allen and Schmitt 2009). HtaA has been characterised from *Corynebacterium diphtheriae* and it is a transmembrane protein containing an N-terminal signal sequence and a transmembrane domain at the C-terminus (Lyman, Peng, and Schmitt 2021). It is proposed to play a role in iron acquisition, in particular from the haemoglobin-haptoglobin complex (Lyman, Peng, and Schmitt 2018). Fig. 3.14 and Fig. 3.15 show that the twin arginines are fully conserved in the HtaA signal sequences, although the AxA cleavage site is more variable. The conservation in the C-tail is also low but shows high levels of hydrophobicity (Fig. 3.15).

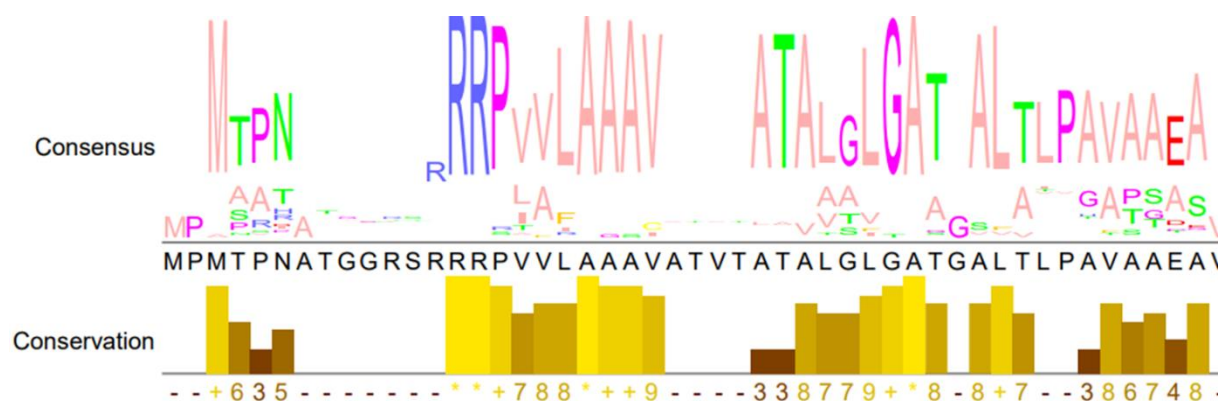


Figure 3.14. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of HtaA domain-containing proteins.

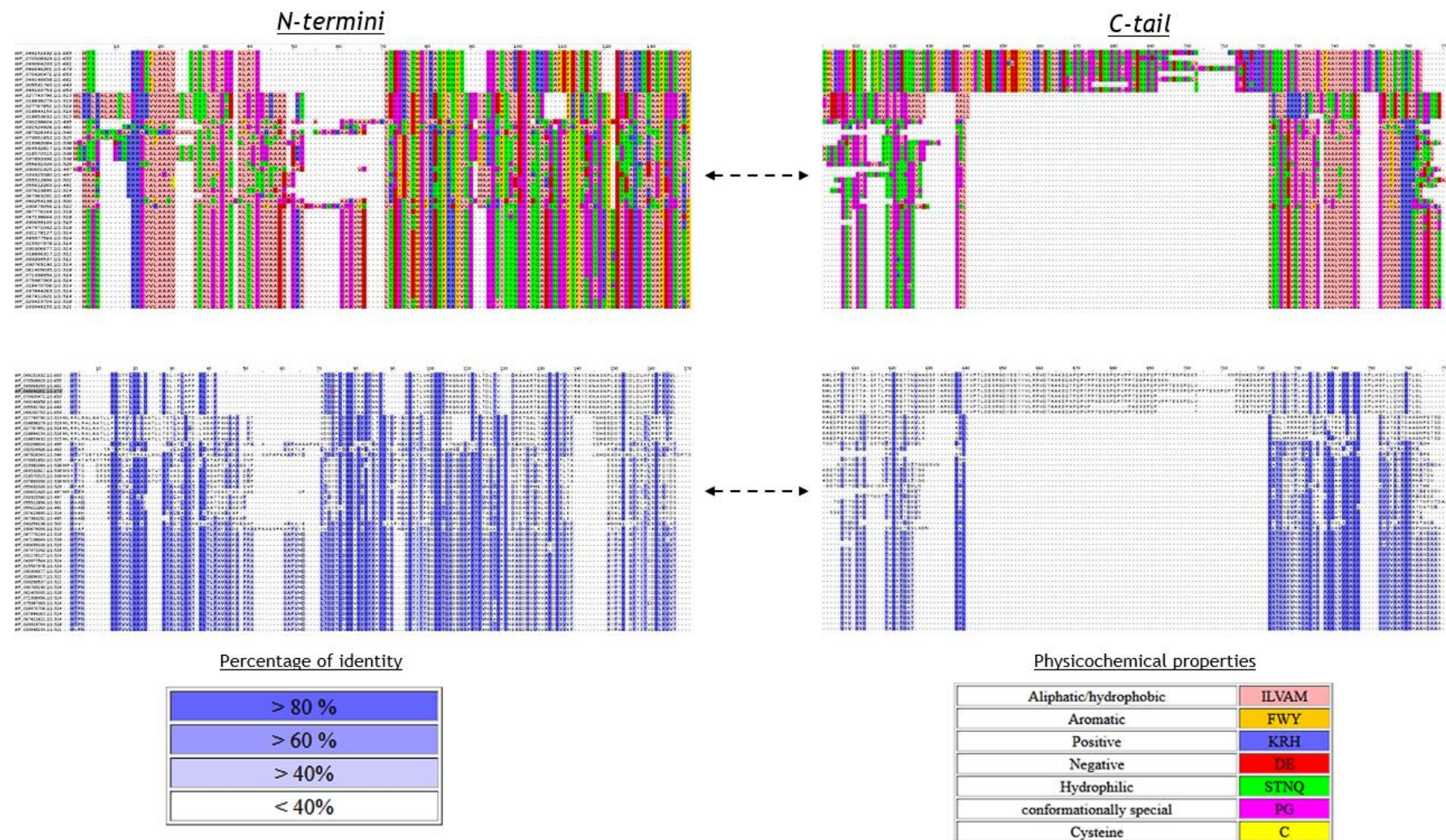


Figure 3.15. Representation of the N-termini and C-tails of the HtaA domain-containing proteins. Sequences are coloured by percentage of identity and physicochemical properties.

3.3.5 LPXTG cell wall anchor domain-containing protein

The LPXTG cell wall anchor domain-containing proteins are a group of proteins that are anchored to the cell wall in Gram positive bacteria. The 'LPXTG' refers to an amino acid sequence motif that is present near the C-termini of such proteins, which are recognised by an enzyme called sortase. Sortases are bacterial transpeptidases that have gained significant attention in protein chemistry due to their ability to catalyse chemoselective ligations of peptides and proteins. This enzymatic reaction, known as sortase-mediated ligation or "sortagging," requires the five amino acid LPXTG "sorting motif" in the peptide chain. Sortase A (SrtA) is known as the "housekeeping sortase", sortase B (SrtB) is important for the cell wall anchoring of proteins involved in iron acquisition and the SrtC family are responsible for catalysing transpeptidation reactions linking pilin subunits in bacteria. (Clancy, Melvin, and McCafferty 2010). Bacteria may have multiple sortase enzymes, for example *Streptomyces coelicolor* possesses seven different sortase genes (Ton-That, Mazmanian, and Schneewind 2001).

Sortase substrates are synthesised with N-terminal signal peptides and to date all sortase substrates that have been characterised use the Sec pathway. After translocation across the membrane and signal peptide cleavage, the sortase substrate remains anchored to the membrane through a C-terminal transmembrane helix that is adjacent to the LPXTG motif (Fig. 3.16). Sortase cleaves between the conserved threonine and glycine of the motif, ligating the carboxyl group of threonine to the amino group of pentaglycine on the cell wall peptidoglycan (Fig. 3.16). Sortases have been classified from A-F, with each recognising slightly different sorting signals (Schmohl and Schwarzer 2014). Sortase substrates are involved in various bacterial processes such as cell adhesion and spore formation (Marraffini, DeDent, and Schneewind 2006).

During my analysis, it became clear that the 'LPXTG cell wall anchor domain-containing protein' was a heterogeneous mixture of proteins. To divide the group by homology, I ran a muscle alignment and visualised it on Jalview. This resulted in three subgroups. The first two subgroups contained the most homogenous sequences, while the third one contained a mixture of proteins that were too different to group. I focused my analysis on the first two groups of LPXTG proteins, which had the most

homogenous members. The first group was found mainly in the *Streptomyces* phylum, the second one was found across *Streptomyces* but also in *Cellulomonas* and *Actinomyces*.

For analysis of group 1, Fig. 3.17 shows a well conserved signal peptide with the twin arginines 100% conserved but with a more variable AxX cleavage site. For the second group the signal peptide showed 92% conservation for the twin arginines and again variability in the cleavage site (Fig. 3.18). Fig. 3.19 shows that absolute sequence conservation in the C-tail for group 1 is relatively low, but hydrophobicity is conserved. The levels of conservation in the C-tail for group 2 is shown in Fig. 3.20, and again although absolute sequence conservation is low, the hydrophobicity is conserved. The C-terminal LPxTG sorting signal (Marraffini, DeDent, and Schneewind 2006) is not visible as a conserved motif in the Fig. 3.20 as the sortase motif can show sequence variation between organisms (Kruger et al. 2004).

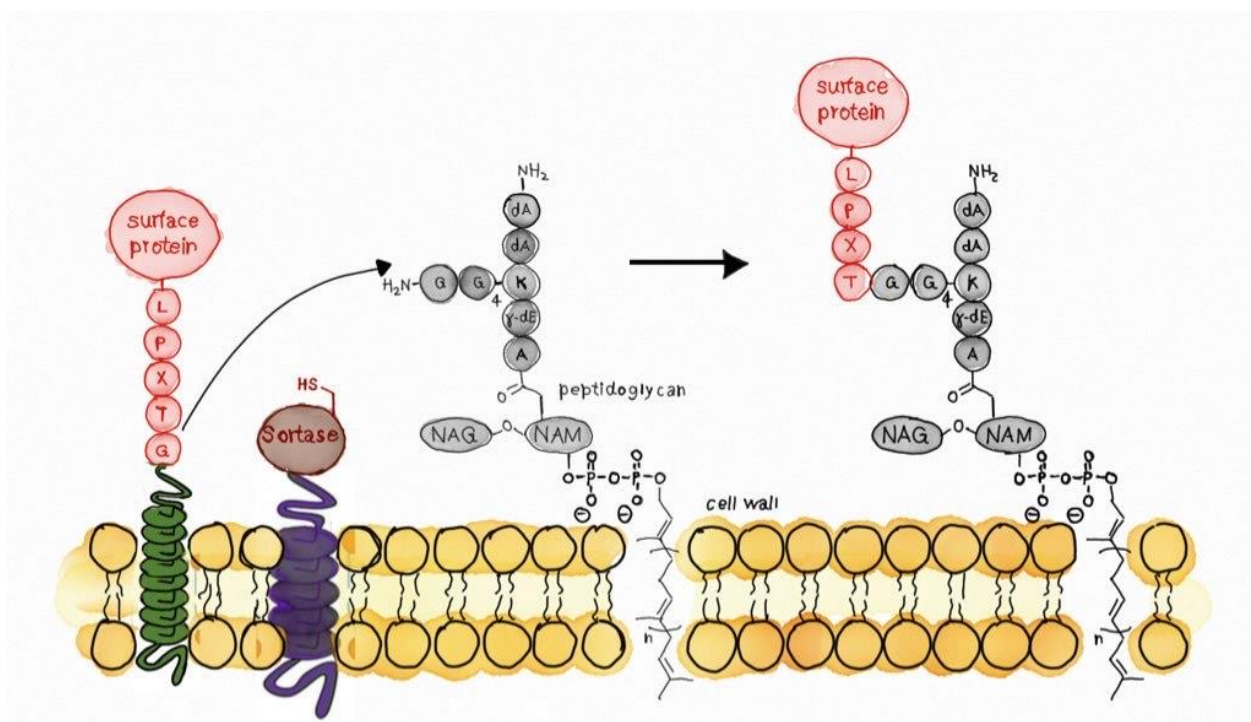


Figure 3.16. Sortase-mediated cell wall ligation. Membrane-bound sortase covalently attaches extracellular proteins to the cross-bridges of peptidoglycan. NAM, N-acetylmuramic acid; NAG, N-acetylglucosamine; dA, D-alanine; γ-dE: γ-D-glutamic acid. Modified from (Schmohl and Schwarzer 2014).



Figure 3.17. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of LXPTG group 1.

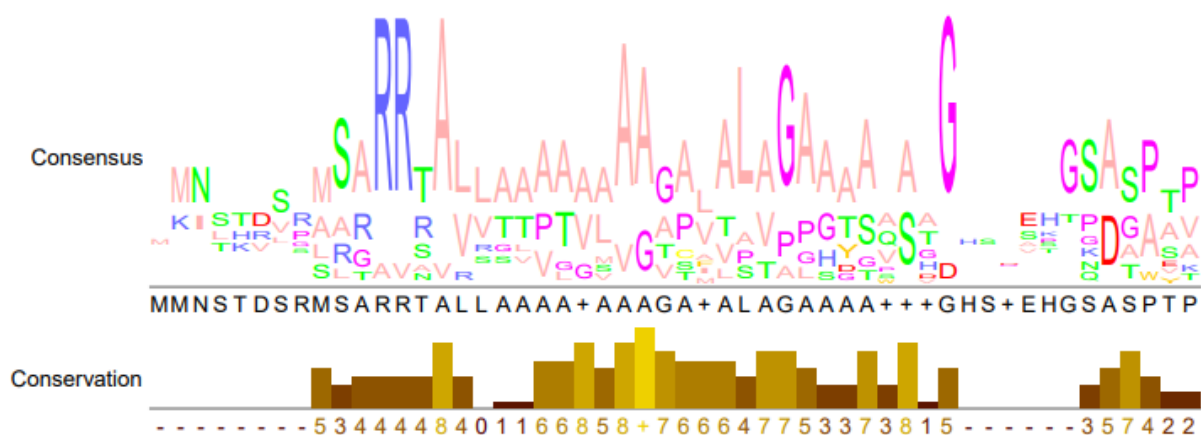


Figure 3.18. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of LXPTG group 2.

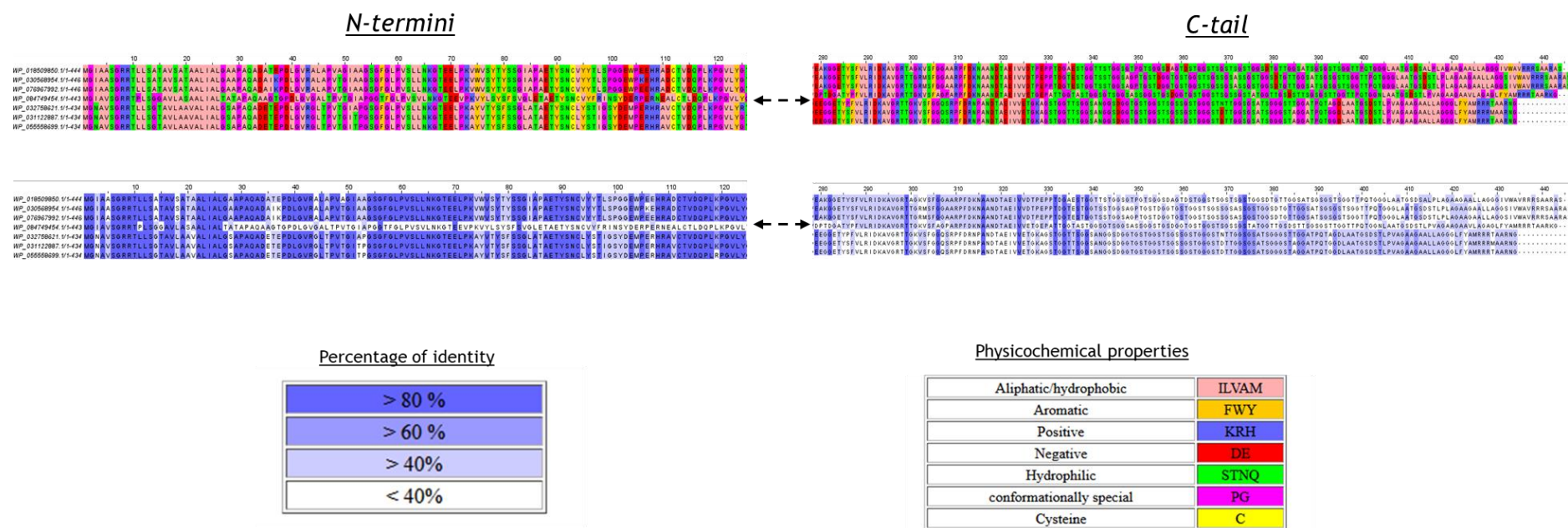


Figure 3.19. Representation of the N-termini and C-tail of the LPXTG family protein group 1. Sequences are coloured by percentage of identity and physicochemical properties.

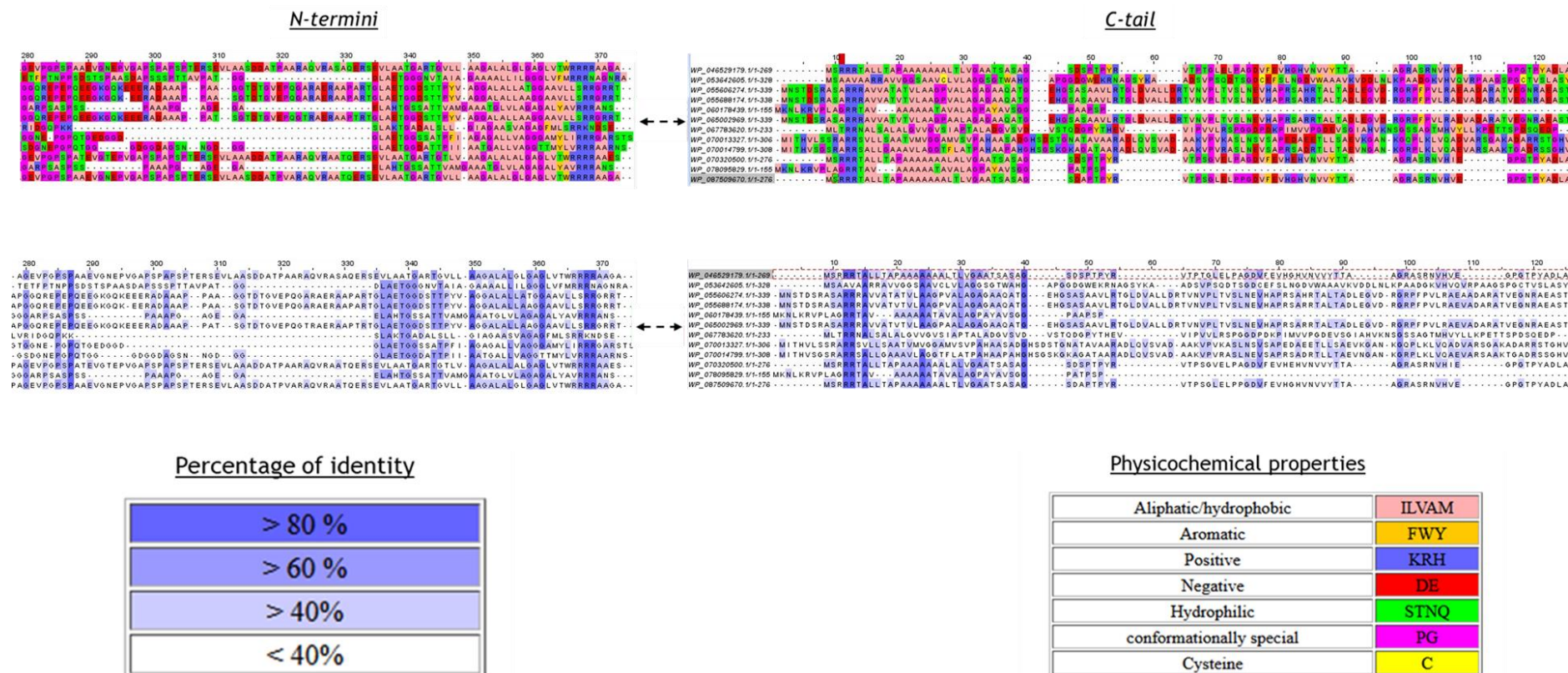


Figure 3.20. Representation of the N-termini and C-tail of the LPXTG family protein group 2. Sequences are coloured by percentage of identity and physicochemical properties.

3.3.6 Terpene cyclase-mutase family protein

The terpene cyclase-mutase protein family are involved in the biosynthesis of terpenes. Terpenes are a large and diverse class of compounds that include many natural products such as essential oils, pigments, and hormones. They make up the majority of secondary metabolites in plants, and have been extensively studied for their potential as antimicrobial, insecticidal, and weed control agents (Ninkuu et al. 2021). Terpenes are also made by some classes of bacteria, in particular the actinomycetes (Reddy et al. 2020).

The signal peptide of the terpene cyclase-mutase protein family shows 100% conservation of the twin arginines, however the positioning of the AxA cleavage site is less conserved (Fig. 3.21). The conservation in the C-tail is also low and does not show high levels of hydrophobicity (Fig. 3.22).

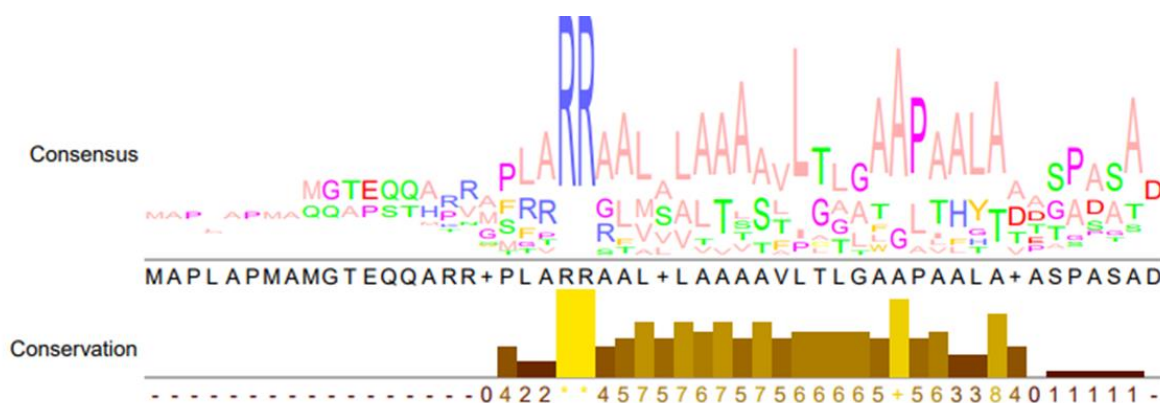


Figure 3.21. Graphical representation of the consensus logo and conservation annotation for sequence alignment analysis of the terpene cyclase-mutase family.

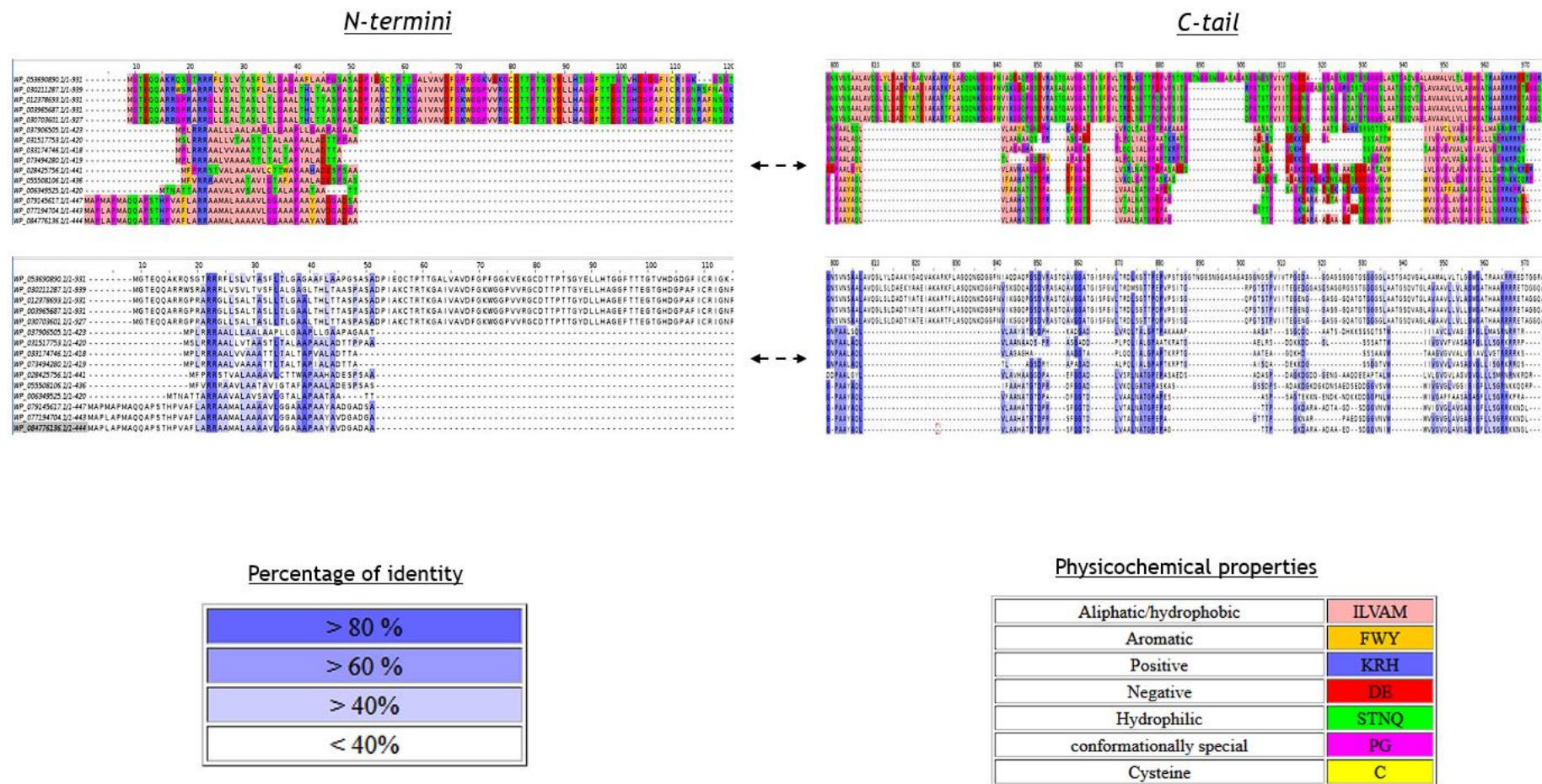


Figure 3.22. Representation of the N-termini and C-tail of the terpene cyclase/mutase family protein. Sequences are coloured by percentage of identity and physicochemical properties.

3.3.7 Others

There were other individual proteins that were also chosen as potential candidates even though they were not assigned to a larger family during the analysis. The presence of a putative C-tail was conserved when a BLAST analysis of the NCBI database was performed, and also, they appeared to be Tat-dependent upon manual inspection of their signal peptide regions. This group is composed of three candidates:

WP_083145723.1 is a twin-arginine translocation signal domain-containing protein found in *Mycolicibacterium parafortuitum* (Fig. 3.23).

```
>WP_083145723.1 twin-arginine translocation signal domain-containing  
protein [Mycolicibacterium parafortuitum]  
MWHVPSRPPTLSRRRVLAGAAVLAALGVAGAGCGTPPPPADLDDLTTALDRSRSDSLAAEVADSAKGKL  
AAVLTDIAAQRAAHAQALADEIVRLTGHQAPTATTEPSAAAPTSAPTPASVPPPTVDDVIGALRTSADSA  
AKSARVLSGYRAGLLASIAASCTAAYTVALAGAGRGR
```

Figure 3.23. Sequence of the protein with hypothesised Signal peptide and C-tail in bold. Twin arginine highlighted in yellow.

WP_056088981.1 is a DUF4349 domain-containing protein found in *Methylobacterium* sp. Leaf99 (Fig. 3.24).

```
>WP_056088981.1 DUF4349 domain-containing protein [Methylobacterium sp.  
Leaf99]  
MSGMAGRRSLVAAALGLAALLGGCSDARPPAPEVASPLPSMQANRSAAAKLAFTHDLTGLGLPPDRVAPHF  
AAARNRCLSDTGLGCVLVSSSLEDGARPSGRPQAQIQVRLPHESVAPYVAFLTAPLPGEAAGDVVLRRQA  
TRADDLTTPLEDGDRRLAQLTAYRARLDELAAKSDTRVEDLIRIAGELSKVQSQVEEGEARQRLQRRVD  
TEIVSVFFHSDGARSGAFAPVEDAWSQAGQTFGASAGDALRFVAVASLPWLPVAAIGLLLVRLWRLRRR  
RARIGRPLTEAQP
```

Figure 3.24. Sequence of the protein with hypothesised Signal peptide and C-tail in bold. Twin arginine highlighted in yellow.

```

>WP_012225357.1 methylamine dehydrogenase (amicyanin) small subunit
[Gluconacetobacter diazotrophicus]
MTRVTETRNDMERNGFDRLTERLARYLAGRSSRRGALARLGGWAASVPLFPLLPVWRGDARAADAPAAPS
AAPSPFAAKAQAKDDTKCDYWRYCAIDGNLCTTCGGGVHSCPPGTHPSPTSWIGTCFNPQDRRSYLIAYR
DCCGQDACNEQNCLGTDGDLPTYRPQANNDIWCFGTGSLLYNCSTAAIVGTAE

```

Figure 3.25. Sequence of the protein with hypothesised Signal peptide and C-tail in bold. Twin arginine highlighted in yellow.

WP_012225357.1 is a methylamine dehydrogenase (MDH) small subunit found in *Gluconacetobacter diazotrophicus* (Fig. 3.25). MDH is a soluble periplasmic enzyme which has been well studied in *Paracoccus denitrificans* (Davidson and Wilmot 2013; Davidson et al. 1997). The enzyme catalyses the oxidation of methylamine to formaldehyde and ammonia, using tryptophan tryptophylquinone (TTQ) as a cofactor. While TTQ-cofactor containing proteins including the small subunit of MDH have been speculated to be Tat substrates (Chang et al. 2011b), it has since been shown that they are not (Datta et al. 2001). Moreover, the vast majority of MDH small subunits were not identified in the original Chandra database because the C-terminal hydrophobicity is not conserved. For these reasons, WP_012225357.1 was not studied any further.

3.4 Discussion

Here a bioinformatics approach has been taken to identify new candidate Tat substrates that have C-tails. The initial search was conducted by Dr Govind Chandra, using standard programmes to identify proteins from the Refseq database with a twin arginine signal sequence and a likely TMH close to the C-terminus. The parameters were carefully set to rule out polytopic membrane proteins which are highly unlikely to be Tat dependent. However, as Tat signal sequences can also be TMHs (for example the monotopic Rieske proteins - Bachmann et al. 2006; De Buck et al. 2007), the search allowed for a total of two TMH within each protein.

From this initial search, which was carried out in 2017, a total of 34,634 candidates were identified. However, the inconsistent annotation of these proteins made it very difficult to group them into similar proteins manually. To this end I

applied bioinformatics tools to sort them into 84 protein families, into which almost all of the proteins could be placed. By far the largest number of proteins identified were the small subunits of periplasmic hydrogenase enzymes, and the iron sulphur protein HybA, validating the approach taken here. Periplasmic hydrogenases are widespread throughout bacteria and because they contain complex cofactors they are strictly Tat-dependent for their export (Rodrigue et al. 1999). All of the other families contained fewer members, although the type VII secretion system (T7SS) component mycosin was also identified multiple times. This may reflect the fact that up to five different copies of the T7SS can be encoded by a single mycobacterial genome, and as these bacteria are clinically relevant, they are over-represented in genome sequence databases. Although mycosin is a known tail anchored protein (Bunduc et al. 2021), my manual analysis indicated that the twin arginines are not conserved in mycosin proteins across the actinobacteria, and I therefore quickly ruled them out as candidate Tat substrates.

Following selection of 27 families for further bioinformatic study, I undertook detailed analysis of seven families, analysing them for conservation of the twin arginine motif and the hydrophobic C-terminus across multiple members (including those that had not been found in the 2016 search because they have been added to the Refseq database more recently). The selected proteins families and individual proteins come from a range of organisms, although the *Streptomyces* genus is heavily represented. It has been noted that these bacteria have the largest proportion of Tat substrates (Joshi et al. 2010; Widdick et al. 2006; Widdick et al. 2008) so it is perhaps not surprising that they may also have the most C-tail anchored Tat candidates.

Tsolis et al. (2018) undertook a comprehensive bioinformatic analysis of subcellular protein topology in *Streptomyces lividans*. From this they identified 30 proteins (from a total of 8,037) that contained a predicted C-tail, of which 16 were candidate sortase substrates. From my manual analysis of these 30 proteins, 21 do not have a twin arginine signal peptide. The remaining nine have plausible twin arginine signal peptides predicted by TatP (<https://services.healthtech.dtu.dk/services/TatP-1.0/>). However, three of these were found in the cell wall fraction of a *tatC* mutant and are therefore unlikely to

be Tat substrates (Widdick et al. 2006). Of the remaining six, one has been predicted as a Tat substrate in previous work (Schaerlaekens et al. 2004), but none have been validated experimentally. None of these 30 proteins were present among the 34,634 proteins identified in Dr Chandra's list.

As mentioned in the introduction, one of the limitations of this work is that it would not identify Tat substrate proteins that have a C-tail but lack a signal peptide because they are exported with Tat-dependent partner proteins. In this context, a previous search for candidate tail-anchored proteins encoded by *Streptomyces coelicolor* identified 20 such proteins that lacked any identifiable signal peptide and could potentially be exported through binding to a Tat targeting partner protein (Craney et al. 2011). Such candidates are difficult to identify bioinformatically as it would require a more complex search where a predicted Tat substrate was encoded in the genetic neighbourhood of a tail anchored protein and is beyond the scope of this thesis.

In conclusion, the work in this Chapter has allowed me to select nine candidate Tat-dependent C-tail proteins, WP_086565138.1, WP_011931836.1, P_019982084.1, WP_049064233.1, WP_031122887.1, WP_030568954.1, WP_046529179.1, WP_031517753.1 and WP_056088981.1 for further analysis. The results of this analysis are reported in Chapter 4.

Chapter 4. Experimental verification of novel Tat dependent tail-anchored proteins

4.1 Introduction

Although bioinformatic prediction programmes provide clues to the presence of signal peptides and TMHs, they can give false positive results. This is particularly true of the Tat signal peptide prediction algorithms TatFind and TatP which can have a substantial false positive hit rate (Gimenez et al. 2018; Joshi et al. 2010; Widdick et al. 2006; Widdick et al. 2008). It is therefore essential that predictions are supported with experimental data to confirm that candidate proteins possess the features that have been identified bioinformatically.

