Molecular Mechanisms of Nutrient Uptake in Marine Microbial Ecosystems

Andrew George McLeish

Master of Research

Department of Chemistry and Biomolecular Sciences
Faculty of Science
Macquarie University
Sydney, NSW, Australia

24th of April 2017

Table of Contents

Abstract

Unicellular cyanobacteria are highly diverse, abundant and ubiquitous in the global oceans. Investigating the range and magnitude of organic compound transport in these organisms will help define pathways of energy and nutrient cycling in marine ecosystems.

ABC membrane transport systems utilise a periplasmic binding protein (PBP) to confer substrate specificity. These PBPs are thus ideal targets to study the ecological significance of predicted amino acid transporters. Twelve PBPs were investigated that group into three orthologous gene clusters with putative ligands: branched-chain amino acids; carbohydrates and amino acids; or amino acids, based on homology predictions.

A high-throughput recombinant expression strategy enabled a protein from each cluster to be successfully heterologously expressed and purified. Differential scanning fluorimetry (DSF) analysis revealed that "SW840" binds a broad range of amino acids, as predicted. "SC634" on the other hand was found to bind with formate and formaldehyde, not the predicted ligand of branched-chain amino acids.

This work provides important evidence of the specific ligand binding characteristics for each transporter. Future work to elucidate the cellular fate of these compounds is proposed. The ability of cyanobacteria, often considered to be obligatory phototrophs, to salvage organic compounds may provide a competitive advantage in marine ecosystems.

Acknowledgements

I would first like to thank my supervisor Dr. Bhumika Shah. The amount of work and effort you put forth for me is truly appreciated and I feel extremely lucky to have had you as my supervisor. Your constant encouragement, guidance, and ability to push me in the right direction really helped me. Thank you so much, you are a fantastic mentor and supervisor and I would not have been able to be where I am today without your help. I also want to extend my thanks to my co-supervisors Prof. Ian Paulsen and Dr. Martin Ostrowski. It is due to their continued support and guidance that I have been able to come this far and learn so much. I would not have been able to do so without any of their help, thank you all so much.

My thanks also goes out to A/Prof. Bridget Mabbutt and her group for the joint collaboration with the Paulsen group for this particular project. I would also like to thank the many members of the Paulsen Laboratory for their continued enthusiasm and support. I wish everyone the best and I hope to give as much support to you as you have shown to me.

Last but certainly not least, a huge thanks to my dearest friends and family, especially my father Robin, my mother Rita, my brother Angus, and my sister Fiona. Thank you all for supporting me all this time and I love you all so much. Thank you!

Abbreviations

3D – Three-dimensional

AA – Amino acid

AA-PBP – Amino acid periplasmic binding protein

ABC – ATP-binding cassette

Amp – Ampicillin

C – Carbon

CCM – CO_2 concentrating mechanism

Chl – Chlorampenicol

CV – Column volume

DMSP – Dimethylsulphoniopropionate

DOM – Dissolved organic matter

DON – Dissolved organic nitrogen

DSF – Differential scanning fluorimetry

EAC – Eastern Australian Current

GCNPR – Guanylate cyclase-atrial natriuretic peptide receptor

GPCR – G-protein coupled receptor

HAAT – Hydrophobic amino acid uptake transporter

HGT – Horizontal gene transfer

His-Tag - Hexahistidine tag

IMAC – Immobilised metal affinity chromatography

IPTG – Isopropyl-β-ᴅ-thio-galactoside

ITC – Isothermal titration calorimetry

Kav – Partition coefficient

LB – Lysogeny broth

LIC – Ligation independent cloning

LIV – Leucine-isoleucine-valine

MS – Mass spectroscopy

N – Nitrogen

NanoSIMS – Nanoscale secondary ion mass spectrometery

NPP – Net primary production

OBP – Oligopeptide-binding protein

OD – Optical density

ORF – Open reading frame

P – Phosphorus

PAAT – Polar amino acid uptake transporter

PBP – Periplasmic binding protein

PCR – Polymerase chain reaction

PDB – Protein Data Bank

PLP – Pyridoxal-5'-phosphate

RT-PCR – Real-time polymerase chain reaction

RuBP – Ribulose bisphosphate

RuMP – Ribulose monophosphate

S – Sulphur

SB – Silver bullets

SDS-PAGE – Sodium dodecyl sulphate-polyacrylamide gel electrophoresis

SPR – Surface plasmon resonance

TEV – Tobacco etch virus

T_m – Melting temperature

TMD – Transmembrane domain

TRAP – Tripartite ATP-independent periplasmic

TT – Tripartite tricarboxylate

V_e – Elution volume

V_o – Void volume

V_t – Total volume

Chapter 1. Introduction

1.1 Global importance and significance of marine microbes

The three main biogeochemical cycles, phosphorus (P), nitrogen (N), and carbon (C) are largely driven by marine microbes[1]. These microbes are intrinsically linked to biogeochemical processes and, in turn, provide many ecosystem services on which humankind depend[1,2]. Around 71% of the Earth's surface is covered by oceans. Within just one millilitre of surface seawater, there may be up to one million bacterial cells and 20,000 species[3]. These microbes form the base of the food web and are responsible for many fundamental processes.

Primary production alone adds an estimated $50 trillion USD per year to the global economy[4]. Open ocean ecosystem services are estimated to add $18 trillion/year[5] and the entire marine environment is estimated at just under $50 trillion/year[4]. Over 80% of the global net primary production (NPP) is generated in the open ocean environment[6]. Marine microbes contribute approximately half[1] of the estimated 44-50 petagrams/year of new production within the oceans[7]. Therefore, to continue benefiting from them, we must understand how these microbes function and interact with their environment as well as predict their response to environmental change.

The phylum cyanobacteria are arguably one of the most important architects of our present day environment as their ancestors were the first organisms to evolve the ability to photosynthesise[8]. These organisms subsequently caused one of the most important events in Earth's history, the 'Great Oxygenation Event' which occurred ~2.3 Ga ago[9]. These photosynthetic marine microbes were able to generate an oxygen rich environment that paved the way for complex lifeforms to evolve.

Currently, the majority of the oceans are dominated by two genera of prokaryotic primary producers, *Synechococcus* and *Prochlorococcus*, collectively known as cyanobacteria (see Figure 1.1). They contribute a significant proportion of chlorophyll biomass and global NPP[10,11], particularly in the open oceans where they can colonise the water column to depths in excess of 200 m. It is remarkable to think that they were only recently discovered, despite their abundance and ubiquity. *Synechococcus* was discovered 37 years ago in 1979[12,13] whilst *Prochlorococcus* was discovered in 1986, just over 3 decades ago[14].

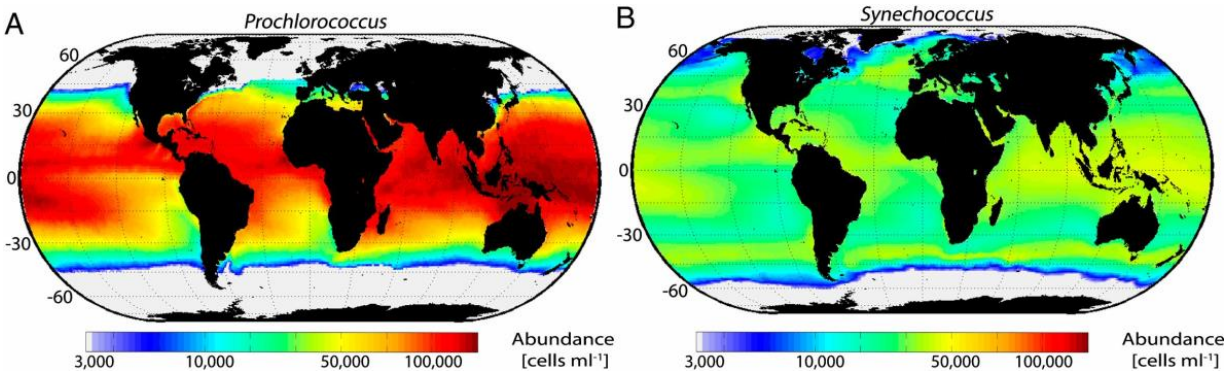

Figure 1.1: Global distribution of marine cyanobacteria. Estimated *Prochlorococcus* (A) and *Synechococcus* (B) mean annual abundance at the sea surface in 2013. Adapted from Flombaum *et al.*, (2013)[15].

Synechococcus are essentially ubiquitous in all upper well-lit marine environments, only being absent in the extremes of the Artic and Southern Oceans[16]. They have an annual mean global abundance of 7.0×10^{26} cells[15] with concentrations ranging from 5.0×10^{2} to 1.5×10^{6} cells ml^{-1} [16]. *Synechococcus* are capable of proliferating across a large range of temperatures, from warm temperate regions of 26°C in the Mediterranean Bay[17], to 7°C in the Gulf of Alaska[18], and were even found to be abundant within areas with temperatures as low as 2°C[19].

In comparison, *Prochlorococcus* are mainly restricted to the latitudinal band 45°N to 40°S (as depicted in Figure 1.1) and are found from surface waters down to depths of ~200 m[15]. However, they are much more abundant than *Synechococcus*, having a mean global abundance of 2.9×10^{27} cells[15] with concentrations of 1.0 to 4.0×10^{5} cells ml^{-1} [16]. *Prochlorococcus* are limited in the range of temperatures they can proliferate, with upper limits of 28°C and minimum temperatures of around 13°C[20]. Interestingly, these are also detected at latitudes as high as 60°N in the North Atlantic with temperatures reaching 10°C[21].

Marine cyanobacteria often co-occur in many areas, with *Prochlorococcus* frequently outnumbering *Synechococcus*, particularly in oligotrophic open ocean environments. *Synechococcus* typically outgrows *Prochlorococcus* in coastal, temperate, and upwelling regions due to higher nutrient concentrations. Regions of low salinities and/or low temperatures also allow *Synechococcus* to thrive over *Prochlorococcus*[15,16].

1.2 Genomic streamlining in cyanobacteria

Ecological genomics and phylogenetics of marine strains have defined clades, sub-clades, clusters and subclusters for *Synechococcus,* whilst *Prochlorococcus* are organised by their ecotype. The environmental distribution of a subcluster/ecotype typically reflects a particular environmental niche of cyanobacterial strains[22-26], tabulated in Table 1.1. Briefly, *Synechococcus* clades I and IV thrive in coastal and temperate mesotrophic open waters while Clades II and III are more suited to oligotrophic open ocean waters. *Prochlorococcus* on the other hand are mainly restricted to the latitudinal band 45°N to 40°S with high-light strains near the surface and low-light strains occupying at depths ~100-200 m. The GC content for *Prochlorococcus* strains range from 31-38% except the LLIV strains (50-51%) that are more closely related to *Synechococcus*. Different strains within a genus are defined by multiple genetic markers, including their intergenic transcribed spacer region of their ribosomal RNA genes[27], protein-encoding marker genes such as *rpoC*[28] and *petB*[24,25].

Synechococcus and *Prochlorococcus* typically contain no plasmids and have a single circular chromosome. Their genomes are very small with sizes ranging from 2.20 to 2.86 Mb and 1.64 to 2.70 Mb respectively, compared to the genomes of other bacteria and cyanobacteria, which have an average genome size of 3.69 ± 1.96 Mb and 5.33 ± 3.69 Mb respectively[11]. This type of genomic reduction is known as genome streamlining. Their finely tuned genomes allow different lineages to collectively proliferate across many oceanic environments[29-31].

There are two main counteractive driving forces which alter bacterial genome sizes, the deletion of non-essential genes, and the procurement of novel genes through horizontal gene transfer (HGT) and/or gene duplication[29]. There is a strong negative environmental selection against non-essential genes because the additional DNA increases the cell's requirements on energy and nutrient usage[31]. This is a huge advantage to have, especially in low nutrient waters. HGT and the random uptake of foreign DNA can lead to population diversity. This increases the chance of adaptation to a particular habitat as well as help with the acclimation to nutrient stress responses[32]. It is important to note that gene transfer between the two genera is thought to be infrequent if any[33] and that HGT is mainly mediated through bacteriophages[34,35]. These processes eventually cause species divergence through the adaptation to a particular environmental niche[36]. The utilisation of various genes across many environmental niches allows these organisms to thrive in nearly all marine environments where there is available light[11].

Table 1.1: Summary of marine cyanobacteria groups, strains, environmental niches, and GC content. Adapted from Scanlan *et al.*, (2009)[11].

Synechococcus

Group	Environmental niche	Strains	GC (%)
Clade I	Coastal/temperate mesotrophic open ocean waters largely above 30°N and below 30°S	CC9311	52
Clade II	Open ocean, offshore, continent shelf, oligotrophic tropical or subtropical waters between 30°N and 30°S	CC9605	59
Clade III	Ultraoligotrophic open ocean waters	WH8102	59
Clade IV	Coastal/temperate mesotrophic open ocean waters largely above 30°N and below 30°S	CC9902	54
Clade V/VI/VII	Relatively wide distribution but in low abundance in various oceanic waters – dominates mesotrophic upwelling regions	WH7803/WH7805/RS9917	60/57/65
Clade CRD1[26]	High nutrient, low chlorophyll regions known to be Fe limited – equatorial upwelling region, seen both at the surface and lower depths	MITS9220	56

Prochlorococcus

Group	Environmental Niche	Strains	GC (%)
High-light I (HLI)	More weakly stratified surface waters – between 35°-40°N and 35°-40°S	MED4	31
High-light II (HLII)	Strongly stratified surface waters – tropical and subtropical regions between 30°N and 30°S	MIT9312	31
Low-light I (LLI)	Occupies an "intermediate" position in the water column in stratified waters at high latitude (above 40°N/30°S) – cells found throughout euphotic zone up to the surface	NAT2A	35
Low-light II/III (LLII/III)	Present in deep waters at very low concentrations	SS120/MIT9211	36/38
Low-light IV (LLIV)	Widely distributed within the 40°N to 35°S latitudinal range but largely restricted to the deep euphotic zone	MIT9303/MIT9313	50/51

1.3 Nutrient metabolism in marine cyanobacteria

Microbes satisfy their elemental requirements from the external environment by acquiring macronutrients (C, N, P, and sulphur (S)) and micronutrients (metals such as magnesium, iron, nickel, manganese, cobalt, selenium, zinc, and copper etc.) for energy and to support various metabolic processes. It was previously thought that a single nutrient, usually nitrogenous, would limit marine phytoplankton productivity (following Liebig's law of the minimum, influenced by agricultural research)[37]. However, in reality it is much more complex than that and has since been seen more of a mosaic of spatial regions, each with their own temporal variations in nutrient levels and limitations[32]. Low concentrations of N, P, silica, and iron are often considered one of the main reasons that limit phytoplanktonic growth[32]. These nutrients can exist in the environment in two forms, inorganic and organic.

1.3.1 Adaptation of cyanobacteria in low nitrogen

Since N is one of the major nutrients required by all organisms, there is fierce competition for the very low concentrations of it present in the open oceans[20,38]. The metabolism of common nitrogen based nutrients are likely to be different in various marine cyanobacteria strains[32]. *Synechococcus* utilise various types of N based compounds such as ammonium, nitrate, nitrite, urea, and amino acids (AAs) as their N source[39-41]. Under N depleted conditions, some *Synechococcus* strains are observed to break down its main light harvesting pigment, phycoerythrin, for use as an internal N source[42]. This flexibility in remobilising N-rich components may be one of the reasons for the ubiquity of *Synechococcus* as well as being able to thrive over *Prochlorococcus* in more mesotrophic environments.

Nitrate is the most abundant N species in the ocean environment. All *Synechococcus* strains have the genes to utilise nitrate and *Prochlorococcus* strains were once thought to lack such genes[11]. Recently, some uncultivated *Prochlorococcus* lineages were discovered that may have the genetic capability to uptake and assimilate nitrate[43 123]. These lineages were identified in the environment using culture-independent techniques based on the sequencing of the 16S rRNA gene and related genomic regions. Metagenomic analysis revealed the genes binned to *Prochlorococcus* were 90% related to the LL MIT9313 and LL MIT9303 strains. The genes associated with nitrate uptake and assimilation are syntenous with those in *Synechococcus* WH8102, suggesting that these genes are likely preserved from a common ancestor.

1.3.2 Organic nutrient uptake observed in marine cyanobacteria

Microbes interact with their environment via membrane transport systems embedded within their cell envelopes. These systems carry out the efflux of wastes and toxins (exporters), as well as capturing nutrients (importers). The presence of multiple organic transport system genes suggests marine cyanobacteria are also capable of transporting and assimilating organic macronutrients, such as AAs, for growth and energy[44-47]. This is consistent with experimental data that demonstrated light-dependent uptake of leucine, methionine, and possibly other AAs by these organisms[48-51]. Marine cyanobacteria are therefore considered mixotrophic or photoheterotrophic in nature.

Additionally, dissolved organic matter (DOM) has been observed to be assimilated by marine cyanobacteria[40,49,50,52,53]. DOM is an important intermediate in some of the major biogeochemical cycles within the ocean[54,55] and, depending on the chemical composition, can be used to satisfy a wide range of nutritional requirements. DOM can be separated into either dissolved organic carbon (DOC) or dissolved organic nitrogen (DON). DOC make up one of the largest reservoirs of organic C in aquatic environments[56] while DON is one of the major forms of nitrogen that is essential for heterotrophic marine bacteria[57]. Even though thousands of organic molecules (mass formulae) have been identified[58,59], it still only represents <10% of marine DOM[60]. Many have been attributed to originate from bacteria such as porins[61], muramic acid[62,63], lipopolysaccharide[64], and D-AAs (specifically D-alanine, D-aspartate, D-glutamate, and D-serine)[63,65]. Viral lysis of bacteria is another process that supplements the DOM pool[66]. AAs make up ~1-3% of DOC and ~6-12% of DON[67,68] which suggests they could be a very important source of preformed monomers for protein synthesis or a supplementary N source for cyanobacteria. The most abundant AAs in DOM are glycine, alanine, aspartic acid, serine, and glutamic acid[67]. Surface waters (<100 m) contain 200-500 nM of total hydrolysable AAs[69].

Environmental assays suggest that dissolved AAs play a vital role in the nutrition of marine cyanobacteria as they may be an energetically preferential source of N[50]. The rate of AA uptake by marine cyanobacteria is directly affected by light[40,49,50,70,71]. Michelou *et al*., (2007)[49] observed the uptake of a 0.5 nM mixture of 15 AAs by marine cyanobacteria compared with other prokaryotes (Figure 1.2). On average, assimilation of AA was higher in light compared to in the dark. Marine cyanobacteria assimilated low concentrations of amino acids at per-cell rates similar to other prokaryotes[49].

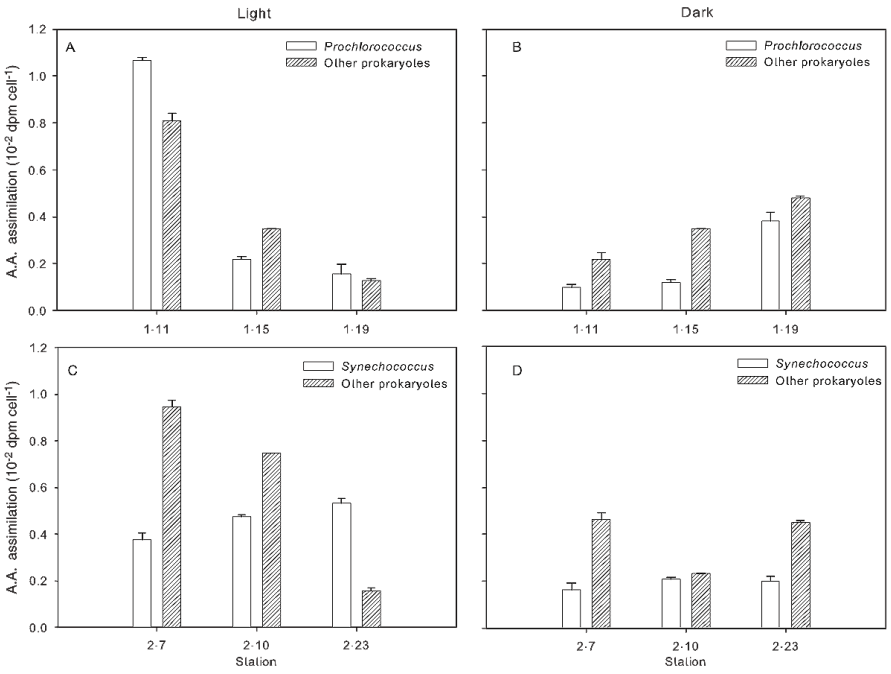

Figure 1.2: Rates of amino acid assimilation per cell for *Prochlorococcus*, *Synechococcus*, and other prokaryotes. Samples were incubated in the dark (B and D) or at 30% surface light irradiance (A and C). Error bars indicate standard errors of three measurements. Taken from Michelou *et al.*, (2007)[49].

Methionine and dimethylsulphoniopropionate (DMSP) are important organic S containing compounds that are taken up by marine cyanobacteria[49,52]. DMSP is most likely the preferred organic source of S because marine cyanobacteria do not need to use energy to reduce it[72]. Leucine is another AA that both cyanobacteria take up[50,70] and, together with methionine, are used to satisfy their N requirements[49]. The rates of methionine uptake were found to be very different between *Synechococcus* and *Prochlorococcus*[50]. In mesotrophic areas, the uptake of methionine by *Synechococcus* was observed to be, on average, 1.5% of the total pool per day[50]. This is very low since other bacterioplankton had an average of 46% daily turnover meaning that *Synechococcus* contributed only 3% of the total turnover[50]. Conversely, *Prochlorococcus* contributed 33% of the total methionine turnover by all bacterioplankton. Methionine is therefore a much more significant nutrient source for *Prochlorococcus* compared to *Synechococcus*[53], at least in the oligotrophic ocean.

Even though there has been evidence of AA uptake by marine cyanobacteria since the late ‘80s[73], there have yet to be any membrane transport system proteins characterised from marine cyanobacteria that show their ligand is indeed AAs. Membrane transport systems, such as ATP-binding cassette (ABC) transporters, are one of various types of transports that mediate the uptake of AAs for assimilation[74]. Characterisation of these transport systems will allow a better understanding of the underlying molecular mechanisms in nutrient uptake.

1.4 Nutrient assimilation facilitated by ABC transport systems

ABC transport systems drive the transfer of ligands through permeases by hydrolysis of ATP. The structure of prokaryotic ABC transporters is made of four domains, two variable integral transmembrane domains (TMDs) that form a channel within the membrane, and two highly conserved periphery domains which interact with ATP for the active intake of ligands (see Figure 1.3)[74]. Approximately 5-6% of the open reading frames (ORFs) in the oceanic strain *Synechococcus* WH8102 are linked with transport genes, which is similar to other bacteria[75]. However, marine cyanobacteria have a heavy bias towards ABC transporters compared to other bacteria (~60% of total transport ORFs)[75].