Many of the genome sequences present in databases are from organisms that cannot be genetically manipulated, or even cultured in the laboratory. In order to test predictions from such organisms, heterologous reporter systems are needed. Several reporters that can be used to assess Tat dependence of signal peptides have been developed. These include cell wall amidases, maltose binding protein, β -lactamase, GFP and agarase (DeLisa et al. 2002; Blaudeck et al. 2003; Gimenez et al. 2018; Tooke et al. 2017; McCann, McDonough, Pavelka, et al. 2007; Thompson et al. 2010; Widdick et al. 2008). While the amidase, maltose binding protein and GFP reporters use *E. coli* as the expression host, mycobacteria can be used as the expression system for β -lactamase and *Streptomyces lividans* is used for agarase. In this chapter the *E. coli* amidase reporter system is used to test Tat-dependence of candidate signal peptides because it is much faster than other assay systems.

Hatzixanthis, Palmer, and Sargent (2003) developed a robust reporter system to examine Tat-dependent C-tails using the SufI protein. SufI is a soluble Tat substrate of *E. coli* but will become anchored to the inner membrane if supplied with a C-terminal TMH. This will be used as the reporter to test whether candidate C-tails result in membrane integration.

4.2 Results

As previously discussed in Chapter 3, I identified nine potential Tat-dependent tail-anchored proteins for further study. These individual candidates were selected from a final list of families outlined in Chapter 3 and are shown in Table 4.1.

Table 4.1. Potential Tat-dependent tail anchored proteins selected for further study.

Candidate	Family	Chosen protein
1	S1 family peptidase	WP_086565138.1 <i>Streptomyces africanus</i>
2	YcnI family protein	WP_011931836.1 <i>Clavibacter michiganensis</i>
3	HtaA domain-containing protein	WP_019982084.1 unclassified <i>Streptomyces</i>
4	HtaA domain-containing protein	WP_049064233.1 <i>Corynebacterium striatum</i>
5	LPXTG cell wall anchor domain-containing protein	WP_031122887.1 <i>Streptomyces</i> sp. NRRL S-623
6	LPXTG cell wall anchor domain-containing protein	WP_030568954.1 <i>Streptomyces cyaneofuscatus</i>
7	LPXTG cell wall anchor domain-containing protein	WP_046529179.1 <i>Cellulomonas</i> sp. FA1
8	terpene cyclase-mutase family protein	WP_031517753.1 <i>Streptomyces</i> sp. NRRL F-5123
9	DUF4349 domain-containing protein	WP_056088981.1 <i>Methylobacterium</i> sp. Leaf99

In order to identify the signal peptide region and C-tail for each chosen candidate, the structures of the full-length and mature proteins without the signal peptide were predicted using RobeTTa fold (Baek and Baker 2022).

4.2.1 WP_086565138.1 S1 family peptidase from *Streptomyces africanus*

The predicted secondary structure of the mature sequence of WP_086565138.1 is shown in Fig. 4.1A. It consists of a globular domain connected to an α -helix, with a proline residue located near a predicted elbow. Fig. 4.1B includes the signal peptide region which is also predicted to be helical.

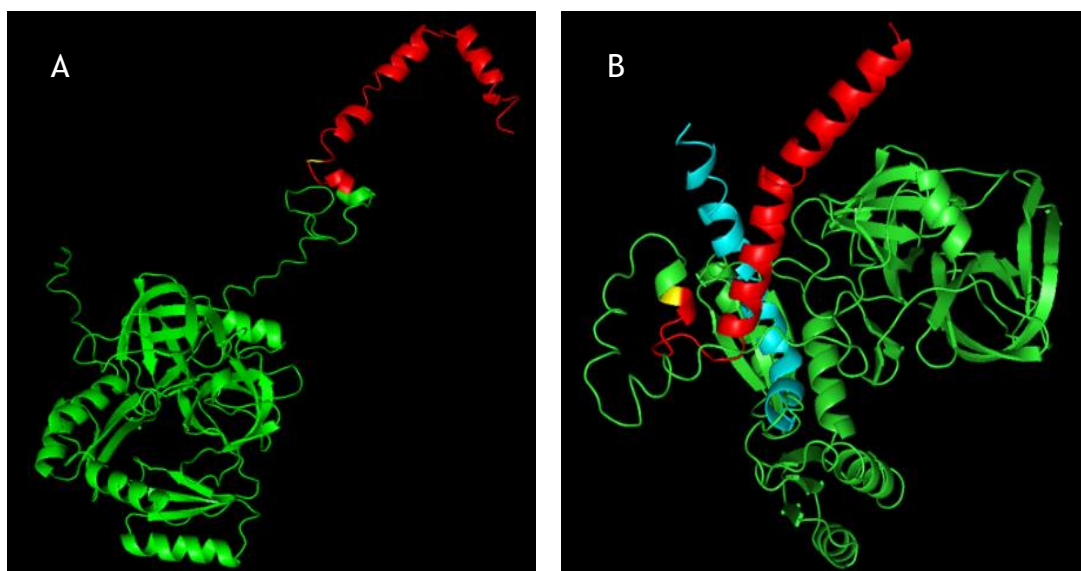


Figure 4.1. Structural predictions for WP_086565138.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 407 is highlighted in yellow.

4.2.2 WP_011931836.1 YcnI family protein from *Clavibacter michiganensis*

The predicted structure of WP_011931836.1 is shown in Fig. 4.2. It indicates that the TMH region (in red) is linked to the globular part of the protein through a disordered linker, with a proline residue immediately preceding the TMH region. BLAST searching indicates that homologues of this protein are present in other Actinomycetales including *Gordonia*, *Rhodococcus*, *Pseudonocardia*, *Streptomyces*, *Microbacterium*, *Kitasatospora*, *Streptacidiphilus* and *Curtobacterium*.

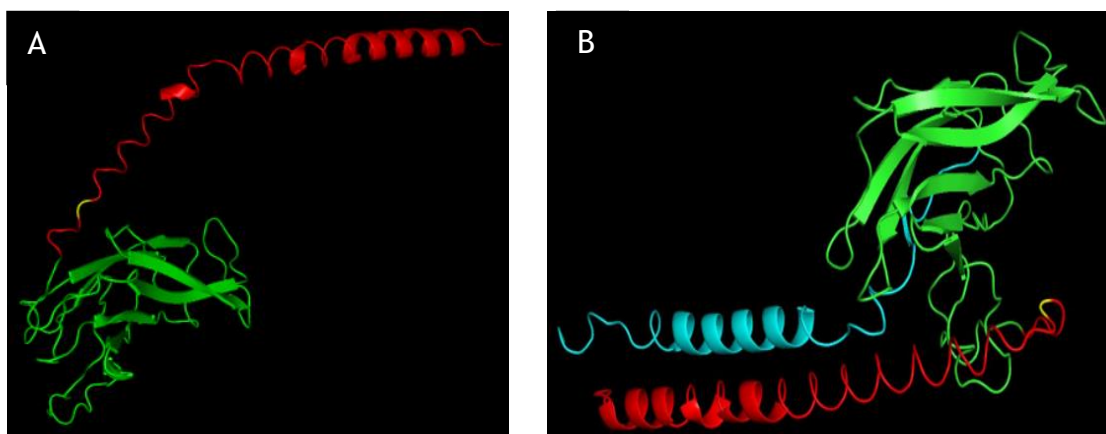


Figure 4.2. Structural predictions for WP_011931836.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 202 is highlighted in yellow.

4.2.3 WP_049064233.1 HtaA domain-containing protein from *Corynebacterium striatum*

The predicted structure of WP_049064233.1 is shown in Fig. 4.3. The protein is predicted to be largely β -sheet, with the α -helical TMH attached through a disordered linker region. Again, a proline residue is located at the end of the linker. Comparative analysis indicates that protein homologues are encoded in multiple subspecies of *Corynebacterium striatum*, as well as *Clavibacter* species.

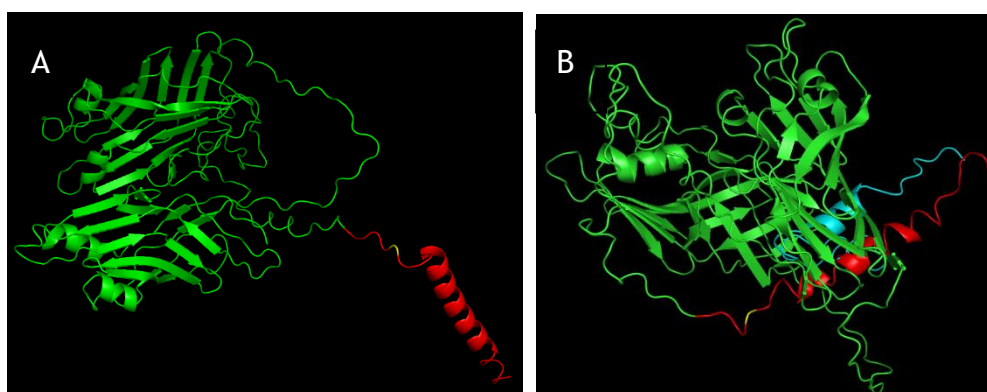


Figure 4.3. Structural predictions for WP_030238604.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 505 is highlighted in yellow.

4.2.4 WP_031122887.1 LPXTG cell wall anchor domain-containing protein from *Streptomyces* sp. NRRL S-623

The predicted structure of WP_031122887.1 is shown in Fig. 4.4. It reveals two β -sheet domains in the mature domain, which are linked to the TMH through a long, disordered region. As this protein is likely to be cell wall anchored following cleavage of the TMH it is possible that this long linker allows surface display of the globular domains. Again, a proline residue immediately precedes the TMH. Homologues of WP_031122887.1 are found only in the streptomycetes.

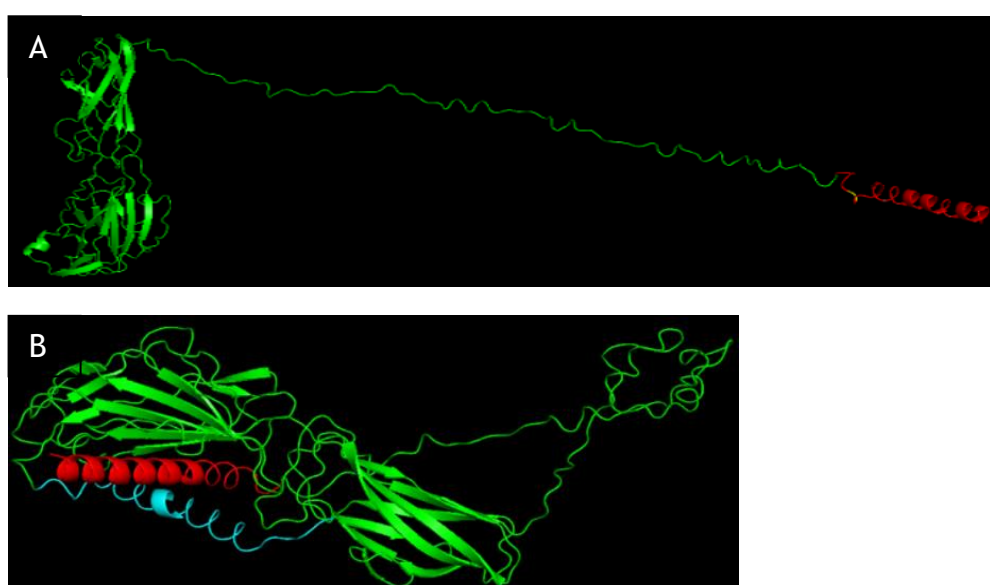


Figure 4.4. Structural predictions for WP_031122887.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 390 is highlighted in yellow.

4.2.5 WP_019982084.1 MULTISPECIES: HtaA domain-containing protein from unclassified *Streptomyces*

Homologues of WP_019982084.1 are found exclusively in the Actinomycetota phylum, within the orders Mycobacteriales, Streptomycetales, and Actinomycetales. These three orders belong to the class Actinobacteria and share several characteristics such as a high G+C content in their DNA and the ability to produce bioactive compounds.

The predicted structure of WP_019982084.1 is shown in Fig. 4.5. It indicates that the globular domain is composed of two lobes, each primarily comprised of β -sheet. An extended linker is predicted to attach the globular domain to the TMH. As the HtaA domain is predicted to bind iron (Allen and Schmitt 2009) it is possible that the long linker allows for surface exposure of the iron binding region.

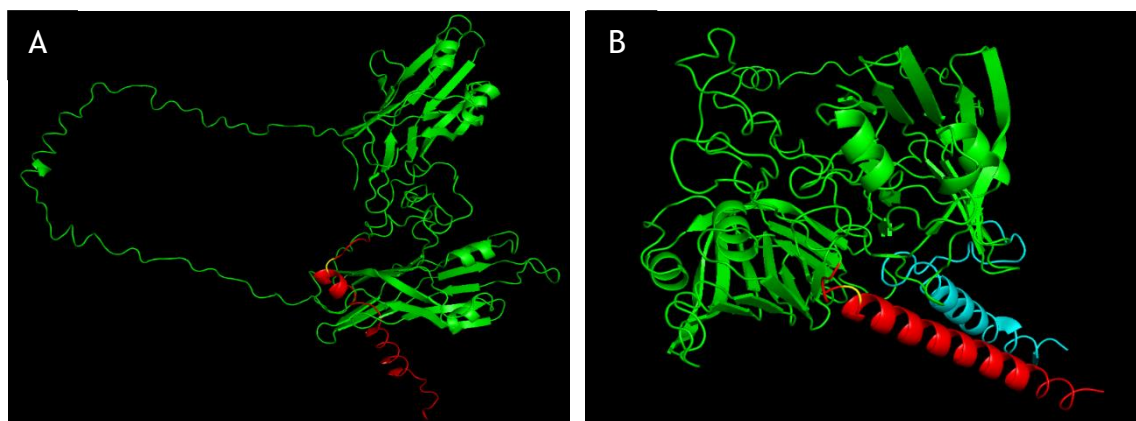


Figure 4.5. Structural predictions for WP_019982084.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 505 is highlighted in yellow.

4.2.6 WP_030568954.1 LPXTG cell wall anchor domain-containing protein from *Streptomyces cyaneofuscatus*

The predicted structure of WP_030568954.1 is shown in Fig. 4.6. It reveals a bi-lobed globular region formed from β -sheets, and a long unstructured linker preceding the TMH. As this protein is predicted to be cell wall anchored, the long linker could facilitate surface display. As with the other candidates, a proline residue is located at the end of the linker. BLAST analysis indicates protein homologues are found only in the *Streptomyces*.

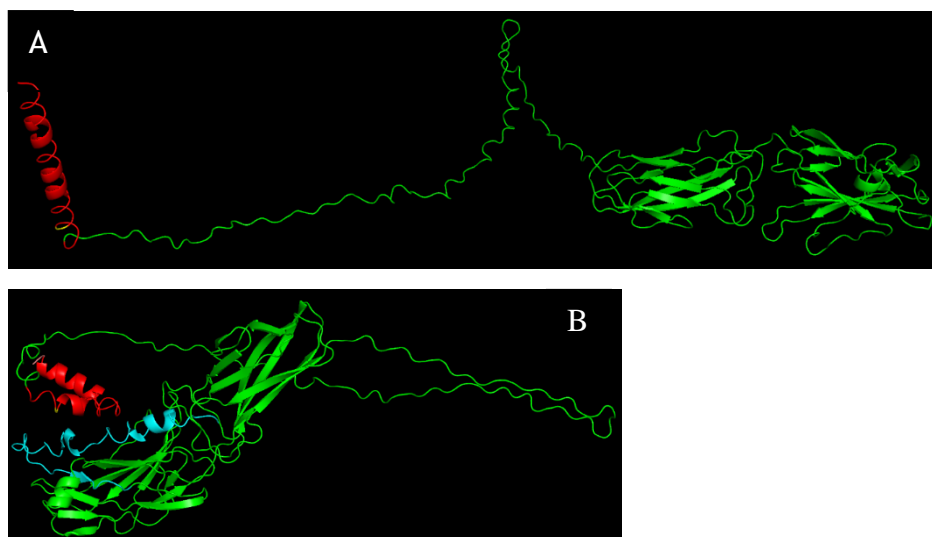


Figure 4.6. Structural predictions for WP_030568954.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 418 is highlighted in yellow.

4.2.7 WP_046529179.1 LPXTG cell wall anchor domain-containing protein from *Cellulomonas* sp. FA1

The predicted structure of WP_046529179.1 is shown in Fig. 4.7. The protein is predicted to comprise a small β -sheet domain attached via a long unstructured linker to the TMH. As the protein is predicted to be a sortase substrate the TMH would be cleaved off and the long linker may facilitate surface display. Again, a proline is found at the start of the TMH.

Homologues of WP_046529179.1 are found only in other *Cellulomonas* species.

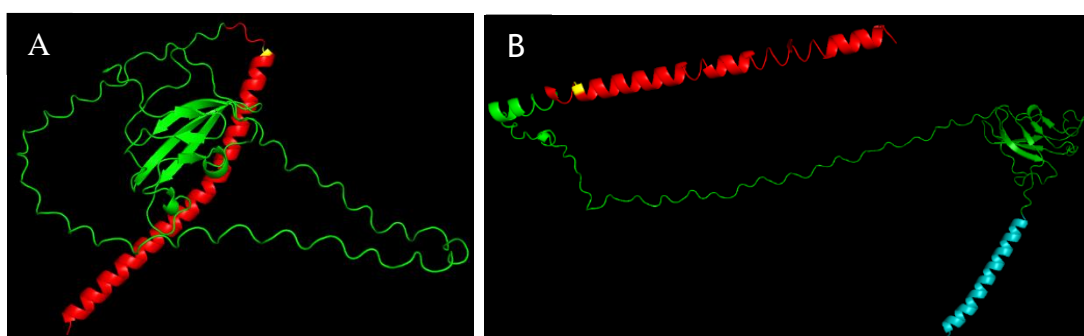


Figure 4.7. Structural predictions for WP_046529179.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 217 is highlighted in yellow.

4.2.8 WP_031517753.1 terpene cyclase/mutase family protein from *Streptomyces sp. NRRL F-5123*

The predicted structure of WP_031517753.1 (Fig. 4.8) suggests it is predominantly α -helical. A relatively short, hydrophobic linker is predicted to attach the mature domain to the TMH. A proline residue is found at the start of the hydrophobic stretch. BLAST analysis indicates that homologues of WP_031517753.1 are only present in the order Streptomycetales.

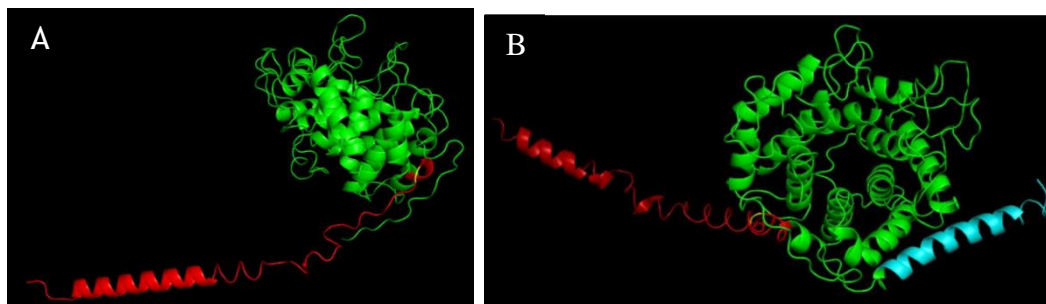


Figure 4.8. Structural predictions for WP_031517753.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 367 is highlighted in yellow.

4.2.9 WP_056088981.1 DUF4349 domain-containing protein from *Methylobacterium sp. Leaf99*

The predicted structure of WP_056088981.1 is shown in Fig. 4.9. The protein is predicted to comprise a small b-sheet domain attached via an unstructured linker to the TMH at the C-terminal side and to several α -helices, potentially a coiled coil structure, at the N-terminal side. A proline, shown in yellow, is predicted to lie in the middle of a long, kinked α -helix, just prior to the hydrophobic stretch. Homologues of WP_046529179.1 are found only *Methylobacterium sp.*

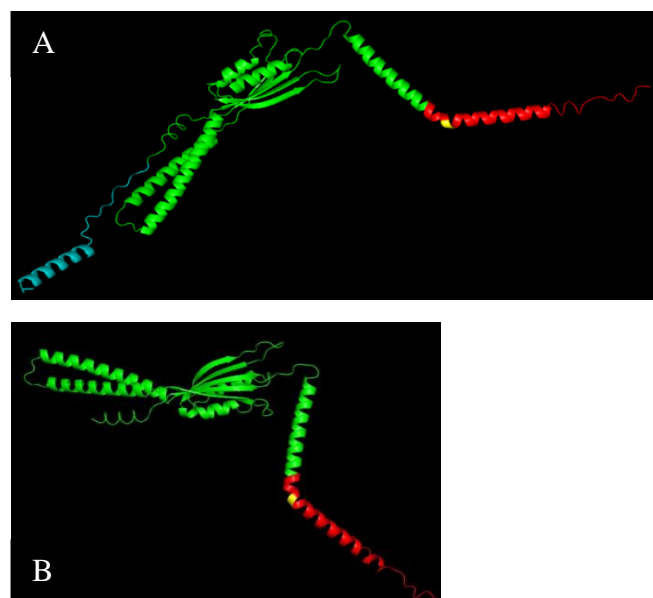


Figure 4.9. Structural predictions for WP_056088981.1 using RobeTTa fold. The protein is shown without the signal peptide (A), or in the precursor form (B). The signal peptide is shown in cyan, the C-tail in red and the rest of the protein in green. In the full protein model (B), a proline residue at position 258 is highlighted in yellow.

4.3 Investigating Tat dependence of the candidate twin-arginine signal peptides.

Based on the analysis above, the signal peptide and C-tail regions selected for each of these proteins are shown in Table 4.2. Synthetic DNA encoding these regions, codon optimised for expression in *E. coli*, was designed and synthesised as described in Chapter 2.

Table 4.2. Signal peptide and C-terminal regions for each protein candidate analysed in this chapter. A proline residue preceding the predicted C-terminal TMH is highlighted in yellow.

Candidate	Proposed Signal Peptide and C-tail.
WP_086565138.1 <i>Streptomyces africanus</i>	Signal peptide: MRHARRRVRRVTRLAAVGGLLGGAMVTNAVA C-tail: RLTDPRNVGPGLLVIAGSLVALVATRWRIRAEQDRKAYRQHYSATWG
WP_011931836.1 <i>Clavibacter michiganensis</i>	Signal peptide: MTTSSPSAPRRRILRSATALVGGVALAVVPLAASAHVRVSPDQAAAGSYSTLTFKVPTESATATT TSVT C-tail: TTAAPDPTTVTTAASDTSATSSAVAVGLGVGGLALGAVALVVAVFALTRVRREGGGQA
WP_019982084.1 unclassified <i>Streptomyces</i>	Signal peptide: MPATTGDRSRRRPLAFAAAVATAAAIGASLAAAPTAAAGG C-tail: GSDAPVAALGTAAALAVAAGAGVVFVAVRRRRGARDAQA
WP_049064233.1 <i>Corynebacterium striatum</i>	Signal peptide: MTSRRGTFLAALVTASLIPLAPPALA C-tail: SAGTPLAVLLGLFAAIAVAVGAIKPLHSFLLQVQRTLGL
WP_031122887.1 <i>Streptomyces sp. NRRL S-623</i>	Signal peptide: MGNAVSGRRTLLSGTAVLAAVALIALGSAPAQA C-tail: PQTGGDLAATGSDSTLPVAGAAGAALLAGGGLFYAMRRRMAARNG
WP_030568954.1 <i>Streptomyces cyaneofuscatus</i>	Signal peptide: MGIAASGRRTLLSATAVSATAALIALGAAPAQADAIAKPDLGVRALA C-tail: DSTLPLAGAAGAALLAGGSIVWAVRRRSAARAS
WP_046529179.1 <i>Cellulomonas sp. FA1</i>	Signal peptide: MSRRRTALLTAPAAAAAALTIVGAATSASA C-tail: DDATPAARAQVRASAQERSEVLAATGARTGVLLAAGALALGLGAGLVTWRRRRAAGA
WP_031517753.1 <i>Streptomyces sp. NRRL F-5123</i>	Signal peptide: MSLRRRAALLVTAASTLTALAAPAALA C-tail: IALGPAATKRATGAELRSDDKKDDGLSSSATTWIIVGVVVFVASAGFGLLLSGRKRRRP
WP_056088981.1 <i>Methylobacterium sp. Leaf99</i>	Signal peptide: MSGMAGRRSLVAAALGLAALLGGCSDARPPAPEVA C-tail: AVASLPWLPVAAIGLLLVRLWRLRRRRRRARIGRPLTEAQP

Following gene synthesis, the synthesised fragments were subjected to PCR amplification and ligation. The signal peptide coding regions were cloned into pSU40 UniAmiA, which encodes the mature sequence of AmiA but lacking the signal peptide. *E. coli* exports two Tat-dependent cell wall amidase enzymes, AmiA and AmiC to the periplasm. When the Tat system is inactivated *E. coli* is unable to grow in the presence of SDS (Fig. 4.10A) because it is unable to export AmiA and AmiC, which leads to incomplete cell separation during division through inability to fully cleave the cell wall septum (Ize et al. 2003b; Bernhardt and de Boer 2003). A *tat*⁺ strain deleted for *amiA* and *amiC* phenocopies the *tat* mutant strain with respect to inability to survive in the presence of SDS (Keller et al. 2012a; Huang and Palmer 2017). If full length AmiA is provided on a multicopy plasmid it is exported by the Tat pathway, restoring the ability to grow with SDS, but if it lacks a functional Tat signal peptide growth is not restored (Fig. 4.10B). Fusing the signal peptide of each candidate in frame with the mature sequence of AmiA therefore provides a facile screen to assess engagement with the Tat pathway.

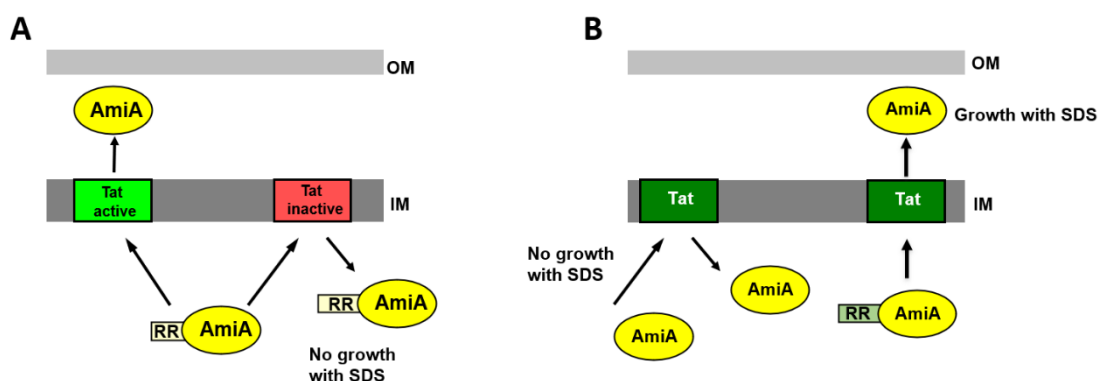


Figure 4.10. Schematic representation of the amidase reporter assay. A) AmiA is translocated to the periplasm in a Tat-dependent manner. In a *tat* mutant strain AmiA is not transported to the periplasm and the peptidoglycan septum is not cleaved during cell division resulting in outer membrane permeability and sensitivity to SDS. B) Fusion of a heterologous Tat signal peptide to the AmiA mature region will restore its translocation and will restore support growth in the presence of SDS. IM - inner membrane, OM - outer membrane.

In order to investigate whether the signal peptides from the candidate Tat dependent C-tail proteins identified above engaged with the Tat pathway, the coding regions were cloned in frame with *amiA*. At the time of writing this thesis I had successfully obtained clones for candidates 1, 2, 4, 8 and 9 (the S1 family peptidase

from *Streptomyces africanus*, the YcnI family protein from *Clavibacter michiganensis*, the LPXTG cell wall anchor domain-containing protein from *Streptomyces* sp. NRRL S-623, the terpene cyclase/mutase family protein from *Streptomyces* sp. NRRL F-5123, and the DUF4349 domain-containing protein from *Methylobacterium* sp. Leaf99).

Each of these constructs was introduced into *E. coli* strain MC4100 Δ amiA Δ amiC which is SDS sensitive because it lacks the Tat-dependent AmiA and AmiC amidases. Figs. 4.11 - 4.15 show the results of the SDS assays obtained with these constructs. Each experiment was performed in triplicate, with the third repeat using a lower amount of inoculum, with the exception of candidate 2, as due to time constraints, only one experiment could be performed.

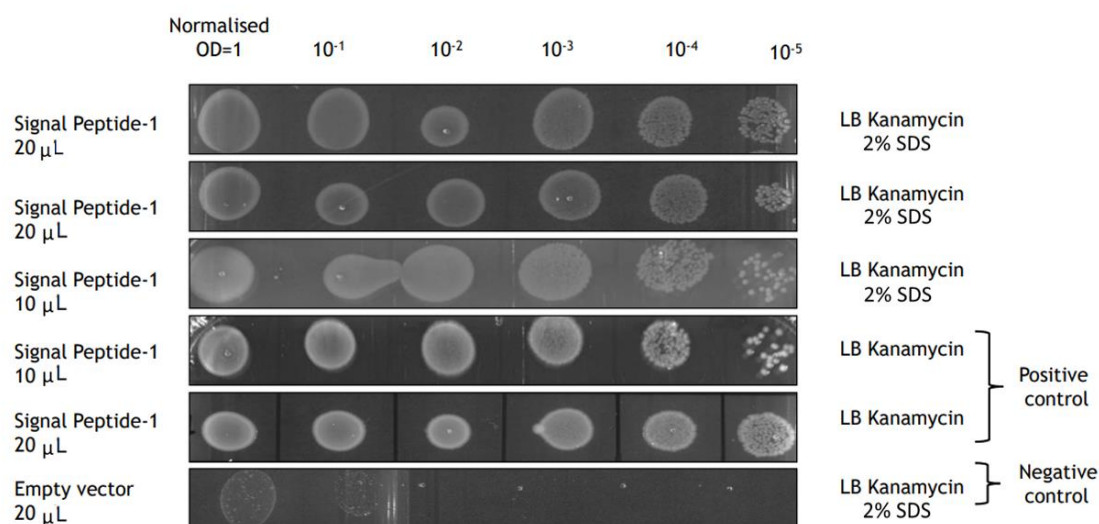


Figure 4.11. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_086565138.1 S1 family peptidase from *Streptomyces africanus* fused to the mature part of AmiA. Strains were grown overnight in liquid medium, serially diluted as indicated and a 10 or 20 μ l aliquot was spotted onto LB agar or LB agar containing 2% (w/v) SDS. Plates were incubated at 37°C for 16 h.

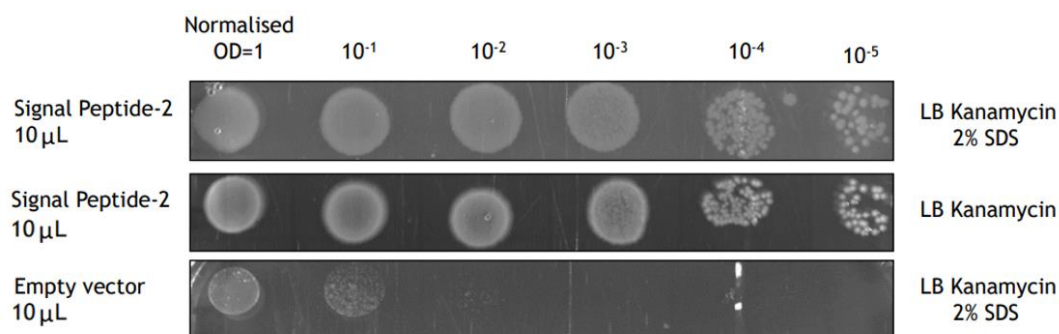


Figure 4.12. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_011931836.1 YcnI family protein from *Clavibacter michiganensis* fused to the mature part of AmiA. Strains were grown overnight in liquid medium, serially diluted as indicated and a 10 or 20 µl aliquot was spotted onto LB agar or LB agar containing 2% (w/v) SDS. Plates were incubated at 37°C for 16 h.

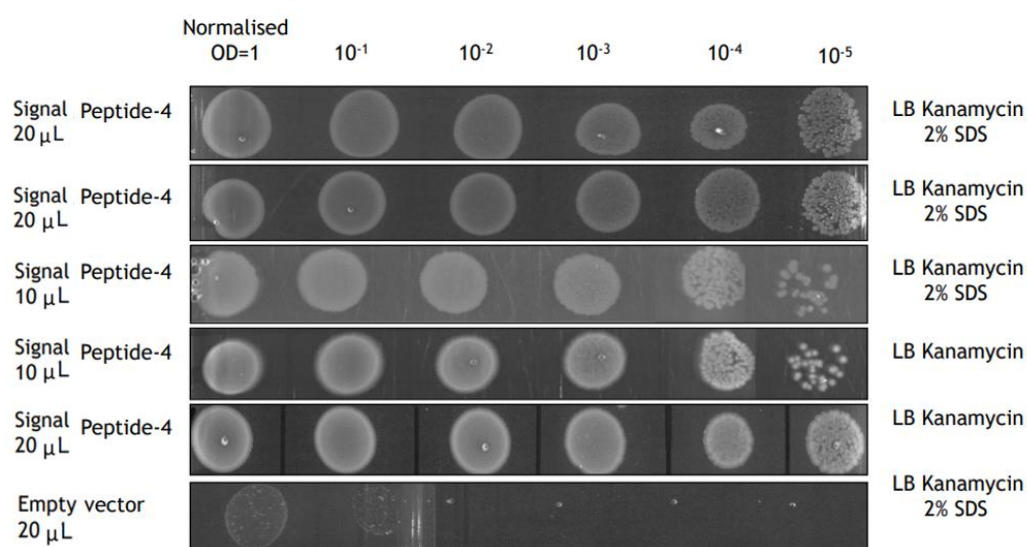


Figure 4.13. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_031122887.1 LPXTG cell wall anchor domain-containing protein from *Streptomyces sp. NRRL S-623* fused to the mature part of AmiA. Strains were grown overnight in liquid medium, serially diluted as indicated and a 10 or 20 µl aliquot was spotted onto LB agar or LB agar containing 2% (w/v) SDS. Plates were incubated at 37°C for 16 h.

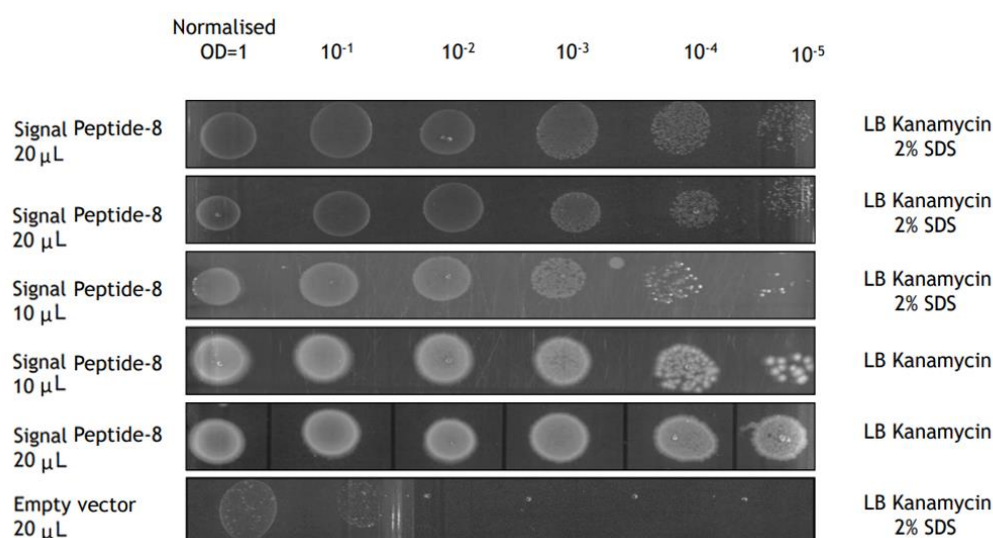


Figure 4.14. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_031517753.1 terpene cyclase/mutase family protein from *Streptomyces* sp. NRRL F-5123 fused to the mature part of AmiA. Strains were grown overnight in liquid medium, serially diluted as indicated and a 10 or 20 µl aliquot was spotted onto LB agar or LB agar containing 2% (w/v) SDS. Plates were incubated at 37°C for 16 h.

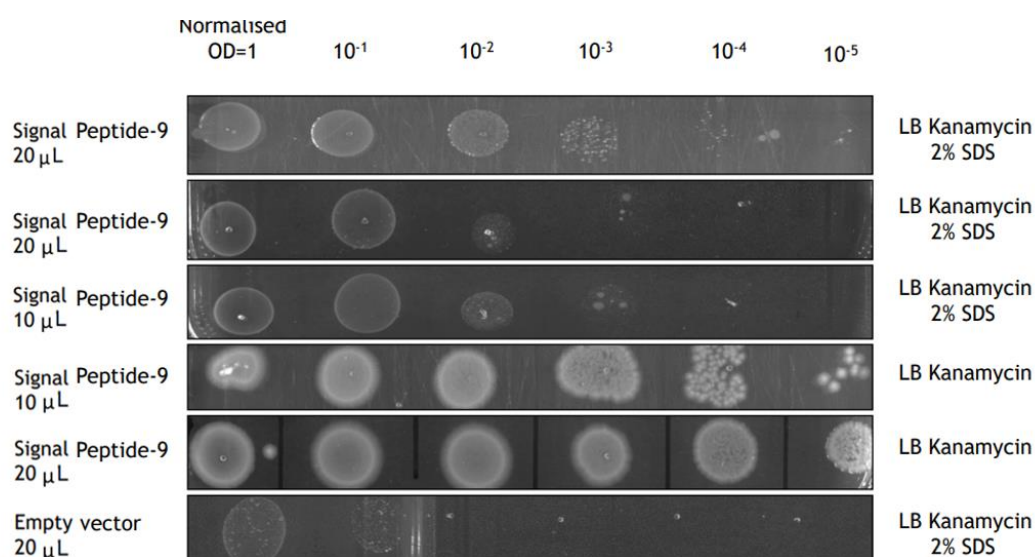
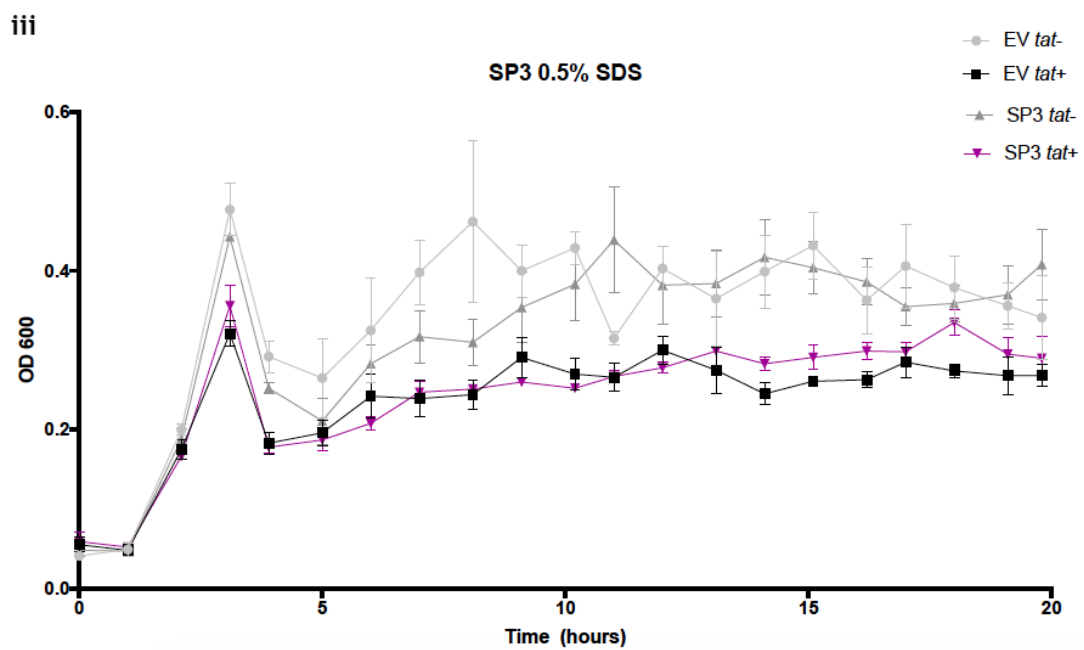
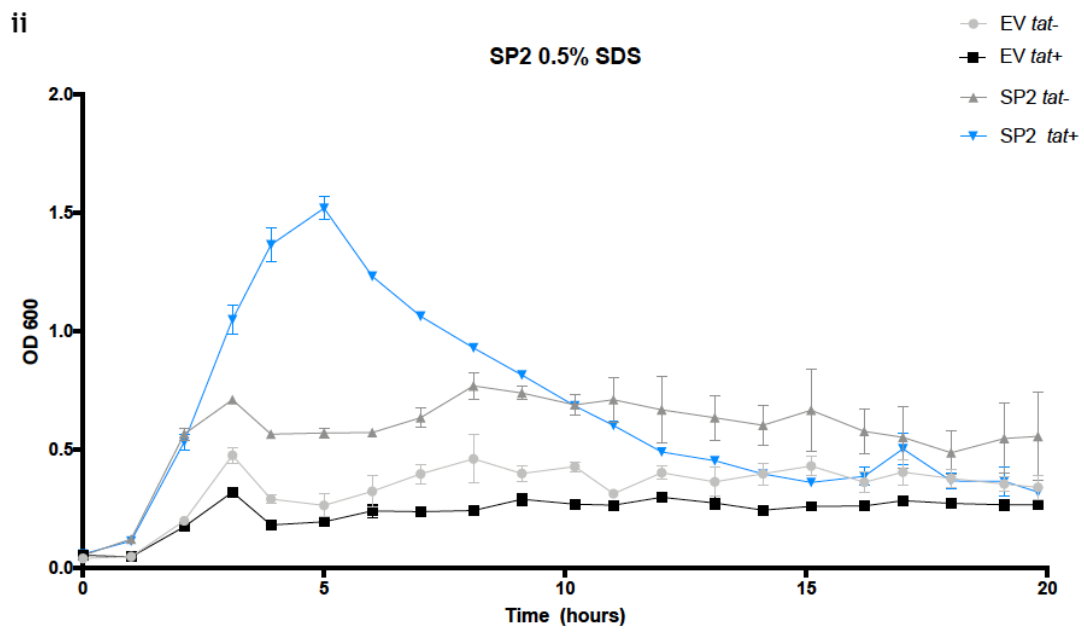
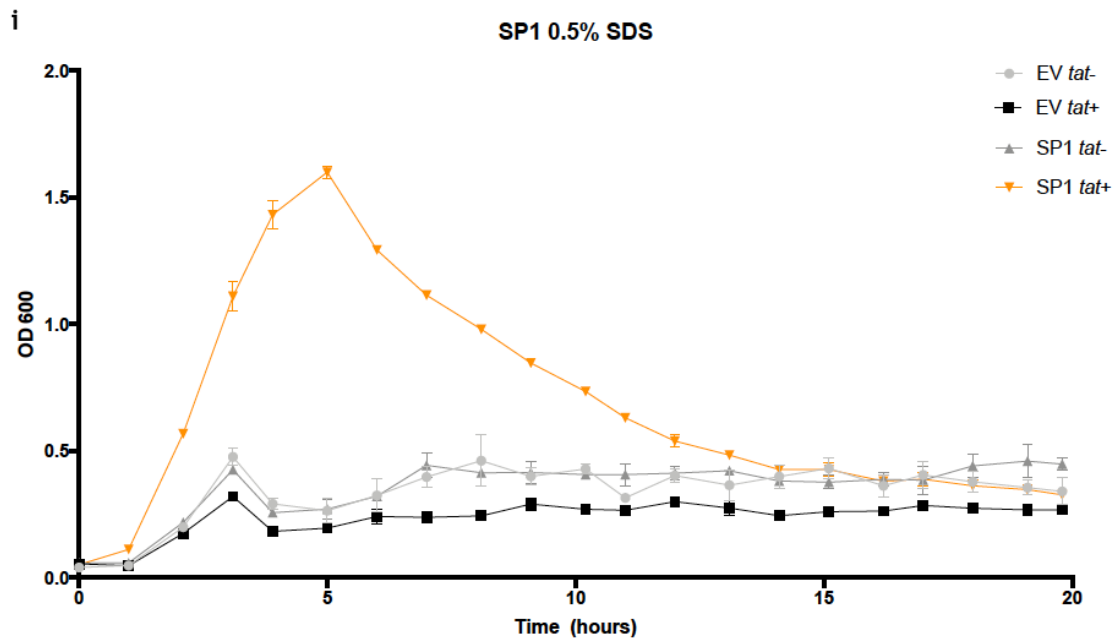


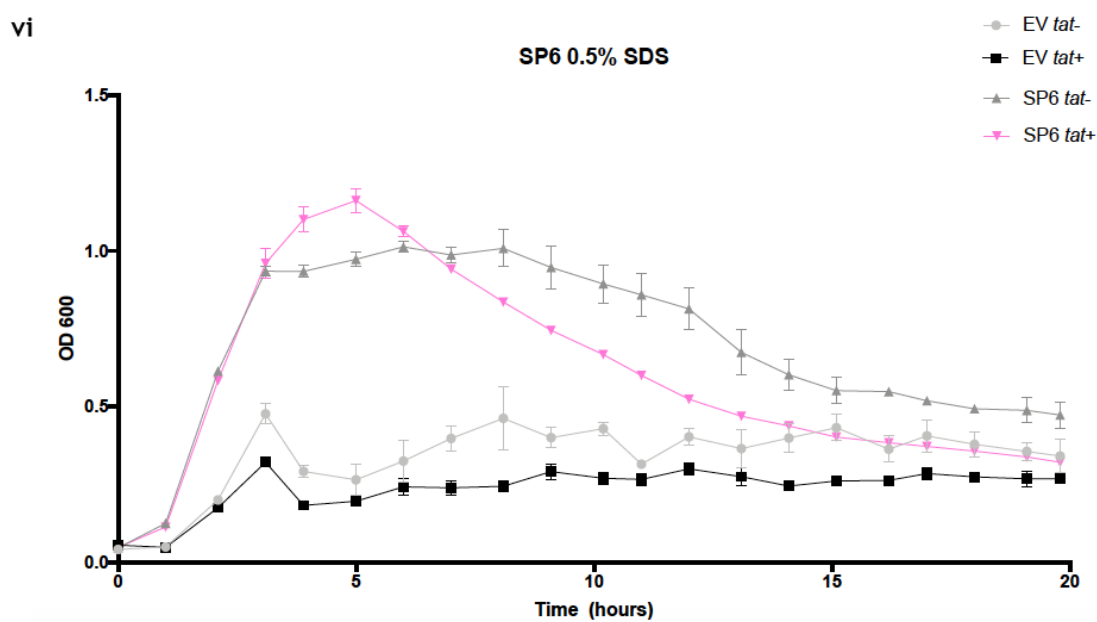
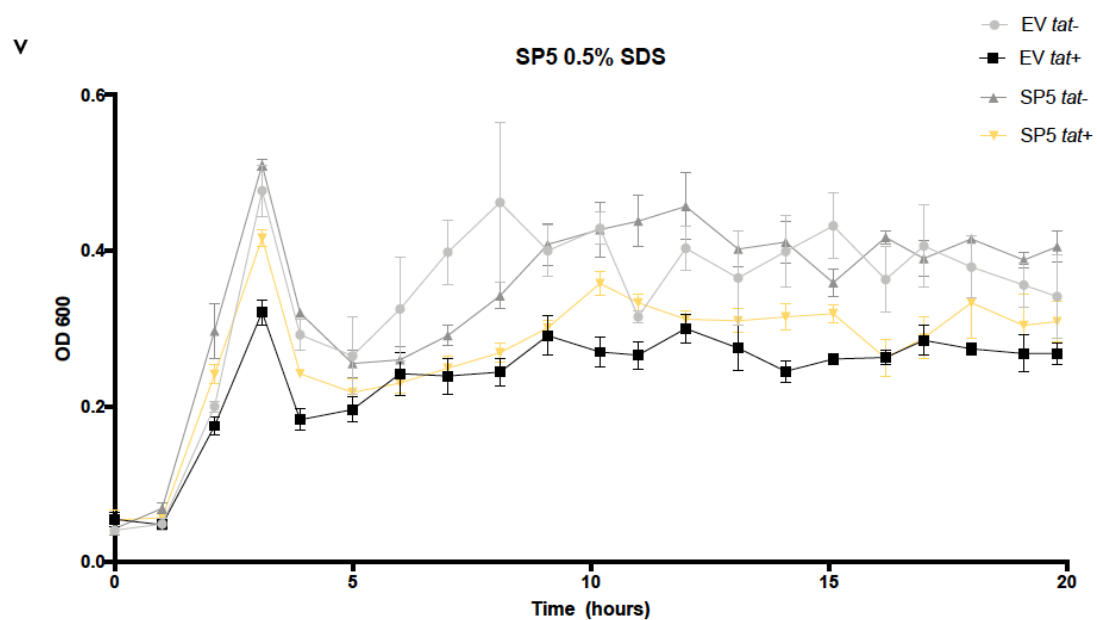
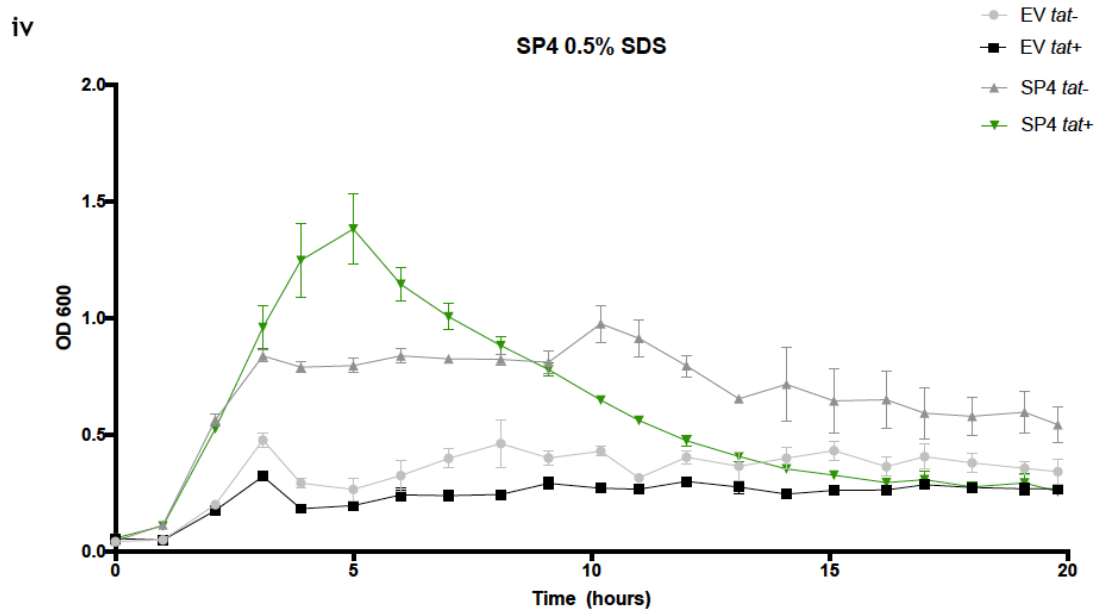
Figure 4.15. Spot dilutions of MC4100 Δ amiA Δ amiC harbouring an empty vector (pSUPROM), or pSUPROM encoding the predicted signal peptide from WP_056088981.1 DUF4349 domain-containing protein from *Methylobacterium* sp. Leaf99 fused to the mature part of AmiA. Strains were grown overnight in liquid medium, serially diluted as indicated and a 10 or 20 µl aliquot was spotted onto LB agar or LB agar containing 2% (w/v) SDS. Plates were incubated at 37°C for 16 h.

Analysis of the SDS growth tests indicate that almost all of the signal peptides are able to mediate export of AmiA as growth on SDS is seen that is stronger than with the empty vector. However, the signal peptide from Candidate 9 (the DUF4349 domain protein; Fig. 4.15) shows only weak export activity. The signal peptide from Candidate 8 (the terpene cyclase-mutase family protein; Fig. 4.14) also did not seem to function particularly well as the colonies on SDS were very small. It should be noted here that due to lack of time I was not able to carry out experiments in a strain deleted both for *amiA* / *amiC* and the *tat* genes. This would have allowed me to conclude that the export activity I am seeing is definitely due to engagement with the Tat pathway. This is important because it has been reported that AmiA can also be (very poorly) exported by Sec (Huang and Palmer 2017).

4.3.1 Liquid growth assays

While I was writing this thesis, Dr Emmanuele Severi from the Palmer group very kindly completed the cloning of the remaining signal peptide candidates, and undertook extensive growth analysis in the presence of SDS, using both the *amiA* / *amiC* mutant and its *tat*⁻ derivative. Dr Severi chose to analyse growth in liquid culture in the presence of 0.5% SDS, which allows a more detailed analysis of growth than simple spot dilutions (Kneuper et al. 2012; Huang and Palmer 2017). These results are presented in Fig. 4.16.





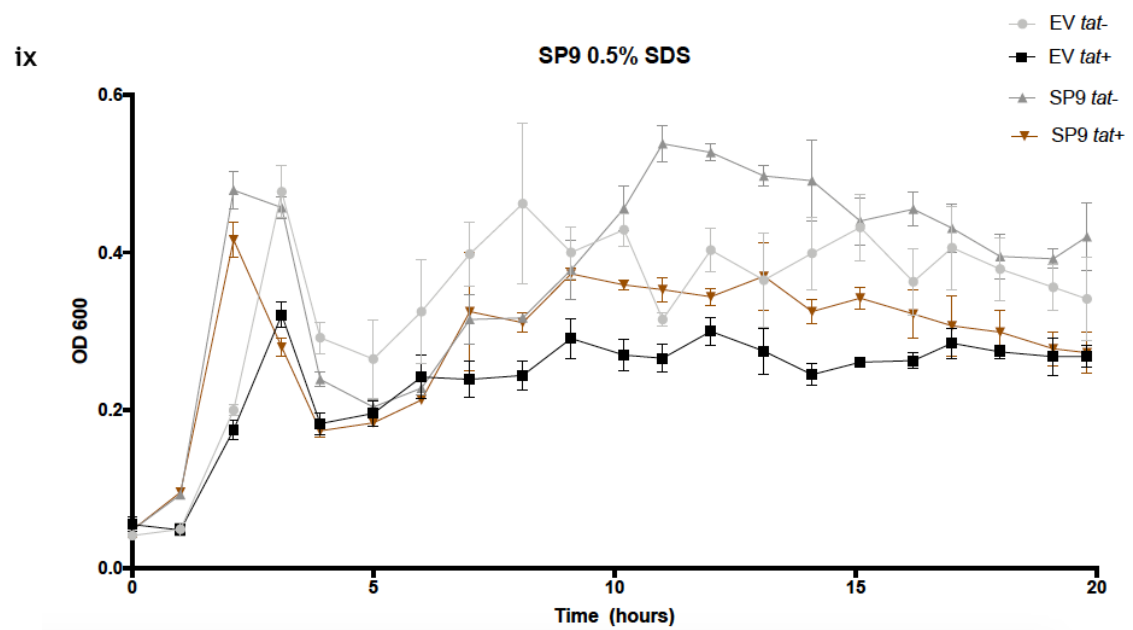
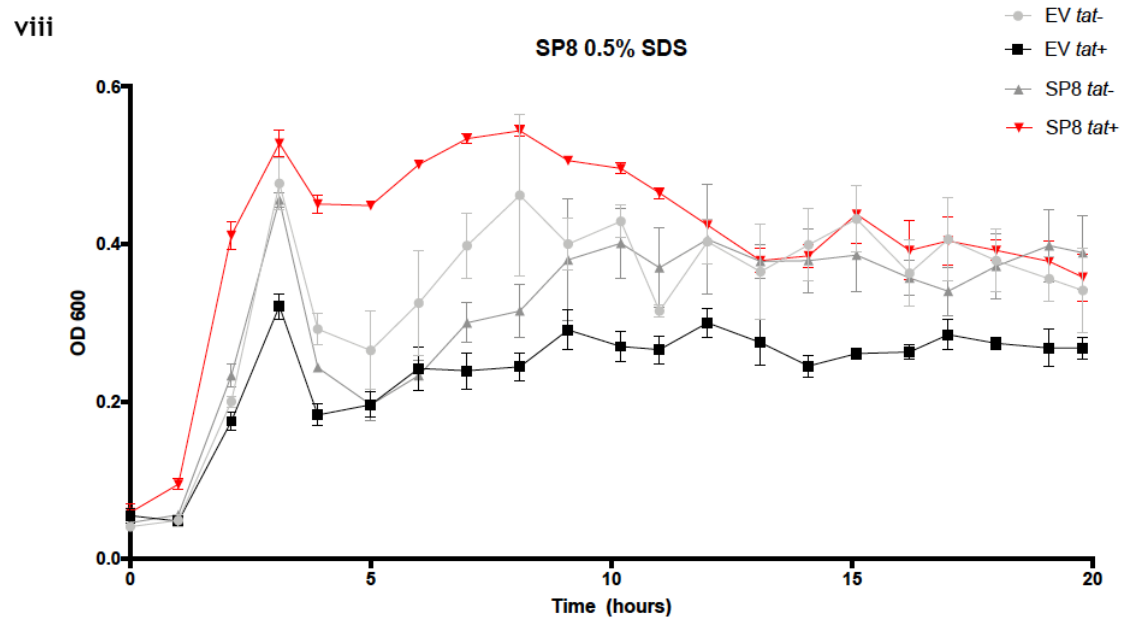
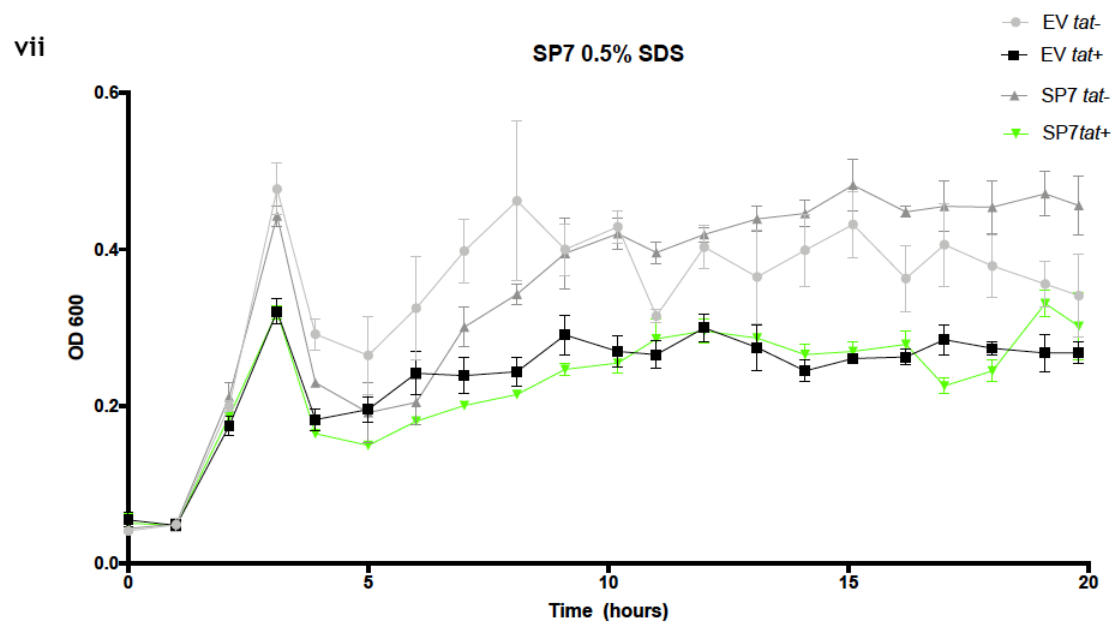


Figure 4.16. Growth of strain MC4100 Δ amiA Δ amiC (*tat*⁺) or MC4100 Δ amiA Δ amiC Δ tatABC (*tat*⁻) harbouring either empty vector (pSUPROM; EV), or pSUPROM encoding the predicted signal peptide from the indicated candidate in the presence of 0.5% SDS. Cultures were grown in a plate reader at 37°C without shaking. Error bars are \pm standard deviations (*n* = 3).

It is clear that signal peptides from Candidates 1, 2 and 4 are able to engage with the Tat pathway. The growth seen in the presence of SDS is significantly higher when the Tat pathway is active (Fig. 4.16 panels i, ii and iv), and control experiments in the absence of SDS (Appendix A) confirm that there is no general growth defect in the *tat* mutant strain. For Candidates 2 and 4 there is clearly also some interaction with the Sec pathway because the *tat* strain producing these signal peptide-AmiA fusions is also able to grow on SDS significantly better than when it carries the empty vector. It should be noted that it is relatively common for Tat signal peptides to show some level of engagement with the Sec pathway in heterologous reporter assays (Tullman-Ercek et al. 2007; Stanley et al. 2002; Cristobal et al. 1999). None-the-less it can be concluded that the S1 family peptidase WP_086565138.1 from *Streptomyces africanus*, the YcnI family protein WP_011931836.1 from *Clavibacter michiganensis* and the HtaA domain-containing protein WP_049064233.1 from *Corynebacterium striatum* are likely to be Tat substrates.

For Candidates 3, 5, 7 and 9 the signal peptide-AmiA fusions support very poor growth on SDS, and the level of growth seen for the *tat*⁺ and *tat*⁻ strain is indistinguishable (Fig. 4.16 panels D, F, H and J). It can be concluded that there is no evidence that any of these proteins (the HtaA domain-containing protein WP_019982084.1 from unclassified *Streptomyces*, the LPXTG cell wall anchor domain-containing protein WP_031122887.1 from *Streptomyces* sp. NRRL S-623, the LPXTG cell wall anchor domain-containing protein WP_046529179.1 from *Cellulomonas* sp. FA1 and the DUF4349 domain-containing protein WP_056088981.1 from *Methylobacterium* sp. Leaf99) are Tat substrates.

For Candidates 6 and 8, the signal peptides may potentially show some level of engagement with the Tat pathway because the growth seen in the *tat*⁺ strain is slightly, but significantly higher than when the Tat system is absent. Therefore, the LPXTG cell wall anchor domain-containing protein WP_030568954.1 from

Streptomyces cyaneofuscatus and the terpene cyclase-mutase family protein WP_031517753.1 from *Streptomyces* sp. NRRL F-5123 may be Tat substrates.

Table 4.3. Summary of the result for the experiments of candidate SP complment with Tat system.

Protein	Family	Specie	Signal Peptide	SP Tat complementation
WP_0865 65138.1	S1 family peptidase	<i>Streptomyces africanus</i>	MRHARRRVVRRVTRLAAVGGLLLG GAMVTNAVA	Yes
WP_0119 31836.1	YcnI family protein	<i>Clavibacter michiganensis</i>	MTTSSPSAPRRRILRSATALVGGVAL AVAVPLAASAHVRVSPDQAAAGSYS TLTFKVPTESATATTTSVT	Yes
WP_0199 82084.1	HtaA domain-containing protein	unclassified <i>Streptomyces</i>	MPATTGDRSRRRPLAFAAAVATAAA IGAASLAAAPTTAAAGG	No
WP_0490 64233.1	HtaA domain-containing protein	<i>Corynebacterium striatum</i>	MTSRRGTFLAALVTASLIPLAPPALA	Yes
WP_0311 22887.1	LPXTG cell wall anchor domain-containing protein	<i>Streptomyces sp. NRRL S-623</i>	MGNAVSGRRRTLLSGTAVLAAVALIAL GSAPAQA	No
WP_0305 68954.1	LPXTG cell wall anchor domain-containing protein	<i>Streptomyces cyaneofuscatus</i>	MGIAASGRRTLLSATAVSATAALIAL GAAPAQADAIKPDLGVRALA	Partial
WP_0465 29179.1	LPXTG cell wall anchor domain-containing protein	<i>Cellulomonas sp. FA1</i>	MSRRRTALLTAPAAAAAALTIVGA ATSASA	No
WP_0315 17753.1	terpene cyclase-mutase family protein	<i>Streptomyces sp. NRRL F-5123</i>	MSLRRAALLVTAASTLTALAAPAAL A	Partial
WP_0560 88981.1	DUF4349 domain-containing protein	<i>Methylobacterium sp. Leaf99</i>	MSGMAGRRSLVAAALGLAALLGGCS DARPPAPEVA	Partial

4.4 Investigating whether the candidate C-tails allow membrane-anchoring of a Tat substrate.

The C-tail encoding regions from the synthetic constructs outlined in Table 4.2 were cloned into pSUPROM *Sufl*. Cloning replaced the stop codon of *Sufl* with the tail region of each candidate as an in-frame fusion. This subsequently allows the assessment of whether a construct is integrated into the membrane following cell fractionation and western blotting using an anti-*Sufl* antibody (Hatzixanthis, Palmer, and Sargent 2003; Buchanan et al. 2002).

Due to time constraints, I was only able to clone the C-tail regions from Candidates 1 (WP_086565138.1 S1 family peptidase from *Streptomyces africanus*), 2 (WP_011931836.1 YcnI family protein from *Clavibacter michiganensis*), 6 (WP_030568954.1 LPXTG cell wall anchor domain-containing protein from *Streptomyces cyaneofuscatus*) and 8 (WP_031517753.1 terpene cyclase/mutase family protein from *Streptomyces* sp). During the writing of this thesis Dr Emmanuele Severi cloned the tail regions from the remaining candidates. He then prepared urea-washed membranes and the cellular soluble fraction from a strain lacking the chromosomal *sufl* gene producing these fusion proteins and undertook western blotting to assess for the presence of the *Sufl* fusions. The results are presented in Fig. 4.17.

The control experiments indicate that plasmid-encoded, native *Sufl* is found almost exclusively in the soluble fraction, consistent with it being a globular periplasmic protein [19]. The known Tat-dependent C-tail from FdnH serves to localise *Sufl* to the membrane fraction, as shown previously (Hatzixanthis, Palmer, and Sargent 2003), although a smaller form (probably formed by proteolysis) is detected in the soluble fraction. For the nine C-tails tested, each of them was capable of localising *Sufl* to the membrane fraction, although in most cases some smaller forms are also detected in the soluble fraction. It can be concluded that each of the hydrophobic stretches, at least when tested in isolation, is capable of being integrated into the membrane.

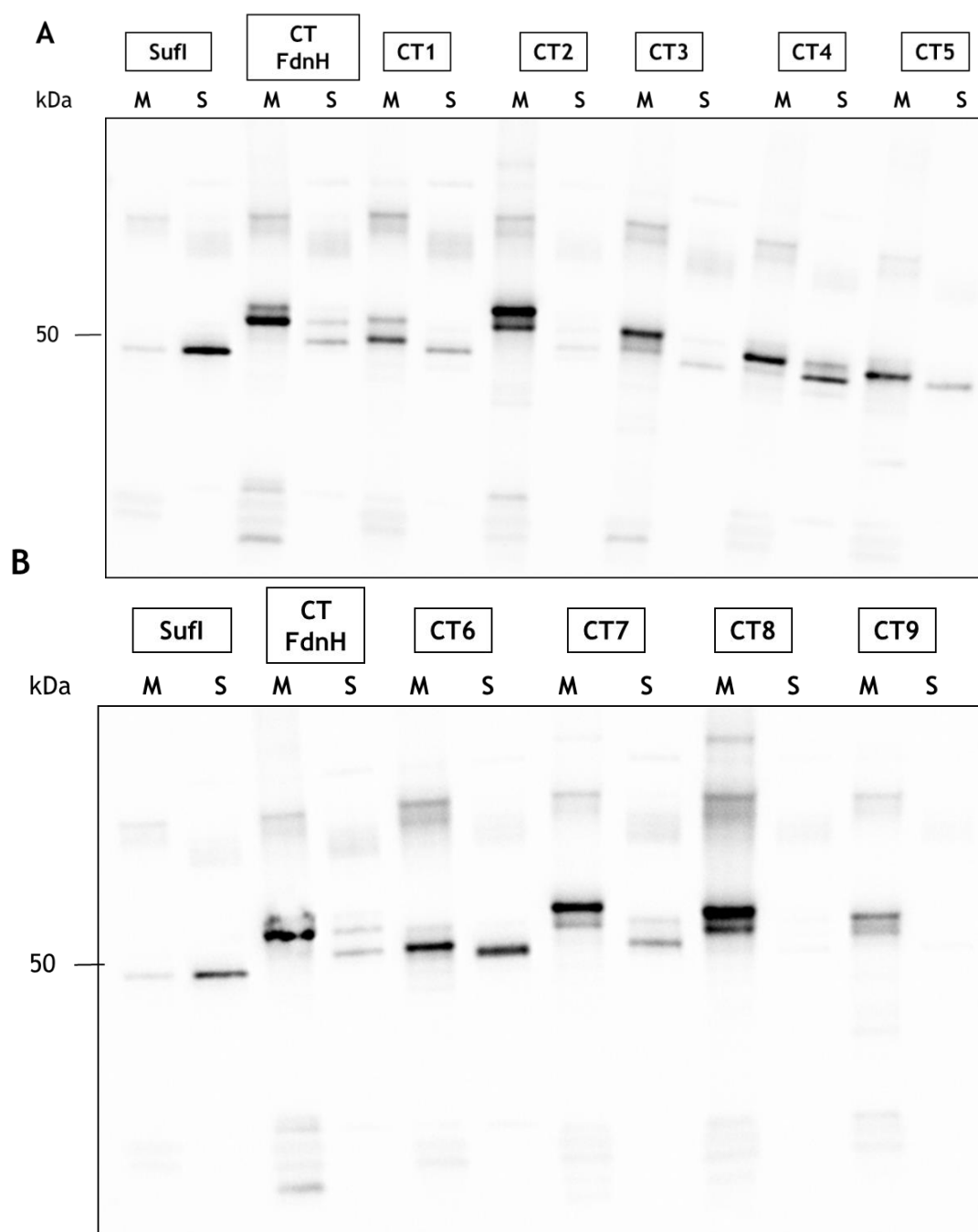


Figure 4.17. Western blot analysis to investigate membrane integration of SufI fused to candidate C-tails. Membranes and soluble fractions were prepared from strain NRS-3 (Δ sufI) producing either native SufI or SufI fused to each of the candidate C-tails as indicated. The membrane fraction from each sample was subsequently washed with 8M urea. Aliquots of the soluble fraction (S) and washed membranes (M) from each sample were analyzed by immunoblotting using anti-SufI antibodies. CT from candidates 1 to 5 are shown in panel A, and the rest in panel B.