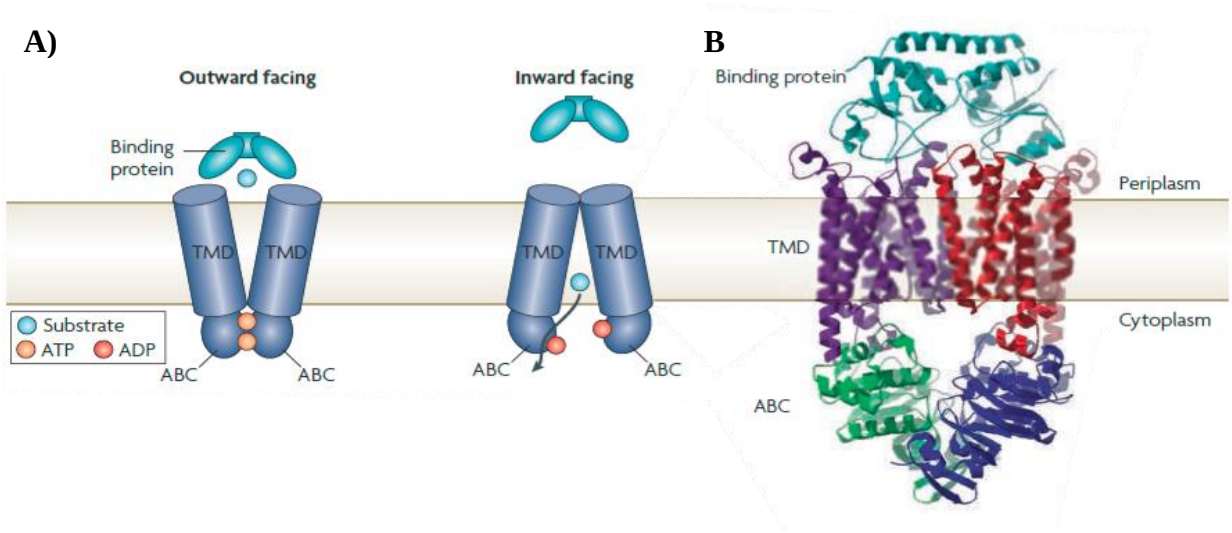

Figure 1.3: Molecular architecture of ABC transporters. A) Cartoon image of the modular organisation of ATP-binding cassette (ABC) transporters, which are composed of two transmembrane domains (TMDs) and two ABC domains. B) The *Escherichia coli* vitamin B12 importer BtuCDF[76] (Protein Data Bank (PDB) code: 2QI9). The core transporter consists of four subunits: the two TMD subunits (purple and red) and the two ABC subunits (green and blue). Adapted from Rees *et al.*, (2009)[74].

There are two types of ABC transporters, importers and exporters[74,77]. While the exporters are found in both eukaryotic and prokaryotic systems, importers seem to be present exclusively in prokaryotes. These importers are further subdivided into type I and type II. Detailed characterisation of sequences and TMD structures classify these importers[74]. The type I importer TMD fold contains 5 TM helices in each subunit whereas type II contains 10 TM helices per subunit[74]. These importers are largely restricted to low molecular weight compounds such as AAs, oligopeptides, mono/di-saccharides, metals, and inorganic nutrients, i.e. sulphate, nitrate, and phosphate.

1.4.1 Amino acid ABC transporter subfamily

The chemical properties of AA side chains vary widely. Therefore, diverse transport systems are required to capture these organic nutrients. The ABC superfamily that specifically transports AAs can be separated into two distinct subfamilies based on their ligands and TMD signatures[78], [1] polar AA uptake transporters (PAATs), and [2] hydrophobic AA uptake transporters (HAATs). The PAAT family is the larger of the two, having eleven types of transporters compared to only one in the HAAT family, the leucine-isoleucine-valine (LIV)/branched-chain AAs systems[78].

The PAAT subfamily is further divided into either narrow or broad solute specific transporters. The narrow subfamily typically transports only a single AA or a group of structurally similar AAs. The broad range transporters facilitate the transport of a range of AAs. The broadest type, such as the general AA permease (AapJQMP) of *Rhizobium leguminosarum*, transports acidic, basic, and aliphatic L-AAs[79]. Structurally, the largest difference between the two subfamilies are the size of the N-terminal integral membrane proteins, as the broad solute transporters have integral proteins that are 30% larger[78]. It is currently unknown whether this N-terminal extension plays a role in the specificity of the broad solute transporters or if it required for regulation. Regardless, there is a 63 residue N-terminal region of similarity that is used for the identification of the PAAT family[80]. This region appears to be important for solute specificity as it has been demonstrated for the His transporter. Deletion of the residues 25-28 (ILAV) within this region caused a shift in solute specificity from histidine to histidinol[81]. The HAAT subfamily on the other hand is not as well characterised and consists of a single group of transporters which uptake branched-chain AAs.

The AA-ABC transporters in two cyanobacteria strains, the freshwater cyanobacteria *Synechocystis* strain PCC 6803 and *Synechococcus* sp. stain PCC 7942 were recently investigated[78,82]. Two ORFs within strain PCC 6803, *natA* and *natB* are characterised to have high-affinity with branched-chain AAs. Mutation of these genes caused an almost total deficiency in the uptake of broad range AAs (i.e. Ala, Gly, Leu, Phe, Pro, and Ser). Furthermore, knock out studies showed a 70% and 30% decrease in the uptake of aspartate and histidine, respectively. Strain PCC 7942 was reported to have low uptake rates of basic AAs (Arg and Lys) which suggests the absence of basic AA uptake systems[82]. Although there was moderate uptake of His, neutral AA transport systems may be responsible for this.

Among other neutral AAs, the nonpolar AAs Ala, Leu, and Phe showed a ~5-fold higher uptake rates compared to polar AAs Asn, Gln, and Ser[82].

1.5 Role of periplasmic binding proteins in substrate transport

All ABC transporter uptake systems in Gram-negative bacteria have at least one or two extracellular solute binding proteins, known as periplasmic binding proteins (PBPs). These are a class of proteins used by organisms for signal transduction or membrane transport systems[83]. They are the first components to interact with the substrate with high affinity for the intermembrane transport. PBPs were first associated with prokaryotic ABC transporters[84]. They have since been linked with other membrane complexes such as, prokaryotic two-component regulatory systems[85], prokaryotic tripartite ATP-independent periplasmic (TRAP)-transporters[86], tripartite tricarboxylate (TT) transporters[87], G-protein coupled receptors (GPCRs) and ligand-gated ion channels[88]. PBPs bind a wide variety of ligands such as monosaccharides, oligosaccharides, AAs, oligopeptides, sulphate, and phosphate in the periplasmic space[89]. Each PBP binds to a specific ligand or a chemically similar group of ligands.

In the presence of a ligand, PBPs exist in a closed conformation as the ligand is trapped between the two domains[89,90], via a "Venus fly-trap" mechanism[91]. There are four postulated states that PBPs can occupy based on their 3D structures: [1] open-unliganded, [2] open-liganded, [3] closed-unliganded, and [4] closed-liganded[83,92,93].

1.5.1 Structural classification of PBPs

PBPs vary in size (~25-70 kDa) and have little sequence homology, yet their three-dimensional (3D) structure remains conserved[83,89]. Based on their 3D structure, six clusters (A-F) of PBPs were identified[83] with a possible seventh cluster (G) recently proposed[94] (shown in Table 1.2). Additionally, based on the topology of the beta-sheet core and the connectivity of secondary structure elements, PBPs are also classified into three different classes[83,89,94]. The overall structure of PBPs contains two α/β domains, a central β-sheet with 5 β-strands fringed by α-helices, connected by a hinge region. These features suggest that PBPs originated from a common ancestor. In absence of a ligand, PBPs exist in an open conformation. This allows the domains to flex and rotate around the hinge region[93].

The central core β-sheets have three classes of topological arrangements. Class I folds have the form of $\beta_2\beta_1\beta_3\beta_4\beta_5$ whereas class II has the form of $\beta_2\beta_1\beta_3\beta_n\beta_4$, where β_n represents the strand just after the first cross-over from the N-terminal domain to the C-terminal domains, and vice versa. Class III folds were more recently identified and were found to have

the form $\beta_3\beta_2\beta_1\beta_4\beta_5$[83,95]. Class III proteins have the distinguishing characteristic of an α-helix acting as the hinge connecting the two domains. This maintains an overall stable structure causing minimal movement upon substrate binding[83]. Class I and III folds are each only found within one cluster of PBPs, cluster B and A respectively. Class II folds encompass all the other clusters (C-G). Classification of clusters, and their respective subclusters, are based on ligand specificity and structural information (Table 1.2).

Analyses of the PBP clusters indicate the presence of AA binding PBP in B-II, F-II, and F-IV clusters (Table 1.2, highlighted in bold). Structures within cluster B are shown to bind diverse ligands such as carbohydrates, autoinducer-2, AAs, aromatic acids, and natriuretic compounds (Table 1.2). The distinguishing structural feature of cluster B PBPs are the three hinge regions connecting the two domains (shown in Figure 1.4) indicating C- and N-terminals of these proteins are located on separate domains. Homologous proteins to this cluster includes the *lac*-repressor type transcription regulators[83] and the ribose-binding protein from *E. coli*[96]. Cluster F proteins are shown to majorly bind trigonal planar anions, AAs, and compatible solutes (betaine, proline betaine, and choline) (Table 1.2). The distinguishing structural feature of cluster F PBPs is the two strands connecting the two domains (Figure 1.4). Sub-cluster F-IV binds with AAs and often have the possibility to bind with other AAs, albeit at an order of magnitude lower affinity.

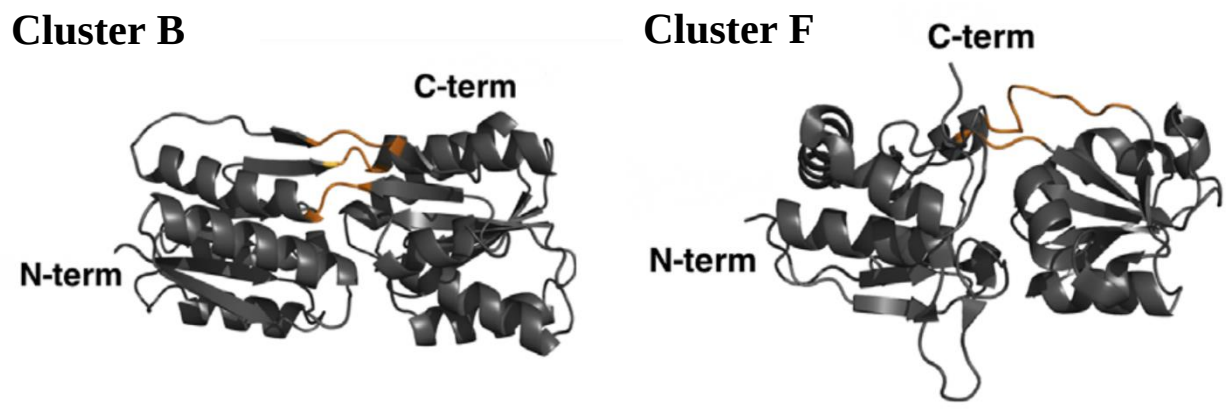

Figure 1.4: PBP structural cluster B and F shown with their distinct structural feature coloured in orange. Cluster B has three interconnecting segments between the two domains. Cluster F has two long hinging segments allowing high flexibility. The two protein used to illustrate the features in cluster B and F were RBP (PDB code: 1DRJ) and HisJ (PDB code: 1HSL), respectively. Adapted from Berntsson *et al*., (2010)[83].

Table 1.2: Summary of PBP structures. Beta-sheet topology with *n* denotes the strand following the first-over from the N_terminal domain to the C-terminal domain, or vice versa. Amino acid associated ABC transporters are highlighted in bold. Adapted from Scheepers *et al*., (2016)[94].

Cluster	Ligand(s)	Transporter type(s)	Class	β-sheet topology	Connective secondary structure elements	Protein examples
A-I	Metal ions	Type II ABC transporters	III	$\beta_3\beta_2\beta_1\beta_4\beta_5$	Single α-helix linking the two α/β domains	Vitamin B_{12} binding protein BtuF[95] and Zinc binding protein TroA[97]
A-II	Siderophores, hemes, vitamin B12, molybdite, tungstate	Type II ABC transporters				
B-I	Sugars, sugar alcohols, autoinducer-2	Type I ABC transporters, two- component sensor kinases, transcription factors	I	$\beta_2\beta_1\beta_3\beta_4\beta_5$	Three connecting strands at the hinge region	Leucine/isoleucine/valine-binding protein[98] and lysine-binding protein (LBP)[99]
B-II	**Amino acids**	**Type I ABC transporters**				
B-III	Aromatic acids	Type I ABC transporters				
B-IV	Natriuretic compounds	GCNPRs				
B-V	Glutamate	GPCRs				
C	Di- and oliopeptides, Arg, cellobiose, nickel, opine	ABC transporters, TRAP transporters, ligand-gated ion channels, and GPCRs	II	$\beta_2\beta_1\beta_3\beta_n\beta_4$	Two connecting strands at the hinge region	Histidine binding protein (HisJ)[100], oligopeptide binding protein (OppA)[101], and Alginate-binding protein[102,103]
D-I	Sugars, sugar alcohols, spermidine/putrescence	Type I ABC transporters				
D-II	Thiamine, spermidine/putrescence	Type I ABC transporters, transcription factors				
D-IIIa	Phosphate, molybdate, tungstate	Type I ABC transporters				
D-IIIb	Molybdate, tungstate	Type I ABC transporters				
D-IV	Iron ions	Type I ABC transporters				
E-I	Organic acids, amino alcohols, dipeptides, glycerol-3-phosphate	TRAP transporters				
E-II	Amino acids	TT transporters				
F-I	Trigonal planar anions	Type I ABC transporters				
F-II	**Methionine**	**Type I ABC transporters**				
F-III	Betaine, proline betaine, and choline	Type I ABC transporters				
F-IV	**Amino acids**	**Type I ABC transporters**, ligand-gated ion channels, GPCRs				
G	Polysaccharide	Pit-dependent ABC transporters				

1.6 Scope of this thesis

Synechococcus and *Prochlorococcus* are ubiquitous and abundant marine phototrophs. These organisms play essential roles within key biogeochemical cycles. They have been observed to actively uptake organic compounds including AAs, which likely serve as important nutrient sources that marine cyanobacteria scavenge for growth and metabolism. PBPs associated with ABC transporters are vital for the active transport of AAs. Direct analysis of PBPs involved in the transport of AAs would allow a deeper understanding behind these organic transport system mechanisms. Recombinant expression and subsequent ligand screening bioassay techniques should aid characterisation of these AA binding PBPs. This will help define pathways of energy and nutrient cycling in marine ecosystems. The environmental significance of these proteins can be studied further using metagenomics analysis.

Previous work in Paulsen Laboratory at Macquarie University selected eight *Synechococcus* strains (namely CC9311, WH8109, CC9605, WH8102, CC9902, BL107, BIOS-U3-1, MITS9220), representing strains adapted to contrasting environmental niches, to undertake a genome-wide analysis. Thus, 111 PBPs (see Figure 1.5) across 23 distinct clusters were identified with gene annotations indicating diverse putative ligands (Paulsen *et al*., unpublished). The latter include iron, manganese, zinc, urea, phosphate, phosphonates, sugars, oligopeptides, AAs, and possibly toluene. The Cyanorak database (abims.sb-roscoff.fr/cyanorak) comprises an expert curation and annotation of clusters of orthologous gene sequences in marine cyanobacteria genomes. Specific to this work, twelve genes encoded within three specific clusters (Cyanorak v1 cluster numbers: CK1309, CK1489, and CK2503) were annotated to putatively bind AAs, depicted as the bright blue coloured clusters in Figure 1.5. As shown in the phylogenetic tree constructed from the secondary structure alignment, all three AA binding clusters are not closely related suggesting distinct binding mechanisms and ligand specificities.

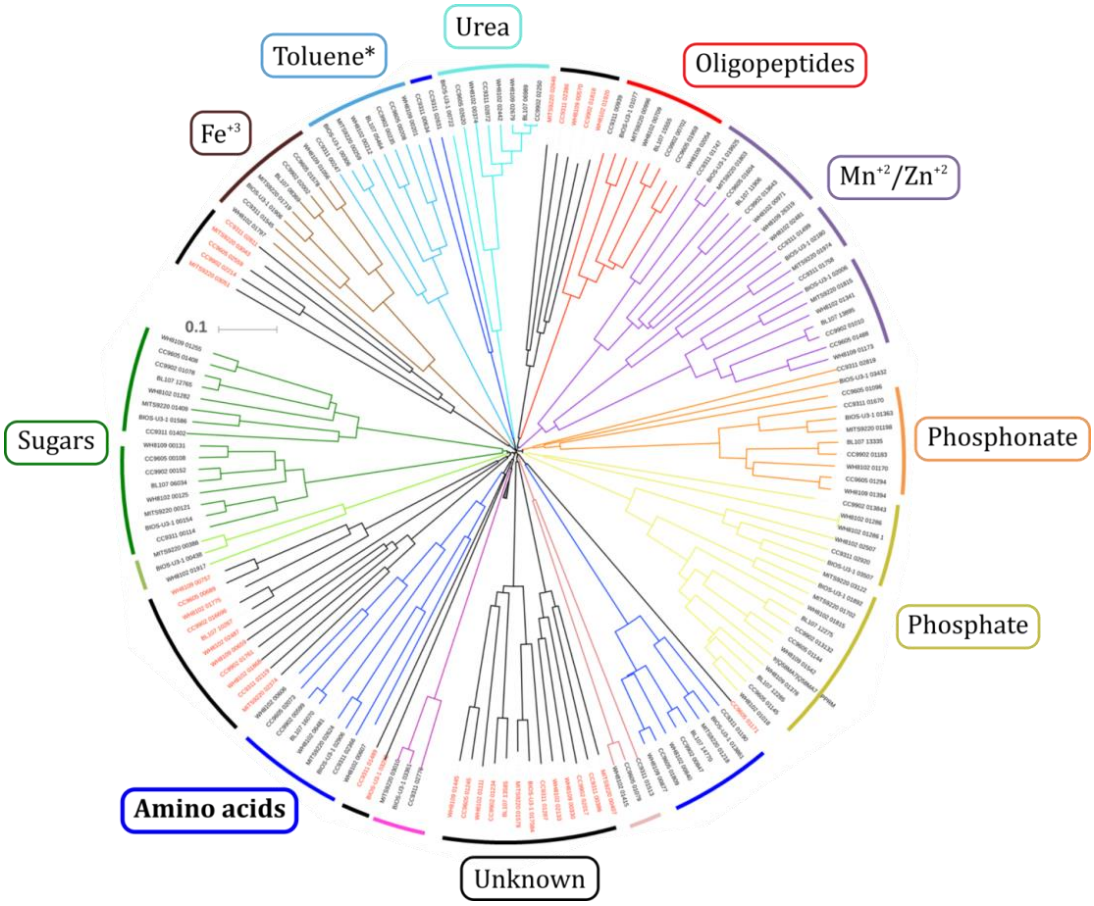

Figure 1.5: Phylogenetic tree depicting the non-relatedness of PBPs identified from genomic analysis of *Synechococcus* strains. Related gene clusters are shown in different colours along with their putative binding partners.

This thesis is part of a larger project that aims to characterise the identified *Synechococcus* PBPs genes involved with ABC transporters. As previously mentioned, there is a heavy bias for ABC transporter genes within marine cyanobacterial genomes. These transporters must play a vital role in their growth and proliferation within their environmental niche. My thesis begins to unravel the molecular characteristics of these *Synechococcus* nutrient transport systems starting with AA transporters.

The specific aims of my project were to:

(a) Clone twelve *Synechococcus* gene targets, predicted to bind AAs in *E. coli* for recombinant expression

(b) Test recombinant protein expression and solubility in a cytoplasmic system (using a pET-15b vector) as well as in a periplasmic system (using a pET-22b vector)

(c) Optimise purification of soluble periplasmic AA binding proteins

(d) Undertake bio-physical characterisation of purified proteins to determine ligand binding specificities

Chapter 2. Experimental Design, Materials, and Methods

Amino acid binding periplasmic proteins selected for this work were from a pool of 111 PBPs present in 23 distinct gene clusters across eight cyanobacterial genomes. The strategy for experimental characterisation of these twelve selected gene products involved several discrete stages as depicted in Figure 2.1:[1] bioinformatic screening of suitable *Synechococcus* PBP genes that putatively bind with amino acids (AAs), [2] custom design of primers for PCR amplification of target genes using extracted *Synechococcus* gDNA, [3] ligation independent cloning (LIC) of target genes into pET vectors, [4] small-scale solubility trials, [5] protein purification of soluble gene products using IMAC and SEC, [6] ligand screening using differential scanning fluorimetry.

2.1 Consideration of *E. coli* cloning vectors

In contrast to the reducing environment of the cytoplasm, the periplasmic space displays an oxidising surrounding. Therefore, recombinant expression of periplasmic binding protein in *E. coli* bacterial host must be carefully considered. Translocation of proteins, either co-translationally or post-translationally, from the cystol to the periplasmic space is determined by the presence of signal peptides. These are characterised by short regions of hydrophobic residues near the N-terminal[104]. Sequence analysis of the twelve-selected putative AA binding proteins indicate that each contains a likely signal sequence. Therefore, two bacterial vectors, pET-15b (5708 bp) leading to a cytoplasmic expression and pET-22b (5493 bp) encompassing *pelB* leader sequence, thus leading to a periplasmic expression were chosen for protein expression for both reducing and oxidising environments. This approach has previously been successful to screen for soluble PBPs[105].

Construct design involved truncation of the native signal peptide sequences at locations identified using the SignalP 4.1 server (www.cbs.dtu.dk/services/SignalP) (Appendix II). As depicted in Figure 2.2, the pET-15b and pET-22b vectors also have a hexahistidine tag (His-Tag) at their N- and C-terminal, respectively. This affinity tag is used for a single step affinity purification method. In contrast to the pET-22b vector, the pET-15b vector also has a thrombin cleavage site incorporated to allow the subsequent removal of its N-terminal His-Tag. The pET-22b vector however does not have a cleavage site to remove its C-terminal His-Tag, therefore a tobacco etch virus (TEV) cleavage site (sequence ENLYFQS) was incorporated for the pET-22b vector. TEV is utilised due to its higher efficacy compared with other cleavage sites[106].