4.5 Discussion

In this Chapter I have undertaken experiments in an attempt to validate bioinformatic findings that identified novel Tat-dependent tail-anchored protein candidates. Nine proteins were selected for further study, the majority of them from Gram-positive actinobacteria.

To assess Tat dependence, I designed codon-optimised DNA fragments encoding the predicted Tat signal peptide from each protein, for expression in *E. coli*. I used AmiA as a reporter for Tat dependence, and from my results and those of Dr Severi it can be concluded that at least three of the signal peptides tested are able to mediate export through the *E. coli* Tat pathway. Two of the other signal peptides also appear to show some level of engagement with the Tat system, but in addition mediate significant Tat-independent export. The four remaining signal peptides did not facilitate export of AmiA. However, based on these results alone it cannot be concluded that these are not Tat signal peptides, and it is possible that they are recognised by Tat system in the native organism. Since many of the signal peptides analysed here are from *Streptomyces*, testing their ability to mediate export of the *Streptomyces coelicolor* Tat-dependent agarase reporter would be an obvious next step to try (Joshi et al. 2010; Widdick et al. 2006; Widdick et al. 2008). Moreover, the signal peptide of candidate 9 has a cysteine residue close at the end of the signal peptide and Signal-P predicts it to be a lipoprotein. Although Tat-dependent lipoprotein signal peptides are compatible with the agarase assay (Thompson et al. 2010), membrane-anchoring of AmiA renders it inactive because it cannot bind to its peptidoglycan substrate (Keller et al. 2012). Testing of this signal peptide should be repeated following site-directed substitution of the cysteine. Finally, it should be noted that the mature domain of a substrate also contributes to the export route, for example if the protein folds quickly in the cytoplasm then it would be incompatible with Sec export even if the signal peptide can engage the Sec machinery (DeLisa, Tullman, and Georgiou 2003; Tottey et al. 2008).

The predicted C-tail of each candidate was tested for the ability to anchor SufI to the membrane. Every one of them resulted in membrane localisation of SufI

that was resistant to urea extraction. This would suggest that all nine candidates have a C-terminal transmembrane segment. However, this conclusion should also be taken with a note of caution. Testing a hydrophobic stretch of amino acids in isolation would be expected to result in membrane interaction, but if that region of the protein forms part of the folded mature domain then it would never be present as a transmembrane segment. For these examples tested here, the presence of a C-terminal TMH is also supported by protein structure prediction, so there can be some confidence that these probably do represent *bona fide* C-tails. Ideally, however, confirmation of membrane association should be carried out for the full length protein in the native organism.

In conclusion, three new Tat-dependent C-tail proteins have been validated experimentally. None of these proteins are predicted to bind redox cofactors and they represent the first such tail-anchored proteins that are not involved in electron transport.

Chapter 5. Developing genetic reporters to assess the assembly of Tat-dependent tail-anchored proteins

5.1 Introduction

As discussed in Chapter 1, tail-anchored proteins are an important class of Tat substrates. However, relatively little is known about how the Tat machinery mediates the integration of these hydrophobic stretches into the membrane. Exploring the membrane integration of native *E. coli* Tat-dependent tail anchored proteins is challenging because each of them contain metal cofactors, are multi-subunit and are mainly expressed only under anaerobic conditions (Palmer, Sargent, and Berks 2005). However, it has been shown previously that these C-tails can be fused to the C-termini of the soluble *E. coli* Tat precursors SufI and TorA rendering them membrane anchored and fully membrane integrated (Hatzixanthis, Palmer, and Sargent 2003). Importantly, integration into the membrane is completely dependent on the Tat system. This finding opens the door to the development of genetic reporters that can be used as a screen for Tat-dependent integration of hydrophobic stretches. This Chapter will describe the analysis of two such reporter fusions.

5.1.1 The use of fusion reporters in protein analysis

In bacteria, fusion proteins are used for a variety of purposes, including the study of protein function, protein-protein interactions, and protein localisation. Reporter fusion proteins can be also used to determine the topology of transmembrane segments (Keller et al. 2012). The protein of interest is fused, usually at the N- or C-terminus with a reporter protein, for example, the well-known green fluorescent protein (GFP), which can be easily visualised in cells or tissues. This allows researchers to track the expression and localisation of the protein of interest. Reporter fusions can also be used to study protein-protein interactions, protein stability, and protein activity. Additionally, multiple reporter fusions can be used simultaneously to study multiple proteins in a single experiment (Hu and Kerppola 2003; Ghim et al. 2010), making it a powerful approach for studying complex biological systems.

Fusion proteins have been extensively exploited as tools to study transport pathways and determine transmembrane protein behaviour (Stanley et al. 2002). The basis of this method is to choose a reporter protein with a specific subcellular compartmental activity, which is then fused to protein fragments containing prospective targeting sequences. A good fusion reporter has a distinct well-defined phenotype; assessing this generally involves plate screens, which specify a growth output or a colour change. Consequently, when this fusion undergoes testing, the targeted compartment is indicated by the activity of the reporter protein (Stanley et al. 2002). There are generally two types of reporter proteins - those that provide a qualitative analysis and those that allow quantitation. In this thesis maltose binding protein (MalE) was used as qualitative reporter and β -lactamase (Bla) was used to provide more quantitative data.

MalE, also known as maltose binding protein (MBP) is the periplasmic component of a membrane bound transporter ATPase binding cassette superfamily, which catalyses the uptake of maltose. It is usually exported to the periplasm via the Sec pathway, with a strong dependence on the chaperone, SecB (Fekkes and Driessen 1999). However, replacement of the native Sec signal peptide with a Tat targeting sequence re-routes MalE to the Tat pathway, indicating that the protein is compatible with Tat export. The Tat-targeted MalE reporter has been used extensively as a genetic reporter for Tat activity (e.g. Blaudeck et al. 2003; Kreutzenbeck et al. 2007; Lausberg et al. 2012; DeLisa et al. 2002; Tullman-Ercek et al. 2007; Keller et al. 2012). As MalE only carries out its biological role in the periplasm, it serves as a reporter for export across the cytoplasmic membrane. When MalE fusions are produced in an *E. coli* strain background lacking chromosomal *malE*, the ability of the strain to metabolise maltose is strictly dependent on export of the MalE fusion protein. Maltose utilisation can be assessed qualitatively by scoring for colony colour on MacConkey indicator medium containing maltose.

Bla enzymes, also known as beta-lactamases, play a crucial role in bacterial resistance mechanisms. They inactivate beta-lactam antibiotics, which are inhibitors of the penicillin binding proteins responsible for a key step in peptidoglycan synthesis. These enzymes cleave the amide bond in the four-membered beta-lactam ring of the incoming antibiotic, rendering it inactive and thus protecting bacterial

cells against lysis (Broome-Smith and Spratt 1986). Bla enzymes belong to several classes, with the TEM enzyme being widely used as a fusion reporter (Kaderabkova et al. 2022). Like all Bla enzymes, TEM-Bla (hereafter referred to as Bla) functions primarily in the periplasm and is typically transported by the Sec system. However, it can also be compatible with the Tat system if a twin-arginine signal peptide replaces the native signal (McCann, McDonough, Pavelka Jr, et al. 2007; Pradel et al. 2009). Bla activity is often assessed semi-quantitatively by examining the antimicrobial susceptibility of bacteria to ampicillin through the Minimum Inhibitory Concentration (MIC) assay, using commercially available MIC strips containing ampicillin in a continuous antibiotic concentration gradient (Tooke et al. 2017).

5.1.2 Modifications in fusion proteins: Linkers

Linkers are sequences of amino acids used to connect two or more proteins in a fusion protein. Linkers can be used to change the distance between the proteins that form a fusion, to increase or decrease their flexibility, to enhance or inhibit their interactions, or to add new functionalities. Linkers can also be used to add tags or epitopes to the protein of interest, which can be useful for purification or detection (Chen, Zaro, and Shen 2013). One of the key functions of a linker is to stabilise the fusion protein by allowing the domains to fold correctly. However, it is important to note that the linker sequence itself may also affect the stability or activity of the fusion protein, and it is therefore necessary to test different linkers and sequences to find the most appropriate one.

5.2 Results

5.2.1 Fusion proteins used in the study.

To assess the integration of Tat-dependent tail anchored proteins, two key fusion proteins, outlined in Fig. 5.1 were used. Both of these are based around the original *Sufl::FdnH_{CT}* construct used by Hatzixanthis, Palmer, and Sargent (2003) (Fig. 5.1(i)). The first of these harbours the full-length Tat substrate *Sufl*, fused to the transmembrane domain (C-tail) of *FdnH* which is in turn fused to the mature sequence of *MalE* (SFM; Fig. 5.1ii). The second is similar but the mature sequence of *Bla* replaces the *MalE* sequence (SFB; Fig. 5.1iii). The *MalE* construct was built in the

pUNIPROM (AmpR) plasmid and the Bla construct was built in pSUPROM (KanR). In each case expression of the fusion protein is under the control of the *E. coli* *tat* promoter present on the plasmids (Jack et al. 2004).

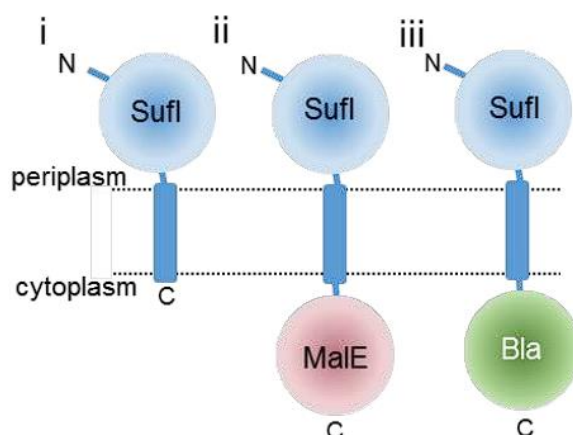


Figure 5.1. Fusion constructs to assess Tat-dependent C-tail insertion. i. C-tail fused to the C-terminus of Sufl anchors it to the membrane. ii. Fusions of maltose binding protein (MalE) or iii. B-lactamase (Bla) after the C-tail allows the development of genetic screens for C-tail integration. The MalE fusion is termed SFM and the Bla fusion SFB. The figure shows the expected topology of these fusions in a *tat*⁺ background.

5.2.2 Phenotypic analysis of the Sufl::FdnH::MalE (SFM) fusion protein

To investigate whether the SFM fusion was integrated into the membrane or exported fully to the periplasm, I assessed for colony colour on MacConkey maltose indicator plates. This medium contains a neutral red pH indicator, and so distinguishes those bacteria that can ferment the sugar (thus acidifying the medium and giving red colonies) from those that cannot. If the Tat system is able to integrate the C-tail of the SFM construct then we expect to see yellow colonies on indicator plates, whereas if C-tail integration is defective and the protein is fully exported to the periplasm we should observe red colonies (Fig. 5.2).

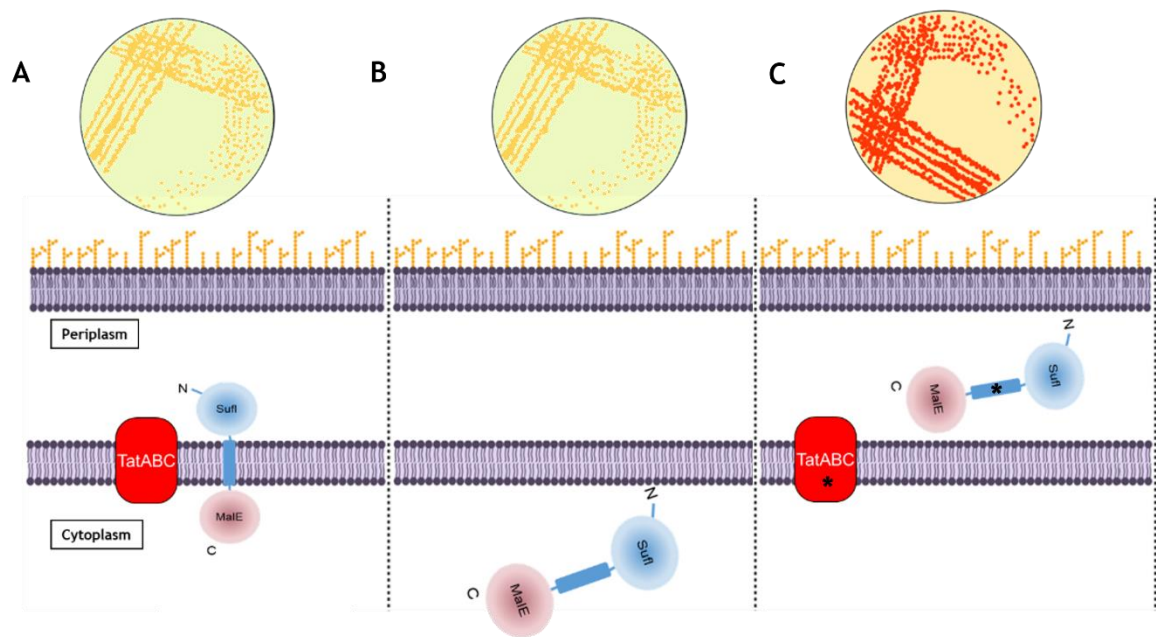


Figure 5.2. MalE as a reporter for C-tail integration. Left and middle panels - schematic illustration of expected subcellular location and topology of the SFM fusion protein in *tat*⁺ and *tat*⁻ backgrounds. In each case the reporter would be expected to result in the production of yellow colonies on MacConkey maltose medium when it is integrated in the membrane (A) or when the Tat system is not present (B). C. The ultimate goal is to use the system to isolate substitutions in the Tat system or in the C-tail (indicated by asterisks) that allow complete export of the fusion protein to the periplasm (resulting in the production of red colonies).

For this assay an *E. coli* strain, ICB5, carrying a chromosomal deletion in the *malE* gene and lacking the *tatABCDE* genes was used, and where required was complemented with a very low copy number plasmid harbouring *tatABC* (pTAT101 (Kneuper et al. 2012)) to generate the *tat*⁺ derivative. The SFM construct was introduced into the *tat*⁺ and *tat*⁻ background and the strains were streaked onto MacConkey maltose medium and colony appearance over time was monitored (Fig. 5.3). As shown in the figure, colonies of both the *tat*⁺ and *tat*⁻ strain harbouring SFM started to turn red after 24h and became increasingly red over time. These results were entirely unexpected; the strain with a functional Tat system should be expected to integrate the fusion protein with the MalE portion facing the cytoplasm and therefore colonies should be yellow. The strain lacking the Tat system should not be able to recognise the Tat signal peptide and therefore the entire protein would be expected to be cytoplasmic, thus also giving yellow colonies (Fig 5.2).

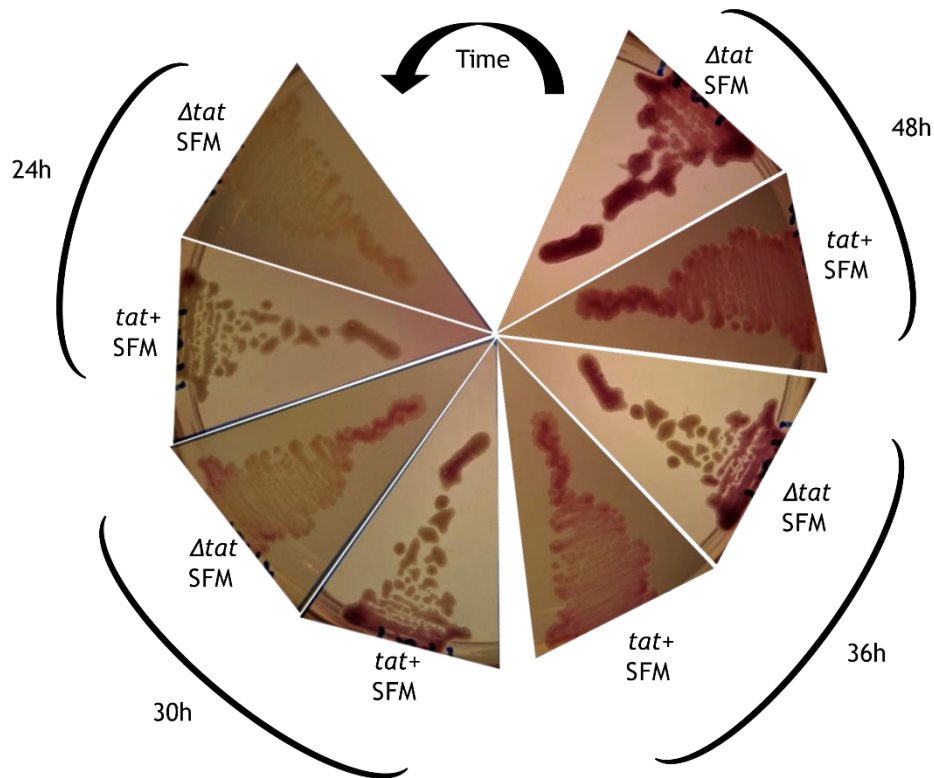


Figure 5.3. The *tat*⁺ and *tat*⁻ strains producing the SFM fusion show red colouration on MacConkey maltose indicator plates. Strain ICB5 harbouring either pThr19 (empty vector) or pTat101 (carrying *tatABC*) alongside pUNIPROM-SFM, were streaked onto MacConkey maltose plates and incubated for the indicated time periods before photographing.

5.2.3 Membrane localisation of the SFM fusion protein.

To further assess the behaviour of the SFM fusion protein, I prepared membrane fractions from the same *tat*⁺ and *tat*⁻ strains and blotted them with an anti-MalE antibody. As shown in Fig. 5.4 (left panel), the full length SFM fusion was detected in the membrane of both strains. Moreover, the protein appeared to be fully integrated into the membrane even in the absence of the Tat machinery because it could not be extracted by urea washing (Fig. 5.4, right panel). This behaviour was completely unexpected and may suggest that some of the fusion protein is engaging with the Sec pathway. It was also noted that several smaller cross-reacting bands were also visible in the membrane fractions from both strains (Fig. 5.4). These could be proteolytic fragments of the full length fusion, or potentially internal translation products. If they represent the latter, the Sec pathway may recognise the FdnH TMH as a signal sequence and integrate the protein with the MalE part facing the

periplasm, providing a potential explanation for the red colouration seen on MacConkey maltose plates.

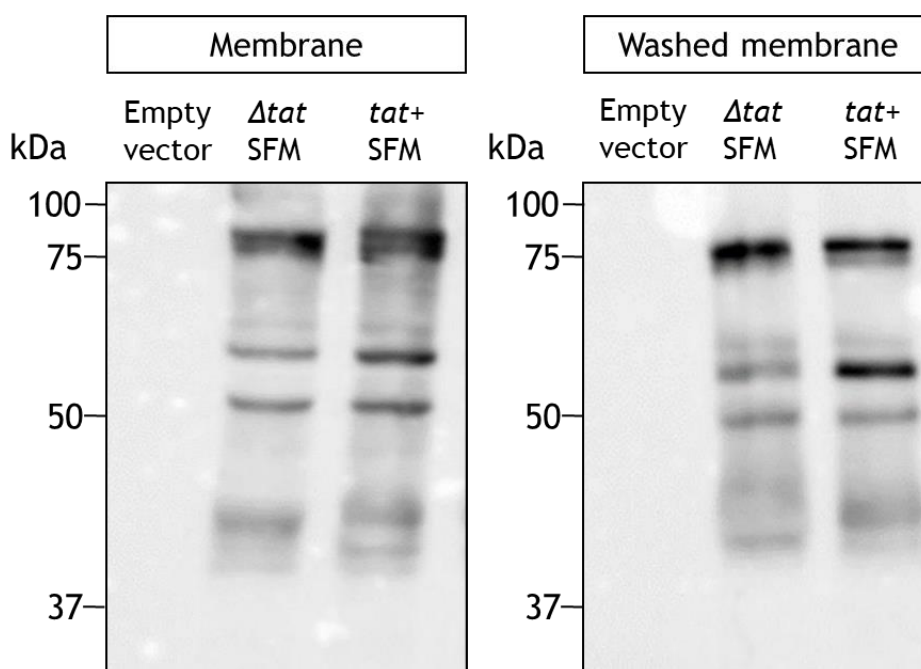


Figure 5.4. Western blot analysis of the SFM fusion protein in membrane fractions of the tat^+ and tat^- strain. Crude membranes were prepared from strain ICB5 harbouring either pThr19 or pTat101 (carrying *tatABC*) alongside pUNIPROM ('Empty vector') or pUNIPROM-SFM (left panel) and these were washed with 4M urea to remove peripherally-bound proteins (right panel). The blots were probed using an α -MalE antibody for detection.

From these experiments it became clear that SFM was not going to serve as a reliable reporter for C-tail integration and therefore it was not studied any further.

5.3 Use of the Suf::FdnH::Bla (SFB) fusion protein to assess Tat-dependence

I next switched my attention to a second fusion protein, SFB (Fig. 5.1) as a reporter for C-tail integration. As discussed previously, the use of Minimum Inhibitory Concentration (MIC) strips containing ampicillin allows a semi-quantitative assessment of Bla export and this approach has been used previously to assess the integration of internal signal sequences by the Tat pathway (Tooke et al. 2017).

In these experiments, MC4100 and DADE (As MC4100, $\Delta tatABCD$, $\Delta tatE$) were used as tat^+ and tat^- strains, respectively. Initially, MIC tests were conducted on each

strain alone, and carrying the empty pSUPROM vector, to determine the basal level of resistance. As shown in Fig. 5.5, the basal resistance for the *tat*⁺ strain alone was 4 µg/mL, and harbouring the pSUPROM empty vector was 4.3 µg/mL. For the *tat* mutant strain DADE, the basal resistance was 0.75 µg/mL, and harbouring the pSUPROM empty vector was also 0.75 µg/mL. The increased sensitivity of *E. coli tat* mutant strains to ampicillin has been reported previously and is likely due to the cell wall defect arising from the inability to export the cell wall amidases AmiA and AmiC (Ize et al. 2004; Stanley et al. 2001).

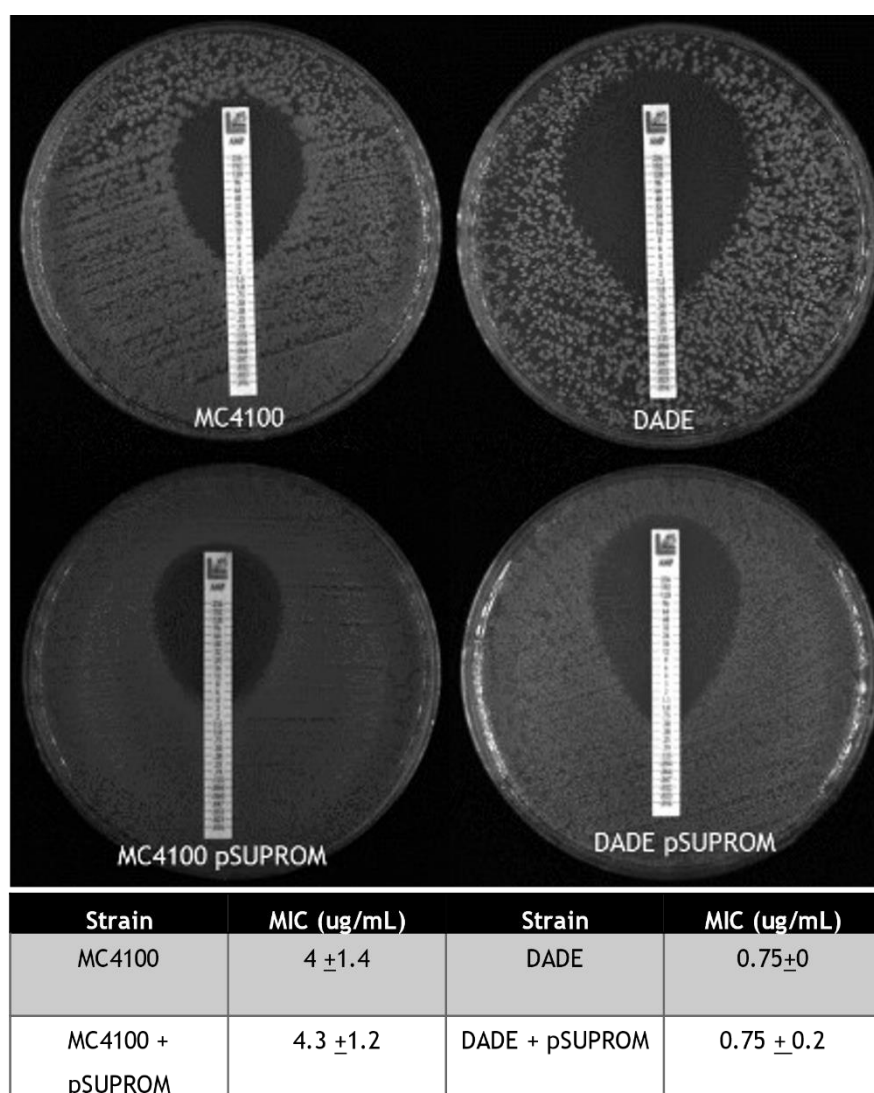


Figure 5.5. MIC assays for strains MC4100 and DADE and the same strains harbouring pSUPROM to determine basal levels of resistance to ampicillin. The Table indicates the average MICs for each strain, ± one standard deviation, n = 3.

Next, the level of resistance conferred by the SFB fusion was assessed in the two strain backgrounds. Unexpectedly, cells producing this fusion protein were fully resistant to ampicillin at the maximum concentration (256 $\mu\text{g/mL}$) regardless of whether the Tat system was present (Fig. 5.6). The very high level of resistance mediated by this fusion was difficult to explain as Tat dependent integration of the fusion protein should lead to ampicillin sensitivity similar to that seen with the empty vector. The *tat* mutant strain should also be expected to show ampicillin sensitivity because SufI, which is present at the N-terminus of the fusion, is not exported by the Sec pathway (Stanley, Palmer, and Berks 2000). None-the-less, these results are consistent with at least the Bla portion of the fusion being exposed at the periplasmic side of the membrane in a Tat-independent manner.

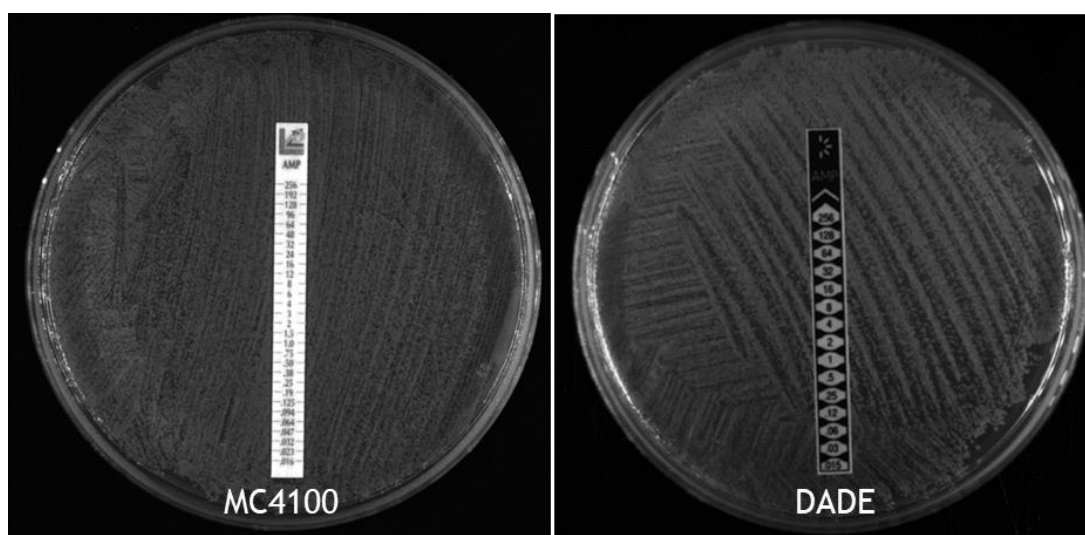


Figure 5.6. MIC assays for strains MC4100 and DADE harbouring pSUPROM SFB showing full resistance to ampicillin.

The FdnH C-tail region of the fusion protein includes the final 59 amino acids of FdnH. This covers the TMH region but also the highly negatively charged C-terminus (Fig. 5.7). The expected topology of the SFB fusion (Fig 5.1) would place this string of negative charges at the cytoplasmic side of the membrane. The distribution of positively charged amino acids in transmembrane proteins can have profound effects on their topology, with such residues being enriched in cytoplasmic regions, the so-called ‘positive inside’ rule of von Heijne (1986). I therefore decided to delete this negatively charged region of the fusion protein and replace it with a strong positive motif consisting of three Lys residues. This new construct, which was designated SFB-

SECA was introduced into the MC4100 and DADE strains and MIC assays were conducted. However, this new construct also resulted in full resistance to ampicillin in a Tat-independent manner (not shown).

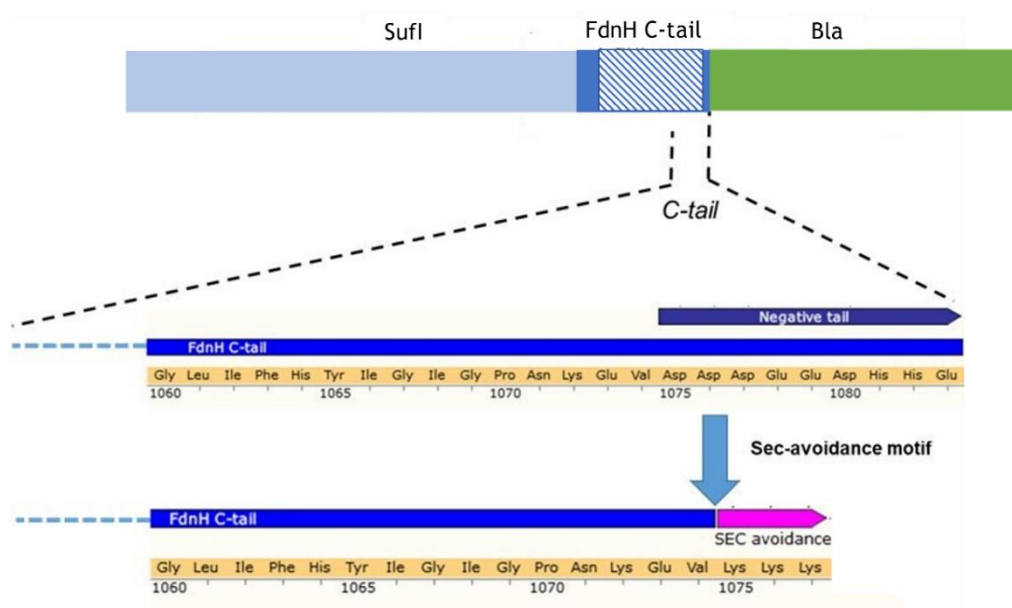


Figure 5.7. The sequence of the last 24 amino acids of the FdnH region in the SFB fusion (top) and the modified sequence in SFB-SECA.

5.3.1 Structural prediction suggests that the SFB fusion may not fold as expected.

Given the unexpected results with the SFB fusion protein, I next used structural prediction programmes to investigate the overall structure of the fusion, with the ultimate aim of introducing modifications to the construct that might confer the expected behaviour.

Initially, I compared the published crystal structure of the C-tail region of FdnH (Jormakka et al. 2002) to the C-tail FdnH models generated by RoseTTAFold and AlphaFold2 tools using the isolated FdnH C-tail amino acid sequence (Fig. 5.8). By doing this I could assess the fidelity of the programmes in simulating the protein structure. In both cases, simulation of the isolated C-tail suggested the formation of an α -helix (Fig. 5.8 B and C). The next step was to simulate the fusion protein previously published by Hatzixanthi, Palmer, and Sargent (2003) as the SFB fusion protein was designed based on their work. In the simulations performed using both

tools, the FdnH C-tail (depicted in blue in Fig. 5.9 A and B) is predicted to maintain its α -helical structure, in agreement with the published findings where the fusion protein was stably integrated into the membrane (Hatzixanthis, Palmer, and Sargent 2003).

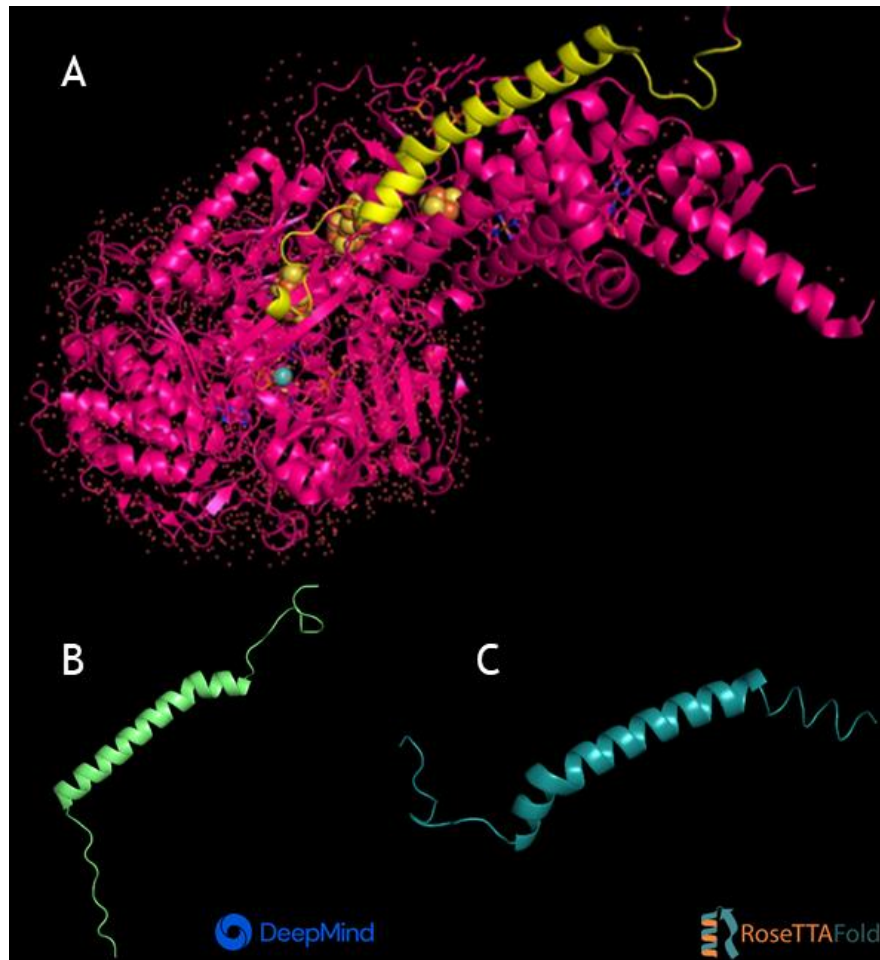


Figure 5.8. A. Crystal structure of formate dehydrogenase with the C-tail in yellow. B. Simulation of the FdnH C-tail using AlphaFold. C. Simulation of the FdnH C-tail using RoseTTAFold.

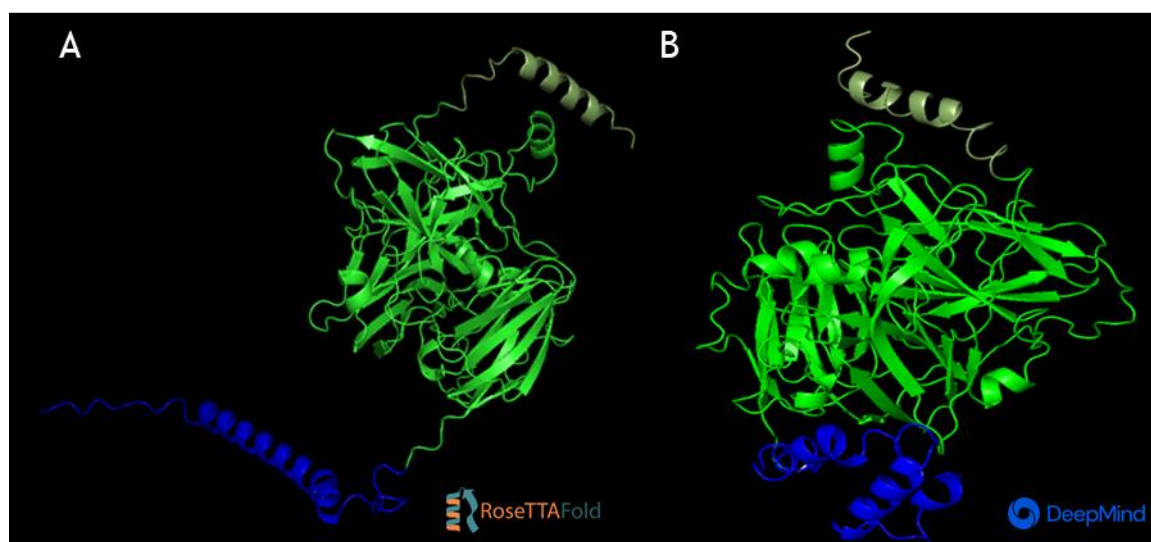


Figure 5.9. Simulation of the SufI::FdnH fusion protein. A. RoseTTAFold simulation; B. AlphaFold simulation. In each model, the SufI part is shown in green and the FdnH C-tail in blue.

Finally, I proceeded to simulate the complete SFB fusion protein. However, in this case, none of the simulations generated the expected folding for the FdnH C-tail, as it was predicted to lack the helical structure and was modelled as a long, mainly disordered, loop (Fig. 5.10). This is consistent with the unexpected behaviour of the SFB fusion protein and may suggest that the aberrant behaviour I observed could arise in part because the protein does not fold as expected and the C-tail is not presented correctly for membrane integration.

It was noted in Chapter 4 that the hydrophobic C-tail regions of candidate Tat-dependent tail-anchored proteins are preceded by a proline residue. The C-tail region of FdnH in the SFB construct used in the experiments described above starts immediately after the proline, and thus this conserved residue is missing. To see whether inclusion of this may improve the folding of the hydrophobic α -helix structure, I repeated the simulation but included this residue. However, extending the FdnH region by this single residue did not result in any substantial difference in structure prediction (Fig 5.10, right panel).

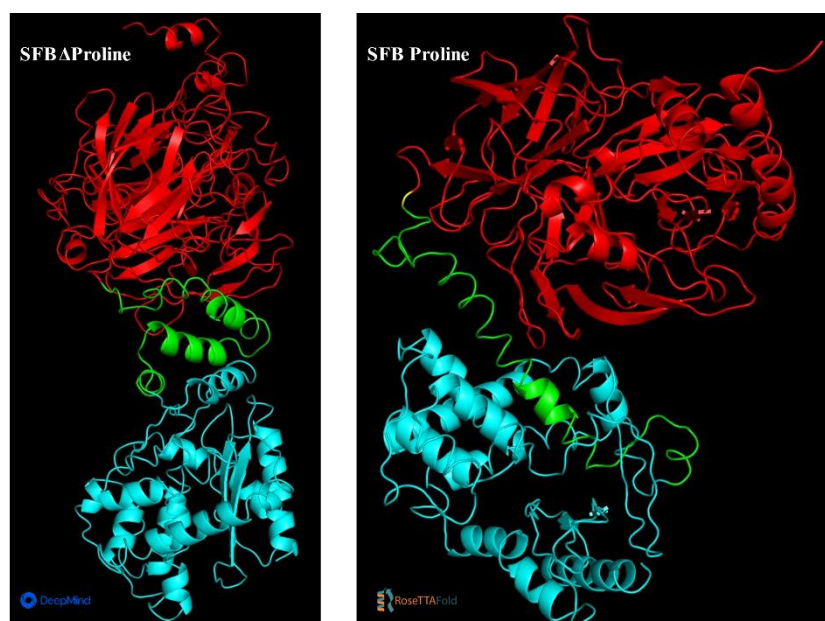


Figure 5.10 RoseTTAFold simulation of the SFB fusion protein (left panel), or the same protein but with inclusion of the proline residue that usually precedes the FdnH C-tail. The Sufl part is shown in red, the FdnH region in green, the proline in yellow and the Bla portion in cyan.

5.3.2 Design of linkers predicted to assist correct folding of the FdnH C-tail in the SFB fusion protein.

As incorrect folding of the FdnH C-tail may be occurring in the SFB fusion, I next undertook rational design of different linker sequences to flank the C-tail region in SFB to facilitate its correct folding. I designed 3 different linkers: The first of these was a rigid linker (EAAAK)₉. This linker is known to form a stiff α -helical structure and has been used extensively to generate a rigid spacer between protein domains (Amet, Lee, and Shen 2009; Chen, Zaro, and Shen 2013). I also designed two flexible linkers (GGGTA[TP]₁₀) (Sun et al. 2021) and (L[GGGGS]₅AAA. Linkers based on the GGGGS sequence in particular have also been heavily used (Chen, Zaro, and Shen 2013). For simulation, two identical copies of each linker were used, one flanking the FdnH sequence on either side. The output is shown in Fig. 5.11.

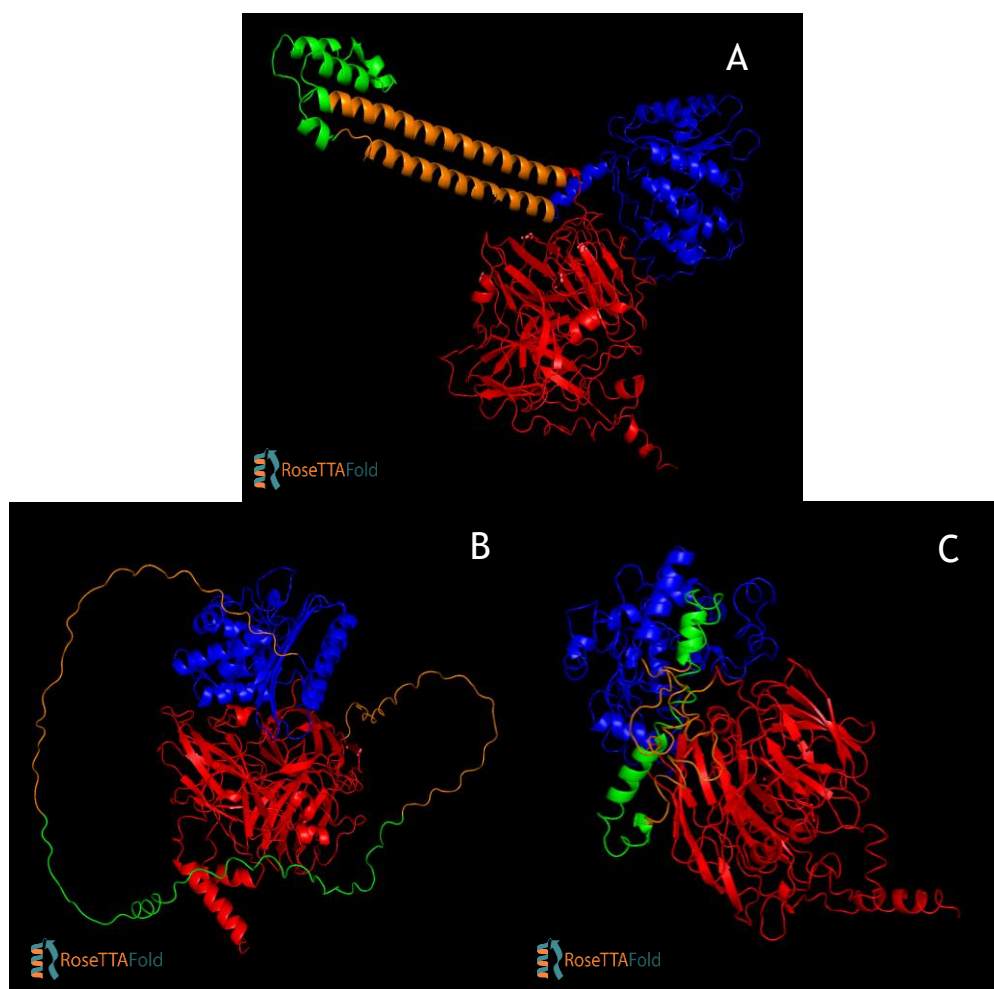


Figure 5.11. RoseTTAFold Simulation of SFB folding after inclusion of different linker sequences flanking the FdnH region. A. Rigid linker B. Flexible linker GGGTA[TP]₁₀ C. Flexible linker L[GGGGS]₅AAA. Linkers are in orange and the FdnH region in green.

From Fig. 5.11, it can be seen that while flanking the FdnH C-tail with the rigid linker or the GGGTA[TP]₁₀ linker was not predicted to result in helical folding of the FdnH sequence, including the L[GGGGS]₅AAA linker did appear to allow the FdnH region to adopt an α -helical structure (Fig 5.11C). Therefore, I continued with this linker, and next I simulated different lengths of this linker to see whether this might further improve the predicted folding of the FdnH C-tail. I tested various numbers of repetitions of the sequence [GGGGS]; [GGGGS]₁, [GGGGS]₂, and [GGGGS]₃ and found that two repetitions appeared to be more effective than one or three in promoting / maintaining the helical secondary structure of FdnH (Fig. 5.12).

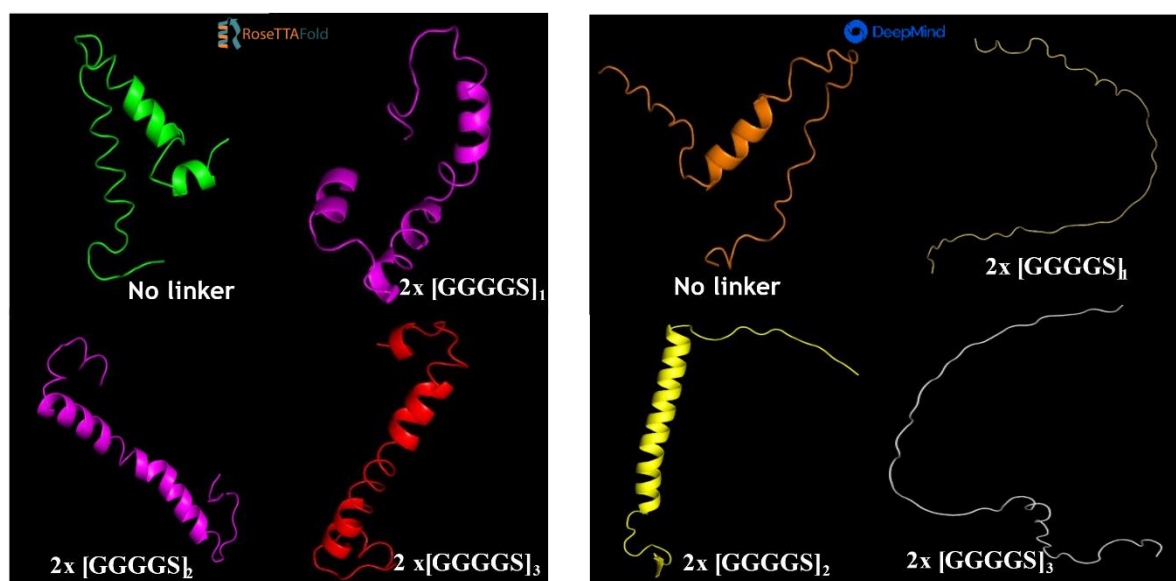


Figure 5.12. Secondary structure prediction of the FdnH region within the SFB fusion protein when flanked by none, one, two, or three iterations of the [GGGGS] linker. Simulations were conducted using both RoseTTAFold (left) and AlphaFold (right) for the entire protein, but only the FdnH region is displayed for clarity.

After determining that two repetitions of the [GGGGS] linker were more effective in promoting/maintaining the helical secondary structure of FdnH (Figure 5.12), my next goal was to ascertain whether the [GGGGS]₂ flanking the FdnH C-tail on both sides (Fig 5.13C and D; 2 x [GGGGS]₂) was predicted to allow better folding than having the linker solely between SufI and FdnH (Fig. 5.13A and B; 1 x [GGGGS]₂). From the simulations it appears that having flanking linkers yielded better results in terms of improving FdnH folding in the output models.

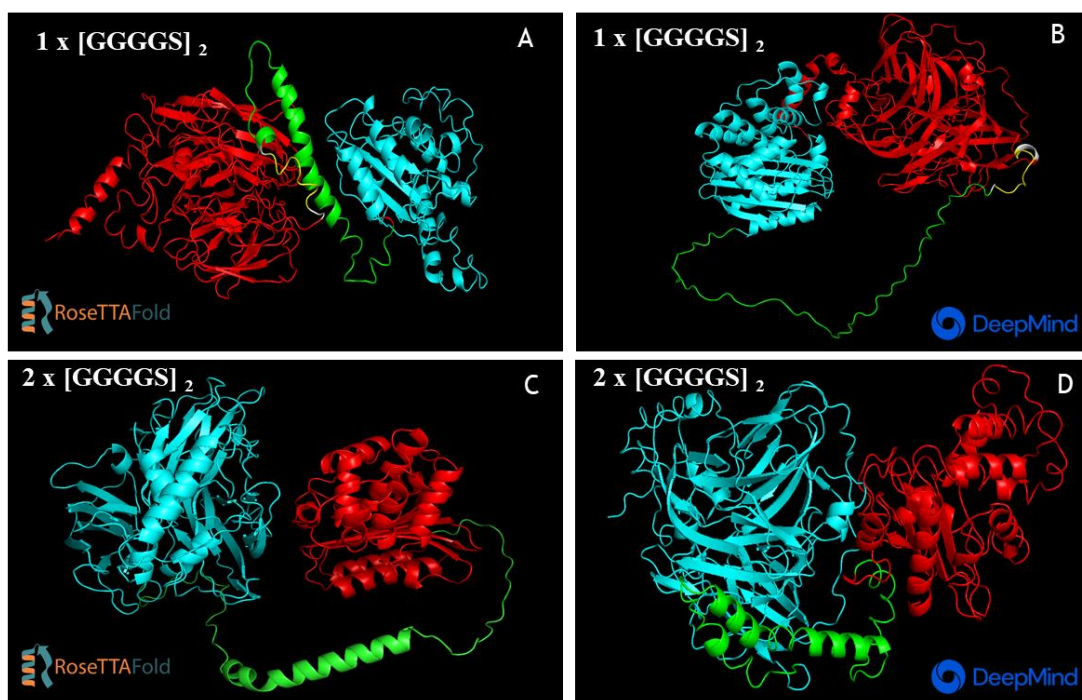


Figure 5.13. A. and B. Simulation of SFB with the L[GGGGS]₂AAA linker before FdnH with A, RoseTTAFold and B, AlphaFold. C. and D. Simulation of SFB with the L[GGGGS]₂AAA linker flanking FdnH with C, RoseTTAFold and D, AlphaFold. Sufl is shown in red, BLA in cyan and FdnH plus linker in green.

5.3.3 Experimental validation of the linker design to improve the behaviour of the SFB fusion.

Based on the modelling results presented above, it appeared that inclusion of a double flexible linker flanking the FdnH region on either side should allow this portion of the SFB fusion to adopt an α -helical structure. To validate this experimentally the SFB construct was modified by cloning to introduce the L[GGGGS]₂AAA flanking linkers. During the design of this construct the proline that preceded the FdnH C-tail was also included. This new construct was designated SFB-DFL.

This construct was introduced into the MC4100 and DADE strains, and MIC assays were performed. The results are shown in Fig. 5.14.

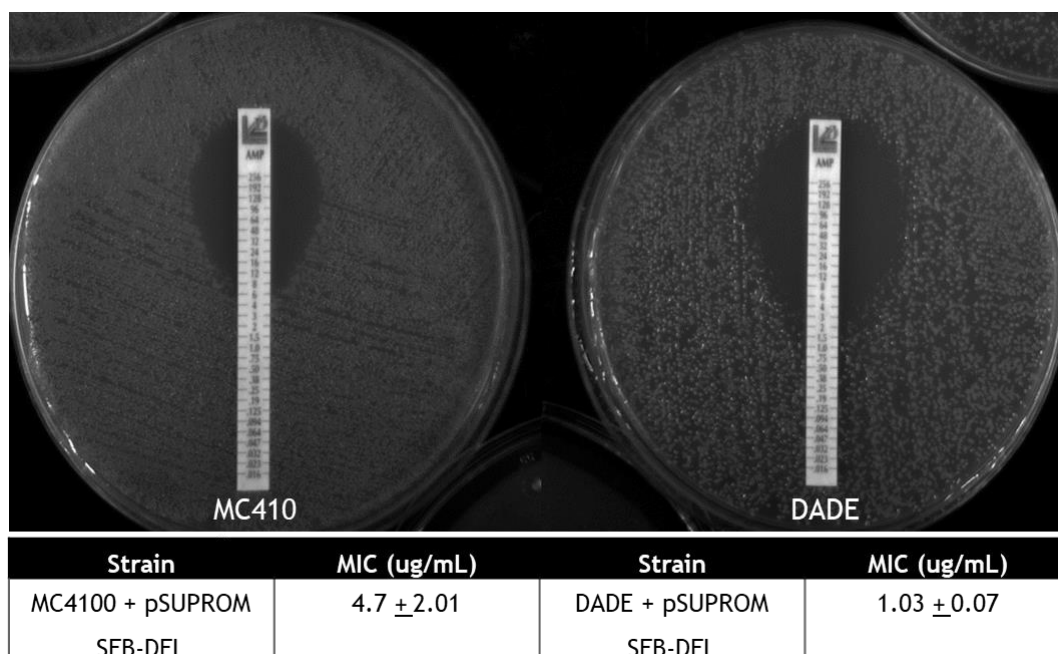


Figure 5.14. MIC assays for strains MC4100 and DADE harbouring pSUPROM SFB-DFL. The Table indicates the average MICs for each strain, \pm one standard deviation, $n = 10$.

Strikingly, inclusion of the linker region resulted in the fusion protein behaving entirely as expected, i.e. no longer conferring significant ampicillin resistance on either the wild type or *tat* mutant strain.

To confirm that the protein was stably produced in the two strain backgrounds and present in the membrane fraction of the *tat*⁺ strain, a cell fractionation experiment was conducted. Western blot analysis showed that the full length fusion protein could be detected in the membrane fraction of the wild type strain (Fig. 5.15) and that this protein was fully integrated because it was not extracted by treatment with urea. Surprisingly, however, the fusion protein was also stably integrated in the membrane of the *tat* mutant strain. A small amount of full length protein was also detected in the soluble fraction of both strains.

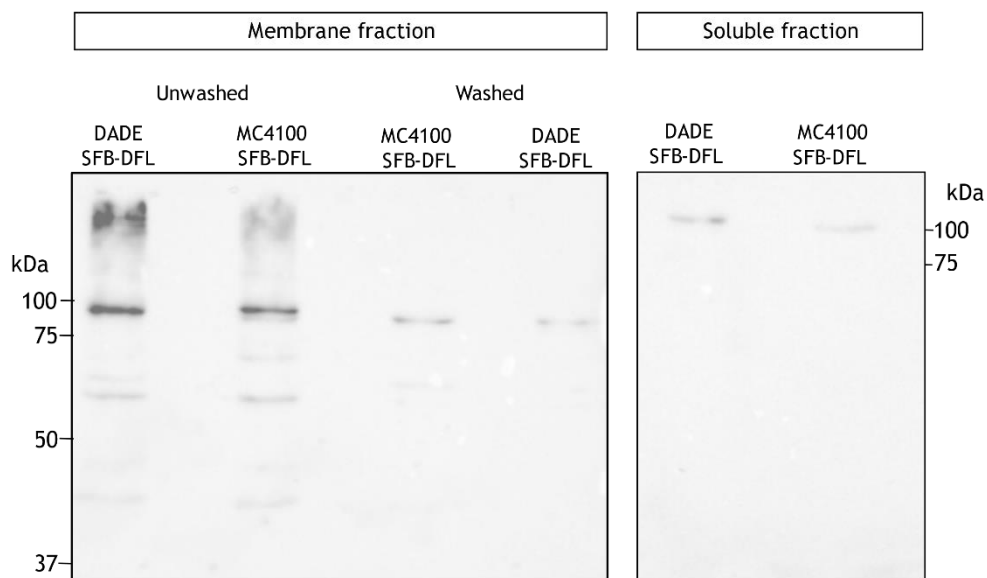


Figure 5.15. Western Blot analysis of membranes, urea-washed membranes and the soluble cell fraction from strains MC4100 and DADE producing SFB-DFL using anti-Bla-antibody. Equivalent amounts of protein were loaded in each lane.

The presence of the SFB-DFL fusion protein in the membranes of the *tat* mutant strain indicates that at least some of the fusion protein is probably being integrated by the Sec pathway. However, given that the MIC for ampicillin is very low it would suggest that the Bla region of the fusion protein is retained at the cytoplasmic side of the membrane.

5.4 Discussion

In this chapter I have attempted to design fusion protein constructs that could serve as reporters for the Tat-dependent integration of C-tails. The fusion proteins were designed based on prior results of Hatzixanthi, Palmer, and Sargent (2003) who showed that fusing the C-tail of FdnH to the Tat substrate SufI resulted in anchoring to the periplasmic side of the membrane. The first of these reporters was MalE, the periplasmic binding protein component of the maltose ABC transporter. MalE was fused directly after the C-tail of FdnH in the SufI-FdnH_{CT} construct and the cellular location of MalE portion was determined indirectly by plating MacConkey maltose indicator plates. Inexplicably, both the *tat*⁺ and *tat*⁻ cells could ferment

maltose suggesting that MalE was reaching the periplasmic side of the membrane. Although I did not investigate this further, my findings with the second fusion protein, SFB, may offer some insight into these results. Aberrant folding of the FdnH helix, as suggested for the analogous Bla construct, may prevent correct integration into the membrane, potentially resulting in some of the SFM fusion protein being fully exported to the periplasm. This could be explored further by fully fractionating the cells and quantifying the levels of SFM in the periplasmic fraction. Structural modelling alongside engineering experiments, similar to that undertaken with the Bla fusion could also be used to improve the behaviour of the MalE fusion. However, given the qualitative nature of the reporter, and the sensitivity of *E. coli tat* mutants to growth inhibition by bile salts present in MacConkey media (Stanley et al. 2001; Ize et al. 2003; Reynolds et al. 2011 - as also seen in Fig 5.3 with the mucoid appearance of the *tat* mutant colonies) it was decided not to pursue this fusion reporter any further.

The second reporter I used here was Bla, which replaced MalE in the tripartite fusion. Bla was deemed a more suitable reporter because its activity can be measured semi-quantitatively using ampicillin-containing MIC strips. However, as before with the MalE fusion, unexpected results were found, with the fusion conferring full resistance to ampicillin in a Tat-independent manner. After making some empirical changes to the fusion sequence by replacing a negatively charged patch with positive charges, which did not alter its behaviour, I turned to protein structure prediction programmes in an attempt to understand why the fusion did not behave as expected. Curiously, both AlphaFold and RoseTTAfold predicted that the presence of Bla at the end of the C-tail would prevent it from adopting the expected α -helical structure. I therefore introduced a series of linker sequences flanking the FdnH segment to examine the predicted folding of the fusion protein containing these modifications. Having identified a flexible linker which reliably appeared to facilitate correct folding of the FdnH helix I engineered this into the SFB fusion. Remarkably, this resulted in the fusion no longer conferring high levels of ampicillin resistance, consistent with the Bla portion now being retained inside the cell.

While my engineering experiments have produced a fusion protein that now retains the C-terminal domain at the cytoplasmic side of the membrane, it is still

not suitable for use as a reporter for Tat-dependent C-tail integration. This is because fractionation experiments indicate that at least some of the fusion protein is integrated into the membrane Tat-independently. Therefore, further engineering of the reporter would be needed, for example by including a strong Sec-avoidance motif in the c-region of the signal peptide, to prevent any interaction with the Sec pathway (Tooke et al. 2017; Cristobal et al. 1999).

Chapter 6. Conclusions and future outlook

At the outset of this PhD project, three families of Tat-dependent tail-anchored proteins had been identified in *E. coli*. These are the small subunits of the periplasmic hydrogenases and formate dehydrogenases, and the electron transport protein HybA (Dubini et al. 2002; Pinske et al. 2011). All three types of protein contain iron-sulphur clusters, and they are widespread across the bacterial kingdom. As a result of the work presented in this thesis, at least three new Tat-dependent tail-anchored proteins have been identified, namely WP_086565138.1 S1 family peptidase from *Streptomyces africanus*, WP_011931836.1 YcnI family protein from *Clavibacter michiganensis*, and WP_031122887.1 LPXTG cell wall anchor domain-containing protein from *Streptomyces* sp. NRRL S-623. None of these proteins are predicted to bind redox cofactors, and they are the first identified Tat-dependent tail-anchored proteins that are not expected to be involved in electron transport.

The selected candidates came from a group of many other proteins that are also potential targets of further study. Beyond the nine proteins analysed here, a further 27 protein families have been identified in Chapter 3 as provide promising avenues for further investigation. It is anticipated that identification of further C-tail anchored Tat substrates will add to our understanding of the biological roles the Tat pathway serves in different organisms.

One class of transiently tail-anchored exported proteins that are abundant in Gram-positive bacteria are the sortase substrates. Here the C-tail plays an important role in their biogenesis, anchoring them to the extracellular face of the cytoplasmic membrane allowing the LPXTG sorting motif to be correctly positioned for recognition by sortase. In this thesis the signal peptides and C-tails of three candidate Tat-dependent sortase substrates were tested for their functionality in *E. coli*. While the C-tails mediated membrane anchoring of the reporter protein, Sufl, none of the three signal peptides appeared to export AmiA in a Tat-dependent manner. Given that these proteins are all from actinobacteria, testing their Tat-dependence using the *Streptomyces* agarose reporter would be an important next step. In this context it should be noted that the bioinformatic analysis of Tsolis *et al.* (Tsolis et al. 2018) identified 30 candidate extracellular C-tail proteins encoded

by the genome of *S. lividans*, 16 of which have sortase motifs. Nine of these have plausible Tat signal peptides - while three of these have been shown to be Tat-independent by Widdick et al. (2006) the other six remain uncharacterised. Future work could focus on analysing the export route of these six candidates as they may also represent novel Tat-dependent C-tail proteins.

Understanding the mechanism of integration of tail-anchored proteins by the Tat pathway is of paramount significance, given the essential roles these proteins play in the respiratory pathways of bacteria such as *E. coli*. Hatzixanthis, Palmer, and Sargent (2003) demonstrated that C-tails act as stop transfer sequences during export by the Tat pathway - this means that they are directly integrated into the membrane during translocation rather than being fully exported to the periplasm and inserting into the membrane from the periplasmic side following export (Hatzixanthis, Palmer, and Sargent 2003). How this is achieved is unclear, but it has been speculated that this may be related to the 'folding quality control' activity exhibited by the Tat system.

It has long been known that the Tat system can distinguish between folded and unfolded substrates. Classic experiments using a Tat signal peptide-alkaline phosphatase (PhoA) fusion showed there was no transport because PhoA, which requires two intra-molecular disulphide bonds for activity and stability, could not fold in the reducing environment of the cytoplasm (Stanley et al. 2002). However, when the same fusion was produced in *E. coli* that had been genetically engineered to have an oxidising cytoplasm, the protein could fold and was subsequently translocated through the Tat pathway (DeLisa, Tullman, and Georgiou 2003). Further experiments testing the ability of the Tat system to transport unfolded proteins surprisingly showed that the *E. coli* Tat machinery could translocate small unstructured proteins as long as they were hydrophilic and less than 100 amino acids in length (Cline and McCaffery 2007; Richter et al. 2007). Unfolded proteins that exceeded this size resulted in Tat transport being blocked at a late stage with release of the polypeptide into the membrane (Cline and McCaffery 2007).

These findings were interpreted to indicate that the Tat system itself does not have an intrinsic quality control mechanism but that proteins which are too large or

have exposed hydrophobic stretches are incompatible with the translocation mechanism and result in transport stalling and translocon disassembly (Richter et al. 2007; Cline 2007; Richter and Bruser 2005). Palmer and Berks proposed the hypothesis that the stalling of unfolded substrates with exposed hydrophobic residues is related to the mechanism of membrane integration of Tat-dependent C-tails (Palmer and Berks 2012).

These findings lead us to speculate on a possible connection between the Tat system's ability to discriminate between folded and unfolded substrates and its role in integrating C-tail-anchored proteins. It is conceivable that the Tat machinery, through its ability to sense substrate folding states, may play a dual role in both translocation and membrane integration. Specifically, we could hypothesize that the Tat system's quality control mechanism, which permits the translocation of properly folded proteins, might also facilitate the integration of C-tail anchors into the membrane.

Moreover, considering the association of the signal peptide h-region with TatB and its role in the Tat complex assembly (Alami et al. 2003; Gerard and Cline 2006; Panahandeh et al. 2008), we could speculate that hydrophobic C-tails might trigger a reversal of this process. Such a reversal could involve TatB relinquishing its binding site on TMH5 of TatC, allowing TatA to occupy it. This hypothetical scenario could suggest a dynamic interplay between C-tail hydrophobicity and Tat component interactions, ultimately influencing the fate of C-tail-anchored proteins.

Furthermore, it is worth considering the functional implications of anchoring enzymes to the membrane via C-tails. Many of the studied proteins are enzymes, including peptidases. While enzymes like peptidases are typically secreted outside the bacterial cell, it is plausible that in some cases, anchoring the enzyme to the membrane could provide benefits. This anchoring might result in enzymatic activity in close proximity to the bacterial membrane, potentially making the products of these enzymes more readily available for the bacterium itself, rather than for other organisms in the environment. This speculation opens up new avenues of inquiry into the potential regulatory and adaptive roles of C-tail-anchored enzymes in bacterial physiology.

Further work could involve directly testing this hypothesis by investigating the minimal hydrophobicity requirements for C-tail integration. This could be achieved by constructing amino acid substitutions at different positions in the tail region to alter its hydrophobicity. The SufI-C-tail reporter used in Chapter 5 could be used to assess the effects of such substitutions, but this would be quite laborious as it would require cell fractionation and membrane washing for each C-tail substitution. A genetic reporter would therefore provide a much more rapid way of screening the effect of amino acid substitutions on C-tail integration. In Chapter 5 I investigated the utility of using the MalE or Bla reporter proteins fused after the C-tail of a SufI-C-tail construct as a reporter for C-tail integration. While neither construct behaved as expected, extensive engineering efforts with the Bla fusion went some way towards design of a construct that could be used for this purpose. Further experimental work would be required to engineer the fusion to avoid interaction with the Sec pathway, for example by inclusion of a strong Sec-avoidance motif in the SufI signal peptide's c-region.

The hydrophobicity of the Tat signal peptide is a key determinant in its interaction with the Tat complex and plays a critical role in the assembly of the active translocon. Cross-linking studies have shown that the signal peptide h-region interacts with the TatB component of the Tat receptor complex (Alami et al. 2003; Gerard and Cline 2006; Panahandeh et al. 2008). In the resting state the receptor TatB occupies a binding site on TMH5 of TatC, however upon interaction with a Tat signal peptide, TatB vacates the TMH5 binding site allowing a molecule of TatA to occupy it (Alcock et al. 2016; Habersetzer et al. 2017). It has been suggested that interaction of TatB with the signal peptide h-region promotes this rearrangement (Huang et al. 2017). One mechanistic possibility is that hydrophobic C-tails promote a reverse of this step, causing a switch of TatB for TatA at the TMH5 binding site and promoting translocase disassembly.

A powerful genetic reporter could be used to probe the role of the Tat components in C-tail integration, for example by isolating substitutions in TatA, TatB or TatC that are unable to integrate C-tails. Alternatively, if substitutions are identified in the C-tail itself that result in a failure to be integrated, suppressor

substitutions could be identified in Tat proteins that restore integration to defective C-tails. Other methods that could be used to explore C-tail integration could include *in vivo* site-specific photoaffinity crosslinking (Farrell et al. 2005; Okuda and Tokuda 2009). Here a genetically-encoded photoreactive *p*-benzoyl-phenylalanine (pBPA) crosslinker would be introduced at site-specific positions within the transmembrane region of the C-tail. Following irradiation of whole cells, western blotting using antibodies to TatA, TatB and TatC with could be used to identify detect crosslinks between the C-tail and Tat components.