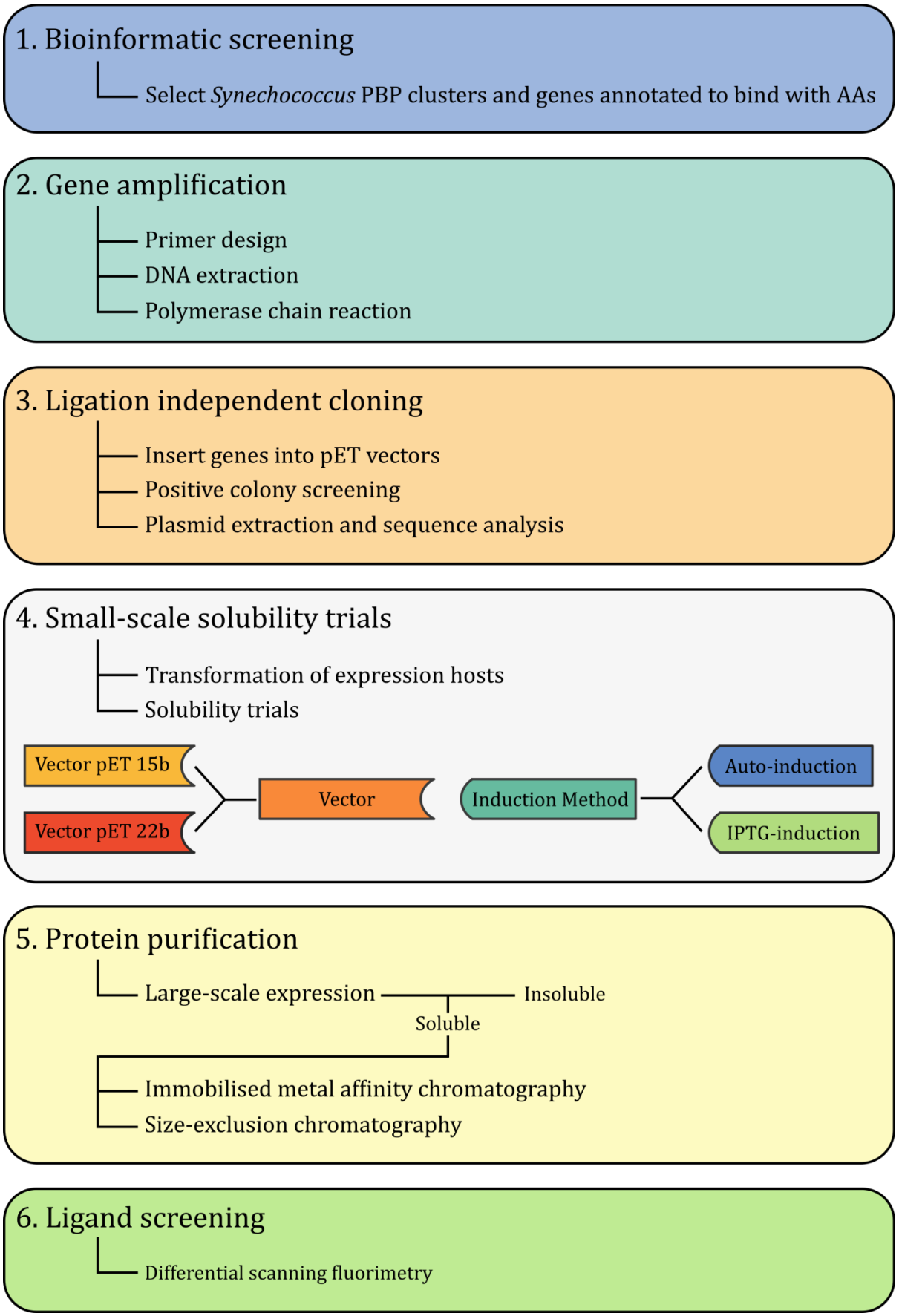

Figure 2.1: Stages of experimental characterisation of recombinant *Synechococcus* amino acid PBPs.

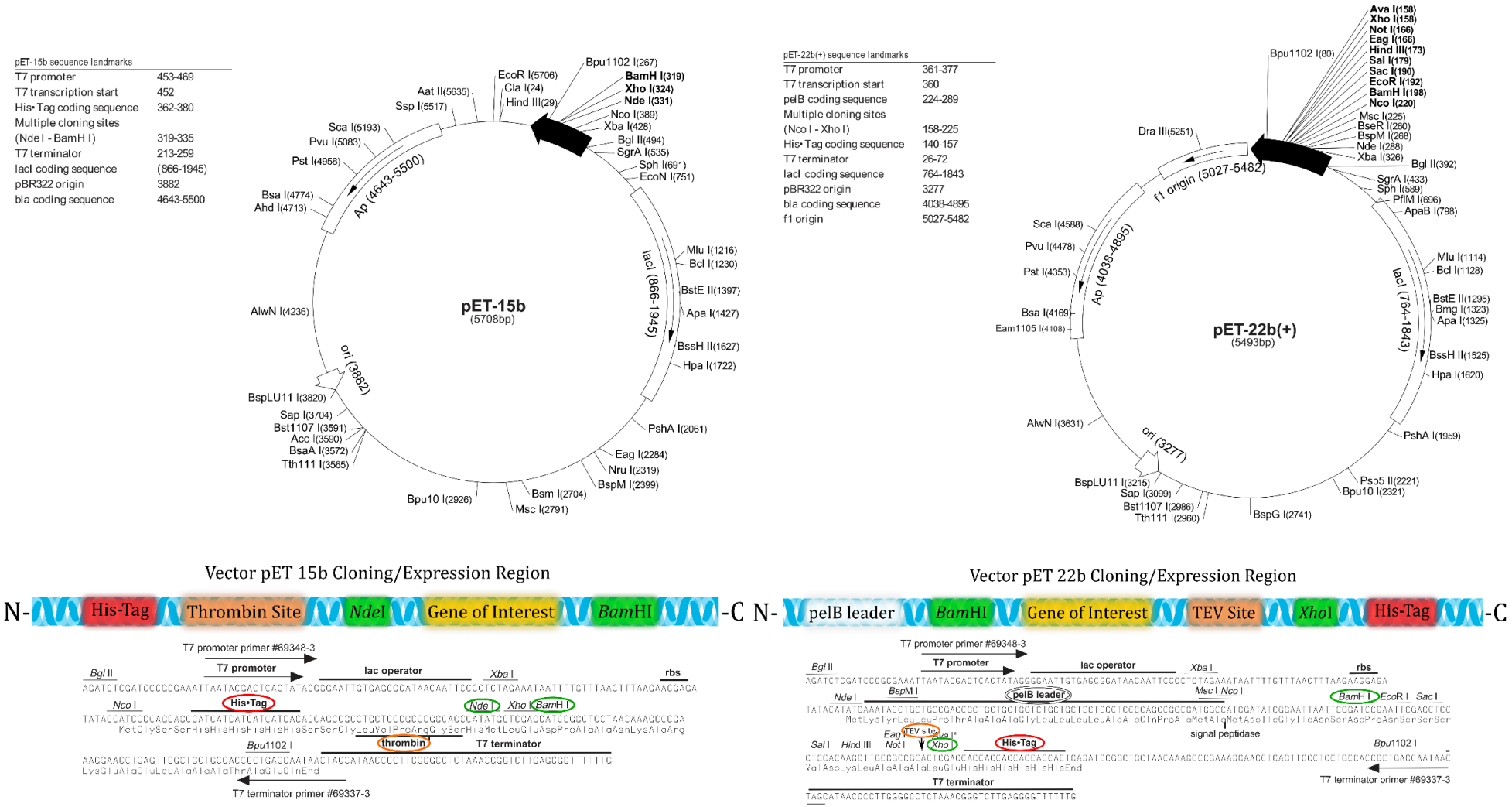

Figure 2.2: Bacterial vectors pET-15b and pET-22b. The overall insert architecture is depicted: His-Tags (red), restriction sites (green), gene of interest (yellow), cleavage sites (orange), and the *pelB* leader sequence (white). Notable components are highlighted in circles in the expression regions. The pET-15b vector has an N-terminal His-Tag followed by a Thrombin cleavage site. The pET-22b vector has a C-terminal His-Tag preceded by an engineered TEV cleavage site. Image adapted from Novagen.

2.2 Gene amplification

Protocols and strategies used for cloning, protein expression, and purification of putative *Synechococcus* AA-PBPs were adapted from procedures previously developed by the Protein Structure Group at Macquarie University[107].

2.2.1 Primer design

To design efficient and effective primers, factors such as length, T_m, GC content, GC clamp, runs, repeats, and secondary structures were taken into consideration[108-110]. Secondary structures such as hairpins, self-dimerisation, and cross-homology were avoided as much as possible as it greatly reduces the efficacy of PCR product generation. Runs (consecutive single nucleotides) and repeats (consecutive di-nucleotides) that occur more than four times were also avoided as they have a high tendency to misprime. The OligoCalc software (http://biotools.nubic.northwestern.edu/OligoCalc.html) was used to calculate various oligonucleotide properties (Appendix I) and screen for secondary structures. As previously mentioned, a tobacco etch virus (TEV) cleavage site was added to the reverse primer of pET-22b for removal of the C-terminal His-Tag tag if required. The vectors also have an ampicillin resistance gene for selection.

2.2.2 Genomic DNA extraction and PCR amplification

The *Synechococcus* strains (CC9902, CC9311, MITS9220, WH8102, WH8109) used for this study were cultivated in artificial seawater at 25°C, 100 rpm, with a day/night cycle. Genomic DNA was extracted from the *Synechococcus* strains using the CTAB method previously described[111]. Genomic DNA is used as template DNA for the primers to amplify the AA-PBP genes using the polymerase chain reaction (PCR).

Briefly, *Synechococcus* culture (50 ml) was harvested by centrifugation (5000 g for 30 min) and resuspended in nuclease-free TE buffer (564 µl, Table 2.2). To the cell suspension, 10% w/v sodium dodecyl sulphate (30 µl) and 20 mg/ml of proteinase K (3 µl) was added before incubation in a water bath (60°C for 4 h). After incubation, 5M NaCl (100 µl) and 10% cetrimonium bromide in 0.7% w/v NaCl (80 µl) was added. The mixture was then incubated again (65°C for 10 min). One volume of chloroform:isoamyl alcohol (24:1 v/v) was added before centrifugation (12000 g for 5 min). The supernatant was collected and 1 volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added before centrifugation (12000 g for 5 min). The previous step was repeated and the supernatant was collected. DNA was precipitated by adding 0.6 volumes of isopropanol. The solution was mixed by gentle inversion and centrifuged (18000 g, 10 min, 4°C). The supernatant was discarded and the

pellet was washed with 70% ice cold ethanol and centrifuged (18000 g, 10 min, 4°C). The supernatant was discarded leaving the pellet to air dry before resuspending in nuclease-free water (30 µl).

The custom designed primers (Appendix I) were ordered from an external source (Integrated DNA Technologies). Genes encoding *Synechococcus* AA-PBPs were PCR amplified from previously extracted genomic DNA. All PCR reactions were performed with Taq DNA polymerase (Qiagen) as follows: 94°C for 5 min; 94°C for 30 s, 55°C for 30 s, 72°C for 1 min 30 s, repeat for 30 cycles; 72°C for 5 min, 4°C hold. The resulting PCR products were visualised on a 0.8% agarose gel. PCR products were purified using the protocol and solutions provided by the Wizard® SV gel and PCR Clean-up System (Promega). The concentration of DNA was determined using a NanoDrop® ND-1000 spectrophotometer and ND-1000 v3.5.2 software (Thermo Fisher Scientific).

2.3 Ligation independent cloning

Ligation independent cloning (LIC)[112,113] was used to clone the amplified *Synechococcus* AA-PBP genes into each vector because it is robust, simple, and inexpensive in comparison to other recombination-based approaches.

LIC was carried out using the In-Fusion™ system (Clontech). Two bacterial expression vectors, pET-15b and pET-22b (Novagen), were linearised by digestion with restriction enzymes highlighted in Figure 2.2. The pET-15b vector (30 µl) was mixed with 10x BamH1 buffer #3 (10 µl) (New England Biolabs), *Bam*HI restriction enzyme (2 µl) (New England Biolabs), and water to a total volume of 98 µl and incubated (37°C, 3 h). After incubation, *Nde*I (2 µl) (New England Biolabs) was added and the mixture was incubated again (37°C, 2 h). For the pET-22b vector, *Xho*I restriction enzyme (New England Biolabs) replaced *Nde*I. Purified digested vector was visualised on a 0.8% agarose gel. Ligation reaction mixture was made that contained the digested vector (50 ng/µl for pET-15b and 35 ng/ml for pET-22b), In-fusion™ enzyme (1 µl, Clontech), fusion buffer (2 µl), PCR insert (2:1 insert/vector molar ratio), and water to make up a total of 10 µl. The mixture was incubated (37°C for 30 min) then placed on ice. Each ligation reaction mixture (2.5 µl) was added to Stellar™ chemically competent *E. coli* cells (25 µl, Novagen) and incubated (30 min). The mixture was heat shocked (42°C for 1 min) and transferred on ice (2 min). SOC media (375 µl) was added before incubation (37°C, 2 h, 250 rpm). The cells (100 µl) were then plated onto LB + Amp plates (Table 2.1) for selection. The plates were incubated overnight (37°C) and stored at 4°C.

Table 2.1: Growth media used. Abbreviations: Amp, ampicillin; Chl, chloramphenicol; LB, lysogeny broth; Me, metal.

Growth media	Components
Lysogeny broth (LB)	Tryptone (10 g/l), yeast extract (5 g/l), NaCl (86 mM)
LB + Amp + Chl	Tryptone (10 g/l), yeast extract (5 g/l), NaCl (86 mM), Amp (50 μg/ml), Chl (25 μg/ml)
LB plates	Tryptone (10 g/l), yeast extract (5 g/l), NaCl (86 mM), bacteriological agar (15 g/l)
LB + Amp plates	Tryptone (10 g/l), yeast extract (5 g/l), NaCl (86 mM), bacteriological agar (15 g/l), Amp (50 μg/ml)
LB + Amp + Chl plates	Tryptone (10 g/l), yeast extract (5 g/l), NaCl (86 mM), bacteriological agar (15 g/l), Amp (50 μg/ml), Chl (25 μg/ml)

2.3.1 Identification and sequencing of recombinant plasmids

Six colonies from each transformation were selected to screen for positive clones using PCR as described in section 2.5 (colonies were used as template DNA). Positive clones were visualised on a 0.8% agarose gel. A colony that gave a positive colony PCR result from each transformation was selected and cultured to enable sequencing of the cloned gene to ensure the correct nucleotide sequence. Each colony was aseptically transferred into separate falcon tubes (50 ml volume) containing 5 ml LB + Amp (Table 2.1) and incubated overnight (37°C, 200 rpm, 12-18 h). From this, an aliquot (1 ml) was taken to make a 25% glycerol stock and stored at -80°C. Plasmids were extracted from the rest of the culture using the Wizard® *Plus* SV Minipreps DNA Purification System (Promega). The purified plasmid products were Sanger sequenced by an external provider (Macrogen). The sequence was manually verified using Serial Cloner v2.6.1 software to ensure the sequence was correct. All twelve genes were successfully cloned into both pET-15b and pET-22b vectors.

2.4 Small-scale expression trials

2.4.1 Transformation into *E. coli* Rosetta™ 2(DE3) cells for protein expression

Transformations was performed using the *E. coli* Rosetta™ 2(DE3) strain (Novagen), which has a chloramphenicol resistance gene for selection. Previously extracted plasmid DNA (2 μl) was added to Rosetta™ 2 competent cells (20 μl, Novagen) and left on ice (30 min). The cells were then heat shocked (42°C for 30 s) and left on ice (2 min). Room

temperature SOC media (400 µl) was then added to the cells and incubated (37°C, 1.5 h, 250 rpm). The cells (100 µl) were then plated onto LB + Amp + Chl plates (Table 2.1) for selection. The plates were incubated overnight (37°C) and stored at 4°C. After transformation, plasmids were extracted for sequence analysis to confirm the presence of the expression cassette.

2.4.2 Small-scale protein solubility trials

Starter cultures were generated from the transformation plates made previously. Colonies were aseptically transferred into separate falcon tubes (50 ml volume) containing 5 ml LB + Amp + Chl (Table 2.1) for selection. Cultures were incubated overnight (30°C, 14-18 h, 250 rpm). After incubation, an aliquot (1 ml) was taken to make a 25% glycerol stock, stored at -80°C. IPTG and auto-induction were both tested at this stage.

For IPTG-induction, a starter culture (50 µl) was added to 5 ml LB + Amp+ Chl (Table 2.1) in a 50 ml falcon tube and incubated (37°C, 250 rpm). The optical density at 600 nm (OD_{600}) was monitored every hour using a Cary® 50 UV-Vis spectrophotometer, Simple Reads software, (Agilent Technologies). When the OD_{600} reached ~0.6, the culture was induced with 0.5 mM IPTG and incubated overnight (18°C, 12-18 h, 250 rpm). For auto-induction, the starter culture (50 µl) was added to 5 ml ZYP-5052 rich medium with Amp (50 µg/ml) + Chl (25 µg/ml) in a 50 ml falcon tube and incubated (37°C, 5-6 h, 250 rpm). After growth, the culture was incubated overnight (25°C, 12-18 h, 250 rpm).

Both IPTG and auto-induction cultures were centrifuged (10,000 g, 15 min, 4°C) and the supernatant discarded. Cells containing pET-22b expression constructs were resuspended in ice cold periplasmic extraction buffer (500 µl, Table 2.2), incubated (4°C, 1.5 h, gentle agitation), and centrifuged (10,000 g, 15 min, 4°C). The resulting supernatant contained the periplasmic fraction and was aliquoted (40 µl) for SDS-PAGE. The remaining supernatant was discarded and the cell pellets were snap frozen in liquid N_2. Cells containing pET-15b expression constructs were also snap frozen in liquid N_2. All frozen cell pellets were thawed, resuspended in cell lysis buffer (500 µl, Table 2.2), and incubated (4°C, 3 h, gentle agitation). An aliquot (40 µl) of the supernatant containing the total lysate fraction was taken for SDS-PAGE. Cells were centrifuged (20,000 g, 10 min, 4°C) and another aliquot of the supernatant (40 µl), containing the soluble fraction, was taken for SDS-PAGE. Periplasmic extraction protocol was adapted from Sockolosky and Szoka (2013)[114].

Table 2.2: Buffers used for protein purification and analysis.

Buffer	Components
TE buffer	10 mM Tris pH 8.0, 0.1 mM EDTA
Binding buffer	50 mM Hepes, 300 mM NaCl, 10 mM imidazole, 5% glycerol, pH 7.4
Wash buffer	50 mM Hepes, 300 mM NaCl, 50 mM imidazole, 5% glycerol, pH 7.4
Elution buffer	50 mM Hepes, 300 mM NaCl, 500 mM imidazole, 5% glycerol, pH 7.4
SEC buffer	50 mM Hepes, 300 mM NaCl, 5% glycerol, pH 7.4
Periplasmic extraction buffer	200 g/l sucrose, 1 mM EDTA, 30 mM Tris-HCl, pH 8.0
Cell lysis buffer	150 mM NaCl, 1 mM Na_2EDTA, 1 mM EGTA, 1% Triton X-100, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM sodium orthovanadate, 1mg/l leupeptin, 20 mM Tris-HCl, pH 7.0
DSF buffer	0.02M HEPES, pH 6.8

2.4.3 SDS-PAGE analysis of small-scale expression trials

Aliquots of protein mixture were collected throughout various stages of the trials, notably total lysate, soluble, and periplasmic fractions. Solutions used for SDS-PAGE are shown in Table 2.3. These aliquots were mixed with loading dye and incubated (95°C, 10 min). The stained proteins were run on a 4-20% gradient gel (Bio-Rad) at 120V for 75 min using a rainbow ladder (GE Healthcare) as a marker. Fixing solution was used on the gels for 10 min and stained with staining solution for 2 min. Destaining solution was then applied until protein bands were visible.

Table 2.3: Solutions used for SDS-PAGE analysis

Solution	Components
5X loading dye	100 mM Tris buffer pH 6.8, 4% w/v sodium dodecyl sulphate, 0.2% w/v bromophenol blue, 20% w/v glycerol, 200 mM dithiothreitol
Fixing solution	50% v/v ethanol, 10% v/v acetic acid
Staining solution	0.25% w/v Coomassie Brilliant Blue, 10% v/v ethanol, 10% v/v acetic acid
Destaining solution	10% v/v acetic acid

2.5 Protein expression and purification

2.5.1 Large-scale protein expression in *E. coli* cells

Starter cultures from the 25% glycerol cell stock (sterile tip) were generated in 5 ml of LB + Amp + Chl (Table 2.1) contained in falcon tubes (50 ml volume). For each construct, 2L cultures were generated. For IPTG-induction, baffled shake flasks (2L volume) containing 500 ml LB + Amp + Chl (Table 2.1) were inoculated with a starter culture. Cells were incubated (37°C, 250 rpm) until OD_{600} ~0.7-0.9. Protein expression was induced by addition of 1 mM IPTG and incubated overnight (18°C, 16-18 h, 250 rpm). For auto-induction, baffled shake flasks (2L volume) containing 500 ml ZYP-5052 rich medium with Amp (50 µg/ml) + Chl (25 µg/ml) was inoculated with a starter culture. Cells were incubated for growth (37°C, 5-6 h, 250 rpm) before incubating overnight (25°C, 12-18 h, 250 rpm) to overexpress recombinant protein.

After overnight incubation, the culture was centrifuged (5000 g, 30 min, 4°C) and the supernatant was discarded. The cell pellet was resuspended in 25 ml binding buffer (Table 2.2), 5 mg/ml DNaseI (20 µl), and 200 mg/ml lysozyme (300 µl). The mixture was snap frozen in liquid N_2 and stored at -80°C.

2.5.2 Protein purification procedures

Protein purification was achieved in a two-step purification process involving immobilised metal affinity chromatography (IMAC) and size-exclusion chromatography (SEC). Briefly, the soluble fraction of protein lysate was filtered and passed through an IMAC column charged with Ni^{2+}, immobilising proteins with a His-Tag through adsorption chemical interactions[115]. Elution of recombinant protein was carried out using high concentrations of imidazole (elution buffer, see Table 2.2) to outcompete interactions with Ni^{2+}. If the purified product still contained impurities, a second chromatography based SEC purification step was utilised. SEC employed the use of cross-linked dextran gels to separate macromolecules based on their molecular size[116].

Frozen cell cultures from the large-scale expression were thawed and a protease inhibitor cocktail (300 µl) (Sigma-Aldrich) was added. The cell cultures were lysed through sonication, 4 pulses (10 s on then 10 s off) at 60% amplitude and incubated on ice (~2 m) before continuing for a further 3 pulses using the same parameters. The soluble solution was harvested through centrifugation (20000 g, 20 min, 4°C) and clarified over a 0.45 µm filter. Purification by IMAC was as follows using a peristaltic pump at room temperature. A 1 ml HisTrap column (GE Healthcare) charged with Ni^{2+} was equilibrated with 5 column volumes

(CV) of binding buffer (Table 2.2). The soluble cell fraction was loaded onto the HisTrap column and washed with wash buffer (Table 2.2) for 50 CV. The HisTrap column was attached to an ÄTKA® Start machine (GE Healthcare) for elution. Bound protein was eluted using four CVs of elution buffer (Table 2.2). Fractions with an absorbance at 280 nm peaking at 1000 mAU were collected. SEC was used for further purification using an ÄTKA® Pure machine (GE Healthcare) coupled with a HiLoad™ 16/60 Superdex™ 200 pg column (GE Healthcare). The column was equilibrated with SEC buffer (Table 2.2) before the IMAC fractions were loaded in. Fractions with an absorbance at 280 nm reaching >20 mAU were collected and concentrated using a Vivaspin® 6 MWCO 10kDa (GE Healthcare) protein concentrator spin column. Analysis at various stages of the purification was carried out using SDS-PAGE 4-20% gradient gels (Bio-Rad) (Figure 3.3). Protein concentrations were measured by absorbance at 280 nm. The extinction co-efficients of recombinant proteins was calculated using the ExPASy ProtParam tool (www.expasy.org).

2.5.3 Analytical SEC experiments

An analytical SEC column Superdex™ 200 increase 10/300 was used to analyse the purity and calculate the MW of the protein products and determine the oligomeric state. Blue dextran and protein standard mixture (ferritin, aldolase, ovalbumin, and DNase I) were separately injected on the column to determine the void volume (V_o) and standard elution volume (V_e), see Figure 2.3. The partition coefficient (Kav) values of the protein standards were calculated using the following equation: Kav = $(V_e - V_o)/(V_t - V_o)$, see Table 2.4. A $\log_{MW}$ vs. Kav graph was subsequently plotted to determine the line of best fit as shown in Figure 2.3.B. As a result, following equation curve allowed the calculation of the nature MW of the protein samples.

$Molecular\ weight\ (Da) = 10^{(b-(-m \cdot Kav))}$ whereby b = 6.0709 and m = -3.2206.

Table 2.4: Summary of SEC standards used for analytical SEC. Column used was a Superdex™ 200 increase 10/300 GL (GE Healthcare). Abbreviations: V_e, elution volume; MW, molecular weight; Kav, partition coefficient.

Protein Standards	V_e (ml)	MW (kDa)	Kav	LogMW
Blue dextran	8.75	2000	-	-
Ferritin	10.83	440	0.14	5.64
Aldolase	12.85	158	0.27	5.20
Ovalbumin	15.46	43	0.44	4.63
RNaseI	17.98	13.7	0.61	4.14

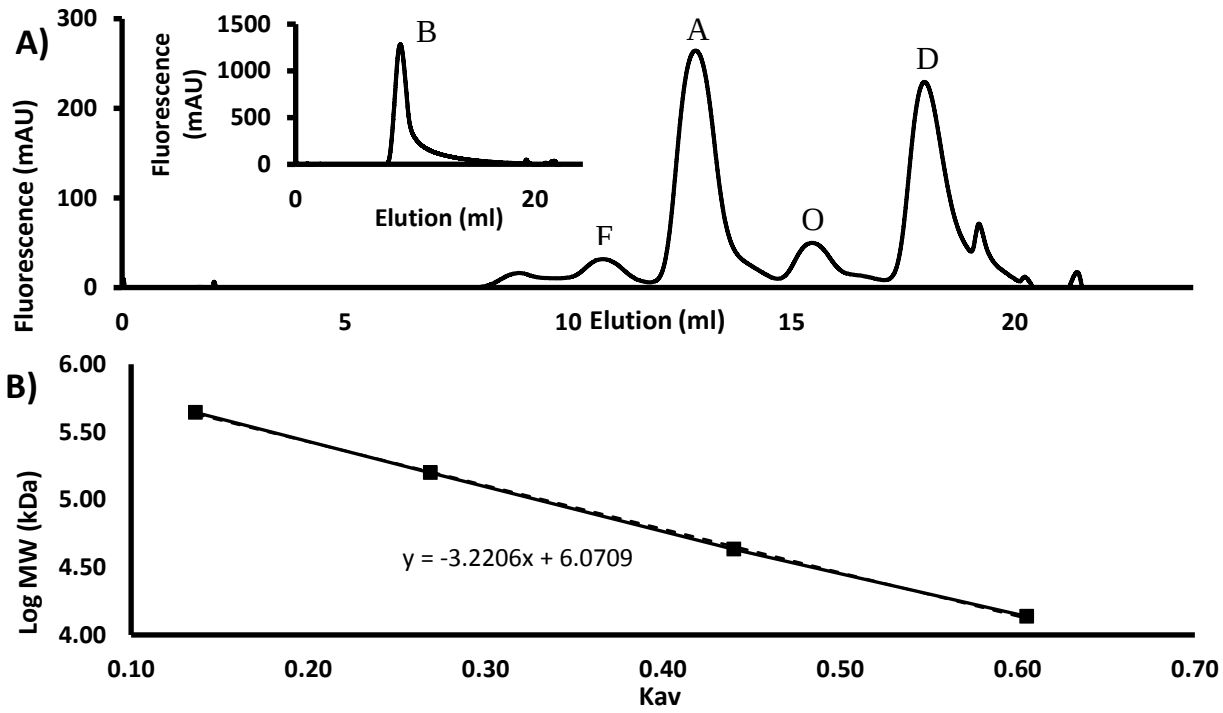

Figure 2.3: Analytical SEC run of protein standards and standard curve. A) Protein standards: blue dextran (2000 kDa, inset), ferritin (440 kDa), aldolase (158 kDa), ovalbumin (43 kDa), and DNase I (13.7 kDa). C) Standard curve with line of best fit (dashed line) and equation shown. All protein standards were run on a Superdex™ 200 increase 10/300 GL (GE Healthcare). Abbreviations: B, blue dextran; F, ferritin; A, aldolase; O, ovalbumin; D, DNase I; MW, molecular weight; Kav, partition coefficient.

2.6 Differential scanning fluorimetry

The Stratagene Mx3005P qPCR System (MxPro QPCR V4.10 software, Agilent Technologies) was used to perform the fluorimetry analysis. Samples were subjected to the heating cycle as described previously described[117]. A protein-dye-buffer mixture containing 7 mg/ml purified protein, SYPRO™ Orange dye (0.2 µL), and SEC buffer (Table 2.2) were added to a Mx3000P® Non-skirted 96-Well Plate (Agilent Technologies) and brought to a volume of 10 µL in water. A SB HR2-096 (Hampton Research) 96 well plate was used as a ligand screening plate. The master stock was diluted at a ratio of 1:10 before adding 10 µL to the protein-dye-buffer mixture. For the control reaction, 10 µL of DSF buffer (Table 2.2) was added to the protein-dye-buffer mixture. Final concentrations of the proteins were 3.5 mg/ml, the amino acids were 0.25% w/v, and the final volume of the reaction mixture was 20 µL. After Optical Caps (Agilent Technologies) were inserted into the wells, the plate was gently vortexed and directly analysed in the RT-PCR instrument. The heating cycle comprised of a 1.0°C increase per minute starting from 25°C ending at 95°C. Data were collected using the calibration setting for SYPRO™ Orange dye detection (λ_{ex} 300/472 nm; λ_{em} 570 nm) installed on the instrument.

The data were analysed as previously described[117]. Transform Agilent Mx3005p.xlsx, DSF Analysis v3.0.2.xlsx, and DSF GraphPad Boltzmann BasicGraphics v3.0.pzf files were utilised (ftp://ftp.sgc.ox.ac.uk/pub/biophysics). The raw data was exported as "horizontally grouped by plot" and transferred into the Transform Excel spreadsheet. The resulting output was placed in the 'Paste in transformed Data' tab in the DSF Analysis Excel spreadsheet. To calculate the Boltzmann equation, data from the 'Processed Data' tab was transferred to the DSF GraphPad file, using the GraphPad Prism v7.02 software. The Boltzmann equation results were transferred back into the 'Processed Data' tab, which then calculated the T_m of the purified target proteins. DSF analysis in triplicate was conducted on all three purified proteins SC634, SW840, and SC599 using the SB HR2-096 plate (Hampton Research) to screen for suitable ligands.

Chapter 3. Expression, purification, and characterisation of cyanobacterial PBPs

3.1 Bioinformatics analysis of *Synechococcus* AA-PBPs

Previous work carried out by the Paulsen group identified ~110 *Synechococcus* PBPs that putatively bind a wide variety of ligands. Three Cyanorak clusters (Cyanorak v1 cluster numbers: CK1309, CK1489, and CK2503) encompassing twelve *Synechococcus* PBP genes are annotated to bind AAs and are thus selected for this study. As shown in Table 3.1, cluster CK1309 genes annotated to bind carbohydrates and AAs, genes in cluster CK1489 are predicted to bind with acidic and neutral polar AAs, whereas cluster CK2503 contains only two PBP genes believed to bind branched-chain AAs based on their annotations. The number of cysteine residues (Table 3.1) also plays a role in the structural integrity of a protein as they form disulphide bridges. These bonds assist in the tertiary and possibly quaternary integrity of proteins.

Synechococcus strains from clades I-IV and CRD1 were identified to have putative AA-PBPs suggesting that both open ocean and coastal strains may possibly be involved in the active uptake of AAs. A BLASTp search was carried out on the putative *Synechococcus* AA-PBP genes to find homologous proteins from other species with structures in the Protein Data Bank (PDB). As shown in Table 3.1 these homologous proteins have an AA sequence identity ranging from 23-59%. The clusters CK1489 and CK2503 were homologous to a tripeptide glutathione (PDB code: 4Z9N) and a probable LIV-binding protein (PDB code: 4KV7), respectively. Compared to cluster CK1309 which has four different homologous proteins, with only one (PDB code: 1II5) that binds with an AA, Glu. The other three proteins (PDB code: 3MPL, 4Q0C, and 5D0U) are a signalling protein, transferase, and an RNA-helicase. SW606 had the highest identity hit (59%) with RNA-helicase from *Chaetomium thermophilum*, a thermophilic fungus (PDB code: 5D0U). Further investigation found a large size discrepancy between the two identified proteins, 8.77 kDa for SW606 and 81.53 kDa for 5D0U which is likely a frameshift mutation present in the genome sequence. Another protein, SynWH8102_00607, also had a lower than predicted MW (11.24 kDa) compared to its homolog (PDB code: 1II5) with a MW of 25.67 kDa. Due to these constraints, both genes are unlikely to produce a functional protein due to their low MW but were still kept in the pipeline. All other target proteins had MWs comparable with those ranging from 27.34 kDa to 42.32 kDa.

Table 3.1: Summary of the twelve target *Synechococcus* proteins. All *Synechococcus* proteins are selected from the Cyanorak database (abims.sb-roscoff.fr/cyanorak). Homologous proteins were retrieved from the Protein Data Bank (http://www.rcsb.org). Abbreviations: Cys, number of cysteine residues present; PDB, protein data bank; ID, identity match.

Cyanorak ORF ID	Cluster	Putative ligand	Clade	Signal P position	Cys	pI	Length (bp)	M.W. (kDa)	Homologous protein	
									PDB Code	ID%
CK_Syn_CC9311_02366 (SC2366)	CK1309	Carbohydrates/amino acids	Ia	22	3	4.93	918	30.22	3MPL	25
CK_Syn_CC9902_00599 (SC599)			IVa	27	2	9.26	858	27.34	4Q0C	23
CK_Syn_MITS9220_02624 (SM2624)			CRD1a	29	3	4.87	921	29.14	4Q0C	26
CK_Syn_WH8102_00606 (SW606)			IIIa	-	1	4.39	258	8.77	5D0U	59
CK_Syn_WH8102_00607 (SW607)			IIIa	21	3	4.56	384	11.24	1II5	35
CK_Syn_CC9311_01190 (SC1190)	CK1489	Acidic and neutral polar amino acids	Ia	-	5	8.78	1146	41.74	4Z9N	42
CK_Syn_CC9902_00847 (SC847)			IVa	41	5	5.47	1104	34.43	4Z9N	44
CK_Syn_MITS9220_01218 (SM1218)			CRD1	22	6	5.87	1053	35.56	4Z9N	42
CK_Syn_WH8102_00840 (SW840)			IIIa	27	5	5.07	1062	34.86	4Z9N	43
CK_Syn_WH8109_00877 (SW877)			IIa	30	5	5.87	1062	34.07	4Z9N	44
CK_Syn_CC9311_00634 (SC634)	CK2503	Branched-chain amino acids	Ia	65	0	9.62	1284	39.89	4KV7	33
CK_Syn_CC9311_02631 (SC2631)			Ia	-	0	7.04	1167	42.32	4KV7	31

3.2 Successful cloning of *Synechococcus* genes in *E. coli*

All twelve target protein genes were successfully amplified for cloning in expression vectors pET-15b and pET22-b. LIC was used to integrate these amplified gene products in cloning vectors. Colonies were successfully obtained for all constructs subjected to LIC cloning. PCR colony screening (six in total for each target) was utilised to screen for positive clones.

Figure 3.1 shows a representative DNA gel picture of the screening process for four constructs. As shown in the figure, at least one of the six colonies yielded a positive PCR result and, was subsequently confirmed by DNA sequencing. Thus all twelve gene targets were successfully amplified and cloned in a pET-15b (cytoplasmic-directed) and pET-22b (periplasmic-directed) expression systems.

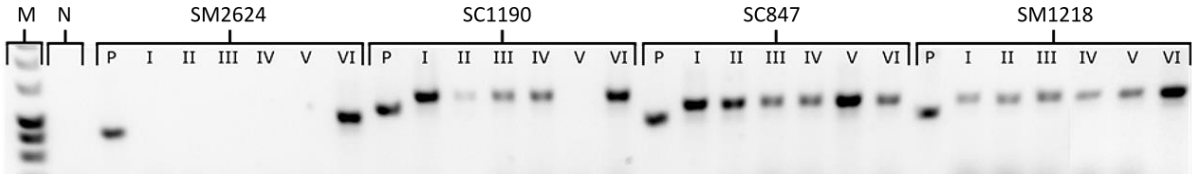

Figure 3.1: 0.8% Agarose gel depicting four pET-22b PCR colony screening. Numerals used to denote colony number. Abbreviations: M, marker; N, negative control; P, positive control.

3.3 Small-scale expression trials identify four soluble proteins

All pET-15b and pET-22b *E. coli* plasmids containing gene of interest were transformed into *E. coli* Rosetta™ 2 for protein expression. Rosetta™ 2 host strains supply tRNAs for 7 rare codons (AGA, AGG, AUA, CUA, GGA, CCC and CGG). The *E. coli* Rosetta™ 2 cells were used to enhance the expression of proteins that contains codons rarely used in *E.coli*[105].

Expression trials were conducted using 5 ml cultures to find out which induction method (IPTG or auto-induction) and vector (pET-15b or pET-22b) was better at producing soluble protein for purification. The location of protein translation, and the size and location of the terminal His-Tag region are the main differences between the proteins produced by each vector. The pET-15b expression construct encodes an N-terminal His-Tag (2.31 kDa), with truncated signal peptide for each gene construct, expressed proteins therefore remain within the cystol.

In contrast, the pET-22b expression construct encoding a C-terminal pET-22b His-Tag (2.09 kDa) encompasses a *pelB* leader sequence. As a result, the expressed protein products are translocated to the periplasm via the Sec system with subsequent cleavage of the leader sequence. For induction of recombinant protein expression, the isopropyl-β-D-thio-

galactoside (IPTG)-dependent T7 *lac*O promoter system[105] integrated within the vectors was utilised. IPTG induction of recombinant protein was carried out once the cells reached mid-to-late log phase. This ensured maximal yield whilst eluding issues pertaining to the stationary phase, most notably the generation of proteases[118]. The auto-induction protocol utilises ZYP-rich medium that contains α-lactose for induction.

The results of the expression trials are summarised in Table 3.2 and depicted in Figure 3.2. Briefly, nine of the twelve pET-15b protein constructs successfully expressed in both, IPTG and auto-induction, media. Comparably, eight pET-22b protein constructs expressed in IPTG medium and nine constructs showed expression in the auto-induction system. Except for protein construct SC2631 (~44kDa) expressing solely in pET-22b and SC634 (~42 kDa) in pET-15b, a very interesting correlation between the two vectors is observed.

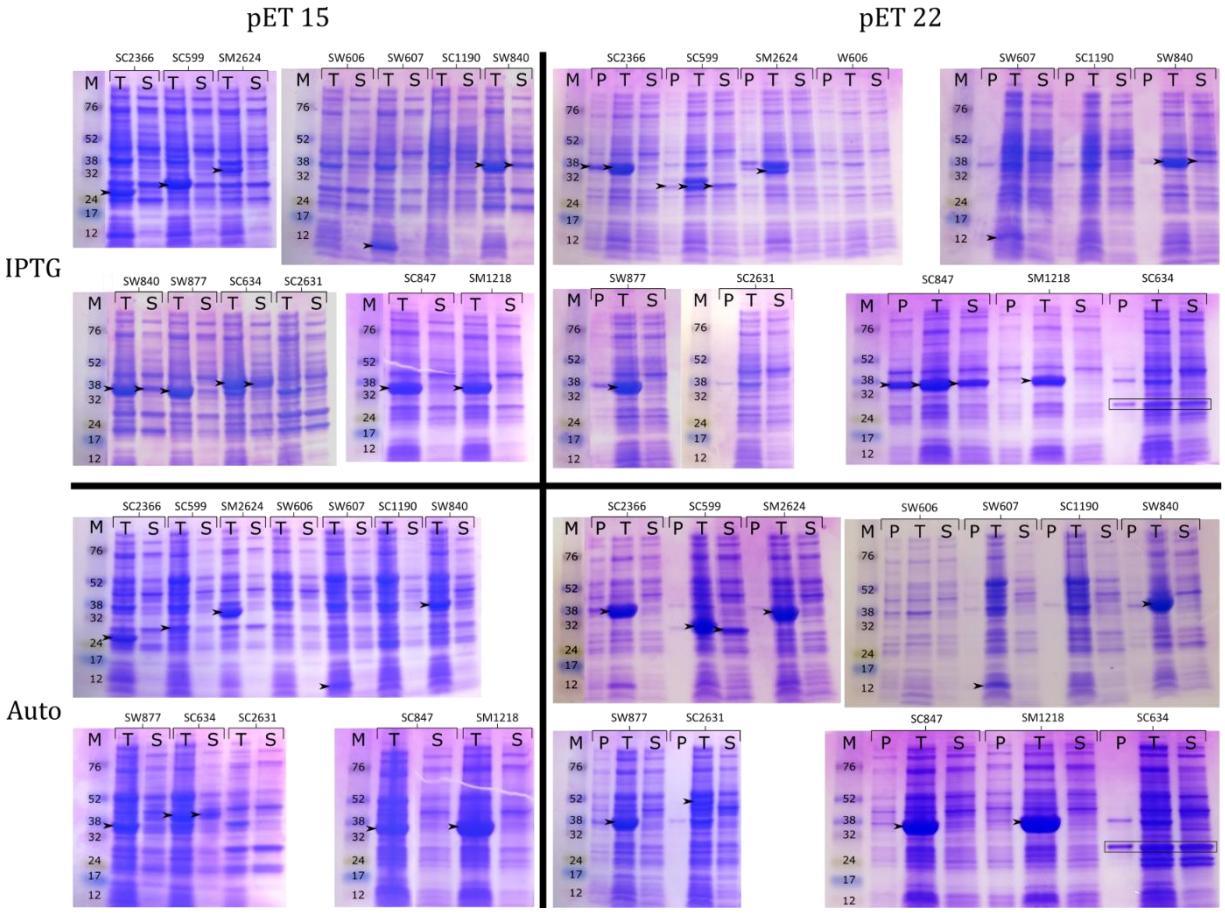

Figure 3.2: SDS-PAGE of small-scale expression trials. The figure is separated into quadrants based on the expression vector (pET-15b or pET-22b) and induction method (IPTG-induction or auto-induction). The black arrows denote the possibility of expressed recombinant protein. Four constructs expressed possible soluble recombinant protein: SW840, SC634, SC599, and SC847. Black box highlights breakdown product in pET-22b SC634. Sizes of standards are in kDa. Abbreviations: M, marker; T, total lysate fraction; S, soluble fraction; P, periplasmic fraction

Table 3.2: Results summary of small-scale expression trials. Detection of the recombinant protein in the total lysate, soluble, and periplasmic fractions for both induction methods are noted using symbols for each vector. Symbol used: • denotes pET-15b vector, ○ denotes pET-22b vector.

Protein Target	Cluster	IPTG-induction			Auto-induction		
		Lysate	Soluble	Periplasm	Lysate	Soluble	Periplasm
SC2366	CK1309	• ○		○	• ○		
SC599		• ○	○	○	• ○	○	
SM2624		• ○			• ○		
SW606							
SW607		• ○			• ○		
SC1190	CK1489						
SC847		• ○	○	○	• ○		
SM1218		• ○			• ○		
SW840		• ○	• ○		• ○		
SW877		• ○			• ○		
SC634	CK2503	•	•		•	•	
SC2631					○		

Out of nine proteins that were expressed within pET-15b constructs, two showed expression in the soluble fraction: SW840 (~37 kDa), through IPTG-induction; and SC634 (~42 kDa), through both IPTG and auto-induction methods. Of the two soluble proteins, SC634 sequence (Table 3.1) shows absence of cysteines indicating the reducing environment of the cystol would not have any effect on the protein folding.

Two pET-22b constructs, SC599 (~29 kDa) and SC847 (~36 kDa) showed expression in periplasmic and soluble fractions. Additionally, protein construct SW840 (~37 kDa) expressed in the soluble fraction but not in periplasmic fraction and protein construct SC2366 showed protein expression in the periplasmic fraction but not the soluble fraction. Notably, SC599 contained marginally smaller amounts of protein in the periplasmic fraction compared to the soluble fraction. It also appears that more protein was expressed using the auto-induction method as opposed to the IPTG-induction method.

The small-scale expression trials reveal no clear bias towards a specific expression strategy (i.e. *pelB* leader sequence or C- versus N-terminal His-Tag) or induction method. Since there is no way to predict beforehand which method is suited for each protein, this approach is an effective way to test different expression strategies. While the expression trials were carried out at 25°C for auto-induction and 18°C for IPTG-induction, reducing the temperature during protein expression is a well-known technique to help keep recombinant protein soluble[119,120]. Systematically examining this variable during small-scale expression trials could help optimise conditions to produce soluble recombinant protein.

Thus, from the small-scale expression trials, four protein constructs: SW840, SC634, SC599, and SC847, are identified as soluble. Based on the results from these trials, the conditions chosen for the large-scale expression of the selected four proteins were as follows: SW840, pET-15b using IPTG-induction; SC634, pET-15b using IPTG-induction; SC599, pET-22b using auto-induction; SC847, pET-22b using IPTG-induction.

3.4 Purification of target proteins

Large-scale expression of recombinant PBPs were conducted using 2L cultures. Three (SC634, SW840, and SC599) of the four proteins selected for large-scale expression were successfully purified using IMAC and SEC. During large-scale expression, SC847 was expressed but remained in the insoluble fraction. The SDS-PAGE analysis of these proteins at various stages of the purification process is depicted in Figure 3.3.

Protein construct SC847 selected for expression in the large-scale system however remained in the insoluble fraction presenting difficulties with the purification process and was therefore discontinued at this stage due to time constraints. Future work should include reduction in the growth temperature to help recombinant protein remain in the soluble fraction[119,120].