Bibliography

- Abby, S. S., J. Cury, J. Guglielmini, B. Néron, M. Touchon, and E. P. C. Rocha. 2016. 'Identification of protein secretion systems in bacterial genomes', *Sci Rep*, 6: 1-14.
- Abby, S. S., and E. P. C. Rocha. 2012. 'The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems', *PLoS Genet*, 8: e1002983.
- Abdallah, A. M., N. C. Gey van Pittius, P. A. DiGiuseppe Champion, J. Cox, J. Luirink, C. M. J. E. Vandenbroucke-Grauls, B. J. Appelmelk, and W. Bitter. 2007. 'Type VII secretion—mycobacteria show the way', *Nat Rev Microbiol*, 5: 883-91.
- Akeda, Y., and J. E. Galán. 2005. 'Chaperone release and unfolding of substrates in type III secretion', *Nature*, 437: 911-15.
- Alami, M., I. Luke, S. Deitermann, G. Eisner, H. G. Koch, J. Brunner, and M. Muller. 2003. 'Differential interactions between a twin-arginine signal peptide and its translocase in *Escherichia coli*', *Mol Cell*, 12: 937-46.
- Alcock, F., M. A. Baker, N. P. Greene, T. Palmer, M. I. Wallace, and B. C. Berks. 2013. 'Live cell imaging shows reversible assembly of the TatA component of the twin-arginine protein transport system', *Proc Natl Acad Sci U S A*, 110: E3650-9.
- Alcock, F., P. J. Stansfeld, H. Basit, J. Habersetzer, M. A. Baker, T. Palmer, M. I. Wallace, and B. C. Berks. 2016. 'Assembling the Tat protein translocase', *Elife*, 5.
- Aldridge, C., X. Ma, F. Gerard, and K. Cline. 2014. 'Substrate-gated docking of pore subunit Tha4 in the TatC cavity initiates Tat translocase assembly', *J Cell Biol*, 205: 51-65.
- Allen, C. E., and M. P. Schmitt. 2009. 'HtaA is an iron-regulated hemin binding protein involved in the utilization of heme iron in *Corynebacterium diphtheriae*', *J Bacteriol*, 191: 2638-48.
- Almagro Armenteros, J. J., K. D. Tsirigos, C. K. Sonderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. 2019. 'SignalP 5.0 improves signal peptide predictions using deep neural networks', *Nat Biotechnol*, 37: 420-23.
- Alvarez-Martinez, C. E., and P. J. Christie. 2009. 'Biological diversity of prokaryotic type IV secretion systems', *Microbiol Mol Biol Rev*, 73: 775-808.
- Amet, N., H.-F. Lee, and W.-C. Shen. 2009. 'Insertion of the designed helical linker led to increased expression of tf-based fusion proteins', *Pharm Res*, 26: 523-28.
- Angelini, S., S. Deitermann, and H.-G. Koch. 2005. 'FtsY, the bacterial signal-recognition particle receptor, interacts functionally and physically with the SecYEG translocon', *EMBO Rep*, 6: 476-81.

- Bachmann, J., B. Bauer, K. Zwicker, B. Ludwig, and O. Anderka. 2006. 'The Rieske protein from *Paracoccus denitrificans* is inserted into the cytoplasmic membrane by the twin-arginine translocase', *Febs J*, 273: 4817-30.
- Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millan, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker. 2021. 'Accurate prediction of protein structures and interactions using a three-track neural network', *Science*, 373: 871-76.
- Baek, M., and D. Baker. 2022. 'Deep learning and protein structure modeling', *Nature Meth*, 19: 13-14.
- Baglieri, J., D. Beck, N. Vasisht, C. Smith, and C. Robinson. 2011. 'Structure of the TatA paralog, TatE, suggests a structurally homogeneous form of Tat protein translocase that transports folded proteins of differing diameter', *J Biol Chem*.
- Bagos, P. G., E. P. Nikolaou, T. D. Liakopoulos, and K. D. Tsirigos. 2010. 'Combined prediction of Tat and Sec signal peptides with hidden Markov models', *Bioinformatics*, 26: 2811-7.
- Barnett, J. P., R. T. Eijlander, O. P. Kuipers, and C. Robinson. 2008. 'A minimal Tat system from a gram-positive organism: a bifunctional TatA subunit participates in discrete TatAC and TatA complexes', *J Biol Chem*, 283: 2534-42.
- Batth, T. S., J. L. Simonsen, C. Hernández-Rollán, S. Brander, J. P. Morth, K. S. Johansen, M. H. H. Nørholm, J. B. Hoof, and J. V. Olsen. 2022. 'A membrane integral methyltransferase catalysing N-terminal histidine methylation of lytic polysaccharide monooxygenases', *bioRxiv*: 2022.10. 03.510680.
- Bechtluft, P. , R. G. H. Van Leeuwen, M. Tyreman, D. Tomkiewicz, N. Nouwen, H. L. Tepper, A. J. M. Driessen, and S. J. Tans. 2007. 'Direct observation of chaperone-induced changes in a protein folding pathway', *Science*, 318: 1458-61.
- Beck, K., G. Eisner, D. Trescher, R. E. Dalbey, J. Brunner, and M. Müller. 2001. 'YidC, an assembly site for polytopic *Escherichia coli* membrane proteins located in immediate proximity to the SecYE translocon and lipids', *EMBO Rep*, 2: 709-14.
- Benabdelhak, H., S. Kiontke, C. Horn, R. Ernst, M. A. Blight, I. B. Holland, and L. Schmitt. 2003. 'A specific interaction between the NBD of the ABC-transporter HlyB and a C-terminal fragment of its transport substrate haemolysin A', *J Mol Biol*, 327: 1169-79.
- Bendtsen, J. D., H. Nielsen, D. Widdick, T. Palmer, and S. Brunak. 2005. 'Prediction of twin-arginine signal peptides', *BMC Bioinform*, 6: 167.

- Benoit, S., H. Abaibou, and M.-A. Mandrand-Berthelot. 1998. 'Topological analysis of the aerobic membrane-bound formate dehydrogenase of *Escherichia coli*', *J Bacteriol*, 180: 6625-34.
- Benoit, S. L., R. J. Maier, R. G. Sawers, and C. Greening. 2020. 'Molecular hydrogen metabolism: a widespread trait of pathogenic bacteria and protists', *Microbiol Mol Biol Rev*, 84: e00092-19.
- Berks, B. C., T. Palmer, and F. Sargent. 2003. 'The Tat protein translocation pathway and its role in microbial physiology', *Adv Microb Physiol*, 47: 187-254.
- Berks, B. C., F. Sargent, and T. Palmer. 2000. 'The Tat protein export pathway', *Mol Microbiol*, 35: 260-74.
- Bernhardt, T. G., and P. A. de Boer. 2003. 'The *Escherichia coli* amidase AmiC is a periplasmic septal ring component exported via the twin-arginine transport pathway', *Mol Microbiol*, 48: 1171-82.
- Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, and G. F. Mayhew. 1997. 'The complete genome sequence of *Escherichia coli* K-12', *Science*, 277: 1453-62.
- Blaudeck, N., P. Kreutzenbeck, R. Freudl, and G. A. Sprenger. 2003. 'Genetic analysis of pathway specificity during posttranslational protein translocation across the *Escherichia coli* plasma membrane', *J Bacteriol*, 185: 2811-9.
- Blount, Z. D. 2015. 'The unexhausted potential of *E. coli*', *Elife*, 4: e05826.
- Blummel, A. S., L. A. Haag, E. Eimer, M. Muller, and J. Frobel. 2015. 'Initial assembly steps of a translocase for folded proteins', *Nat Commun*, 6: 7234.
- Bolhuis, A., J. E. Mathers, J. D. Thomas, C. M. Barrett, and C. Robinson. 2001. 'TatB and TatC form a functional and structural unit of the twin-arginine translocase from *Escherichia coli*', *J Biol Chem*, 276: 20213-9.
- Bowman, L., T. Palmer, and F. Sargent. 2013. 'A regulatory domain controls the transport activity of a twin-arginine signal peptide', *FEBS Lett*, 587: 3365-70.
- Bowman, L., L. Flanagan, P. K. Fyfe, A. Parkin, W. N. Hunter, and F. Sargent. 2014. 'How the structure of the large subunit controls function in an oxygen-tolerant [NiFe]-hydrogenase', *Biochem J*, 458: 449-58.
- Bowman, L., and T. Palmer. 2021. 'The type VII secretion system of *Staphylococcus*', *Annu Rev Microbiol*, 75: 471-94.
- Braun, V., and K. Hantke. 2020. 'Novel Tat-dependent protein secretion', *J Bacteriol* 202: e00058-20.
- Brodin, P., I. Rosenkrands, P. Andersen, S. T. Cole, and R. Brosch. 2004. 'ESAT-6 proteins: protective antigens and virulence factors?', *Trends Microbiol*, 12: 500-08.

- Broome-Smith, J. K., and B. G. Spratt. 1986. 'A vector for the construction of translational fusions to TEM β -lactamase and the analysis of protein export signals and membrane protein topology', *Gene*, 49: 341-49.
- Bruser, T., and C. Sanders. 2003. 'An alternative model of the twin arginine translocation system', *Microbiol Res*, 158: 7-17.
- Buchanan, G., E. de Leeuw, N. R. Stanley, M. Wexler, B. C. Berks, F. Sargent, and T. Palmer. 2002. 'Functional complexity of the twin-arginine translocase TatC component revealed by site-directed mutagenesis', *Mol Microbiol*, 43: 1457-70.
- Bumba, L. , J. Masin, P. Macek, T. Wald, L. Motlova, I. Bibova, N. Klimova, L. Bednarova, V. Veverka, and M. Kachala. 2016. 'Calcium-driven folding of RTX domain β -rolls ratchets translocation of RTX proteins through type I secretion ducts', *Mol Cell*, 62: 47-62.
- Bunduc, C. M., D. Fahrenkamp, J. Wald, R. Ummels, W. Bitter, E. N. G. Houben, and T. C. Marlovits. 2021. 'Structure and dynamics of a *mycobacterial* type VII secretion system', *Nature*, 593: 445-48.
- Büttner, D., C. Lorenz, E. Weber, and U. Bonas. 2006. 'Targeting of two effector protein classes to the type III secretion system by a HpaC-and HpaB-dependent protein complex from *Xanthomonas campestris* pv. *vesicatoria*', *Mol Microbiol*, 59: 513-27.
- Caldelari, I., T. Palmer, and F. Sargent. 2008. '*Escherichia coli* tat mutant strains are able to transport maltose in the absence of an active malE gene', *Arch Microbiol*, 189: 597-604.
- Casadaban, M. J., and S. N. Cohen. 1979. 'Lactose genes fused to exogenous promoters in one step using a Mu-lac bacteriophage: in vivo probe for transcriptional control sequences', *Proc Natl Acad Sci U S A*, 76: 4530-3.
- Cascales, E., and P. J Christie. 2003. 'The versatile bacterial type IV secretion systems', *Nat Rev Microbiol*, 1: 137-49.
- Cao, Z., M. G. Casabona, H. Kneuper, J. D. Chalmers, and T. Palmer. 2016. The type VII secretion system of *Staphylococcus aureus* secretes a nuclease toxin that targets competitor bacteria. *Nat Microbiol*, 2: 16183.
- Chaddock, A. M., A. Mant, I. Karnauchov, S. Brink, R. G. Herrmann, R. B. Klosgen, and C. Robinson. 1995. 'A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase', *EMBO J*, 14: 2715-22.
- Chance, R. E., and B. H. Frank. 1993. 'Research, development, production, and safety of biosynthetic human insulin', *Diabet Care*, 16: 133-42.
- Chang, C. Y., L. Hobley, R. Till, M. Capeness, M. Kanna, W. Burt, P. Jagtap, S. Aizawa, and R. E. Sockett. 2011a. 'The *Bdellovibrio bacteriovorus* twin-arginine transport system has roles in predatory and prey-independent growth', *Microbiology*, 157: 3079-93.

- Chang, Y.-W., L. A. Rettberg, D. R. Ortega, and G. J. Jensen. 2017. 'In vivo structures of an intact type VI secretion system revealed by electron cryotomography', *EMBO Rep*, 18: 1090-99.
- Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, and S. J. Hultgren. 2002. 'Role of *Escherichia coli* curli operons in directing amyloid fiber formation', *Science*, 295: 851-55.
- Chatzi, K. E., M. F. Sardis, A. Economou, and S. Karamanou. 2014. 'SecA-mediated targeting and translocation of secretory proteins', *Biochim Biophys Acta -Mol Cell Res*, 1843: 1466-74.
- Chen, X., J. L. Zaro, and W.-C. Shen. 2013a. 'Fusion protein linkers: property, design and functionality', *Adv Drug Deliv Rev* 65: 1357-69.
- Chen, X., J. L. Zaro, and W.-C. Shen. 2013b. 'Fusion protein linkers: effects on production, bioactivity, and pharmacokinetics', *Fusion protein technologies for biopharmaceuticals: applications and challenges*: 57-73.
- Chen, Y., S. K. Shanmugam, and R. E. Dalbey. 2019. 'The principles of protein targeting and transport across cell membranes', *Protein J*, 38: 236-48.
- Chernyatina, A. A., and H. H. Low. 2019. 'Core architecture of a bacterial type II secretion system', *Nat Commun*, 10: 5437.
- Clancy, K. W., J. A. Melvin, and D. G. McCafferty. 2010. 'Sortase transpeptidases: insights into mechanism, substrate specificity, and inhibition', *Pep Sci*, 94: 385-96.
- Clantin, B., A.-S. Delattre, P. Rucktooa, N. Saint, A. C. Méli, C. Loch, F. Jacob-Dubuisson, and V. Villeret. 2007. 'Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily', *Science*, 317: 957-61.
- Cleon, F., J. Habersetzer, F. Alcock, H. Kneuper, P. J. Stansfeld, H. Basit, M. I. Wallace, B. C. Berks, and T. Palmer. 2015. 'The TatC component of the twin-arginine protein translocase functions as an obligate oligomer', *Mol Microbiol*, 98: 111-29.
- Cline, K., and M. McCafferty. 2007. 'Evidence for a dynamic and transient pathway through the TAT protein transport machinery', *Embo J*, 26: 3039-49.
- Cline, K., Theg, S.M. 2007. 'The Sec and Tat protein translocation pathways in chloroplasts.' in R. E. Dalbey, Koehler, C.M., Tamanoi, F. (ed.), *Molecular Machines Involved in Protein Transport across Cellular Membranes* (Elsevier: London).
- Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaiia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. 1998. 'Deciphering the biology of

- Mycobacterium tuberculosis* from the complete genome sequence', *Nature*, 396: 190-90.
- Collier, D. N, V. A. Bankaitis, J. B. Weiss, and P. J. Bassford Jr. 1988. 'The antifolding activity of SecB promotes the export of the *E. coli* maltose-binding protein', *Cell*, 53: 273-83.
- Costa, A. C., C. Figueiredo, and E. Touati. 2009. 'Pathogenesis of *Helicobacter pylori* infection', *Helicobacter*, 14: 15-20.
- Costa, T. R. D., C. Felisberto-Rodrigues, A. Meir, M. S. Prevost, A. Redzej, M. Trokter, and Gabriel Waksman. 2015. 'Secretion systems in Gram-negative bacteria: G. and mechanistic insights', *Nat Rev Microbiol*, 13: 343-59.
- Craney, A., K. Tahlan, D. Andrews, and J. Nodwell. 2011. 'Bacterial transmembrane proteins that lack N-terminal signal sequences', *PLoS One*, 6: e19421.
- Cristobal, S., J. W. de Gier, H. Nielsen, and G. von Heijne. 1999. 'Competition between Sec- and TAT-dependent protein translocation in *Escherichia coli*', *EMBO J*, 18: 2982-90.
- Crystal-Ornelas, R., B. P. M. Edwards, K. Hébert, E. J. Hudgins, L L. Sánchez Reyes, E. R. Scott, M. J. Grainger, V. Foroughirad, A. D. Binley, and C. B. Brookson. 2022. "Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution." In.: Manubot.
- Dabney-Smith, C., H. Mori, and K. Cline. 2006. 'Oligomers of Tha4 organize at the thylakoid Tat translocase during protein transport', *J Biol Chem*, 281: 5476-83.
- Dalbey, R. E., and A. Kuhn. 2012. 'Protein traffic in Gram-negative bacteria-how exported and secreted proteins find their way', *FEMS Microbiol Rev*, 36: 1023-45.
- Damle, M. S, A. N. Singh, S. C. Peters, V. A. Szalai, and O. S. Fisher. 2021. 'The YcnI protein from *Bacillus subtilis* contains a copper-binding domain', *J Biol Chem*, 297: 101078.
- Datta, S., Y. Mori, K. Takagi, K. Kawaguchi, Z.-W. Chen, T. Okajima, S. Kuroda, T. Ikeda, K. Kano, and K. Tanizawa. 2001. 'Structure of a quinoxinoprotein amine dehydrogenase with an uncommon redox cofactor and highly unusual crosslinking', *Proc Natl Acad Sci U S A*, 98: 14268-73.
- Dautin, N. 2021. 'Folding control in the path of type 5 secretion', *Toxins*, 13: 341.
- Dave, K. 2012. 'A Python Book: Beginning Python, Advanced Python, and Python Exercises', *Section 1.1. Archived from the original (PDF) on*.
- Davidson, V. L., L. H. Jones, M. E. Graichen, F. S. Mathews, and J. P. Hosler. 1997. 'Factors which stabilize the methylamine dehydrogenase- amicyanin electron transfer protein complex revealed by site-directed mutagenesis', *Biochemistry*, 36: 12733-38.

- Davidson, V. L., and C. M. Wilmot. 2013. 'Posttranslational biosynthesis of the protein-derived cofactor tryptophan tryptophylquinone', *Ann Rev Biochem*, 82: 531-50.
- De Buck, E., L. Vranckx, E. Meyen, L. Maes, L. Vandersmissen, J. Anne, and E. Lammertyn. 2007. 'The twin-arginine translocation pathway is necessary for correct membrane insertion of the Rieske Fe/S protein in *Legionella pneumophila*', *FEBS Lett*, 581: 259-64.
- De Gier, J.-W. L., P. A. Scotti, A. Sääf, Q. A. Valent, A. Kuhn, J. Luirink, and G. Von Heijne. 1998. 'Differential use of the signal recognition particle translocase targeting pathway for inner membrane protein assembly in *Escherichia coli*', *Proc Natl Acad Sci U S A*, 95: 14646-51.
- De Keersmaecker, S., L. Van Mellaert, E. Lammertyn, K. Vrancken, J. Anne, and N. Geukens. 2005. 'Functional analysis of TatA and TatB in *Streptomyces lividans*', *Biochem Biophys Res Commun*, 335: 973-82.
- DeLisa, M. P., P. Samuelson, T. Palmer, and G. Georgiou. 2002. 'Genetic analysis of the twin arginine translocator secretion pathway in bacteria', *J Biol Chem*, 277: 29825-31.
- DeLisa, M. P., D. Tullman, and G. Georgiou. 2003. 'Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway', *Proc Natl Acad Sci U S A*, 100: 6115-20.
- Depluvere, S., S. Devos, and B. Devreese. 2016. 'The role of bacterial secretion systems in the virulence of gram-negative airway pathogens associated with cystic fibrosis', *Front Microbiol*, 7: 1336.
- Dilks, K., M. I. Gimenez, and M. Pohlschroder. 2005. 'Genetic and biochemical analysis of the twin-arginine translocation pathway in halophilic archaea', *J Bacteriol*, 187: 8104-13.
- Dilks, K., R. W. Rose, E. Hartmann, and M. Pohlschroder. 2003. 'Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey', *J Bacteriol*, 185: 1478-83.
- Drake, Z. C., J. T. Seffernick, and S. Lindert. 2022. 'Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling', *Nat Comm*, 13: 7846.
- Driessen, A. J. 1992. 'Bacterial protein translocation: kinetic and thermodynamic role of ATP and the protonmotive force', *Trends Biochem Sci*, 17: 219-23.
- Dubini, A., and F. Sargent. 2003. 'Assembly of Tat-dependent [NiFe] hydrogenases: identification of precursor-binding accessory proteins', *FEBS Lett*, 549: 141-6.
- Dubini, A., R. L. Pye, R. L. Jack, T. Palmer, and F. Sargent. 2002. 'How bacteria get energy from hydrogen: a genetic analysis of periplasmic hydrogen oxidation in *Escherichia coli*', *Int J Hydr Energ*, 27: 1413-20.
- Eddy, S. R. 2011. 'Accelerated Profile HMM Searches', *PLoS Comp Biol*, 7: e1002195.

- Edgar, R. C. 2022. 'Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny', *Nat Commun*, 13: 6968.
- Edgar, R. C. 2004. 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Res*, 32: 1792-97.
- Egea, P. F, and R. M. Stroud. 2010. 'Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes', *Proc Natl Acad Sci U S A*, 107: 17182-87.
- El Rayes, J., R. Rodríguez-Alonso, and J.-F. Collet. 2021. 'Lipoproteins in Gram-negative bacteria: New insights into their biogenesis, subcellular targeting and functional roles', *Curr Opin Microbiol*, 61: 25-34.
- Evans, M. L., E. Chorell, J. D. Taylor, J. Åden, A. Götheson, F. Li, M. Koch, L. Sefer, S. J. Matthews, and P. Wittung-Stafshede. 2015. 'The bacterial curli system possesses a potent and selective inhibitor of amyloid formation', *Mol Cell*, 57: 445-55.
- Farrell, I. S., R. Toroney, J. L. Hazen, R. A. Mehl, and J. W. Chin. 2005. 'Photo-cross-linking interacting proteins with a genetically encoded benzophenone', *Nat Methods*, 2: 377-84.
- Fekkes, P., and A. J. M. Driessen. 1999. 'Protein targeting to the bacterial cytoplasmic membrane', *Microbiol Mol Biol Rev*, 63: 161-73.
- Fekkes, P., C. van der Does, and A. J. M. Driessen. 1997. 'The molecular chaperone SecB is released from the carboxy-terminus of SecA during initiation of precursor protein translocation', *EMBO J*, 16: 6105-13.
- Felmlee, T., and R. A. Welch. 1988. 'Alterations of amino acid repeats in the *Escherichia coli* hemolysin affect cytolytic activity and secretion', *Proc Natl Acad Sci U S A*, 85: 5269-73.
- Felmlee, T., S. Pellett, and R. A. Welch. 1985. 'Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin', *J Bacteriol*, 163: 94-105.
- Filloux, A., A. Hachani, and S. Bleves. 2008. 'The bacterial type VI secretion machine: yet another player for protein transport across membranes', *Microbiology*, 154: 1570-83.
- Gannon, P. M., P. Li, and C. A. Kumamoto. 1989. 'The mature portion of *Escherichia coli* maltose-binding protein (MBP) determines the dependence of MBP on SecB for export', *J Bacteriol*, 171: 813-18.
- Gao, L., Z. Guan, P. Gao, W. Zhang, Q. Qi, and X. Lu. 2020. '*Cytophaga hutchinsonii* gldN, encoding a core component of the type IX secretion system, is essential for ion assimilation, cellulose degradation, and cell motility', *Appl Env Microbiol*, 86: e00242-20.
- Gawarzewski, I., S. H. J. Smits, L. Schmitt, and J. Jose. 2013. 'Structural comparison of the transport units of type V secretion systems', *Biol Chem*, 394: 1385-98.

- Gaytán, M. O., V. I. Martínez-Santos, E. Soto, and B. González-Pedrajo. 2016. 'Type three secretion system in attaching and effacing pathogens', *Front Cell Infect Microbiol*, 6: 129.
- Gerard, F., and K. Cline. 2006. 'Efficient twin arginine translocation (Tat) pathway transport of a precursor protein covalently anchored to its initial cpTatC binding site', *J Biol Chem*, 281: 6130-5.
- Ghim, C.-M., S. K. Lee, S. Takayama, R. J. Mitchell. 2010. 'The art of reporter proteins in science: past, present and future applications'.
- Gibson, D. G., L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison III, and H. O. Smith. 2009. 'Enzymatic assembly of DNA molecules up to several hundred kilobases', *Nat Meth*, 6: 343-45.
- Gimenez, M. R., G. Chandra, P. Van Overvelt, R. Voulhoux, S. Bleves, and B. Ize. 2018. 'Genome wide identification and experimental validation of *Pseudomonas aeruginosa* Tat substrates', *Sci Rep*, 8: 11950.
- Glorigrijević, V., P. D. Renfrew, T. Kosciolk, J. Koehler Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, and H. Vlamakis. 2021. 'Structure-based protein function prediction using graph convolutional networks', *Nat Comm*, 12: 3168.
- Goosens, V. J., and J. M. van Dijl. 2017. 'Twin-arginine protein translocation', *Protein and Sugar Export and Assembly in Gram-positive Bacteria*: 69-94.
- Gorasia, D. G., P. D. Veith, and E. C. Reynolds. 2020. 'The type IX secretion system: advances in structure, function and organisation', *Microorganisms*, 8: 1173.
- Green, E. R., and J. Meccas. 2016. 'Bacterial secretion systems: an overview', *Virulence mechanisms of bacterial pathogens*: 213-39.
- Grenier, D., and V. Dang La. 2011. 'Proteases of *Porphyromonas gingivalis* as important virulence factors in periodontal disease and potential targets for plant-derived compounds: a review article', *Curr Drug Tar*, 12: 322-31.
- Griessl, M. H, B. Schmid, K. Kassler, C. Braunsmann, R. Ritter, B. Barlag, Y.-D. Stierhof, K. U. Sturm, C. Danzer, and C. Wagner. 2013. 'Structural insight into the giant Ca²⁺-binding adhesin SiiE: implications for the adhesion of *Salmonella enterica* to polarized epithelial cells', *Structure*, 21: 741-52.
- Grossman, A. S., T. J. Mauer, K T. Forest, and H. Goodrich-Blair. 2021. 'A widespread bacterial secretion system with diverse substrates', *MBio*, 12: e01956-21.
- Habersetzer, J., K. Moore, J. Cherry, G. Buchanan, P. J. Stansfeld, and T. Palmer. 2017. 'Substrate-triggered position switching of TatA and TatB during Tat transport in *Escherichia coli*', *Open Biol*, 7: 170091.
- Hallgren, J., K. D. Tsirigos, M.D. Pedersen, J. J. Almagro Armenteros, P. Marcatili, H. Nielsen, A. Krogh, and O. Winther. 2022. 'DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks', *Biorxiv*, <https://doi.org/10.1101/2022.04.08.487609>.

- Hatzixanthis, K., T. Palmer, and F. Sargent. 2003. 'A subset of bacterial inner membrane proteins integrated by the twin-arginine translocase', *Mol Microbiol*, 49: 1377-90.
- Heidrich, C., M. F. Templin, A. Ursinus, M. Merdanovic, J. Berger, H. Schwarz, M. A. de Pedro, and J. V. Holtje. 2001. 'Involvement of N-acetylmuramyl-L-alanine amidases in cell separation and antibiotic-induced autolysis of *Escherichia coli*', *Mol Microbiol*, 41: 167-78.
- Hicks, M. G., E. de Leeuw, I. Porcelli, G. Buchanan, B. C. Berks, and T. Palmer. 2003. 'The *Escherichia coli* twin-arginine translocase: conserved residues of TatA and TatB family components involved in protein transport', *FEBS Lett*, 539: 61-7.
- Hirooka, K., T. Edahiro, K. Kimura, and Y. Fujita. 2012. 'Direct and Indirect Regulation of the Operon Involved in Copper Uptake through Two Transcriptional Repressors, YcnK and CsoR, in *Bacillus subtilis*', *J Bacteriol*, 194: 5675-87.
- Ho, B. T., T. G. Dong, and J. J. Mekalanos. 2014. 'A view to a kill: the bacterial type VI secretion system', *Cell Host Microbe*, 15: 9-21.
- Hu, C.-., and T. K. Kerppola. 2003. 'Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis', *Nat. Biotech*, 21: 539-45.
- Huang, Q., F. Alcock, H. Kneuper, J. C. Deme, S. E. Rollauer, S. M. Lea, B. C. Berks, and T. Palmer. 2017. 'A signal sequence suppressor mutant that stabilizes an assembled state of the twin arginine translocase', *Proc Natl Acad Sci U S A*, 114: E1958-E1967.
- Huang, Q., and T. Palmer. 2017. 'Signal Peptide Hydrophobicity Modulates Interaction with the Twin-Arginine Translocase', *MBio*, 8: e00909-17.
- Idalia, V.-M. N., and F. Bernardo. 2017. '*Escherichia coli* as a model organism and its application in biotechnology', *Recent Adv. Physiol. Pathog. Biotechnol. Appl. Tech Open Rij. Croat*: 253-74.
- Ize, B., N. R. Stanley, G. Buchanan, and T. Palmer. 2003. 'Role of the *Escherichia coli* Tat pathway in outer membrane integrity', *Mol Microbiol*, 48: 1183-93.
- Ize, B., I. Porcelli, S. Lucchini, J. C. Hinton, B. C. Berks, and T. Palmer. 2004. 'Novel phenotypes of *Escherichia coli* tat mutants revealed by global gene expression and phenotypic analysis', *J Biol Chem*, 279: 47543-54.
- Jack, R. L., G. Buchanan, A. Dubini, K. Hatzixanthis, T. Palmer, and F. Sargent. 2004. 'Coordinating assembly and export of complex bacterial proteins', *Embo J*, 23: 3962-72.
- Johnson, T. J, and L. K. Nolan. 2009. 'Pathogenomics of the virulence plasmids of *Escherichia coli*', *Microbiol Mol Biol Rev*, 73: 750-74.
- Jongbloed, J. D., H. Antelmann, M. Hecker, R. Nijland, S. Bron, U. Airaksinen, F. Pries, W. J. Quax, J. M. van Dijl, and P. G. Braun. 2002. 'Selective contribution

of the twin-arginine translocation pathway to protein secretion in *Bacillus subtilis*', *J Biol Chem*, 277: 44068-78.

- Jørgensen, T. K., L. H. Bagger, J. Christiansen, G. H. Johnsen, J. R. Faarbæk, L. Jørgensen, and B. S. Welinder. 1998. 'Quantifying biosynthetic human growth hormone in *Escherichia coli* with capillary electrophoresis under hydrophobic conditions', *J Chromatog A*, 817: 205-14.
- Jormakka, M., S. Tornroth, B. Byrne, and S. Iwata. 2002a. 'Molecular basis of proton motive force generation: structure of formate dehydrogenase-N', *Science*, 295: 1863-8.
- Joshi, M. V., S. G. Mann, H. Antelmann, D. A. Widdick, J. K. Fyans, G. Chandra, M. I. Hutchings, I. Toth, M. Hecker, R. Loria, and T. Palmer. 2010. 'The twin arginine protein transport pathway exports multiple virulence proteins in the plant pathogen *Streptomyces scabies*', *Mol Microbiol*, 77: 252-71.
- Juhas, M., J. R. Van Der Meer, M. Gaillard, R. M. Harding, D. W. Hood, and D. W. Crook. 2009. 'Genomic islands: tools of bacterial horizontal gene transfer and evolution', *FEMS Microbiol Rev*, 33: 376-93.
- Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.
- Kaderabkova, N., M. Bharathwaj, R. C. D. Furniss, D. Gonzalez, T. Palmer, and D. A. I. Mavridou. 2022. 'The biogenesis of beta-lactamase enzymes', *Microbiology*, 168: doi: 10.1099/mic.0.001217
- Karnholz, A., C. Hoefler, S. Odenbreit, W. Fischer, D. Hofreuter, and R. Haas. 2006. 'Functional and topological characterization of novel components of the *comB* DNA transformation competence system in *Helicobacter pylori*', *J Bacteriol*, 188: 882-93.
- Keller, R., J. de Keyser, A. J. Driessen, and T. Palmer. 2012a. 'Co-operation between different targeting pathways during integration of a membrane protein', *J Cell Biol*, 199: 303-15.
- Kharade, S. S., and M. J. McBride. 2014. '*Flavobacterium johnsoniae* chitinase ChiA is required for chitin utilization and is secreted by the type IX secretion system', *J Bacteriol* 196: 961-70.
- Kikuchi, Y., M. Date, H. Itaya, K. Matsui, and L. F. Wu. 2006. 'Functional analysis of the twin-arginine translocation pathway in *Corynebacterium glutamicum* ATCC 13869', *Appl Environ Microbiol*, 72: 7183-92.
- Kneuper, H., B. Maldonado, F. Jager, M. Krehenbrink, G. Buchanan, R. Keller, M. Muller, B. C. Berks, and T. Palmer. 2012. 'Molecular dissection of TatC defines

- critical regions essential for protein transport and a TatB-TatC contact site', *Mol Microbiol*, 85: 945-61.
- Korotkov, K. V., and M. Sandkvist. 2019. 'Architecture, function, and substrates of the type II secretion system', *EcoSal Plus*, 8.
- Kouwen, T. R., R. van der Ploeg, H. Antelmann, M. Hecker, G. Homuth, U. Mader, and J. M. van Dijl. 2009. 'Overflow of a hyper-produced secretory protein from the *Bacillus* Sec pathway into the Tat pathway for protein secretion as revealed by proteogenomics', *Proteomics*, 9: 1018-32.
- Kreutzenbeck, P., C. Kroger, F. Lausberg, N. Blaudeck, G. A. Sprenger, and R. Freudl. 2007. '*Escherichia coli* twin arginine (Tat) mutant translocases possessing relaxed signal peptide recognition specificities', *J Biol Chem*, 282: 7903-11.
- Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes', *J Mol Biol*, 305: 567-80.
- Kruger, R. G., B. Otvos, B. A. Frankel, M. Bentley, P. Dostal, and D. G. McCafferty. 2004. 'Analysis of the substrate specificity of the *Staphylococcus aureus* sortase transpeptidase SrtA', *Biochemistry*, 43: 1541-51.
- Kuhn, P., B. Weiche, L. Sturm, E. Sommer, F. Drepper, B. Warscheid, V. Sourjik, and H.-G. Koch. 2011. 'The bacterial SRP receptor, SecA and the ribosome use overlapping binding sites on the SecY translocon', *Traffic*, 12: 563-78.
- Kuzmanović, M., C. Fagorzi, A. Mengoni, F. Lassalle, and G. C. diCenzo. 2022. 'Taxonomy of *Rhizobiaceae* revisited: proposal of a new framework for genus delimitation', *International J Systemat Evolut Microbiol* 72: 005243.
- Laemmli, U. K. 1970. 'Cleavage of structural proteins during the assembly of the head of bacteriophage T4', *Nature*, 227: 680-5.
- Lara-Tejero, M., J. Kato, S. Wagner, X. Liu, and J. E. Galán. 2011. 'A sorting platform determines the order of protein secretion in bacterial type III systems', *Science*, 331: 1188-91.
- Lasica, A. M., M. Ksiazek, M. Madej, and J. Potempa. 2017. 'The type IX secretion system (T9SS): highlights and recent insights into its structure and function', *Front Cell Infect Microbiol*, 7: 215.
- Lausberg, F., S. Fleckenstein, P. Kreutzenbeck, J. Frobel, P. Rose, M. Muller, and R. Freudl. 2012. 'Genetic evidence for a tight cooperation of TatB and TatC during productive recognition of twin-arginine (Tat) signal peptides in *Escherichia coli*', *PLoS One*, 7: e39867.
- Lawson, M. V. 2003. *Finite automata* (CRC Press).
- Lazarus, O., T. W. Woolerton, A. Parkin, M. J. Lukey, E. Reisner, J. Seravalli, E. Pierce, S. W. Ragsdale, F. Sargent, and F. A. Armstrong. 2009. 'Water-gas shift reaction catalyzed by redox enzymes on conducting graphite platelets', *J Am Chem Soc*, 131: 14154-5.

- Leake, M. C., N. P. Greene, R. M. Godun, T. Granjon, G. Buchanan, S. Chen, R. M. Berry, T. Palmer, and B. C. Berks. 2008. 'Variable stoichiometry of the TatA component of the twin-arginine protein transport system observed by in vivo single-molecule imaging', *Proc Natl Acad Sci U S A*, 105: 15376-81.
- Leiman, P. G., and M. M. Shneider. 2012. 'Contractile tail machines of bacteriophages', *Adv Exp Med Biol*, 726: 93-114.
- Leung, H. 2010. 'Regular Languages and Finite Automata', *AMC*, 10: 12.
- Linhartová, I., L. Bumba, J. Mašín, M. Basler, R. Osíčka, J. Kamanová, K. Procházková, I. Adkins, J. Hejnová-Holubová, and L. Sadílková. 2010. 'RTX proteins: a highly diverse family secreted by a common mechanism', *FEMS Microbiol Rev*, 34: 1076-112.
- Llosa, M., C. Roy, and C. Dehio. 2009. 'Bacterial type IV secretion systems in human disease', *Mol Microbiol*, 73: 141-51.
- Lukjancenko, O., T. M. Wassenaar, and D. W. Ussery. 2010. 'Comparison of 61 sequenced *Escherichia coli* genomes', *Microb Ecol*, 60: 708-20.
- Lyman, L. R., E. D. Peng, and M. P. Schmitt. 2018. '*Corynebacterium diphtheriae* iron-regulated surface protein HbpA is involved in the utilization of the hemoglobin-haptoglobin complex as an iron source', *J Bacteriol*, 200: e00676-17.
- Lyman, L. R., E. D. Peng, and M. P. Schmitt. 2021. 'The *Corynebacterium diphtheriae* HbpA hemoglobin-binding protein contains a domain that is critical for hemoprotein binding, cellular localization, and function', *J Bacteriol*, 203: e00196-21.
- Marraffini, L. A., A. C. DeDent, and O. Schneewind. 2006. 'Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria', *Microbiol Mol Biol Rev*, 70: 192-221.
- Martens, E. C., K. Heungens, and H. Goodrich-Blair. 2003. 'Early colonization events in the mutualistic association between *Steinernema carpocapsae* nematodes and *Xenorhabdus nematophila* bacteria', *J Bacteriol*, 185: 3147-54.
- Martinson, J. N. V., and S. T. Walk. 2020. '*Escherichia coli* residency in the gut of healthy human adults', *EcoSal Plus*, 9.
- McBride, M. J. 2019. 'Bacteroidetes gliding motility and the type IX secretion system', *Microbiol Spectr*, 7: 363-74.
- McBride, M. J., and D. Nakane. 2015. '*Flavobacterium gliding* motility and the type IX secretion system', *Curr Opin Microbiol*, 28: 72-77.
- McCallum, M., L. L. Burrows, and P. L. Howell. 2019. 'The dynamic structures of the type IV pilus', *Microbiol Spectr*, 7: 7.2. 02.
- McCann, J. R., J. A. McDonough, M. S. Pavelka, and M. Braunstein. 2007. 'Beta-lactamase can function as a reporter of bacterial protein export during

- Mycobacterium tuberculosis* infection of host cells', *Microbiology*, 153: 3350-9.
- Mekasha, S., and D. Linke. 2021. 'Secretion systems in gram-negative bacterial fish pathogens', *Front Microbiology*, 12: 782673.
- Mendel, S., A. McCarthy, J. P. Barnett, R. T. Eijlander, A. Nenninger, O. P. Kuipers, and C. Robinson. 2008. 'The *Escherichia coli* TatABC system and a *Bacillus subtilis* TatAC-type system recognise three distinct targeting determinants in twin-arginine signal peptides', *J Mol Biol*, 375: 661-72.
- Mendler, K., H. Chen, D. H. Parks, B. Lobb, L. A. Hug, and A. C. Doxey. 2019. 'AnnoTree: visualization and exploration of a functionally annotated microbial tree of life', *Nucleic Acids Res*, 47: 4442-48.
- Menetret, J.-F., J. Schaletzky, W. M. Clemons, A. R. Osborne, S. S. Skånland, C. Denison, S. P. Gygi, D. S. Kirkpatrick, E. Park, and S. J. Ludtke. 2007. 'Ribosome binding of a single copy of the SecY complex: implications for protein translocation', *Mol Cell*, 28: 1083-92.
- Meuskens, I., A. Saragliadis, J. C. Leo, and D. Linke. 2019. 'Type V secretion systems: an overview of passenger domain functions', *Front Microbiol*, 10: 1163.
- Mitkov, R., H. L. An, and N. Karamanis. 2006. 'A computer-aided environment for generating multiple-choice test items', *Nat Lang Engineer*, 12: 177-94.
- Molik, S., I. Karnauchov, C. Weidlich, R. G. Herrmann, and R. B. Klosgen. 2001. 'The Rieske Fe/S protein of the cytochrome *b6/f* complex in chloroplasts: missing link in the evolution of protein transport pathways in chloroplasts?', *J Biol Chem*, 276: 42761-6.
- Mori, H., and K. Cline. 2002. 'A twin arginine signal peptide and the pH gradient trigger reversible assembly of the thylakoid [Delta]pH/Tat translocase', *J Cell Biol*, 157: 205-10.
- Mori, H., E. J. Summer, and K. Cline. 2001. 'Chloroplast TatC plays a direct role in thylakoid (Delta)pH-dependent protein transport', *FEBS Lett*, 501: 65-8.
- Mougous, J. D., M. E. Cuff, S. Raunser, A. Shen, M. Zhou, C. A. Gifford, A. L. Goodman, G. Joachimiak, C. L. Ordoñez, and S. Lory. 2006. 'A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus', *Science*, 312: 1526-30.
- Mould, R. M., and C. Robinson. 1991. 'A proton gradient is required for the transport of two luminal oxygen-evolving proteins across the thylakoid membrane', *J Biol Chem*, 266: 12189-93.
- Narita, Y., K. Sato, H. Yukitake, M. Shoji, D. Nakane, K. Nagano, F. Yoshimura, M. Naito, and K. Nakayama. 2014. 'Lack of a surface layer in *Tannerella forsythia* mutants deficient in the type IX secretion system', *Microbiology*, 160: 2295.
- Naskar, S., M. Hohl, M. Tassinari, and H. H. Low. 2021. 'The structure and mechanism of the bacterial type II secretion system', *Mol Microbiol*, 115: 412-24.

- Nenninger, A. A., L. S. Robinson, and S. J. Hultgren. 2009. 'Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF', *Proc Natl Acad Sci U S A*, 106: 900-05.
- Ninkuu, V., L. Zhang, J. Yan, Z. Fu, T. Yang, and H. Zeng. 2021. 'Biochemistry of terpenes and recent advances in plant protection', *Int J Mol Sci*, 22: 5710.
- Okuda, S., and H. Tokuda. 2009. 'Model of mouth-to-mouth transfer of bacterial lipoproteins through inner membrane LolC, periplasmic LolA, and outer membrane LolB', *Proc Natl Acad Sci U S A*, 106: 5877-82.
- Palmer, T., and B. C. Berks. 2012. 'The twin-arginine translocation (Tat) protein export pathway', *Nat Rev Microbiol*, 10: 483-96.
- Palmer, T., F. Sargent, and B. C. Berks. 2005. 'Export of complex cofactor-containing proteins by the bacterial Tat pathway', *Trends Microbiol*, 13: 175-80. 2010. 'The Tat Protein Export Pathway', *EcoSal Plus*, 4.
- Palmer, T., A. J. Finney, C. K. Saha, G. C. Atkinson, and F. Sargent. 2021. 'A holin/peptidoglycan hydrolase-dependent protein secretion system', *Mol Microbiol*, 115: 345-55.
- Palmer, T., and P. J. Stansfeld. 2020. 'Targeting of proteins to the twin-arginine translocation pathway', *Mol Microbiol*, 113: 861-71.
- Panahandeh, S., C. Maurer, M. Moser, M. P. DeLisa, and M. Muller. 2008. 'Following the path of a twin-arginine precursor along the TatABC translocase of *Escherichia coli*', *J Biol Chem*, 283: 33267-75.
- Passmore, I. J., J. M. Dow, F. Coll, J. Cuccui, T. Palmer, and B. W. Wren. 2020. 'Ferric citrate regulator FecR is translocated across the bacterial inner membrane via a unique twin-arginine transport-dependent mechanism', *J Bacteriol*, 202: e00541-19.
- Paulson, T. 2013. 'Epidemiology: a mortal foe', *Nature*, 502: S2-S3.
- Pett, W., and D. V. Lavrov. 2013. 'The twin-arginine subunit C in *Oscarella*: origin, evolution, and potential functional significance', *Integr Comp Biol*, 53: 495-502.
- Pickering, B. S., and I. J. Oresnik. 2010. 'The twin arginine transport system appears to be essential for viability in *Sinorhizobium meliloti*', *J Bacteriol*, 192: 5173-80.
- Pineau, C., N. Guschinskaya, X. Robert, P. Gouet, L. Ballut, and V. E. Shevchik. 2014. 'Substrate recognition by the bacterial type II secretion system: more than a simple interaction', *Mol Microbiol*, 94: 126-40.
- Pinske, C., S. Krüger, B. Soboh, C. Ihling, M. Kuhns, M. Braussemann, M. Jaroschinsky, C. Sauer, F. Sargent, and A. Sinz. 2011. 'Efficient electron transfer from hydrogen to benzyl viologen by the [NiFe]-hydrogenases of *Escherichia coli* is dependent on the coexpression of the iron-sulfur cluster-containing small subunit', *Arch Microbiol*, 193: 893-903.

- Potter, S. C., A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn. 2018. 'HMMER web server: 2018 update', *Nucleic Acids Res*, 46: W200-W04.
- Pradel, N, J Delmas, LF Wu, CL Santini, and R Bonnet. 2009. 'Sec-and Tat-dependent translocation of B-lactamases across the *Escherichia coli* inner membrane', *Antimicrobial agents and chemotherapy*, 53: 242-48.
- Procter, J. B., G. M. Carstairs, B. Soares, K. Mourao, T. C. Ofoegbu, D. Barton, L. Lui, A. Menard, N. Sherstnev, D. Roldan-Martinez, S. Duce, D. M. A. Martin, and G. J. Barton. 2021. 'Alignment of Biological Sequences with Jalview', *Methods Mol Biol*, 2231: 203-24.
- Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2007. 'NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res*, 35: D61-5.
- Pukatzki, S., A. T. Ma, A. T. Revel, D. Sturtevant, and J. J. Mekalanos. 2007. 'Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin', *Proc Natl Acad Sci U S A*, 104: 15508-13.
- Pukatzki, S., A. T. Ma, D. Sturtevant, B. Krastins, D. Sarracino, W. C. Nelson, J. F. Heidelberg, and J. J. Mekalanos. 2006. 'Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system', *Proc Natl Acad Sci U S A*, 103: 1528-33.
- Punginelli, C., B. Maldonado, S. Grahl, R. Jack, M. Alami, J. Schroder, B. C. Berks, and T. Palmer. 2007. 'Cysteine scanning mutagenesis and topological mapping of the *Escherichia coli* twin-arginine translocase TatC Component', *J Bacteriol*, 189: 5482-94.
- Ramasamy, S., R. Abrol, C. J. Suloway, and W. M. Clemons, Jr. 2013. 'The glove-like structure of the conserved membrane protein TatC provides insight into signal sequence recognition in twin-arginine translocation', *Structure*, 21: 777-88.
- Rawlings, N. D. 2020. 'Twenty-five years of nomenclature and classification of proteolytic enzymes', *Biochim Biophys Acta Prot Proteom*, 1868: 140345.
- Reddy, G. K., N. G. H. Leferink, M. Umemura, S. T. Ahmed, R. Breitling, N. S. Scrutton, and E. Takano. 2020. 'Exploring novel bacterial terpene synthases', *PLoS One*, 15: e0232220.
- Renshaw, P. S., K. L. Lightbody, V. Veverka, F. W. Muskett, G. Kelly, T. A. Frenkiel, S. V. Gordon, R. G. Hewinson, B. Burke, and J. Norman. 2005. 'Structure and function of the complex formed by the tuberculosis virulence factors CFP-10 and ESAT-6', *EMBO J*, 24: 2491-98.
- Reynolds, M. M., L. Bogomolnaya, J. Guo, L. Aldrich, D. Bokhari, C. A. Santiviago, M. McClelland, and H. Andrews-Polymenis. 2011. 'Abrogation of the twin arginine transport system in *Salmonella enterica* serovar Typhimurium leads to colonization defects during infection', *PLoS One*, 6: e15800.
- Richter, S., and T. Bruser. 2005. 'Targeting of unfolded PhoA to the TAT translocon of *Escherichia coli*', *J Biol Chem*, 280: 42723-30.

- Richter, S., U. Lindenstrauß, C. Lucke, R. Bayliss, and T. Bruser. 2007. 'Functional Tat transport of unstructured, small, hydrophilic proteins', *J Biol Chem*, 282: 33257-64.
- Riquelme, E., S. Omarova, B. Ize, and D. O'callaghan. 2023. 'Analysis of the *Brucella suis* Twin Arginine Translocation System and Its Substrates Shows That It Is Essential for Viability', *Infect Immun*, 91: e00459-22.
- Rivera-Calzada, A., N. Famelis, O. Llorca, and S. Geibel. 2021. 'Type VII secretion systems: structure, functions and transport models', *Nat Rev Microbiol*, 19: 567-84.
- Rodrigue, A., A. Chanal, K. Beck, M. Muller, and L. F. Wu. 1999. 'Co-translocation of a periplasmic enzyme complex by a hitchhiker mechanism through the bacterial tat pathway', *J Biol Chem*, 274: 13223-8.
- Rodriguez, F., S. L. Rouse, C. E. Tait, J. Harmer, A. De Riso, C. R. Timmel, M. S. Sansom, B. C. Berks, and J. R. Schnell. 2013. 'Structural model for the protein-translocating element of the twin-arginine transport system', *Proc Natl Acad Sci U S A*, 110: E1092-101.
- Rollauer, S. E., M. J. Tarry, J. E. Graham, M. Jaaskelainen, F. Jager, S. Johnson, M. Krehenbrink, S. M. Liu, M. J. Lukey, J. Marcoux, M. A. McDowell, F. Rodriguez, P. Roversi, P. J. Stansfeld, C. V. Robinson, M. S. Sansom, T. Palmer, M. Høgbom, B. C. Berks, and S. M. Lea. 2012. 'Structure of the TatC core of the twin-arginine protein transport system', *Nature*, 492: 210-4.
- Rose, R. W., T. Bruser, J. C. Kissinger, and M. Pohlschroder. 2002. 'Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway', *Mol Microbiol*, 45: 943-50.
- Saint-Joanis, B., C. Demangel, M. Jackson, P. Brodin, L. Marsollier, H. Boshoff, and S. T. Cole. 2006. 'Inactivation of Rv2525c, a substrate of the twin arginine translocation (Tat) system of *Mycobacterium tuberculosis*, increases beta-lactam susceptibility and virulence', *J Bacteriol*, 188: 6669-79.
- Salacha, R., F. Kovačić, C. Brochier-Armanet, S. Wilhelm, J. Tommassen, A. Filloux, R. Voulhoux, and S. Bleves. 2010. 'The *Pseudomonas aeruginosa* patatin-like protein PlpD is the archetype of a novel Type V secretion system', *Env Microbiol*, 12: 1498-512.
- Sambrook, J., and D. W. Russell. 2001. *Molecular cloning : a laboratory manual* (Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y.).
- Samuelson, J. C., M. Chen, F. Jiang, I. Møller, M. Wiedmann, A. Kuhn, G. J. Phillips, and R. E. Dalbey. 2000. 'YidC mediates membrane protein insertion in bacteria', *Nature*, 406: 637-41.
- Santin, Y. G., T. Doan, L. Journet, and E. Cascales. 2019. 'Cell width dictates type VI secretion tail length', *Curr Biol* 29: 3707-13. e3.
- Sargent, F. 2007. 'Constructing the wonders of the bacterial world: biosynthesis of complex enzymes', *Microbiology*, 153: 633-51.

- Sargent, F., S. P. Ballantine, P. A. Rugman, T. Palmer, and D. H. Boxer. 1998. 'Reassignment of the gene encoding the *Escherichia coli* hydrogenase 2 small subunit--identification of a soluble precursor of the small subunit in a hypB mutant', *Eur J Biochem*, 255: 746-54.
- Sargent, F., E. G. Bogsch, N. R. Stanley, M. Wexler, C. Robinson, B. C. Berks, and T. Palmer. 1998. 'Overlapping functions of components of a bacterial Sec-independent protein export pathway', *EMBO J*, 17: 3640-50.
- Sargent, F., N. R. Stanley, B. C. Berks, and T. Palmer. 1999. 'Sec-independent protein translocation in *Escherichia coli*. A distinct and pivotal role for the TatB protein', *J Biol Chem*, 274: 36073-82.
- Sato, K., M. Naito, H. Yukitake, H. Hirakawa, M. Shoji, M. J. McBride, R. G. Rhodes, and K. Nakayama. 2010. 'A protein secretion system linked to bacteroidete gliding motility and pathogenesis', *Proc Natl Acad Sci U S A*, 107: 276-81.
- Sato, K., H. Yukitake, Y. Narita, M. Shoji, M. Naito, and K. Nakayama. 2013. 'Identification of *Porphyromonas gingivalis* proteins secreted by the Por secretion system', *FEMS Microbiol Letts* 338: 68-76.
- Schaerlaekens, K., M. Schierova, E. Lammertyn, N. Geukens, J. Anne, and L. Van Mellaert. 2001. 'Twin-arginine translocation pathway in *Streptomyces lividans*', *J Bacteriol*, 183: 6727-32.
- Schaerlaekens, K., L. Van Mellaert, E. Lammertyn, N. Geukens, and J. Anne. 2004. 'The importance of the Tat-dependent protein secretion pathway in *Streptomyces* as revealed by phenotypic changes in *tat* deletion mutants and genome analysis', *Microbiology*, 150: 21-31.
- Schmohl, L., and D. Schwarzer. 2014. 'Sortase-mediated ligations for the site-specific modification of proteins', *Curr Opin Chem Biol*, 22: 122-28.
- Schulze, R. J., J. Komar, M. Botte, W. J. Allen, S. Whitehouse, V. A. Gold, A. Nijeholt J. A. Lycklama, K. Huard, I. Berger, C. Schaffitzel, and I. Collinson. 2014. 'Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC', *Proc Natl Acad Sci U S A*, 111: 4844-9.
- Scotti, P. A., M. L. Urbanus, J. Brunner, J. W. de Gier, G. von Heijne, C. van der Does, A. J. Driessen, B. Oudega, and J. Luirink. 2000. 'YidC, the *Escherichia coli* homologue of mitochondrial Oxa1p, is a component of the Sec translocase', *EMBO J*, 19: 542-9.
- Seers, C. A., N. Slakeski, P. D. Veith, T. Nikolof, Y.-Y. Chen, S. G. Dashper, and E. C. Reynolds. 2006. 'The RgpB C-terminal domain has a role in attachment of RgpB to the outer membrane and belongs to a novel C-terminal-domain family found in *Porphyromonas gingivalis*', *J Bacteriol* 188: 6376-86.
- Severi, E., M. B. Batista, A. Lannoy, P. J. Stansfeld, and T. Palmer. 2023. 'Characterization of a TatA/TatB binding site on the TatC component of the *Escherichia coli* twin arginine translocase', *Microbiology*, 169: 001298.

- Shneider, M. M., S. A. Buth, B. T. Ho, M. Basler, J. J. Mekalanos, and P. G. Leiman. 2013. 'PAAR-repeat proteins sharpen and diversify the type VI secretion system spike', *Nature*, 500: 350-53.
- Shoji, M., K. Sato, H. Yukitake, Y. Kondo, Y. Narita, T. Kadowaki, M. Naito, and K. Nakayama. 2018. 'Por Secretion System-Dependent Secretion and Glycosylation of *Porphyromonas gingivalis* Hemin-Binding Protein 35', *PLoS One*, 13: e0203154.
- Simone, D., D. C. Bay, T. Leach, and R. J. Turner. 2013. 'Diversity and evolution of bacterial twin arginine translocase protein, TatC, reveals a protein secretion system that is evolving to fit its environmental niche', *PLoS One*, 8: e78742.
- Singhania, R. R., P. Dixit, A. K. Patel, B. S. Giri, C. H. Kuo, C.-W. Chen, and C. D. Dong. 2021. 'Role and significance of lytic polysaccharide monooxygenases (LPMOs) in lignocellulose deconstruction', *Biores Tech* 335: 125261.
- Smith, T. J., M. E. Font, C. M. Kelly, H. Sondermann, and G. A. O'Toole. 2018. 'An N-terminal retention module anchors the giant adhesin LapA of *Pseudomonas fluorescens* at the cell surface: a novel subfamily of type I secretion systems', *J Bacteriol*, 200: e00734-17.
- Snavely, E. A., M. Kokes, J. D. Dunn, H. A. Saka, B. D. Nguyen, R. J. Bastidas, D. G. McCafferty, and R. H. Valdivia. 2014. 'Reassessing the role of the secreted protease CPAF in *Chlamydia trachomatis* infection through genetic approaches', *Pathog Dis*, 71: 336-51.
- Spitz, O., I. N. Erenburg, T. Beer, K. Kanonenberg, I. B. Holland, and L. Schmitt. 2019. 'Type I secretion systems—one mechanism for all?', *Microbiol Spectr*, 7: 7.2. 12.
- Stanley, N. R., K. Findlay, B. C. Berks, and T. Palmer. 2001. '*Escherichia coli* strains blocked in Tat-dependent protein export exhibit pleiotropic defects in the cell envelope', *J Bacteriol*, 183: 139-44.
- Stanley, N. R., T. Palmer, and B. C. Berks. 2000. 'The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in *Escherichia coli*', *J Biol Chem*, 275: 11591-6.
- Stanley, N. R., F. Sargent, G. Buchanan, J. Shi, V. Stewart, T. Palmer, and B. C. Berks. 2002. 'Behaviour of topological marker proteins targeted to the Tat protein transport pathway', *Mol Microbiol*, 43: 1005-21.
- Stephenson, M., and L. H. Stickland. 1931. 'Hydrogenase: a bacterial enzyme activating molecular hydrogen: The properties of the enzyme', *Biochem J*, 25: 205-14.
- Strauch, E. M., and G. Georgiou. 2007. '*Escherichia coli* *tatC* mutations that suppress defective twin-arginine transporter signal peptides', *J Mol Biol*, 374: 283-91.
- Sun, C., G. Li, H. Li, Y. Lyu, S. Yu, and J. Zhou. 2021. 'Enhancing flavan-3-ol biosynthesis in *Saccharomyces cerevisiae*', *J of Agr Food Chem*, 69: 12763-72.

- Taj, M. K., Z. Samreen, J. X. Ling, I. Taj, T. M. Hassan, and W. Yunlin. 2014. 'Escherichia coli as a model organism', *Int J Eng Res Sci Techn*, 3: 1-8.
- Thomas, J. R., and A. Bolhuis. 2006. 'The tatC gene cluster is essential for viability in halophilic archaea', *FEMS Microbiol Lett*, 256: 44-9.
- Thompson, B. J., D. A. Widdick, M. G. Hicks, G. Chandra, I. C. Sutcliffe, T. Palmer, and M. I. Hutchings. 2010. 'Investigating lipoprotein biogenesis and function in the model Gram-positive bacterium *Streptomyces coelicolor*', *Mol Microbiol*, 77: 943-57.
- Tjalsma, H., A. Bolhuis, J. D. Jongbloed, S. Bron, and J. M. van Dijl. 2000. 'Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome', *Microbiol Mol Biol Rev*, 64: 515-47.
- Ton-That, H., S. K. Mazmanian, and O. Schneewind. 2001. 'An embarrassment of sortases-a richness of substrates? Response', *Trends Microbiol*, 9: 101-02.
- Tooke, F. J., M. Babot, G. Chandra, G. Buchanan, and T. Palmer. 2017. 'A unifying mechanism for the biogenesis of membrane proteins co-operatively integrated by the Sec and Tat pathways', *Elife*, 6: e26577.
- Totter, S., K. J. Waldron, S. J. Firbank, B. Reale, C. Bessant, K. Sato, T. R. Cheek, J. Gray, M. J. Banfield, C. Dennison, and N. J. Robinson. 2008. 'Protein-folding location can regulate manganese-binding versus copper- or zinc-binding', *Nature*, 455: 1138-42.
- Tseng, T.-T., B. M. Tyler, and J. C. Setubal. 2009. 'Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology', *BMC Microbiol*, 9: 1-9.
- Tsirigotaki, A., J. De Geyter, N. Šoštarić, A. Economou, and S. Karamanou. 2017. 'Protein export through the bacterial Sec pathway', *Nat Rev Microbiol*, 15: 21-36.
- Tsolis, K. C., E. P. Tsare, G. Orfanoudaki, T. Busche, K. Kanaki, R. Ramakrishnan, F. Rousseau, J. Schymkowitz, C. Ruckert, J. Kalinowski, J. Anne, S. Karamanou, M. I. Klapa, and A. Economou. 2018. 'Comprehensive subcellular topologies of polypeptides in *Streptomyces*', *Microb Cell Fact*, 17: 43.
- Tullman-Ercek, D., M. P. DeLisa, Y. Kawarasaki, P. Iranpour, B. Ribnicky, T. Palmer, and G. Georgiou. 2007. 'Export pathway selectivity of *Escherichia coli* twin arginine translocation signal peptides', *J Biol Chem*, 282: 8309-16.
- Ulfig, A., J. Frobel, F. Lausberg, A. S. Blummel, A. K. Heide, M. Muller, and R. Freudl. 2017. 'The h-region of twin arginine signal peptides supports productive binding of bacterial Tat precursor proteins to the TatBC receptor complex', *J Biol Chem*, 292: 10865-82.
- Ulhuq, F. R., M. C. Gomes, G. M. Duggan, M. Guo, C. Mendonca, G. Buchanan, J. D. Chalmers, Z. Cao, H. Kneuper, S. Murdoch, S. Thomson, H. Strahl, M. Trost, S. Mostowy, and T. Palmer. 2020. 'A membrane-depolarizing toxin substrate

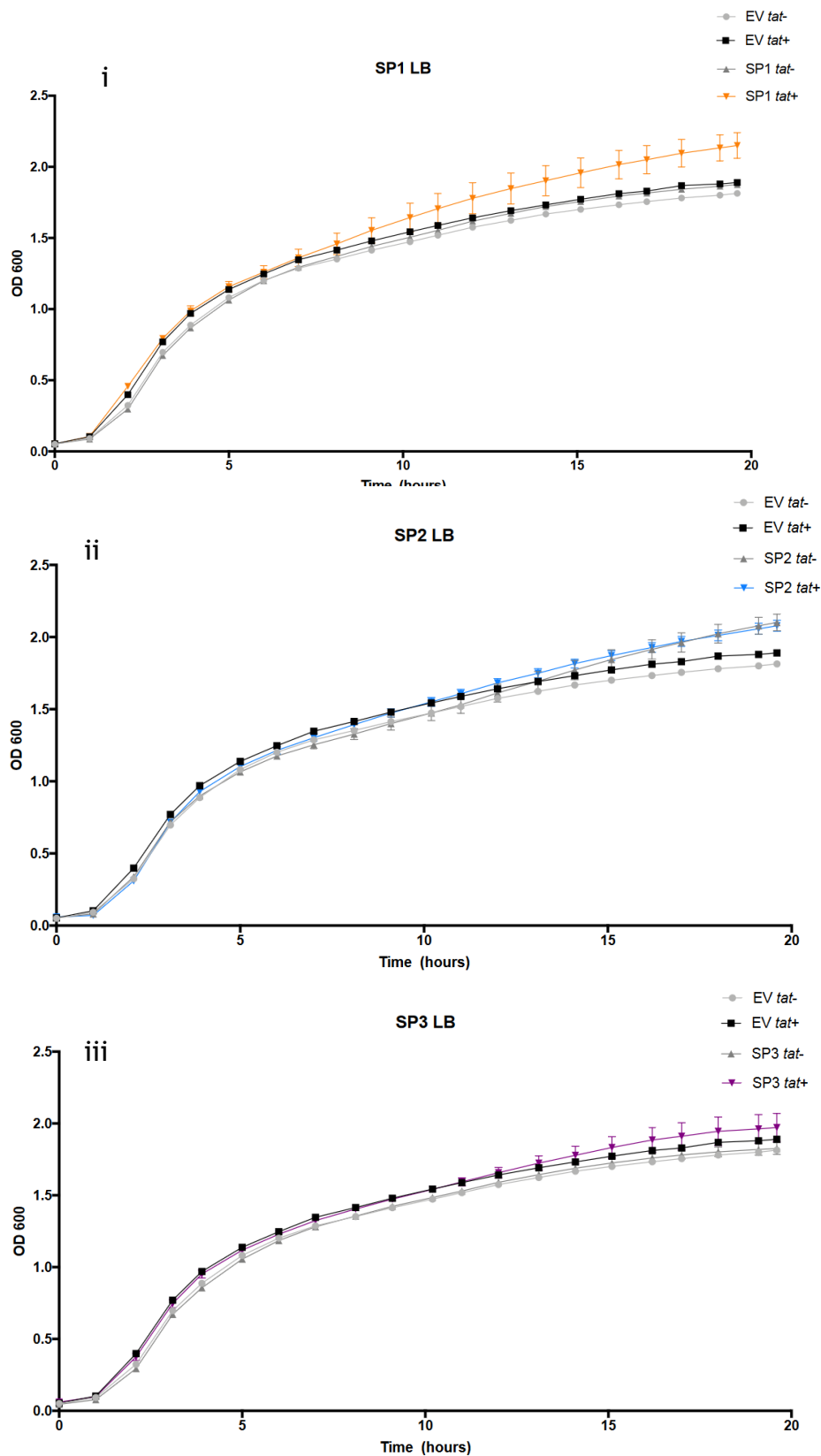
- of the *Staphylococcus aureus* type VII secretion system mediates intraspecies competition', *Proc Natl Acad Sci U S A*, 117: 20836-47.
- Unnikrishnan, M., C. Constantinidou, T. Palmer, and M. J. Pallen. 2017. 'The Enigmatic Esx Proteins: Looking Beyond Mycobacteria', *Trends Microbiol*, 25: 192-204.
- Urbanus, M. L., P. A. Scotti, L. Froderberg, A. Saaf, J. W. de Gier, J. Brunner, J. C. Samuelson, R. E. Dalbey, B. Oudega, and J. Luirink. 2001. 'Sec-dependent membrane protein insertion: sequential interaction of nascent FtsQ with SecY and YidC', *EMBO Rep*, 2: 524-9.
- van der Laan, M., N. P. Nouwen, and A. J. M. Driessen. 2005. 'YidC-an evolutionary conserved device for the assembly of energy-transducing membrane protein complexes', *Curr Opin Microbiol*, 8: 182-87.
- Van Rossum, G., and F.L. Drake Jr. 1995. *Python Tutorial*.
- Van Ulsen, P., L. Van Alphen, J. Ten Hove, F. Fransen, P. Van Der Ley, and J. Tommassen. 2003. 'A Neisserial autotransporter NalP modulating the processing of other autotransporters', *Mol Microbiol*, 50: 1017-30.
- Veith, P. D., M. D. Glew, D. G. Gorasia, and E. C. Reynolds. 2017. 'Type IX secretion: the generation of bacterial cell surface coatings involved in virulence, gliding motility and the degradation of complex biopolymers', *Mol Microbiol*, 106: 35-53.
- Veith, P. D., N. A. Muhammad, S. G. Dashper, V. A. Likic, D. G. Gorasia, D. Chen, S. J. Byrne, D. V. Catmull, and E. C. Reynolds. 2013. 'Protein substrates of a novel secretion system are numerous in the Bacteroidetes phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification, and cell-surface attachment', *J Proteome Res*, 12: 4449-61.
- Vignais, P. M., and B. Billoud. 2007. 'Occurrence, classification, and biological function of hydrogenases: an overview', *Chem Rev*, 107: 4206-72.
- Volbeda, A., C. Darnault, A. Parkin, F. Sargent, F. A. Armstrong, and J. C. Fontecilla-Camps. 2013. 'Crystal structure of the O₂-tolerant membrane-bound hydrogenase 1 from *Escherichia coli* in complex with its cognate cytochrome b', *Structure*, 21: 184-90.
- Von Heijne G. 1986. 'The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology' *EMBO J*, 5: 3021-7.
- Wang, J., M. Brackmann, D. Castano-Diez, M. Kudryashev, K. N. Goldie, T. Maier, H. Stahlberg, and M. Basler. 2017. 'Cryo-EM structure of the extended type VI secretion system sheath-tube complex', *Nat Microbiol*, 2: 1507-12.
- Weiner, J. H., P. T. Bilous, G. M. Shaw, S. P. Lubitz, L. Frost, G. H. Thomas, J. A. Cole, and R. J. Turner. 1998. 'A novel and ubiquitous system for membrane targeting and secretion of cofactor-containing proteins', *Cell*, 93: 93-101.