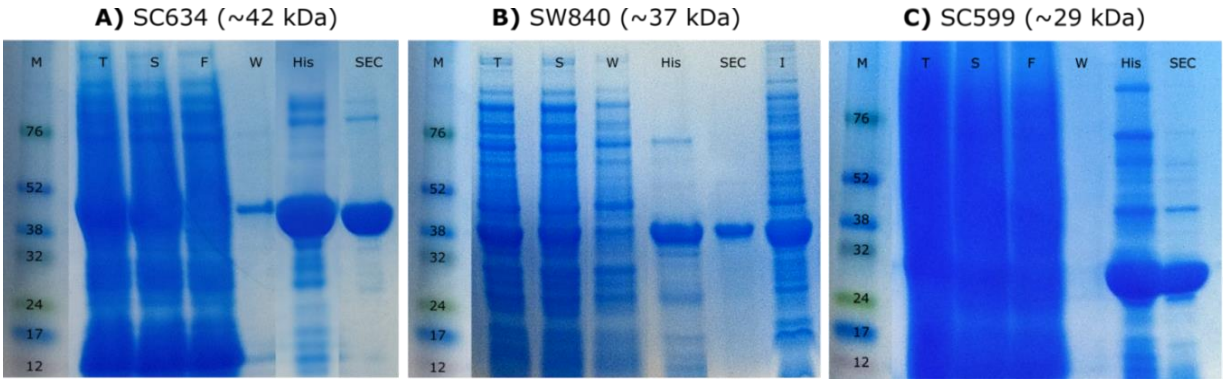

Figure 3.3: SDS-PAGE analysis at various stages of the purification process. A) CK_Syn_CC9311_00634 (SC634), B) CK_Syn_WH8102_00840 (SW840), and C) CK_Syn_CC9902_00599 (SC599). The MW of each target protein is shown for reference. Sizes of standards are in kDa. Abbreviations: M, molecular weight ladder; T, total lysate; S, soluble fraction; F, flowthrough fraction; W, wash fraction; His, eluted His-Trap fraction; SEC, collected high mAU SEC fractions; I, insoluble fraction.

3.4.1 Purification of CK_Syn_CC9311_00634 (SC634)

As depicted in Figure 3.3.A and Figure 3.4, protein construct SC634 (MW ~42 kDa) was successfully purified to homogeneity using IMAC and SEC steps. The preparative SEC trace showed a protein peak at V_e 85.70 ml (Figure 3.4.A). Fractions were taken from V_e 84 ml to 88 ml and concentrated to 12.61 mg/ml using a protein concentrator spin column. SDS-PAGE analysis of concentrated SC634 shows a thick single band at ~43 kDa mark (Figure 3.4.A, inset) indicating a highly pure protein product

To further confirm protein purity as well as determine oligomeric state of SC634 protein, an analytical SEC was carried out (Figure 3.4.B). The results showed a single peak at 1341.96 mAU at V_e 16.26 ml. The MW of SC634 was calculated to be ~30 kDa based on the standard SEC run (Figure 2.3) and equations given in section 2.8.3. This data suggests the protein exists in a monomeric form.

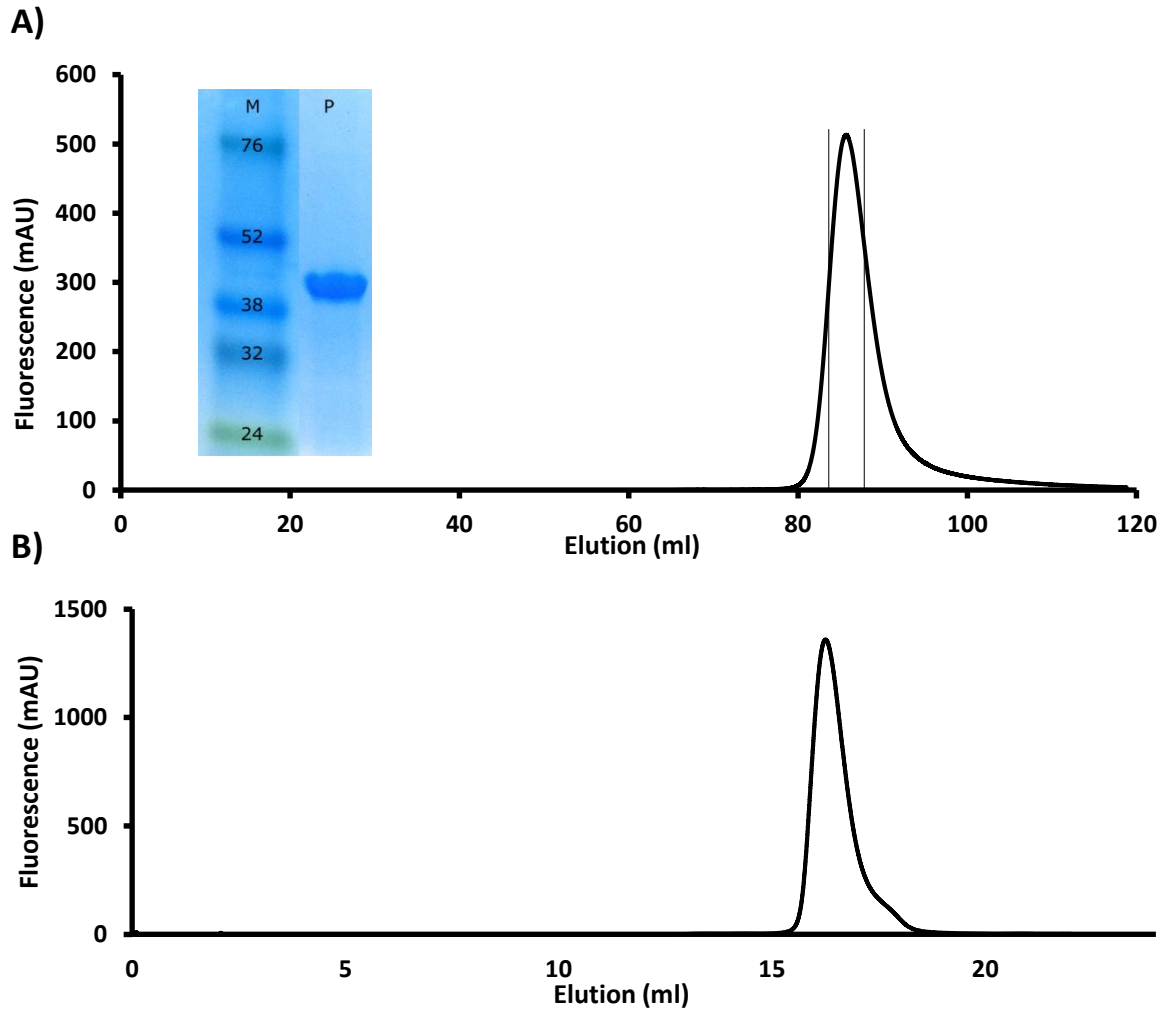

Figure 3.4: Purification of SC634. A) Inset shows SDS-PAGE analysis of purified SC634 taken from V_e 84 ml to 88 ml (black bars) and concentrated to 12.61 mg/ml. Preparative SEC column used was a HiLoad™ 16/60 Superdex™ 200 pg (GE Healthcare). B) Analytical SEC column used was a Superdex™ 200 increase 10/300 GL (GE Healthcare).Molecular weight marker is in kDa. Abbreviations: M, molecular weight marker; P, purified protein.

3.4.2 Purification of CK_Syn_WH8102_00840 (SW840)

SDS-PAGE analysis of recombinant protein expression and purification of SW840 (~37 kDa) is depicted in Figure 3.3.B. As shown in the figure, the wash step successfully removed non-specifically bound proteins. However, the 500 mM imidazole eluted fraction contained other contaminating proteins. IMAC fractions were therefore pooled for a second preparative SEC purification.

The preparative SEC trace is depicted in Figure 3.5. As shown in the SEC trace, the protein eluted between elution volume (V_e) 77 ml to 83 ml. To ensure collection of homogeneous protein material, fractions were taken from V_e 78 ml to 82 ml and concentrated to 6.30 mg/ml using a protein concentrator spin column. SDS-PAGE analysis of purified SW840 protein product shows a thick band at ~37 kDa mark (Figure 3.5, inset).

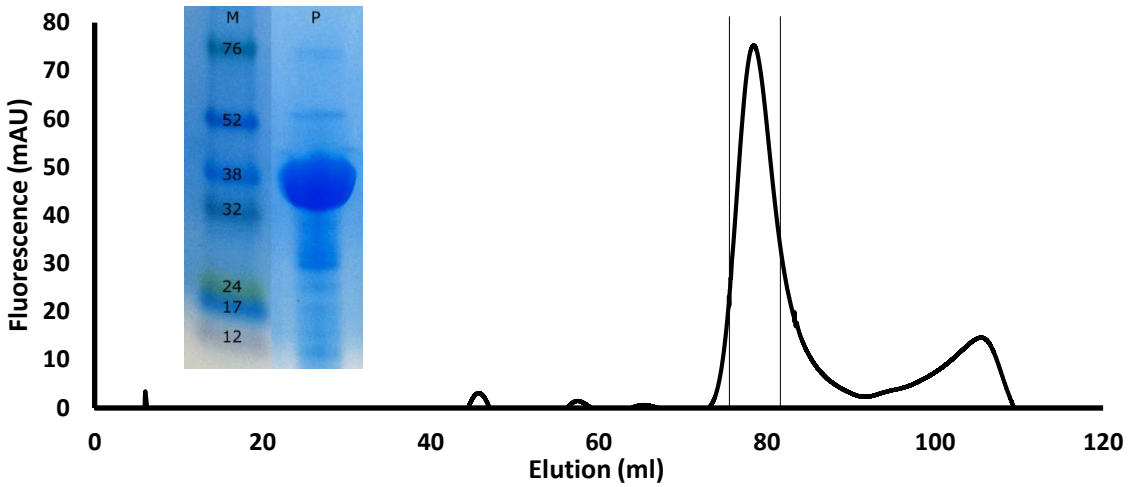

Figure 3.5: Preparative SEC trace and SDS-PAGE of SW840. Inset shows SDS-PAGE analysis of purified SW840 taken from V_e 76 ml to 82 ml (black bars) and concentrated to 6.30 mg/ml. Column used was a HiLoad™ 16/60 Superdex™ 200 pg (GE Healthcare). Molecular weight marker is in kDa. Abbreviations: M, molecular weight marker; P, purified protein.

3.4.3 Purification of CK_Syn_CC9902_00599 (SC599)

SC599 (~29 kDa) was the only protein taken to the large-scale purification to be induced through auto-induction. When eluted off the IMAC His-Trap column, several contaminant proteins eluted with the target proteins (Figure 3.3.C). Interestingly, the eluted protein product was observed to have a yellow tinge. The protein may have eluted with a bound ligand. Pyridoxal-5'-phosphate (PLP) may explain this observation as it is yellow in colour and is known to act as a cofactor of many enzymes of AA metabolism[121].

To further purify and analyse the sample, a preparative SEC step was initiated. Following preparative SEC of the high mAU His-Trap fractions, the trace showed a peak of 207.72 mAU at V_e 80.69 ml (Figure 3.6.A). Fractions were taken from V_e 78 ml to 84 ml and concentrated to 10.05 mg/ml. SDS-PAGE analysis of the concentrated protein showed a single band at ~30 kDa mark (Figure 3.6.A, inset). Furthermore, the SEC trace depicted additional peaks at the end of the preparative SEC run (Figure 3.6.A). These are suggestive of a very small molecule being eluted. It may be possible that SC599 was bound to a ligand that was released when passed through the preparative SEC column. This may also explain why the solution had a yellow colour prior to being loaded onto the preparative SEC column.

To confirm the purity and oligomeric form of SC599, an analytical SEC was carried out (Figure 3.6,B). The results showed two peaks eluting at V_e 16.47 ml and V_e 20.97 ml. With reference to a standard protein analytical SEC run (Figure 2.3) and equations given in section 2.8.3, the MW of SC599 was calculated to be 27.58 kDa (Table 3.3) indicating the protein exists in monomeric form nature.

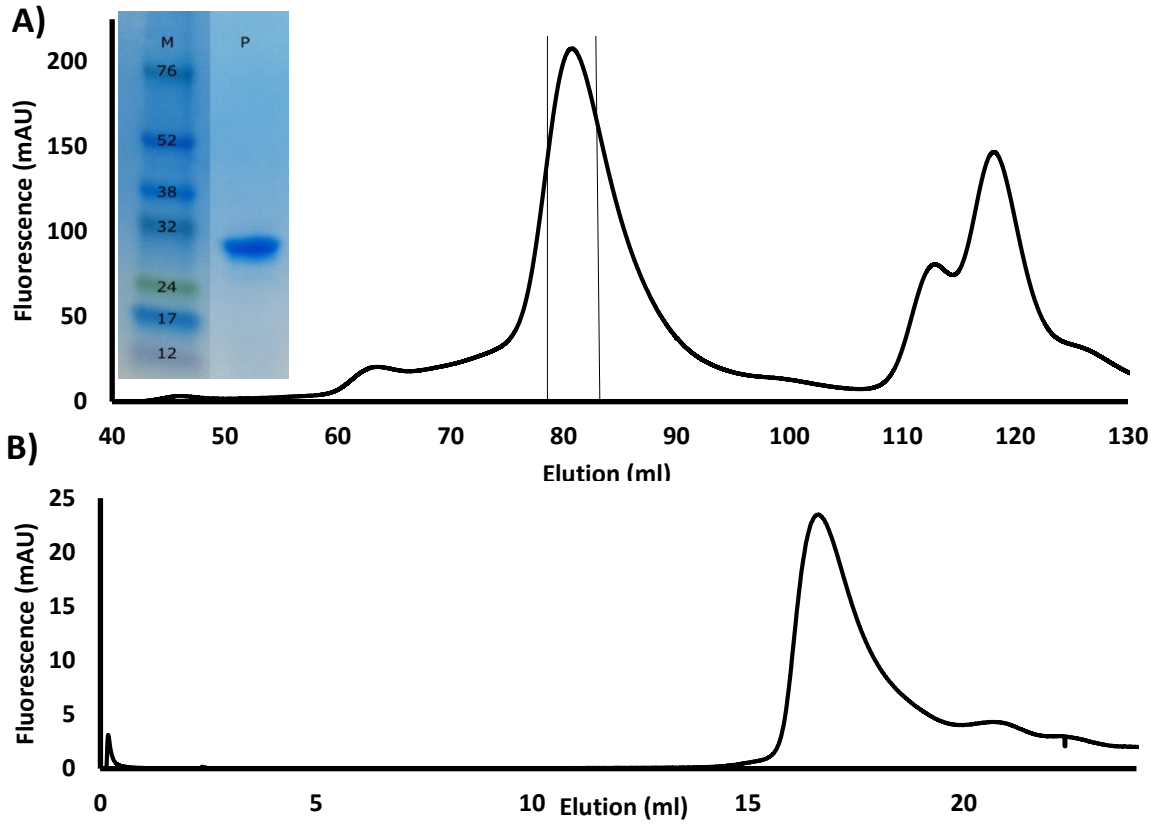

Figure 3.6: Protein purification of SC599. A) Inset shows SDS-PAGE analysis of purified SC599 taken from V_e 78 ml to 84 ml (black bars) and concentrated to 10.05 mg/ml. Preparative column used was a HiLoad™ 16/60 Superdex™ 200 pg (GE Healthcare). B) Analytical column used was a Superdex™ 200 increase 10/300 GL (GE Healthcare). Molecular weight marker is in kDa. Abbreviations: M, molecular weight marker; P, purified protein.

3.5 Ligand screening by differential scanning fluorimetry

Differential scanning fluorimetry (DSF) employs the use of a dye (typically SYPRO™ orange) that binds to the hydrophobic regions of a protein. A real-time polymerase chain reaction (RT-PCR) instrument is used to measure the fluorescence whilst the temperature slowly increases. As proteins reach their T_m, hydrophobic regions become exposed. The dye begins to bind to these regions that result in an increase in fluorescence that is detected by the RT-PCR instrument up to a maximum level when the protein is completely denatured. In presence of a suitable ligand, the T_m of the target protein should increase significantly due to the increased stability caused by the binding of the protein to its ligand compound[117]. For efficient screening of the protein, we utilised a commercially available 'Silver Bullets' (SB) HR2-096 (Hampton Research) 96-well plate. Each well with the screen contain a cocktail of 2-20 small molecules (average 5-6 compounds/well) statistically known to bind proteins. Thus, ~1100 compounds were screened for each protein sample, including all 20 standard AAs.

3.5.1 DSF analysis of SC634 (CK2503) reveals potential binding with formate and formaldehyde

DSF analysis on purified SC634 was carried out in triplicates to probe for a suitable ligand(s) using SB HR2-096 plate for screening. A summary of the DSF analysis results of SC634 with SB HR2-096 is shown below in Figure 3.7 and Table 3.3.

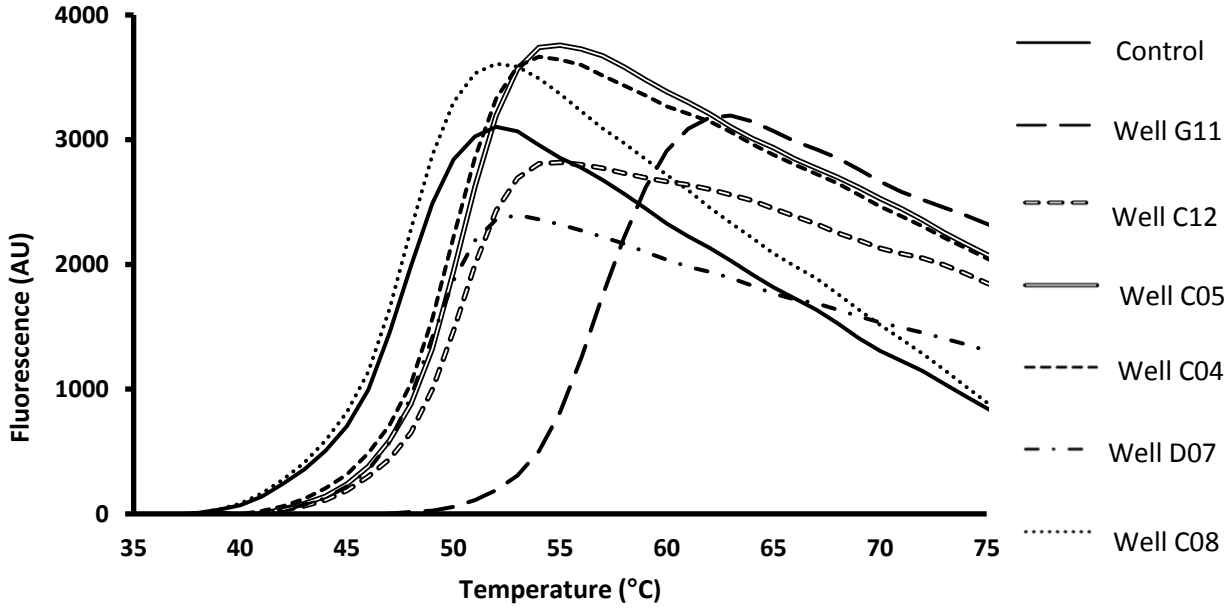

Figure 3.7: DSF analysis of SC634 with SB HR2-096. Well components and T_m is shown in Table 3.3.

Table 3.3: Summary of SC634 DSF results with SB HR2-096. Wells that caused an increase in T_m are shown including well C08, which contained all 20 standard amino acids except cysteine. All wells are buffered with 0.02M HEPES pH 6.8. Abbreviations: Con, negative control well.

Well	Components	Average T_m(°C)	ΔT_m (°C)
Con	0.02M HEPES sodium pH 6.8	47	-
G11	10% v/v - Tacsimate™ pH 7.0 (1.8305 M Malonic acid, 0.25 M Ammonium citrate tribasic, 0.12 M Succinic acid, 0.3 M DL-Malic acid, 0.4 M Sodium acetate trihydrate, 0.5 M Sodium formate, and 0.16 M Ammonium tartrate dibasic)	57	10
C12	0.25% w/v 5-Sulphoisophthalic acid monosodium salt, Anthraquinone-2,6-disulphonic acid disodium salt, N-(2-acetamido)-2-aminoethanesulphonic acid, Tetrahydroxy-1,4-benzoquinone hydrate	50	3
C05	0.16% w/v Ala-ala, Aspartame, Gly-tyr, Leu-gly-gly, Ser-glu, Tyr-ala	50	3
C07	0.16% w/v Ala-ala, Gly-asp, Gly-gly, Gly-phe, Gly-ser, Ser-tyr	50	3
C04	0.2% w/v Aspartame, Gly-asp, Gly-ser, Ser-tyr, Tyr-phe	49	2
C08	0.05% w/v L-Ala, L-Arg, L-Asn monohydrate, L-Asp, L-Gln, L-Glu, Gly, L-His, L-Ile, L-Leu, L-(+)- Lys, L-Met, L-Phe, L-Pro, L-Ser, L-(-)-Thr, L-Trp, L-Tyr, L-Val	47	0

The highest ΔT_m was caused by well G11 (contains Tacsimate™) with an increase of ~10°C (Figure 3.7). Tacsimate™ is typically used as a protein crystallisation solution that contains 1.8305 M Malonic acid, 0.25 M Ammonium citrate tribasic, 0.12 M Succinic acid, 0.3 M DL-Malic acid, 0.4 M Sodium acetate trihydrate, 0.5 M Sodium formate, and 0.16 M Ammonium tartrate dibasic. Each of these individual components were investigated further using DSF (Figure 3.8). A summary of the DSF analysis results of SC634 with individual Tacsimate™ reagents is shown in Table 3.4. The only compound that caused a shift in T_m was sodium formate, by ~11°C. Since sodium formate is a single-carbon (C1) compound, two other C1 compounds (methanol and formaldehyde) were trialled. While, methanol did not change the T_m of SC634, formaldehyde showed an increase by ~6°C (Table 3.4). DSF analysis thus indicates that the SC634 protein binds with C1 compounds, such as sodium formate and formaldehyde.

Furthermore, four other wells displayed an increase in the T_m of SC634 (Figure 3.7). Well C12 (components in Table 3.3) increased the T_m by ~3°C. Wells C04, C05, and C07 all contained dipeptides and increased the T_m of SC634 by ~2-3°C (Table 3.3). If dipeptides are binding to SC634, then it may be a possible periplasmic oligopeptide-binding protein OppA of an oligopeptide permease. OppA in found in *Salmonella typhimurium* binds peptides of two to five amino acids with no regard to sequence, yet excludes single AAs[101]. Although single AAs did not change the T_m of SC634, dipeptides are still a possible ligand. However, due to only a small ΔT_m and no ΔT_m observed in other wells containing oligopeptides (A3, C3, C6, E6, E7, F5, H2, and H12) it is unlikely that oligopeptides are the binding partner of SC634. Notably, SC634 gene product belongs to a cluster CK2503 annotated to bind with branched-chain AAs. However, there was no ΔT_m in well C08 containing all 20 standard amino acids except Cysteine (Table 3.3).