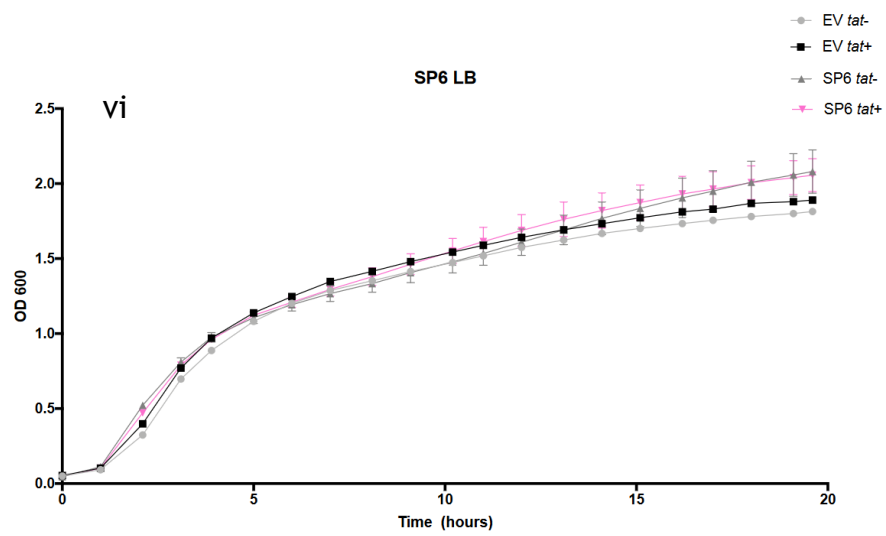
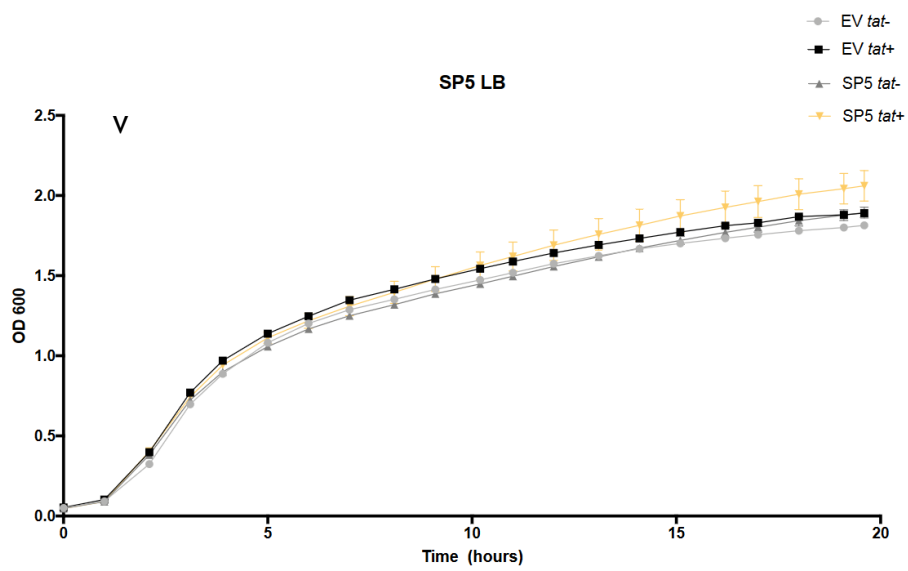
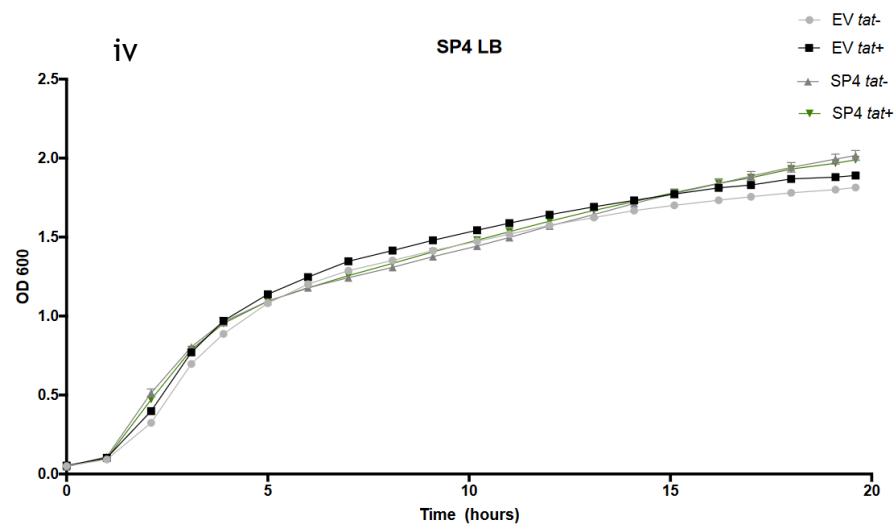
- Wexler, M., F. Sargent, R. L. Jack, N. R. Stanley, E. G. Bogsch, C. Robinson, B. C. Berks, and T. Palmer. 2000. 'TatD is a cytoplasmic protein with DNase activity. No requirement for TatD family proteins in sec-independent protein export', *J Biol Chem*, 275: 16717-22.
- Widdick, D. A., K. Dilks, G. Chandra, A. Bottrill, M. Naldrett, M. Pohlschroder, and T. Palmer. 2006. 'The twin-arginine translocation pathway is a major route of protein export in *Streptomyces coelicolor*', *Proc Natl Acad Sci U S A*, 103: 17927-32.
- Widdick, D. A., R. T. Eijlander, J. M. van Dijl, O. P. Kuipers, and T. Palmer. 2008. 'A facile reporter system for the experimental identification of twin-arginine translocation (Tat) signal peptides from all kingdoms of life', *J Mol Biol*, 375: 595-603.
- Wu, L. F., B. Ize, A. Chanal, Y. Quentin, and G. Fichant. 2000. 'Bacterial twin-arginine signal peptide-dependent protein translocation pathway: evolution and mechanism', *J Mol Microbiol Biotechnol*, 2: 179-89.
- Yahr, T. L., and W. T. Wickner. 2001. 'Functional reconstitution of bacterial Tat translocation in vitro', *Embo J*, 20: 2472-9.
- Yen, M. R., Y. H. Tseng, E. H. Nguyen, L. F. Wu, and M. H. Saier, Jr. 2002. 'Sequence and phylogenetic analyses of the twin-arginine targeting (Tat) protein export system', *Arch Microbiol*, 177: 441-50.
- Yoshida, N., E. M. Frickel, and S. Mostowy. 2017. 'Macrophage-Microbe Interactions: Lessons from the Zebrafish Model', *Front Immunol*, 8: 1703.
- Zhang, Y., L. Wang, Y. Hu, and C. Jin. 2014. 'Solution structure of the TatB component of the twin-arginine translocation system', *Biochim Biophys Acta*, 1838: 1881-8.
- Zhu, D., H. Xiong, J. Wu, C. Zheng, D. Lu, L. Zhang, and X. Xu. 2022. 'Protein targeting into the thylakoid membrane through different pathways', *Front Physiol* 12: 2396.
- Zhu, L., C. Klenner, A. Kuhn, and R. E. Dalbey. 2012. 'Both YidC and SecYEG are required for translocation of the periplasmic loops 1 and 2 of the multispinning membrane protein TatC', *J Mol Biol*, 424: 354-67.
- Zoufaly, S., J. Frobel, P. Rose, T. Flecken, C. Maurer, M. Moser, and M. Muller. 2012. 'Mapping precursor-binding site on TatC subunit of twin arginine-specific protein translocase by site-specific photo cross-linking', *J Biol Chem*, 287: 13430-41.

Appendices

Appendix A:

Control experiments in the absence of SDS to confirm that there is no general growth defect in the *tat* mutant strain





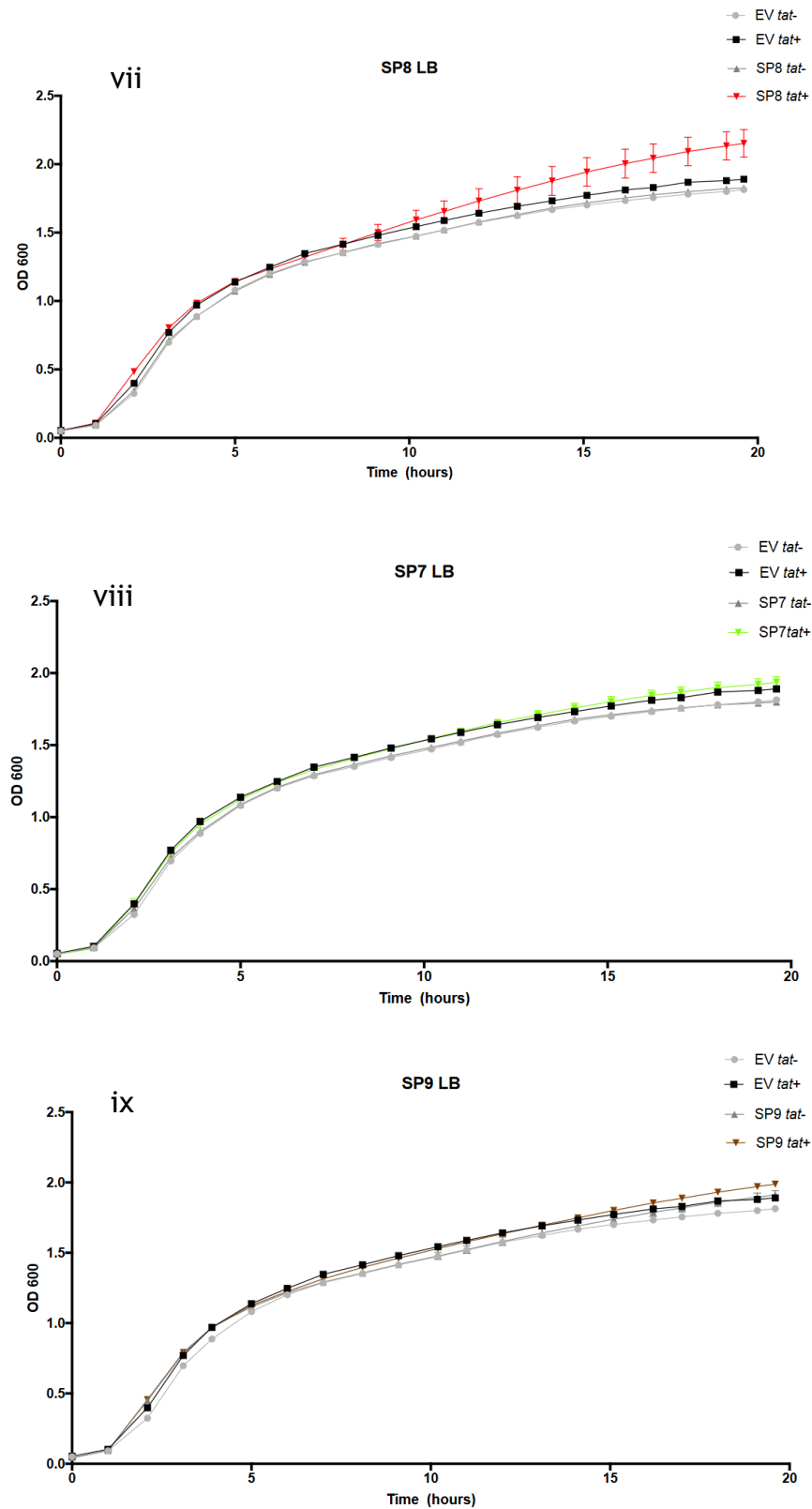


Figure Ap.1. Growth of strain MC4100 Δ amiA Δ amiC (*tat*+) or MC4100 Δ amiA Δ amiC Δ tatABC (*tat*-) harbouring either empty vector (pSUPROM; EV), or pSUPROM encoding the predicted signal peptide from the indicated candidate in LB. Cultures were grown in a plate reader at 37°C without shaking. Error bars are \pm standard deviations ($n = 3$).

Appendix B

Included in this appendix are all the scripts used in this thesis.

cala.py

This script provides an interactive way for the user to handle protein IDs from an HTML file called "chandra.html". The user is given two options: they can either view the protein ID in the console, or they can export a list of protein IDs to a CSV file. If the user opts to view the protein ID in the console, the script displays the filename, the protein count for "*Acidobacterium ailaui*", and all protein names. If the user opts to export the list to a CSV file, the script prompts the user for a filename, and then exports the list to the specified file. The file will contain one column, labeled "ID", which contains all the protein names.

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
```

This script allows the user to interactively choose between two options: displaying protein ID in the console or exporting protein IDs to a CSV file.

```
Created on Mon Jul 13 19:30:19 2020
__author__ = 'Jose Jesus Gallego-Parrilla'
__license__ = "GPL"
__maintainer__ = "Jose Jesus Gallego-Parrilla"
__email__ = "J.J.Gallego-Parrilla2@newcastle.ac.uk"
"""
```

```
from ProteinHtmlID import ProteinHtmlID
import pandas as pd
```

```
# Interactively ask the user to choose an option
while True:
    try:
        answer = int(input("Press 1 to see protein ID in console \nPress 2 to export protein
        CSV list \nChoose="))
        if answer == 1 or answer == 2:
            break
        except ValueError:
            pass
        print("Sorry, not what I was expecting \nTry again")

    if answer == 1:
        # If the user chooses option 1, display the protein ID in the console
```

```

protein = ProteinHtmlID("chandra.html") #name of the file you want check
name = protein.getFileName()
print(name)
count = protein.searchProtein("Acidobacterium ailaui")
print(count)
found = protein.findAllProteinNames()
print(found)
elif answer== 2:
# If the user chooses option 2, export the protein IDs to a CSV file
wp_num = ProteinHtmlID("chandra.html")
found = wp_num.findAllProteinNames()
wp_num = []
for elem in found:
wp_num.append(elem)
data = {'ID' : wp_num}
dataframe = pd.DataFrame(data)
input_console = input("Give the file a name that ends in .csv:\n")
dataframe.to_csv(input_console, index=False) #name for exported document
print(dataframe)

```

get: fasta L.py

This script fetches protein sequences from the NCBI database given a list of accession numbers. The accession numbers are read from a CSV file, whose name is passed as a command-line argument, and the fetched sequences are written to a text file, whose name is also passed as a command-line argument. If the necessary command-line arguments are not provided, the script prints a usage message and exits.

```
#!/usr/bin/env python
```

```
# coding: utf-8
```

```
__author__ = 'Jose Jesus Gallego-Parrilla'  
__license__ = "GPL"  
__maintainer__ = "Jose Jesus Gallego-Parrilla"  
__email__ = "J.J.Gallego-Parrilla2@newcastle.ac.uk"
```

```
# Import necessary modules.  
# sys module is used to work with command-line arguments.  
# csv module is used to read from and write to csv files.  
# Entrez from Bio is used to fetch data from NCBI databases.  
import sys  
import csv  
from Bio import Entrez
```

```
# Set the email address to be used by NCBI when you're accessing their  
database.  
Entrez.email = "your@email.com"
```

```
# Check if the correct number of command-line arguments has been  
provided.  
# If not, print a usage message and exit.  
if len(sys.argv) < 3:  
    print("It needs input and output arguments.")  
    print("Ej: python get_fasta_L.py efetch.csv efetch_output.txt")  
    sys.exit(0)
```

```
# The first command-line argument is the name of the csv file containing  
the accession numbers.  
csv_file = sys.argv[1]
```

```
# The second command-line argument is the name of the output text file  
where the fetched sequences will be written.  
txt_file = sys.argv[2]
```



```

# Initialize an empty list to store the accession numbers.
list_of_accession = []

# Open the csv file and read the accession numbers, appending them to
the list.
with open(csv_file, 'r', encoding='utf-8-sig') as csvfile:
    efetchin = csv.reader(csvfile, delimiter=',')
    for row in efetchin:
        list_of_accession.append(str(row[0]))

# Open the output text file in write mode.
with open(txt_file, mode='w') as efetch_output:

# Fetch the protein sequences in FASTA format from NCBI database.
input_handle = Entrez.efetch(db="protein", id=list_of_accession,
    rettype="fasta")

# Open the output text file in append mode.
output_handle = open(txt_file, "a")

# Write the fetched sequences to the output file.
for line in input_handle:
    output_handle.write(line)

# Close the handles.
input_handle.close()
output_handle.close()

print('program finished')

```

reptile2.py

This script reads in a tab-separated file and processes it to remove entries where the 'Protein Name' is 'hypothetical protein'. It then writes out the processed data to a new file. It also creates a second output file where only the first occurrence of each unique 'Protein' is kept.

```
#!/usr/bin/env python
# coding: utf-8
```

```
__author__ = 'Dr Giusy Mariano'
__email__ = 'giusy.mariano@ncl.ac.uk'
__license__ = "GPL"
__modification__ = 'Jose Jesus Gallego-Parrilla'
__email__ = 'J.J.Gallego-Parrilla2@newcastle.ac.uk'
__license__ = "GPL"
```

```
# Import necessary modules.
# sys module is used to work with command-line arguments.
# pandas is used for handling data in structured format (dataframes).
import sys
import pandas as pd
```

```
# The first command-line argument is the name of the input file.
file_name = sys.argv[1]
```

```
# The second command-line argument is the base name of the output
files.
file_name_output = sys.argv[2]
```

```
# Read the input file into a pandas dataframe.
# low_memory=False is used to eliminate a warning that can occur when
inferring data types.
# names provides column names for the dataframe.
df = pd.read_csv(file_name, sep="\t", low_memory=False,
names=['ID', 'Source', 'Nucleotide Accession', 'Protein', 'Protein Name',
'Start',
'Stop', 'Strand', 'Organism', 'Strain', 'Assembly'])
```

```
# Get names of indexes for which rows have to be dropped
# Rows where the 'Protein Name' is 'hypothetical protein' will be dropped.
indexNames = df[df['Protein Name'] == 'hypothetical protein'].index
```

```
# Delete these row indexes from dataframe
df.drop(indexNames, inplace=True)
```

```
# Write out the dataframe to a tsv file.
df.to_csv(file_name_output + ".tsv", sep="\t", index=False)

# Create a new dataframe where only one representative for each
'Protein' is kept.
# Rows are sorted by 'Protein', and for each unique 'Protein', only the first
occurrence is kept.
df2 = df.sort_values(by="Protein", axis=0, ascending=True,
inplace=False).drop_duplicates(subset=['Protein'],keep='first')

# Write out this new dataframe to a tsv file.
df2.to_csv(file_name_output + "_one_WP_per_assembly.tsv", sep="\t",
index=False)

print ('program finished')
```

Main_proteins.py

In this script, you're leveraging a class called ProteinHtml that presumably is defined in another Python file, ProteinHtml.py. This class seems to provide methods for extracting protein-related information from an HTML file, including the names of all unique proteins found in the file and the number of occurrences of a specific protein.

```
#!/usr/bin/env python3
```

```
# -*- coding: utf-8 -*-  
"""
```

```
Created by: Jorge Camarero Vera, Jose Jesus Gallego-Parrilla  
Maintained by: Jose Jesus Gallego-Parrilla  
Email: J.J.Gallego-Parrilla2@newcastle.ac.uk  
License: GPL
```

```
This script uses the ProteinHtml class to extract specific protein information from a  
provided HTML file.  
"""
```

```
# Import the ProteinHtml class from the ProteinHtml.py module  
from ProteinHtml import ProteinHtml
```

```
# This is the main entry point of the script. It checks whether this script is being run  
directly
```

```
# (as opposed to being imported as a module), in which case it executes the code  
within this block.
```

```
if __name__ == "__main__":
```

```
# Create an instance of the ProteinHtml class, initializing it with the name of an HTML  
file
```

```
protein = ProteinHtml("chandra.html") #name of the file you want check
```

```
# Call the getFileName method of the protein object, which returns the name of the  
file it was initialized with,
```

```
# and print this file name to the console
```

```
name = protein.getFileName()
```

```
print(name)
```

```
# Call the searchProtein method of the protein object, which returns the number of  
occurrences of a specified
```

```
# protein in the HTML file, and print this count to the console
```

```
count = protein.searchProtein("Acidobacterium ailaui")
```

```
print(count)
```

```
# Call the findAllProteinNames method of the protein object, which returns a list of all  
unique protein names found  
# in the HTML file, and print this list to the console  
found = protein.findAllProteinNames()  
print(found)
```

ProteinHtml.py

The ProteinHtml class contains three methods. The getFileName method returns the name of the file that was provided when the class instance was created. The searchProtein method returns the number of occurrences of a specific protein in the HTML file. The findAllProteinNames method returns a dictionary where the keys are the names of all unique proteins found in the HTML file and the values are the counts of their occurrences

```
#!/usr/bin/env python3
```

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Fri May 1 12:12:05 2020
```

```
This module contains the ProteinHtml class, which is used to analyze protein  
information in an HTML file.
```

```
"""
```

```
# Import the re module for regular expression operations  
import re
```

```
class ProteinHtml:
```

```
# Private class variables to hold the name of the file and its content
```

```
__fileName = None
```

```
__contentFile = None
```

```
# Constructor method that initializes a new instance of the class
```

```
def __init__(self, fileName):
```

```
self.__fileName = fileName # Assign the provided file name to the __fileName variable
```

```
with open(fileName) as f: # Open the file and assign its content to the __contentFile  
variable
```

```
self.__contentFile = f.read()
```

```
# Method to get the name of the file
```

```
def getFileName(self):
```

```
return self.__fileName
```

```
# Method to count the number of occurrences of a specific protein in the file
```

```
def searchProtein(self, proteinName):
```

```
# Construct a regular expression that matches the specified protein in the HTML  
structure
```

```

regexBegin = "<tr style = \"background:#[\\w\\d]+\"><td>WP_[\\d]+\\.1</td><td
colspan = 5>(.\""
regexEnd = ".*)</td></tr>"
regex = regexBegin + proteinName + regexEnd

# Count the number of matches of the regex in the content of the file
count = sum(1 for match in re.finditer(r"{}".format(regex),
self.__contentFile))
return count

# Method to find all unique protein names in the file
def findAllProteinNames(self):
allProteins = {} # Dictionary to hold each protein name and its count of occurrences

# Construct a regular expression that matches protein names in the HTML structure
regex = "<tr style = \"background:#[\\w\\d]+\"><td>WP_[\\d]+\\.1</td><td colspan =
5>.*\\[(.*)\\]</td></tr>"

# Iterate over all matches of the regex in the content of the file
for match in re.finditer(r"{}".format(regex), self.__contentFile):
protein = match.group(1) # Extract the protein name from the match

# If the protein is already in the dictionary, increment its count;
# otherwise, add it to the dictionary with a count of 1
if (protein in allProteins):
allProteins[protein] += 1
else:
allProteins.update({protein: 1})

return allProteins

```

Protein_sorter.py

This script creates an instance of the ProteinHtml class with a specific HTML file, uses the findAllProteinNames method to get a dictionary of all unique protein names and their counts of occurrences in the file, and writes this information into a CSV file. The script also prints the DataFrame that contains the protein names and their counts. The resulting CSV file can be used for further analysis of the protein data.

```
#!/usr/bin/env python3
```

```
# -*- coding: utf-8 -*-  
"""
```

This script uses the ProteinHtml class to find all unique protein names in an HTML file and write them into a CSV file along with their counts of occurrences.

```
__author__ = 'Jorge Camarero Vera, Jose Jesus Gallego-Parrilla'  
__license__ = "GPL"  
__maintainer__ = "Jose Jesus Gallego-Parrilla"  
__email__ = "J.J.Gallego-Parrilla2@newcastle.ac.uk"  
"""
```

```
# Import the pandas library for data manipulation  
import pandas as pd
```

```
# Import the ProteinHtml class from the ProteinHtml module  
from ProteinHtml import ProteinHtml
```

```
# Create a new instance of the ProteinHtml class with the "chandra.html" file  
protein = ProteinHtml("chandra.html")
```

```
# Find all unique protein names in the file  
found = protein.findAllProteinNames()
```

```
# Create two lists to hold the protein names and their counts of occurrences  
names = []  
places = []
```

```
# Iterate over all proteins found in the file  
for elem in found:  
    names.append(elem) # Append the protein name to the names list  
    places.append(found[elem]) # Append the count of occurrences to the places list
```

```
# Create a dictionary to hold the names and places lists  
data = {'Name' : names, 'Place' : places}
```



```
# Create a DataFrame from the dictionary
dataframe = pd.DataFrame(data)
df = pd.DataFrame(data)

# Write the DataFrame to a CSV file
df.to_csv('all_proteins.csv', index=False)

# Print the DataFrame
print(dataframe)
```

Incomplete protein_counter.py

This script reads the content of an HTML file and uses regular expressions to find and count occurrences of specific protein names or other protein-related terms. It then prints these counts to the console. The proteins and terms that are searched for are hardcoded into the script, and each term is searched for separately. The counts are not stored or exported, they are simply printed to the console. The script provides a quick and easy way to count occurrences of specific terms in a text file, but it does not offer much flexibility or functionality beyond that.

```
#!/usr/bin/env python3
```

```
# -*- coding: utf-8 -*-
```

```
"""
```

This script reads an HTML file and counts the occurrences of specific proteins or protein-related terms.

```
__author__ = 'Jorge Camarero Vera, Jose Jesus Gallego-Parrilla'
```

```
__license__ = "GPL"
```

```
__maintainer__ = "Jose Jesus Gallego-Parrilla"
```

```
__email__ = "J.J.Gallego-Parrilla2@newcastle.ac.uk"
```

```
"""
```

```
# Define a function to count the occurrences of a specific word in a file
```

```
def countWords(address, word):
```

```
# Open the file in read mode
```

```
logfile = open(address, "r")
```

```
# Initialize the word count to 0
```

```
wordcount = 0
```

```
# Iterate over each line in the file
```

```
for lines in logfile:
```

```
# If the word is in the line, increment the word count
```

```
if word in lines.split():
```

```
wordcount += 1
```

```
# Return the final word count
```

```
return wordcount
```

```
# Import the regular expressions (re) module
```

```
import re
```

```
# Open the "chandra.html" file and read its contents
```

```
with open("chandra.html") as f:
```

```
contents = f.read()
```

```
# For each protein or protein-related term, count its occurrences in the file and print
the count
# The \b word boundary is used to ensure whole words are matched
# Note: Some of the regular expressions include specific formatting or additional
words for more precise matching
count = sum(1 for match in re.finditer(r"\bhydrogenase", contents))
print("Hydrogenases:", count)
# Repeat this process for each protein or protein-related term of interest
# ...
```

Arcana.py

This script performs a search for Hidden Markov Models (HMMs) on protein sequences and generates a CSV file with the identified hits. It is mainly composed of two parts. In the first part, it uses HMMER to conduct the search, and then it parses the results, selecting only the hits that are within a specified length range defined by the user. The second part of the script uses NCBI's Entrez service to fetch the protein information from the identified hits, and creates a final table with the proteins' details. This table can be used as an input for the FLAG software. The script also cleans and organizes the data by removing duplicates, sorting the entries, and extracting relevant information to create the final output files.

```
__author__ = 'Dr Giusy Mariano'
__email__ = 'giusy.mariano@ncl.ac.uk'
__license__ = "GPL"
#!/usr/bin/env python
# coding: utf-8
import sys
import subprocess as sp
import csv
from Bio import SearchIO
import pandas as pd
from Bio import Entrez

print ('Can go for cup of coffee')
print ('Come back later to input min and max lenght for your parse...')

#initialize search function
def hmm_search(hmm_directory, protein_directory):

hmmsearch = "hmmsearch" + " " + "-T 30" + " " + "--incT 30" + " " + "-o"
log" + " --domtblout" + " " + "HMM_output.txt" + " " + hmm_directory + "
" + protein_directory

sp.run(hmmsearch, shell=True)
if __name__ == '__main__':
hmm_search(sys.argv[1],sys.argv[2]) #makes it take 2 args from
command line
x = int(input('min lenght'))
y = int(input('max lenght'))

with open('HMM_output.txt', newline='') as input:
for qresult in SearchIO.parse(input, 'hmmscan3-domtab'):
query_id = qresult.id
```

```

query_len = qresult.seq_len
hits = qresult.hits
num_hits = len(hits)
hits_len = qresult.seq_len
if num_hits > 0 :
    with open('parsed_output.csv', mode='w') as parsed_output: #write table
    for Suppl.data with hitsID, scores and lenght
    parsed_output = csv.writer(parsed_output, delimiter=',', quotechar='"',
    quoting=csv.QUOTE_MINIMAL)
    parsed_output.writerow(['hit_accession', 'hit_description', 'E-value', 'Bit
    Score', 'Hits_Lenght'])
    for i in range(0,num_hits):
    hit_evalue = hits[i].evalue
    hit_bit_Score = hits[i].bitscore
    hit_accession= hits[i].id
    hit_length = hits[i].seq_len
    hit_description = hits[i].description
    if hit_length > x and hit_length < y:
    with open('parsed_output.csv', mode='a') as parsed_output: #write table
    for Suppl.data with hitsID, scores and lenght. Need mode 'a' to append to
    existing table
    parsed_output = csv.writer(parsed_output, delimiter=',', quotechar='"',
    quoting=csv.QUOTE_MINIMAL)
    parsed_output.writerow([hit_accession, hit_description,
    hit_evalue, hit_bit_Score, hit_length])
    with open('efetch_input.csv', mode='a') as efetch_input: #write table with
    Hits IDs for FLAGS
    efetch_input = csv.writer(efetch_input, delimiter=',', quotechar='"',
    quoting=csv.QUOTE_MINIMAL)
    efetch_input.writerow([hit_accession])
    print ('HMMer_parse_complete')
    print ('efetch_start')
    # get from WPs accession, corresponding assembly, NC IDs, strains
    names.
    # Write a csv table with all these as final data tablee,
    # + a table with WPs and Assembly IDs for inputting in FLAG

Entrez.email = "your@gmail.com"

# get from WPs accession, corresponding assembly, NC IDs, strains
names. Write a csv table with all these as final data tablee,
# + a table with WPs and Assembly IDs for inputting in FLAG

list_of_accession = []
with open (sys.argv[1], 'r') as csvfile:
    efetchin=csv.reader(csvfile, delimiter = ',')
    for row in efetchin:
    list_of_accession.append(str(row[0]))
    with open('efetch_output.tsv', mode = 'w') as efetch_output:

```

```

efetch_output = csv.writer(efetch_output, delimiter='\t', quotechar='',
quoting=csv.QUOTE_MINIMAL)
efetch_output.writerow(['ID', 'Source', 'Nucleotide Accession', 'Start',
'Stop', 'Strand', 'Protein', 'Protein Name', 'Organism', 'Strain',
'Assembly'])
input_handle = Entrez.efetch(db="protein", id= list_of_accession,
rettype="ipg", retmode="tsv")
output_handle = open("efetch_output.tsv", "wb")
for line in input_handle:
output_handle.write(line)
output_handle.close()
input_handle.close()
#process file in pandas
file_name = "efetch_output.tsv"
file_name_output = "final_output.tsv"
df = pd.read_csv(file_name, sep="\t", low_memory=False)
# Get names of indexes for which rows have to be dropped
indexNames = df[ df['Source'] == 'INSDC'].index
# Delete these row indexes from dataFrame
df.drop(indexNames , inplace=True)
#rearrange table columns
df = df[['ID', 'Source', 'Nucleotide Accession', 'Protein', 'Protein Name',
'Start', 'Stop', 'Strand', 'Organism', 'Strain', 'Assembly']]
#Sort table on Assembly number ignoring GCF_
df['sort'] = df['Assembly'].str.extract('(\d+)', expand=False).astype(float)
df.sort_values('sort',inplace=True, ascending=True)
df = df.drop('sort', axis=1)
#drop all duplicates that're similar in indicated subset fields
df3=df.drop_duplicates(subset=['Start', 'Stop', 'Strand', 'Organism',
Strain', 'Assembly'],keep='first')
#sorts dataframe alphabetically by Organism and writes to csv
df3.sort_values(by = "Organism", axis=0, ascending=True,
inplace=False).to_csv("final_parsed_output.tsv", "\t", index=False)
#get WP_X and GFC_X IDs in a tsv to input in FLAGS
new_dataframe1 = df3[['Assembly', 'Protein']]
new_dataframe2 = df3[['Organism', 'Strain', 'Assembly', 'Protein']]
new_dataframe1.sort_values(by = "Protein", axis=0, ascending=True,
inplace=False).to_csv('flags_input.tsv', '\t', header=False, columns =
['Assembly', 'Protein'])
new_dataframe2.sort_values(by = "Organism", axis=0, ascending=True,
inplace=False).to_csv('flags_input_wstrains.tsv', '\t', header=False,
columns = ['Organism', 'Strain', 'Assembly', 'Protein'])
print ('program finished')

```

Etna2.py

This script is a standalone version of the second part of the Arcana.py script. It fetches protein information from a list of accessions (provided as an input file) using NCBI's Entrez service, and it outputs a final table with the details of the proteins. Just like in Arcana.py, this table can be used as an input for the FLAG software. It also performs the same cleaning and organizing steps as Arcana.py, including removing duplicates, sorting the entries, and extracting relevant information to create the final output files.

```
__author__ = 'Dr Giusy Mariano'
__email__ = 'giusy.mariano@ncl.ac.uk'
__license__ = "GPL"
__modification__ = "Jose Jesus Gallego-Parrilla"
__email__ = "J.J.Gallego-Parrilla2@newcastle.ac.uk"
# coding: utf-8

# In[ ]:

import sys
import csv
import pandas as pd
from Bio import Entrez
Entrez.email = "your@gmail.com"

# get from WPs accession, corresponding assembly, NC IDs, strains
names. Write a csv table with all these as final data tablee,
#+ a table with WPs and Assembly IDs for inputting in FLAG

list_of_accession = []
with open(sys.argv[1], 'r') as csvfile:
    efetchin=csv.reader(csvfile, delimiter = ',')
    for row in efetchin:
        list_of_accession.append(str(row[0]))
    with open('efetch_output.tsv', mode = 'w') as efetch_output:
        efetch_output = csv.writer(efetch_output, delimiter='\t', quotechar='"',
        quoting=csv.QUOTE_MINIMAL)
        efetch_output.writerow(['ID', 'Source', 'Nucleotide Accession', 'Start',
        'Stop', 'Strand', 'Protein', 'Protein Name', 'Organism', ' Strain',
        'Assembly'])
    input_handle = Entrez.efetch(db="protein", id= list_of_accession,
    rettype="ipg", retmode="tsv")
    output_handle = open("efetch_output.tsv", "wb")
    for line in input_handle:
        output_handle.write(line)
```

```

output_handle.close()
input_handle.close()
#process file in pandas
file_name = "efetch_output.tsv"
file_name_output = "final_output.tsv"
df = pd.read_csv(file_name, sep="\t", low_memory=False)
# Get names of indexes for which rows have to be dropped
indexNames = df[ df['Source'] == 'INSDC'].index
# Delete these row indexes from dataFrame
df.drop(indexNames , inplace=True)
#rearrange table columns
df = df[['ID', 'Source', 'Nucleotide Accession', 'Protein', 'Protein Name',
'Start', 'Stop', 'Strand', 'Organism', 'Strain', 'Assembly']]
#Sort table on Assembly number ignoring GCF_
df['sort'] = df['Assembly'].str.extract('(\d+)', expand=False).astype(float)
df.sort_values('sort',inplace=True, ascending=True)
df = df.drop('sort', axis=1)
#drop all duplicates that're similar in indicated subset fields
df3=df.drop_duplicates(subset=['Start', 'Stop', 'Strand', 'Organism', '
Strain', 'Assembly'],keep='first')
#sorts dataframe alphabetically by Organism and writes to csv
df3.sort_values(by = "Organism", axis=0, ascending=True,
inplace=False).to_csv("final_parsed_output.tsv", "\t", index=False)
#get WP_X and GFC_X IDs in a tsv to input in FLAGS
new_dataframe1 = df3[['Assembly', 'Protein']]
new_dataframe2 = df3[['Organism', 'Strain', 'Assembly', 'Protein']]
new_dataframe1.sort_values(by = "Protein", axis=0, ascending=True,
inplace=False).to_csv('flags_input.tsv', '\t', header=False, columns =
['Assembly', 'Protein'])
new_dataframe2.sort_values(by = "Organism", axis=0, ascending=True,
inplace=False).to_csv('flags_input_wstrains.tsv', '\t', header=False,
columns = ['Organism', 'Strain', 'Assembly', 'Protein'])
print ('program finished')

```