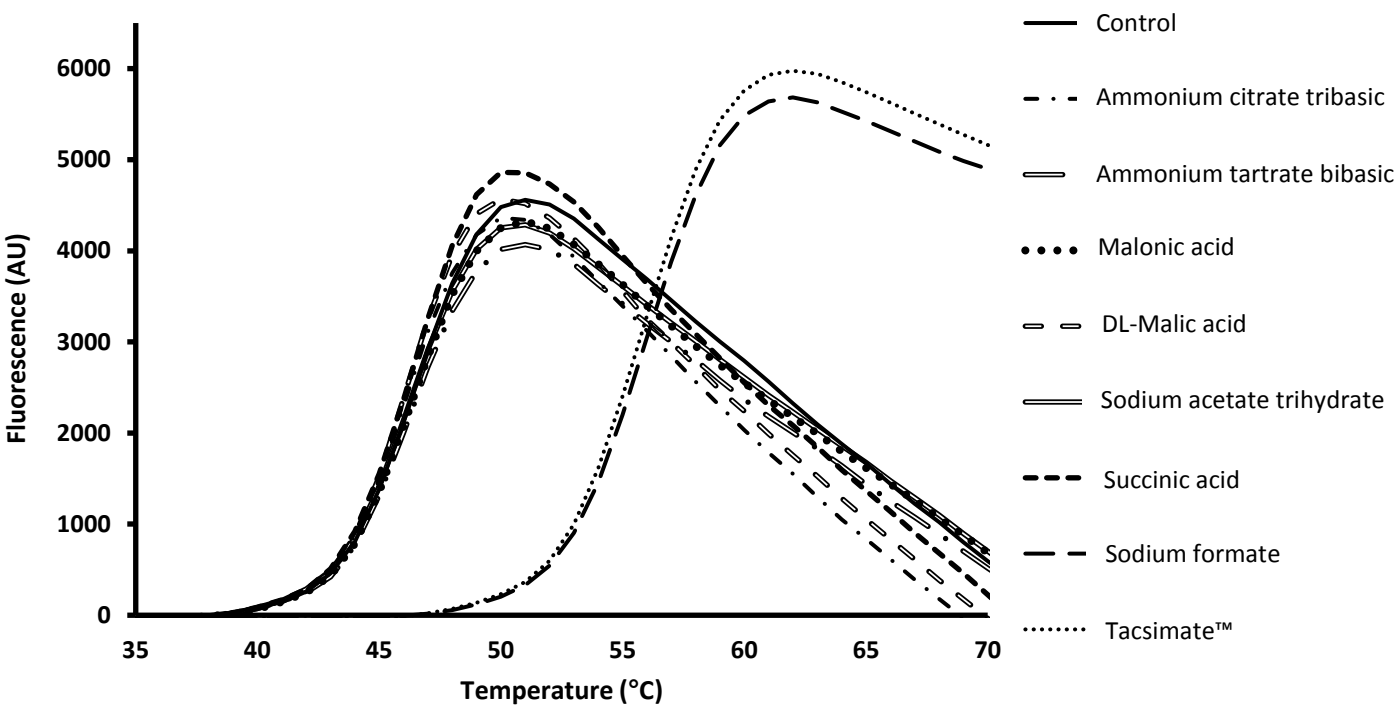

Figure 3.8: DSF analysis of SC634 with individual Tacsimate™ components. Control well just contains buffer (0.02M HEPES pH 6.8). T_m is shown in Table 3.4.

Table 3.4: Summary of SC634 DSF results with individual Tacsimate™ components and some additional C1 compounds, formaldehyde and methanol. All wells are buffered with 0.02M HEPES pH 6.8. Tacsimate™ was also used as a positive control. Abbreviations: Con, negative control well.

Well	Components	Average T_m(°C)	ΔT_m (°C)
Con	0.02M HEPES sodium pH 6.8	46	-
	10% v/v - Tacsimate™ pH 7.0 (1.8305 M Malonic acid, 0.25 M Ammonium citrate tribasic, 0.12 M Succinic acid, 0.3 M DL-Malic acid, 0.4 M Sodium acetate trihydrate, 0.5 M Sodium formate, and 0.16 M Ammonium tartrate dibasic)	55	9
	0.5 M Sodium formate	56	10
	0.12 M Succinic acid	46	0
	0.3 M DL-Malic acid	46	0
	0.25 M Ammonium citrate tribasic	46	0
	0.16 M Ammonium tartrate bibasic	46	0
	1.8305 M Malonic acid	46	0
	0.4 M Sodium acetate trihydrate	46	0
	0.5 M Formaldehyde	52	6
	0.5 M Methanol	46	0

3.5.2 DSF analysis of SW840 (CK1489) shows potential binding with amino acids

DSF analysis on purified SW840 was carried out in triplicate to probe for a suitable ligand (Figure 3.9). The control well comprised buffer solution, 0.02M HEPES pH 6.8. Seven wells displayed a positive ΔT_m in SW840 and all of them contained single AAs and/or oligopeptides (see Table 3.5). Wells D08 and C08 displayed the highest increase in ΔT_m, ~17°C and ~16°C respectively. Well D08 contains two AAs, L-Glu and L-Pro, while well C08 encompasses all 20 standard AAs except Cysteine. L-Cysteine is present in the SB HR2-096 wells C1 and H11, however no significant ΔT_m was observed (Table 3.5). This makes L-Cysteine an unlikely substrate. Wells E03, E06, and E07 all contained an assortment of digested peptides and increased the T_m of SW840 by ~12°C, ~7°C, and ~13°C, respectively. Well F04 contained L-Arg and other organic compounds, mainly non-proteinogenic AAs or AA derivatives, and increased the T_m of SW840 by ~5°C. These organic compounds may be contributing to the stability of SW840 due to similarities in ligand chemical properties and structure. Finally, well B03 contained several organic compounds and increased the T_m by ~5°C.

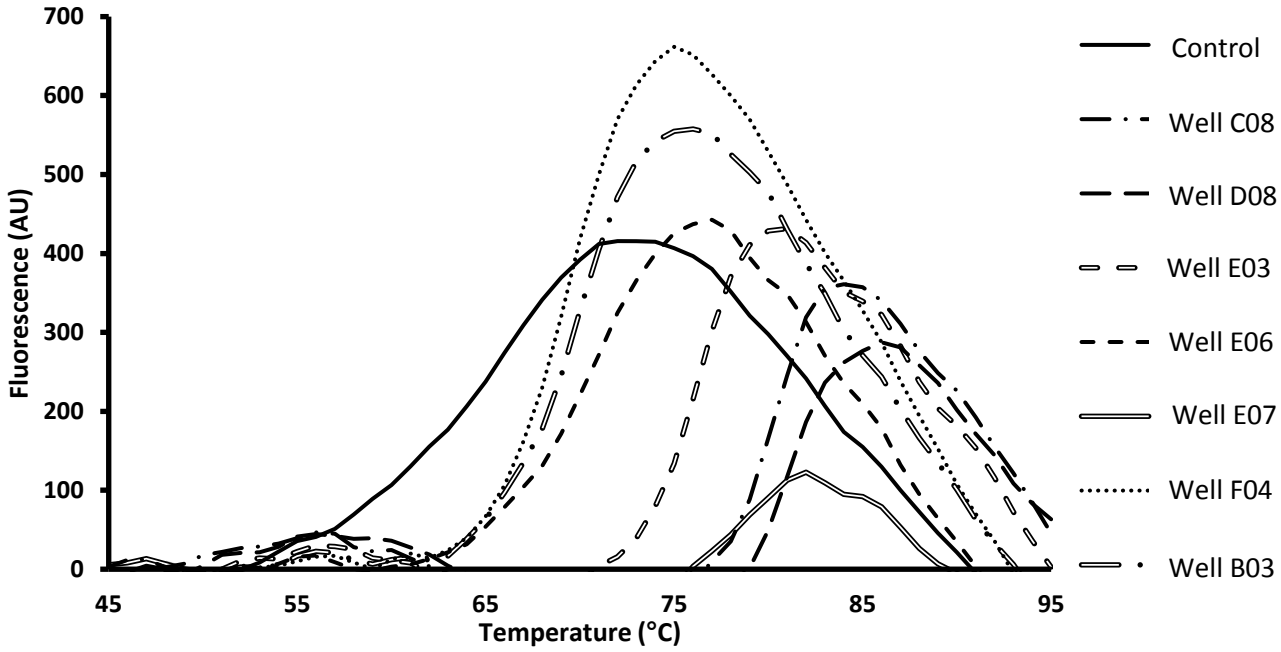

Figure 3.9: Positive ΔT_m DSF results of SW840 with SB HR2-096. Well components and T_m are shown in Table 3.5.

Table 3.5: Summary of SW840 DSF results with SB HR2-096. All wells are buffered with 0.02M HEPES pH 6.8. Abbreviations: Con, negative control well. Well A9* is unreliable due to a non-sigmodal DSF curve.

Well	Components	Average T_m(°C)	ΔT_m (°C)
Con	0.02M HEPES sodium pH 6.8	64	-
D08	0.2% w/v Betaine anhydrous, L-Glu, L-Pro, Taurine, Trimethylamine N-oxide dehydrate	81	17
C08	0.05% w/v L-Ala, L-Arg, L-Asn monohydrate, L-Asp, L-Gln, L-Glu, Gly, L-His, L-Ile, L-Leu, L-(+)- Lys, L-Met, L-Phe, L-Pro, L-Ser, L-(-)-Thr, L-Trp, L-Tyr, L-Val	80	16
E07	1% w/v Ovalbumin, 0.005% w/v Pepsin, Proteinase K, Trypsin	77	13
E03	1% w/v Tryptone	76	12
E06	0.5% w/v Casein, Haemoglobin, 0.005% w/v Pepsin, Proteinase K, Trypsin	71	7
F04	0.2% w/v L-Arg, L-Canavanine, L-Carnitine hydrochloride, L-Citrulline, Taurine	69	5
B03	0.25% w/v 5-Sulphoisophthalic acid monosodium salt, Cystathionine, Dithioerythritol, L-Citrulline	68	4
H11	0.07% w/v Barbituric acid, Benzidine, Cystathionine, L-Canavanine, L-Carnitine hydrochloride, L-Cystine, Mellitic acid	64	0
C01	0.25% w/v Benzamidine hydrochloride, L-Carnitine hydrochloride, L-Cystine, L-Ornthine hydrochloride	63	-1
A9*	0.16% w/v L-His, L-Ile, L-Leu, L-Phe, L-Typ, L-Tyr	32	-32

The organic AA compounds may be contributing to the stability of SW840 due to similarities in ligand chemical properties and structure. L-Carnitine and L-canavanine are unlikely since they are present in well H11 which conferred no ΔT_m. Hence, the components that are likely stabilising SW840 in well F04 are taurine (an organic sulphonic acid) and L-citrulline (an α-AA) as they are present in other wells that caused an increase in T_m, namely wells D08 and B03 respectively. While taurine it is not considered an AA, it does share chemical similarities in the form of amino and carboxylic acid groups. Both taurine and L-citrulline have low pKa's, <0 and 2.508, respectively and are likely able to bind and stabilise SW840 due to chemical similarities with acidic AAs. Further DSF investigation in the presence of the 20 standard amino acids is required to confirm specific AAs binding SW840. Well A9 was an interesting outlier as it contained six AAs: L-His, L-Ile, L-Leu, L-Phe, L-Trp, L-Tyr and seemed to decrease the T_m of SW840 by ~32°C (Table 3.5). Further probing revealed that the melt curve of the A9 well proved to be non-sigmoidal and is therefore unreliable.

3.5.3 DSF results of SC599 (CK1309) shows that further probing is required

DSF analysis on purified SC599 was carried out in triplicate to probe for a suitable ligand. The results with highest change in T_m are depicted in Figure 3.10. SC599 protein is classified within a cluster, CK1309, annotated to bind carbohydrates and AAs. Two wells, A11 and E02, showed increase in T_m of SC599 by ~2°C (Table 3.6) indicating increased protein stability. Components within well A11 seem unlikely the be the ligand. Well E02, however, does contain AAs and oligopeptides that SC599 could bind. However, with such a low change in T_m it is difficult to determine if it does bind with oligopeptides. Nonetheless, single AAs are unlikely binding partners for SC599 due to no ΔT_m within well C08 (Table 3.6). Further DSF investigation in the presence of individual components of well A11 and well E02 is warranted.

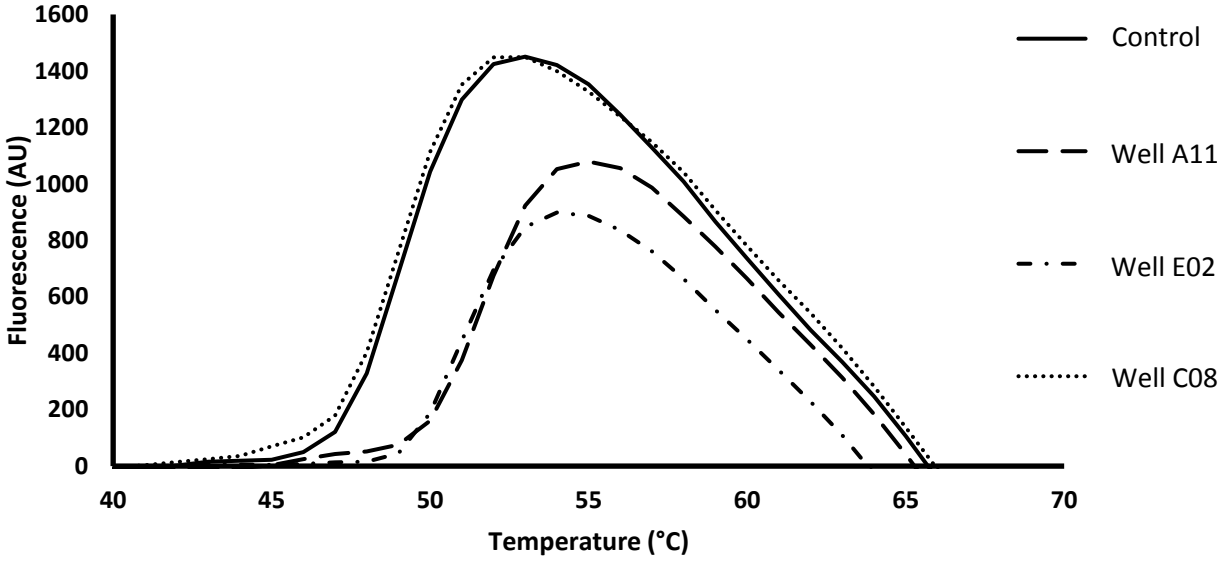

Figure 3.10: DSF of SC599 with SB HR2-096. Well components and T_m are shown in Table 3.6.

Table 3.6: Summary of SC599 DSF results with SB HR2-096. All wells are buffered with 0.02M HEPES pH 6.8. Abbreviations: Con, negative control well.

Well	Components	Average T_m(°C)	ΔT_m (°C)
Con	0.02M HEPES sodium pH 6.8	49	-
A11	0.33% w/v 2,5-Pyridinedicarboxylic acid, 4-Nitrobenzoic acid, Mellitic acid	51	2
E02	1% w/v Dextran sulphate sodium salt, 0.005% w/v Dextranase, α-Amylase	51	2
C08	0.05% w/v L-Ala, L-Arg, L-Asn monohydrate, L-Asp, L-Gln, L-Glu, Gly, L-His, L-Ile, L-Leu, L-(+)- Lys, L-Met, L-Phe, L-Pro, L-Ser, L-(-)-Thr, L-Trp, L-Tyr, L-Val	49	0

Chapter 4. Discussion

4.1 Strategies for characterising gene functions

To determine an optimal method for soluble recombinant protein expression, various expression strategies are utilised in this study. Twelve *Synechococcus* gene targets selected from the Cyanorak database (Table 3.1) were cloned into *E. coli* expression hosts with a 100% success rate. From the gene products, ~75% showed successful recombinant expression in trials with both IPTG- and auto-induction methods and both expression vectors, pET-15b and pET-22b, except for one gene-product that was only expressed in pET-15b (Table 3.2). Out of those that expressed protein, ~44% were in the soluble fraction: SC599, pET-22b in either induction method; SC847, pET-22b via IPTG-induction; SW840, both vectors using IPTG-induction; and SC634, pET-15b in either induction method. Three proteins were successfully purified, two using the pET-15b vector coupled with IPTG-induction (SW840 and SC634) and one using the pET-22b vector via auto-induction (SC599). These results showed that trialling for both vectors and induction methods was a sound strategy since some constructs only expressed soluble protein with one vector and/or induction method.

Notably, a high percentage of expressed proteins were observed in the insoluble fraction indicating presence of inclusion bodies. As previously mentioned (section 3.4), lowering the expression temperature further may help improve the solubility of other recombinant proteins[119,120]. Other methods involve re-solubilising inclusion bodies to recover bioactive proteins and co-expressing the target protein with the *E. coli* maltose binding protein[120]. Inclusion bodies may maintain biological activity when expressed at low temperatures[122,123]. A typical procedure involves four steps: [1] isolation of purified inclusion bodies, [2] solubilisation of inclusion bodies, [3] refolding of solubilised proteins, and [4] purification of refolded proteins by various chromatographic techniques[123]. Moreover, an increase in the concentration of salt may also allow the proteins to be eluted in the soluble fraction.

4.2 Purified SC634 (cluster CK2503) found to possibly bind C1 compounds

Proteins classified within cluster CK2503 are annotated to putatively bind branched-chain amino acids (such as Leu, Ile, and Val) in the Cyanorak database. I successfully

purified a protein, SC634 (section 3.4.1) enclosed in this cluster using a gene isolated from *Synechococcus* strain CC9311. Interestingly, as shown in section 3.5.1, SC634 protein showed a significant increase in the thermal stability in presence of two C1 compounds, formate and formaldehyde. The protein is predicted to bind branched amino acids, however DSF results show no binding evidence for any single AAs and only limited evidence is observed for dipeptide binding. Stabilisation of the protein truly occurred and was not just a result of the crystallisation liquor as this behaviour was not observed in the other two proteins investigated.

A search in the PDB, to identify a characterised branched-chain AA binding protein detected a SH1 protein isolated from *Rhodopirellula baltica* (PDB code: 4KV7). The SH1 protein from *R. baltica* is also predicted to bind branched amino acids and exhibits 33% sequence similarity to SC634 protein. However, a closer inspection of the crystal structure (see Figure 4.1) indicates the protein crystallised with formic acid bound. This interaction is stabilised by three residues, Tyr 101, Thr 128, and Tyr 171. It is currently unclear whether the interaction in SH1 is a result of protein-substrate binding or if formic acid is simply promoting stable protein solution[124]. Remarkably, the three amino acids interacting with formic acid in SH1 are highly conserved across ~95% of the proteins enclosed within cluster CK2503 (Figure 4.2). Future work involving site-directed mutagenesis on these residues would provide more evidence on the formate binding property of SC634.

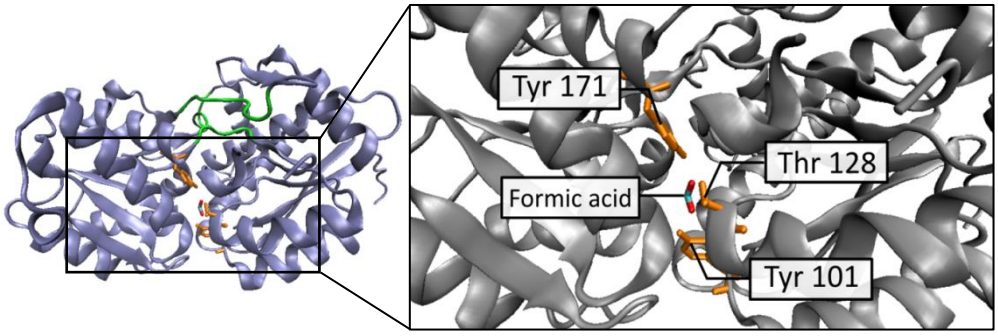

Figure 4.1: 3D crystal structure of the SH1 protein (PDB code: 4KV7) with the active site enlarged. Formic acid (cyan and red) is shown interacting with three active site residues (orange), Tyr 101, Thr 128 and Tyr 171. Three strands (green) connect the two domains indicating that it is part of the cluster B PBPs.

Figure 4.2: The conserved regions of 19 proteins in CK2503 and the SH1 protein (PDB code: 4KV7). The three conserved residues (Tyr 101, Thr 128 and Tyr 171) are highlighted by a black box. Sequences were aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) and the sequence logo was generated using WebLogo (http://weblogo.berkeley.edu/logo.cgi).

The biological importance of formate and formaldehyde can be traced to C1 metabolism via the ribulose monophosphate (RuMP), serine cycle, and ribulose bisphosphate (RuBP) pathways (Figure 4.3)[125]. Oxidation of ^{14}C-labelled formate was observed by natural microbial communities from the SAR 11 clade from the Sargasso Sea[126]. The OM43 clade, commonly found in coastal ecosystems, is also shown to grow on C1 compounds such as methanol, formaldehyde, formate, and methylamine[127].

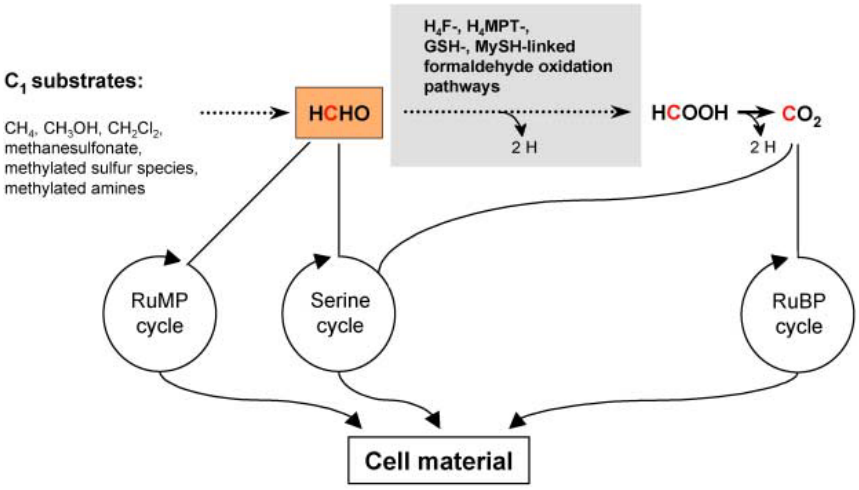

Figure 4.3: C1 substrate metabolic pathways. Taken from Vorholt (2002)[125].

A C1 compound (e.g. methanol) destined for the C1 metabolic pathways shown above (Figure 4.3) is oxidised three times to generate formaldehyde, formic acid, and CO_2. These processes are carried out by the enzymes methanol dehydrogenase, formaldehyde dehydrogenase, and formate dehydrogenase, respectively. The Cyanorak database revealed genes annotated to encode these enzymes in cyanobacteria: a possible methanol dehydrogenase beta subunit cluster CK3860 found in five strains of *Prochlorococcus* (MIT0701, MIT0702, MIT0703, MIT9303, MIT9313), two formaldehyde dehydrogenase genes found in *Cyanobium* sp. PCC700, and a glutathione-dependent formaldehyde dehydrogenase (cluster: CK50658) in *Synechococcus* RS9917 and WH7803.

However, all these annotations are putative and none are present within strain CC9311. Instead, a formate/nitrate transporter gene in CC9311 (CK_Syn_CC9311_02899) was identified in the Cyanorak database. This gene is sandwiched between a ferredoxin/nitrate reductase (SC112898), and a possible homeobox domain (SC2900). This operon is likely associated with the ferredoxin metabolic system as the ferredoxin/nitrate reductase enzyme is used to catalyse the following chemical reaction[128]:

$$Nitrite + H_2O + 2\ oxidised\ ferredoxin \leftrightharpoons nitrate + 2\ reduced\ ferredoxin + 2\ H^+$$

There are 20 genes associated with ferredoxin in CC9311 (Cyanorak), thus the use of this pathway is likely and formate may play role in this pathway.

Future work should include an isothermal titration calorimetry (ITC) assay and surface plasmon resonance (SPR) with SC634 protein in the presence of formate and formaldehyde to understand the binding affinity kinetics of these ligands. In addition, experiments involving uptake assays in cultures, potentially coupled with gene knockouts of cluster CK2503 in *Synechococcus* would also provide quantitative information about the transport rates of the transport system coupled to this PBP cluster *in vivo*. Two-hybrid screening and pull down methods could also be used to detect any protein-protein interactions, for example, to identify which other ABC transport protein components interact with this PBP to form an active transporter. Furthermore, heavy isotopes of formate can also be used to inoculate *Synechococcus* sp. CC9311 to visualise any possible transport with the use of a nanoscale secondary ion mass spectrometry (nanoSIMS). Formate uptake assays in whole cells could be undertaken with *Synechococcus* cultures with the addition of radiolabelled formate, followed by filtration and scintillation counting. The active site of 4KV7 (Figure 4.1) also shows binding with formic acid via 3 residues (Tyr 101, Thr 128, and Tyr 171). These residues were found to be highly conserved within the CK2503 cluster, and site-directed mutagenesis on the equivalent three residues may provide additional binding information.

4.3 SW840 (cluster CK1489) binds amino acids

The CK1489 cluster is annotated to putitavely bind with acidic and neutral polar AAs. The SW840 protein from this cluster was successfully purified (section 3.4.2) from *Synechococcus* strain WH8102. DSF analysis revealed a significant rise in the thermal stability of SW840 within a well containing 19 standard AAs (~16°C) and another well containing L-Glu (~17°C) (section 3.5.2). These results show promise that CK1489 is correctly annotated to bind with a range of acidic and neutral polar AAs.

The protein products classified within this cluster is manually annotated to be homologous with the *natF* protein from *Anabaena* strain PCC 7120, a filamentous cyanobacteria[129]. The rate of acidic AA uptake in *Synechococcus* has been observed to be similar in regard to other cyanobacteria such as *Anabaena* and *Nostoc*[82]. The AA uptake rates and the associated genes in *Anabaena* strain PCC 7120 has been closely investigated[129-133], providing important pointers into CK1489 protein function.

At least five transporters have been identified to carry out AA uptake systems in *Anabaena* sp. PCC 7120, namely, a high-affinity active transporter (basic AA), a low-affinity passive system (basic AA), an acidic AA transporter, and the ABC transporters N-I (hydrophobic AAs) and N-II (acidic and neutral polar AAs)[129,133,134]. The two ABC transporters, N-I and N-II, are part of the HAAT and PAAT family, respectively[129]. Pernil *et al*., (2008)[129] studied the uptake rate of mutant strains of *Anabaena* strain PCC 7120. Inactivation of genes associated with the *natF* gene caused an increase in the uptake of proline and hydrophobic AAs. Whereas inactivation of the N-I transporter (*natA* and *natB*) allowed increased transport of N-II AAs such as Asp and Glu. The *natF* gene is thus likely to bind acidic AAs and mediate its passage, in conjunction with an ABC transporter, into the cell. The N-II transporter is predicted to be a product of 4 genes, the PBP (*natF*), two transmembrane proteins (*natG* and *natH*), and an ATPase subunit (*bgtA*).

The purified protein SW840 is encoded within *Synechococcus* strain WH8102. Further investigation into the genome revealed that the SW840 gene is located within an operon (Figure 4.4). Upstream from SW840 are SYNW838 and SYNW839. These are annotated as a hypothetical FAD domain-containing protein and a putative sulphate permease, respectively. Downstream from SW840 (*natF*) are three genes that code for two N-II transmembrane domains (*natG* and *natH*), and a bgtA-like ATPase subunit. Analogous to *Anabaena* strain PCC 7120 described above, SW840 protein is highly likely to be the PBP component of a *Synechococcus* N-II ABC transporter.

Once the specific AAs that are binding with SW840 are defined, the binding affinity of individual ligands can be investigated using ITC. Gene-knockouts and subsequent substrate uptake assays will also provide more evidence that this cluster is indeed transporting acidic and neutral polar AAs *in vivo*.

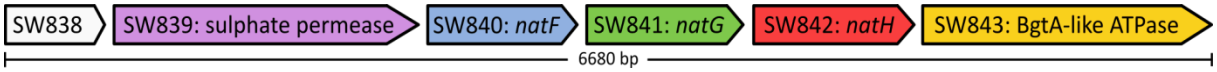

Figure 4.4: Putative operon that SW840 is associated with in *Synechococcus* strain WH8102. Upstream from the SW840 target gene (blue) are two genes coding for a hypothetical protein (white) and a putative sulphate permease (purple). There are three genes downstream from SW840 that code for two ABC transmembrane domains (*natG*, green and *natH*, red), and a bgtA-like ABC-type uptake transporter for acidic and neutral polar AAs (N-II), ATPase subunit (yellow). The length of the operon is 6680 bp.

4.4 Further probing of SC599 (cluster CK1309) is required

SC599 is enclosed in cluster CK1309, annotated to putatively bind carbohydrates and AAs. During the purification process, the protein solution showed yellow colour after elution from the His-Trap. Subsequent SEC purification detected an eluent towards the very end of the run (Figure 3.6.A). It is therefore possible that SC599 was already bound to a suitable ligand picked up from the auto-induction media. Mass spectroscopy (MS) analysis of this small molecular weight fraction is therefore proposed in future to determine its molecular characteristics.

The SEC data indicates that the SC599 protein may exist in a bound-closed conformation could also explain only a minor increase (~2°C) in the T_m of the protein during fluorimetry analysis (section 3.5.3). Future work involving on-column protein unfolding and re-folding to remove any bound ligands[135] with a subsequent DSF analysis on the apo/unbound form of SC599 could provide more insights into its ligand binding characteristics. The presence of a yellow colour also indicates the possibility that the enzyme cofactor PLP may be bound to the protein, further investigation in this is required. Additionally, the possibility of the absence of natural ligand for SC599 within the SB HR2-096 tray cannot be negated. Further investigation of SC599 protein binding characteristics is thus required.

Ten genes were identified downstream from SC599 gene product located within the *Synechococcus* strain CC9902 genome (Figure 4.5). These genes were searched in the Cyanorak database and most were found to be poorly characterised hypothetical proteins (SC598, SCSC597, SC596, SC595, SC593, and SC591). Two genes, SC594 and SC592, were found to be annotated as a transcription elongation factor NusA and a translation initiation factor IF-2, respectively. The PBP could be binding to a substrate (likely a carbohydrate or AA) and interacting with a receptor. Unfortunately, this is only proposed as more research into marine cyanobacterial genes is needed to better understand these systems.

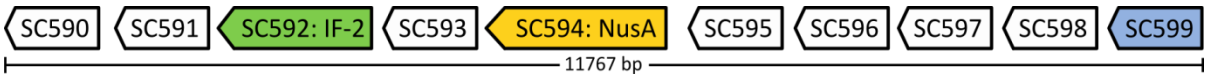

Figure 4.5: Genes downstream from SC599 in *Synechococcus* strain CC9902. The SC599 target gene (blue) is upstream from 7 genes coding for hypothetical proteins (white), a transcription elongation factor NusA gene (yellow), and a translation initiation factor IF-2 gene (green). The length of the operon is 11767 bp.

4.5 Preliminary analysis on the ecological distribution of organic nutrient transport clusters

A preliminary analysis of marine metagenomes was undertaken to determine the relative abundance of each orthologous group in different marine environments around Australia. The dataset consisted of an annual cycle of monthly samples taken from two depths at seven National references Stations (115 samples) as well as 86 samples taken from 5 research voyages in ocean regions in the North and East of Australia. The sample set covers a latitudinal range from the subtropics (12.4°S) to temperate waters (42.6°S) off Tasmania.

Figure 4.6 shows the relative abundance in wild cyanobacteria populations of the three clusters studied in in this thesis. Cluster CK2503, which included SC634 that was discovered to interact with formate and formaldehyde, was generally found to be present at sites lower in latitude and temperature. In the east coast, the CK2503 cluster is completely absent in subtropical populations of *Synechococcus* but present in populations closer to the coast. In comparison, the two other clusters (CK1489 and CK1309) were much more abundant and covered a larger range of latitude. This shows that these three clusters are present in environmental populations of cyanobacteria. In addition, there are biogeographical factors influencing the distribution of cyanobacterial ABC transporter genes suggesting that analysis of the metagenomic data could yield interesting results. Further characterisation of these clusters are required to understand the ecological implications of these data

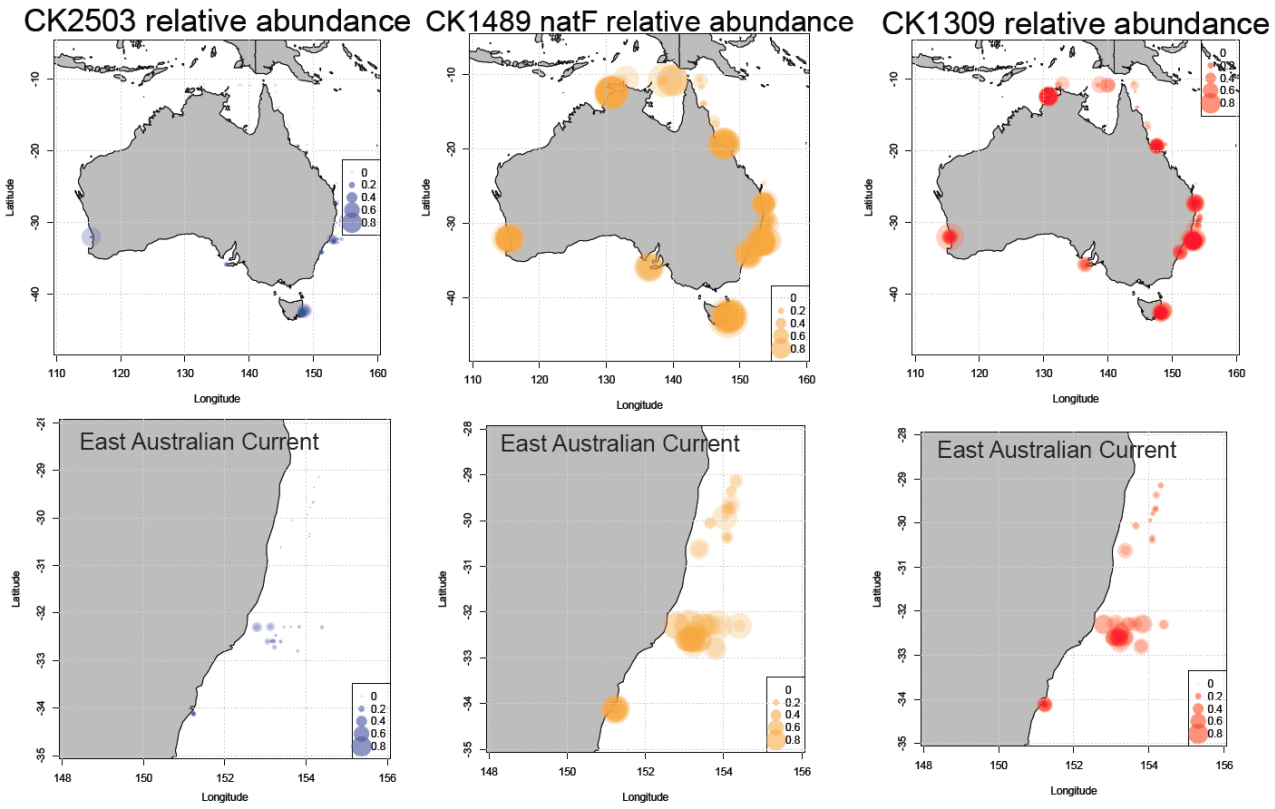

Figure 4.6: Relative abundance of organic nutrient transporters within wild cyanobacterial populations. Relative abundance in 200 regional metagenomes was compiled by read-mapping against complete or near complete cyanobacterial genomes using BLASTn. Filtered results were normalised against the abundance of 520 single-copy housekeeping genes. The Eastern Australian Current (EAC) panels display a north to south drift within the EAC followed by a west to east transect of the Tasman Sea downstream.

Chapter 5. Conclusion

Marine cyanobacterial genomes hold the secret to their adaptability that enables them to flourish across diverse marine environments. A biotechnological approach to understand the function of gene products can help connect the knowledge gap between the vast quantities of genomic information and molecular characterisation of cellular machinery.

Understanding how transport systems, particularly ABC transporters, function will be a means to begin to understand the range of compounds that can be assimilated by marine cyanobacteria. Previous studies also demonstrate AAs to likely be a valuable resource that marine cyanobacteria can potentially scavenge for direct use in protein synthesis or catabolised through various metabolic pathways[136-138]. Thus, the aim of this thesis was to characterise AA-PBPs associated with ABC transporters via the molecular approach of cloning, expressing, purifying, and ligand screening of recombinant protein.

A total of 110 PBP ABC transporter *Synechococcus* genes spanning 23 clusters were identified for research in the larger scope/project (Figure 1.5). This thesis investigated 12 genes from 3 clusters that are associated with AA transport. Recombinant expression, purification and DSF ligand screening techniques were used to characterise *Synechococcus* AA-PBPs. Out of the 12 genes investigated, all were successfully cloned into *E. coli* expression hosts with three proteins, one from each cluster, successfully purified. DSF analysis revealed a possible formate and formaldehyde PBP (SC634, CK2503) from strain CC9311 as well as an acidic and neutral polar AA-PBP (SW840, CK1489) from strain WH8102. The third protein (SC599, CK1309) is hypothesised to be in a closed conformation thus requiring further probing to elucidate any binding mechanisms.

Future investigation to determine the binding interactions between the PBPs and predicted substrates can be carried out using various experimental approaches such as measuring the binding affinity using ITC, substrate uptake assays coupled with gene knockouts, MS, SPR, visualisation of heavy isotopic substrates with the use of nanoSIMS, site-directed mutagenesis of the active site residues, and x-ray crystallography.

Ultimately, more research into the annotated cyanobacterial genes involving metabolic pathways within the Cyanorak database is required to further our understanding of how these microbes influence the biogeochemical cycles.

References

1 Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437**, 349-355 (2005).

2 Ducklow, H. Microbial services: challenges for microbial ecologists in a changing world. *Aquatic Microbial Ecology* **53**, 13-19 (2008).

3 Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12115-12120 (2006).

4 Costanza, R. *et al.* Changes in the global value of ecosystem services. *Global Environmental Change-Human and Policy Dimensions* **26**, 152-158 (2014).

5 de Groot, R. *et al.* Global estimates of the value of ecosystems and their services in monetary units. *Ecosystem Services* **1**, 50-61 (2012).

6 Longhurst, A., Sathyendranath, S., Platt, T. & Caverhill, C. An Estimate of Global Primary Production in the Ocean from Satellite Radiometer Data. *Journal of Plankton Research* **17**, 1245-1271 (1995).

7 Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography* **42**, 1-20 (1997).

8 Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307-315 (2014).

9 Anbar, A. D. *et al.* A whiff of oxygen before the great oxidation event? *Science* **317**, 1903-1906 (2007).

10 Garcia-Pichel, F., Belnap, J., Neuer, S. & Schanz, F. Estimates of global cyanobacterial biomass and its distribution. *Algological Studies* **109**, 213-227 (2003).

11 Scanlan, D. J. *et al.* Ecological Genomics of Marine Picocyanobacteria. *Microbiology and Molecular Biology Reviews* **73**, 249-299 (2009).

12 Johnson, P. W. & Sieburth, J. M. Chroococcoid cyanobacteria in the sea: a ubiquitous and diverse phototrophic biomass. *Limnology and Oceanography* **24**, 928-935 (1979).

13 Waterbury, J. B., Watson, S. W., Guillard, R. R. L. & Brand, L. E. Wide-spread occurrence of a unicellular, marine planktonic, cyanobacterium. *Nature* **277**, 293-294 (1979).

14 Chisholm, S. W. *et al.* A Novel Free-Living Prochlorophyte Abundant in the Oceanic Euphotic Zone. *Nature* **334**, 340-343 (1988).

15 Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 9824-9829 (2013).

16 Partensky, F., Blanchot, J. & Vaulot, D. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters: a review. *Bulletin de l'Institut océanographique de Monaco*, 457-476 (1999).

17 Agawin, N. S. R., Duarte, C. M. & Agusti, S. Growth and abundance of Synechococcus sp. in a Mediterranean Bay: seasonality and relationship with temperature. *Marine Ecology Progress Series* **170**, 45-53 (1998).

18 Neuer, S. Growth Dynamics of Marine Synechococcus Spp in the Gulf of Alaska. *Marine Ecology Progress Series* **83**, 251-262 (1992).

19 Shapiro, L. P. & Haugen, E. M. Seasonal Distribution and Temperature Tolerance of Synechococcus in Boothbay Harbor, Maine. *Estuarine Coastal and Shelf Science* **26**, 517-525 (1988).

20 Partensky, F., Hess, W. R. & Vaulot, D. Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews* **63**, 106-127 (1999).

21 Buck, K. R., Chavez, F. P. & Campbell, L. Basin-wide distributions of living carbon components and the inverted trophic pyramid of the central gyre of the North Atlantic Ocean, summer 1993. *Aquatic Microbial Ecology* **10**, 283-298 (1996).

22 Ahlgren, N. A. & Rocap, G. Culture isolation and culture-independent clone libraries reveal new marine Synechococcus ecotypes with distinctive light and N physiologies. *Applied and Environmental Microbiology* **72**, 7193-7204 (2006).

23 Dufresne, A. *et al.* Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology* **9**, R90 (2008).

24 Farrant, G. K. *et al.* Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3365-3374 (2016).

25 Mazard, S., Ostrowski, M., Partensky, F. & Scanlan, D. J. Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environmental Microbiology* **14**, 372-386 (2012).

26 Sohm, J. A. *et al.* Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* **10**, 333-345 (2016).

27 Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344**, 416-420 (2014).

28 Tai, V. & Palenik, B. Temporal variation of Synechococcus clades at a coastal Pacific Ocean monitoring site. *ISME J* **3**, 903-915 (2009).

29 Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology* **6,** R14 (2005).

30 Kettler, G. C. *et al.* Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *Plos Genetics* **3**, 2515-2528 (2007).

31 Strehl, B., Holtzendorff, J., Partensky, F. & Hess, W. R. A small and compact genome in the marine cyanobacterium Prochlorococcus marinus CCMP 1375: lack of an intron in the gene for tRNA(Leu)(UAA) and a single copy of the rRNA operon. *FEMS Microbiology Letters* **181**, 261-266 (1999).

32 Palenik, B. Molecular Mechanisms by Which Marine Phytoplankton Respond to Their Dynamic Chemical Environment. *Annual Review of Marine Science, Vol 7* **7**, 325-340 (2015).

33 Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of Prochlorococcus and Synechococcus (Cyanobacteria). *Journal of Molecular Evolution* **46**, 188-201 (1998).

34 McDaniel, L. D. *et al.* High frequency of horizontal gene transfer in the oceans. *Science* **330**, 50 (2010).

35 Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature* **424**, 1047-105 (2003).

36 Lawrence, J. G. & Roth, J. R. Genomic flux: genome evolution by gene loss and acquisition. *American Society of Microbiology*, 263-289 (1999).

37 Mills, E. L. *Biological oceanography: An early history, 1870-1960*. (University of Toronto Press, 2012).

38 Scanlan, D. J. in *Ecology of Cyanobacteria II* 503-533 (Springer, 2012).

39 Moore, L. R., Post, A. F., Rocap, G. & Chisholm, S. W. Utilization of different nitrogen sources by the marine cyanobacteria Prochlorococcus and Synechococcus. *Limnology and Oceanography* **47**, 989-996 (2002).

40 Paerl, H. W. Ecophysiological and trophic implications of light-stimulated amino Acid utilization in marine picoplankton. *Applied and Environmental Microbiology* **57**, 473-479 (1991).

41 Collier, J. L., Brahamsha, B. & Palenik, B. The marine cyanobacterium Synechococcus sp. WH7805 requires urease (urea amidohydrolase, EC 3.5.1.5) to utilize urea as a nitrogen source: molecular-genetic and biochemical analysis of the enzyme. *Microbiology* **145**, 447-459 (1999).

42 Wyman, M., Gregory, R. P. F. & Carr, N. G. Novel Role for Phycoerythrin in a Marine Cyanobacterium, Synechococcus Strain Dc2. *Science* **230**, 818-820 (1985).

43 Astorga-Elo, M., Ramirez-Flandes, S., DeLong, E. F. & Ulloa, O. Genomic potential for nitrogen assimilation in uncultivated members of Prochlorococcus from an anoxic marine zone. *ISME Journal* **9**, 1264-1267 (2015).

44 Moore, L. R. More mixotrophy in the marine microbial mix. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 8323-8324 (2013).

45 Zubkov, M. V. Photoheterotrophy in marine prokaryotes. *Journal of Plankton Research* **31**, 933-938 (2009).

46 Palenik, B. *et al.* The genome of a motile marine Synechococcus. *Nature* **424**, 1037-1042 (2003).

47 Rocap, G. *et al.* Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042-1047 (2003).

48 Church, M. J., Ducklow, H. W. & Karl, D. A. Light dependence of [H-3]leucine incorporation in the oligotrophic North Pacific ocean. *Applied and Environmental Microbiology* **70**, 4079-4087 (2004).

49 Michelou, V. K., Cottrell, M. T. & Kirchman, D. L. Light-stimulated bacterial production and amino acid assimilation by cyanobacteria and other microbes in the North Atlantic ocean. *Applied and Environmental Microbiology* **73**, 5539-5546 (2007).

50 Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H. & Amann, R. High rate of uptake of organic nitrogen compounds by Prochlorococcus cyanobacteria as a key to their dominance in oligotrophic oceanic waters. *Applied and Environmental Microbiology* **69**, 1299-1304 (2003).

51 Zubkov, M. V., Tarran, G. A. & Fuchs, B. M. Depth related amino acid uptake by Prochlorococcus cyanobacteria in the Southern Atlantic tropical gyre. *FEMS Microbiology Ecology* **50**, 153-161 (2004).

52 Malmstrom, R. R., Kiene, R. P., Vila, M. & Kirchman, D. L. Dimethylsulfoniopropionate (DMSP) assimilation by Synechococcus in the Gulf of Mexico and northwest Atlantic Ocean. *Limnology and Oceanography* **50**, 1924-1931 (2005).

53 Zubkov, M. V. & Tarran, G. A. Amino acid uptake of Prochlorococcus spp. in surface waters across the South Atlantic Subtropical Front. *Aquatic Microbial Ecology* **40**, 241-249 (2005).

54 Azam, F. & Godson, R. Size distribution and activity of marine microheterotrophs. *Limnology and Oceanography* **22**, 492-501 (1977).

55 Pomeroy, L. R. The ocean's food web, a changing paradigm. *Bioscience* **24**, 499-504 (1974).

56 Amon, R. M. W. & Benner, R. Bacterial utilization of different size classes of dissolved organic matter. *Limnology and Oceanography* **41**, 41-51 (1996).

57 Yamashita, Y. & Tanoue, E. Distribution and alteration of amino acids in bulk DOM along a transect from bay to oceanic waters. *Marine Chemistry* **82**, 145-160 (2003).

58 Hertkorn, N. *et al.* Characterization of a major refractory component of marine dissolved organic matter. *Geochimica Et Cosmochimica Acta* **70**, 2990-3010 (2006).

59 Koch, B. P., Witt, M. R., Engbrodt, R., Dittmar, T. & Kattner, G. Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Geochimica Et Cosmochimica Acta* **69**, 3299-3308 (2005).

60 Jiao, N. *et al.* Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology* **8**, 593-599 (2010).

61 Yamada, N. & Tanoue, E. Detection and partial characterization of dissolved glycoproteins in oceanic waters. *Limnology and Oceanography* **48**, 1037-1048 (2003).

62 Benner, R. & Kaiser, K. Abundance of amino sugars and peptidoglycan in marine particulate and dissolved organic matter. *Limnology and Oceanography* **48**, 118-128 (2003).

63 Kaiser, K. & Benner, R. Major bacterial contribution to the ocean reservoir of detrital organic carbon and nitrogen. *Limnology and Oceanography* **53**, 99-112 (2008).

64 Wakeham, S. G., Pease, T. K. & Benner, R. Hydroxy fatty acids in marine dissolved organic matter as indicators of bacterial membrane material. *Organic Geochemistry* **34**, 857-868 (2003).

65 McCarthy, M. D., Hedges, J. I. & Benner, R. Major bacterial contribution to marine dissolved organic nitrogen. *Science* **281**, 231-234 (1998).

66 Middelboe, M. & Lyck, P. G. Regeneration of dissolved organic matter by viral lysis in marine microbial communities. *Aquatic Microbial Ecology* **27**, 187-194 (2002).

67 Hubberten, U., Lara, R. J. & Kattner, G. Refractory Organic-Compounds in Polar Waters - Relationship between Humic Substances and Amino-Acids in the Arctic and Antarctic. *Journal of Marine Research* **53**, 137-149 (1995).

68 Keil, R. G. & Kirchman, D. L. Contribution of Dissolved Free Amino-Acids and Ammonium to the Nitrogen Requirements of Heterotrophic Bacterioplankton. *Marine Ecology Progress Series* **73**, 1-10 (1991).

69 Druffel, E. R. M., Williams, P. M., Bauer, J. E. & Ertel, J. R. Cycling of Dissolved and Particulate Organic-Matter in the Open Ocean. *Journal of Geophysical Research-Oceans* **97**, 15639-15659 (1992).

70 Chen, T. H., Chen, T. L., Hung, L. M. & Huang, T. C. Circadian-Rhythm in Amino-Acid-Uptake by Synechococcus Rf-1. *Plant Physiology* **97**, 55-59 (1991).

71 Mary, I. *et al.* Light enhanced amino acid uptake by dominant bacterioplankton groups in surface waters of the Atlantic Ocean. *FEMS Microbiol Ecol* **63**, 36-45 (2008).

72 Vila-Costa, M. *et al.* Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science* **314**, 652-654 (2006).

73 Kana, T. M. & Glibert, P. M. Effect of Irradiances up to 2000 Mu-E M-2s-1 on Marine Synechococcus Wh7803 .1. Growth, Pigmentation, and Cell Composition. *Deep-Sea Research Part a-Oceanographic Research Papers* **34**, 479-495 (1987).

74 Rees, D. C., Johnson, E. & Lewinson, O. ABC transporters: the power to change. *Nature Reviews Molecular Cell Biology* **10**, 218-227 (2009).

75 Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Microbial genome analyses: Comparative transport capabilities in eighteen prokaryotes. *Journal of Molecular Biology* **301**, 75-100 (2000).

76 Hollenstein, K., Frei, D. C. & Locher, K. P. Structure of an ABC transporter in complex with its binding protein. *Nature* **446**, 213-216 (2007).

77 Hollenstein, K., Dawson, R. J. & Locher, K. P. Structure and mechanism of ABC transporter proteins. *Current Opinion in Structural Biology* **17**, 412-418 (2007).

78 Hosie, A. H. F. & Poole, P. S. Bacterial ABC transporters of amino acids. *Research in Microbiology* **152**, 259-270 (2001).

79 Walshaw, D. L. & Poole, P. S. The general L-amino acid permease of Rhizobium leguminosarum is an ABC uptake system that also influences efflux of solutes. *Molecular Microbiology* **21**, 1239-1252 (1996).

80 Walshaw, D. L., Lowthorpe, S., East, A. & Poole, P. S. Distribution of a sub-class of bacterial ABC polar amino acid transporter and identification of an N-terminal region involved in solute specificity. *FEBS Letters* **414**, 397-401 (1997).

81 Payne, G. M., Spudich, E. N. & Ames, G. F. A mutational hot-spot in the hisM gene of the histidine transport operon in Salmonella typhimurium is due to deletion of repeated sequences and results in an altered specificity of transport. *Molecular and General Genetics* **200**, 493-496 (1985).

82 Montesinos, M. L., Herrero, A. & Flores, E. Amino acid transport in taxonomically diverse cyanobacteria and identification of two genes encoding elements of a neutral amino acid permease putatively involved in recapture of leaked hydrophobic amino acids. *Journal of Bacteriology* **179**, 853-862 (1997).

83 Berntsson, R. P. A., Smits, S. H. J., Schmitt, L., Slotboom, D. J. & Poolman, B. A structural classification of substrate-binding proteins. *FEBS Letters* **584**, 2606-2617 (2010).

84 Berger, E. A. Different mechanisms of energy coupling for the active transport of proline and glutamine in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 1514-1518 (1973).

85 Zschiedrich, C. P., Keidel, V. & Szurmant, H. Molecular Mechanisms of Two-Component Signal Transduction. *Journal of Molecular Biology* **428**, 3752-3775 (2016).

86 Kelly, D. J. & Thomas, G. H. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol Rev* **25**, 405-424 (2001).

87 Winnen, B., Hvorup, R. N. & Saier, M. H., Jr. The tripartite tricarboxylate transporter (TTT) family. *Research in Microbiology* **154**, 457-465 (2003).

88 Armstrong, N. & Gouaux, E. Mechanisms for activation and antagonism of an AMPA-Sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. *Neuron* **28**, 165-181 (2000).

89 Fukami-Kobayashi, K., Tateno, Y. & Nishikawa, K. Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. *Journal of Molecular Biology* **286**, 279-290 (1999).

90 Quiocho, F. A. & Ledvina, P. S. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Molecular Microbiology* **20**, 17-25 (1996).

91 Mao, B., Pear, M. R., McCammon, J. A. & Quiocho, F. A. Hinge-bending in L-arabinose-binding protein. The "Venus's-flytrap" model. *Journal of Biological Chemistry* **257**, 1131-1133 (1982).

92 Bermejo, G. A., Strub, M. P., Ho, C. & Tjandra, N. Ligand-free open-closed transitions of periplasmic binding proteins: the case of glutamine-binding protein. *Biochemistry* **49**, 1893-1902 (2010).

93 Tang, C., Schwieters, C. D. & Clore, G. M. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* **449**, 1078-1082 (2007).

94 Scheepers, G. H., Lycklama, A. N. J. A. & Poolman, B. An updated structural classification of substrate-binding proteins. *FEBS Letters* **590**, 4393-4401 (2016).

95 Karpowich, N. K., Huang, H. H., Smith, P. C. & Hunt, J. F. Crystal structures of the BtuF periplasmic-binding protein for vitamin B12 suggest a functionally important reduction in protein mobility upon ligand binding. *Journal of Biological Chemistry* **278**, 8429-8434 (2003).

96 Felder, C. B., Graul, R. C., Lee, A. Y., Merkle, H. P. & Sadee, W. The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. *The AAPS Journal* **1**, E2 (1999).

97 Lee, Y. H., Deka, R. K., Norgard, M. V., Radolf, J. D. & Hasemann, C. A. Treponema pallidum TroA is a periplasmic zinc-binding protein with a helical backbone. *Nature Structural and Molecular Biology* **6**, 628-633 (1999).

98 Sack, J. S., Saper, M. A. & Quiocho, F. A. Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. *Journal of Molecular Biology* **206**, 171-191 (1989).

99 Sack, J. S., Trakhanov, S. D., Tsigannik, I. H. & Quiocho, F. A. Structure of the L-leucine-binding protein refined at 2.4 A resolution and comparison with the Leu/Ile/Val-binding protein structure. *Journal of Molecular Biology* **206**, 193-207 (1989).

100 Oh, B. H. *et al.* The bacterial periplasmic histidine-binding protein. structure/function analysis of the ligand-binding site and comparison with related proteins. *Journal of Biological Chemistry* **269**, 4135-4143 (1994).

101 Tame, J. R. *et al.* The structural basis of sequence-independent peptide binding by OppA protein. *Science* **264**, 1578-1581 (1994).

102 Mishima, Y., Momma, K., Hashimoto, W., Mikami, B. & Murata, K. Crystal structure of AlgQ2, a macromolecule (alginate)-binding protein of Sphingomonas sp. A1, complexed with an alginate tetrasaccharide at 1.6-A resolution. *Journal of Biological Chemistry* **278**, 6552-6559 (2003).

103 Momma, K., Mishima, Y., Hashimoto, W., Mikami, B. & Murata, K. Direct evidence for Sphingomonas sp. A1 periplasmic proteins as macromolecule-binding proteins associated with the ABC transporter: molecular insights into alginate transport in the periplasm. *Biochemistry* **44**, 5053-5064 (2005).

104 Marsden, R. L. & Orengo, C. A. Target selection for structural genomics: an overview. *Methods in Molecular Biology* **426**, 3-25 (2008).

105 Vetting, M. W. *et al.* Experimental Strategies for Functional Annotation and Metabolism Discovery: Targeted Screening of Solute Binding Proteins and Unbiased Panning of Metabolomes. *Biochemistry* **54**, 909-931 (2015).

106 Blommel, P. G. & Fox, B. G. A combined approach to improving large-scale production of tobacco etch virus protease. *Protein Expression and Purification* **55**, 53-68 (2007).

107 Robinson, A. *et al.* Structural genomics of the bacterial mobile metagenome: an overview. *Methods in Molecular Biology* **426**, 589-595 (2008).

108 Borah, P. Primer designing for PCR. *Science Vision* **11**, 134-136 (2011).

109 Schochetman, G., Ou, C. Y. & Jones, W. K. Polymerase Chain-Reaction. *Journal of Infectious Diseases* **158**, 1154-1157 (1988).

110 Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J. & Wallace, R. B. The Effect of Temperature and Oligonucleotide Primer Length on the Specificity and Efficiency of Amplification by the Polymerase Chain-Reaction. *DNA and Cell Biology* **10**, 233-238 (1991).

111 Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **8**, 4321-4325 (1980).

112 Aslanidis, C. & de Jong, P. J. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Research* **18**, 6069-6074 (1990).

113 Haun, R. S. & Moss, J. Ligation-independent cloning of glutathione S-transferase fusion genes for expression in Escherichia coli. *Gene* **112**, 37-43 (1992).

114 Sockolosky, J. T. & Szoka, F. C. Periplasmic production via the pET expression system of soluble, bioactive human growth hormone. *Protein Expression and Purification* **87**, 129-135 (2013).

115 Porath, J. & Olin, B. Immobilized metal ion affinity adsorption and immobilized metal ion affinity chromatography of biomaterials. Serum protein affinities for gel-immobilized iron and nickel ions. *Biochemistry* **22**, 1621-1630 (1983).

116 Nozaki, Y., Schechter, N. M., Reynolds, J. A. & Tanford, C. Use of gel chromatography for the determination of the Stokes radii of proteins in the presence and absence of detergents. A reexamination. *Biochemistry* **15**, 3884-3890 (1976).

117 Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* **2**, 2212-2221 (2007).

118 Graslund, S. *et al.* Protein production and purification. *Nature Methods* **5**, 135-146 (2008).

119 Schein, C. H. Production of Soluble Recombinant Proteins in Bacteria. *Nature Biotechnology* **7**, 1141-1147 (1989).

120 Sorensen, H. P. & Mortensen, K. K. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. *Microbial Cell Factories* **4**, 1 (2005).

121 Hughes, R. C., Jenkins, W. T. & Fischer, E. H. The site of binding of pyridoxal-5'-phosphate to heart glutamic-aspartic transaminase. *Proceedings of the National Academy of Sciences of the United States of America* **48**, 1615-1618 (1962).

122 Jevsevar, S. *et al.* Production of nonclassical inclusion bodies from which correctly folded protein can be extracted. *Biotechnology Progress* **21**, 632-639 (2005).

123 Singh, A., Upadhyay, V., Upadhyay, A. K., Singh, S. M. & Panda, A. K. Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microbial Cell Factories* **14**, 41 (2015).

124 Um, I. C., Kweon, H. Y., Lee, K. G. & Park, Y. H. The role of formic acid in solution stability and crystallization of silk protein polymer. *International Journal of Biological Macromolecules* **33**, 203-213 (2003).

125 Vorholt, J. A. Cofactor-dependent pathways of formaldehyde oxidation in methylotrophic bacteria. *Archives Microbiology* **178**, 239-249 (2002).

126 Sun, J. *et al.* One carbon metabolism in SAR11 pelagic marine bacteria. *PLoS One* **6**, e23973 (2011).

127 Giovannoni, S. J. *et al.* The small genome of an abundant coastal ocean methylotroph. *Environmental Microbiology* **10**, 1771-1782 (2008).

128 Mikami, B. & Ida, S. Purification and Properties of Ferredoxin Nitrate Reductase from the Cyanobacterium Plectonema-Boryanum. *Biochimica Et Biophysica Acta* **791**, 294-304 (1984).

129 Pernil, R., Picossi, S., Mariscal, V., Herrero, A. & Flores, E. ABC-type amino acid uptake transporters Bgt and N-II of Anabaena sp. strain PCC 7120 share an ATPase subunit and are expressed in vegetative cells and heterocysts. *Molecular Microbiology* **67**, 1067-1080 (2008).

130 Flores, E. & Herrero, A. in *The molecular biology of cyanobacteria* 487-517 (Springer, 1994).

131 Flores, E. & Muropastor, M. I. Uptake of Glutamine and Glutamate by the Dinitrogen-Fixing Cyanobacterium Anabaena Sp Pcc7120. *FEMS Microbiology Letters* **56**, 127-130 (1988).

132 Ping, X. & Mcauley, P. J. Uptake of Amino-Acids by the Cyanobacterium Anabaena Atcc-27893. *New Phytologist* **115**, 581-585 (1990).

133 Montesinos, M. L., Herrero, A. & Flores, E. Amino acid transport systems required for diazotrophic growth in the cyanobacterium Anabaena sp. strain PCC 7120. *Journal of Bacteriology* **177**, 3150-3157 (1995).

134 Herrero, A. & Flores, E. Transport of Basic-Amino-Acids by the Dinitrogen-Fixing Cyanobacterium Anabaena Pcc 7120. *Journal of Biological Chemistry* **265**, 3931-3935 (1990).

135 Clifton, B. E. & Jackson, C. J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chemical Biology* **23**, 236-245 (2016).

136 Christensen, J. E., Dudley, E. G., Pederson, J. A. & Steele, J. L. Peptidases and amino acid catabolism in lactic acid bacteria. *Antonie van Leeuwenhoek* **76**, 217-246 (1999).

137 Massey, L. K., Sokatch, J. R. & Conrad, R. S. Branched-chain amino acid catabolism in bacteria. *Bacteriological Reviews* **40**, 42-54 (1976).

138 Neis, E. P., Dejong, C. H. & Rensen, S. S. The role of microbial amino acid metabolism in host metabolism. *Nutrients* **7**, 2930-2946 (2015).

Appendix I: PCR amplification primers

Abbreviations: F, forward; R, reverse.

Gene	pET Vector	Primer	Length	T_m (°C)	GC %
CK_Syn_CC9311_02366	15b	F – GCCGCGCGGCAGCCATatgtgccaaaagtctgctcc	36	72.4	64
		R – GTTAGCAGCCGGATCCtcaggaacctttggataattc	37	66.7	49
	22b	F – GATATCGGAATTAATTCGatgtgccaaaagtctgctcc	38	64.5	42
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCggaacctttggataattc	54	72.9	50
CK_Syn_CC9902_00599	15b	F – GCCGCGCGGCAGCCATatgggacctttgcgaaaag	36	71.3	61
		R – GTTAGCAGCCGGATCCtcagttgggtgatacggac	35	69.1	57
	22b	F – GATATCGGAATTAATTCGatgggacctttgcgaaaag	38	63.4	39
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCgttgggtgatacggac	52	74.8	56
CK_Syn_MITS9220_02624	15b	F – GCCGCGCGGCAGCCATatgggacctttgcgaaaag	34	72.9	68
		R – GTTAGCAGCCGGATCCtcaggactgcggattgaattc	37	68.9	54
	22b	F – GATATCGGAATTAATTCGatggcggataactccagc	36	64.4	44
		R –GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCggactgcggattgaattc	54	74.5	54
CK_Syn_WH8102_00606	15b	F – GCCGCGCGGCAGCCATttgttggaaagcgtggttcc	36	72.4	64
		R – GTTAGCAGCCGGATCCttaggcctgcatcatttggc	36	69.0	56
	22b	F – GATATCGGAATTAATTCGttgttggaaagcgtggttcc	38	64.5	42
		R – GTGGTGGTGCTCGAG*GGATTGGAAGTACAGGTTCTC*ggcctgcatcatttggc	53	75.4	57
CK_Syn_WH8102_00607	15b	F – GCCGCGCGGCAGCCATatgcaacccgaatcatccaag	37	72.2	62
		R – GTTAGCAGCCGGATCCtcaggtcaacaaccgggttc	36	70.1	58
	22b	F – GATATCGGAATTAATTCGatgcaacccgaatcatccaag	39	64.5	41
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCggtcaacaaccgggttc	53	75.4	57
CK_Syn_CC9311_01190	15b	F – GCCGCGCGGCAGCCATttgttgcggaaactaaaccagttc	40	71.7	58
		R – GTTAGCAGCCGGATCCtcattggaacgggggtgag	35	70.3	60
	22b	F – GATATCGGAATTAATTCGttgttgcggaaactaaaccagttc	42	64.5	38
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCttggaacgggggtgag	52	75.6	58
CK_Syn_CC9902_00847	15b	F – GCCGCGCGGCAGCCATttgagtgatggcactagttc	36	71.3	61
		R – GTTAGCAGCCGGATCCtcaagtgaacggcggag	33	69.4	61
	22b	F – GATATCGGAATTAATTCGttgagtgatggcactagttc	38	63.4	39
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCagtgaacggcggag	50	75.2	58
CK_Syn_MITS9220_01218	15b	F – GCCGCGCGGCAGCCATatgacccttggagatggc	34	72.9	68
		R – GTTAGCAGCCGGATCCtcattggaatggtggcgc	34	69.2	59
	22b	F – GATATCGGAATTAATTCGatgacccttggagatggc	36	64.4	44
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCttggaatggtggcgc	51	75.0	57
CK_Syn_WH8102_00840	15b	F – GCCGCGCGGCAGCCATatggacgaggccggtc	32	74.6	75
		R – GTTAGCAGCCGGATCCtcaggtgaatgggggcg	33	70.6	64
	22b	F – GATATCGGAATTAATTCGatggacgaggccggtc	34	65.6	50
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCggtgaatgggggcg	50	76.1	60
CK_Syn_WH8109_00877	15b	F – GCCGCGCGGCAGCCATatggctggctcgcgtc	32	74.6	75
		R – GTTAGCAGCCGGATCCtcaggtgaacggtgggg	33	70.6	64
	22b	F – GATATCGGAATTAATTCGatggctggctcgcgtc	34	65.6	50
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCggtgaacggtgggg	50	76.1	60
CK_Syn_CC9311_00634	15b	F – GCCGCGCGGCAGCCATatggcccaacaaaaagacacc	37	72.2	62
		R – GTTAGCAGCCGGATCCttatggctcccattgttgtgc	37	68.9	54
	22b	F – GATATCGGAATTAATTCGatggcccaacaaaaagacacc	39	64.5	41
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCtggctcccattgttgtgc	54	75.2	56
CK_Syn_CC9311_02631	15b	F – GCCGCGCGGCAGCCATatgatgacgggggcgttg	34	74.1	71
		R – GTTAGCAGCCGGATCCtcagggttcccagggatc	34	70.4	62
	22b	F – GATATCGGAATTAATTCGatgatgacgggggcgttg	36	65.6	47
		R – GTGGTGGTGCTCGAGGGATTGGAAGTACAGGTTCTCgggttcccagggatc	51	75.8	59

Appendix II: Cyanorak ORF signal peptide truncated gene sequences

Cyanorak ORF ID	Signal peptide truncated AA sequence
CK_Syn_CC9311_02366	CQKSAPTSEKEASGQTGNVFETGKLRAVVVDDVLPMVDKKDGKYEGLSFVVLDAIRDQL KSAPENKSDDIVIEPVSIKSAQDGLNKIRSGEADIACGVAFTWERQRTLTYSLPFATSG TRVLAPKGNDGTPDSLKGKTIGVVKDTAAAAVLAKSVDDAQFQFFATPTEALAGLKDGT IEFLGGDTLWLKASRKDTAPDADLVPTFPYARSSVGCVIADTTPHLLNYSNLAIGRMLT AYVDDNKDVRTAVNKWIGPDSQVGLSENMIGDFFTIVLATTAELSKGS
CK_Syn_CC9902_00599	GPFAKRTSLKAVIFEDVKPLYQKSGDTYEGFGVDMLNLIKDQAGRRTLKFIPTTSAEDG MKAITSGKADIACGVAFDWGRAEKVSYTIPFAIGGTRLLTKTTINGTPSSLRGRTVGVV KDSSSAKILETVVPSASLQAYATPAEAFAAYDSGKISTLAGGTLWLAANSDAKNSDLVP IRPYGRTGIGCIVKQNNGKLLAAANNAIGQTMQEYVNGNAATREMVNRWIGPGSNVMLP ESVITALYSLLLSTTSEMSVSPN
CK_Syn_MITS9220_02624	ADNSSQSSSAKSVYDTGKLKAVVIGDSLPMVKKDGDNYDGLSFVVLEAIRDQINVSPKK KDLDVSIEPVAASSARDGLDMIRSGAADIACGVAFTWQRQRSLTYTLPFSVGGVRLLAP AGIDGTPKSLNGKTVGVVKDSMAANVLAASVDDANFQFFDTPDQALAAVKDGSVEILGG DSLWLRANKAATAPEASLVPDRPYARSGVGCVVADTTPHLLNISNLGIGRLLSGYINDD DGVRSAINTWIGTGSTVGLKDEQINRFFTIVLSTAAEFNPQS
CK_Syn_WH8102_00606	LESVVPEATFKAFDTPAEALSAYNDGDVPILGGGTLWLAANSSTDTTALLPFRPYGRSG IGCVVGQNNGKLLSQANIAIGQMMQA
CK_Syn_WH8102_00607	QPESSKTSAEEPAAAELKAVIFEDVKPLFSKTDAGYEGFGVDVLEQVRMQSGRSEVTYQ VASSITDGIDAVATGKADIACGVAFTWGRASQLSYSLPFGVGGTRLLT
CK_Syn_CC9311_01190	LRKLNQFFELTRTISRDERGLTHLFLTRDKQMRVHLLALPLVSFLFILNGCATLGEGGA SRLDLIKRRAELRCGVSGKIPGFSFLRRDGSFAGLDVDICRAFAAALTGSPDQVQYRSL TAPERFTALRTGEIDLLSRNTTFNLSRDAVGGNGVSFAPVVFHDGQGLLVKRQSGVSDL NDLQGKTICVGSGTTTEQNLNDAFQARGIDYKPIKYQDLNQVIAGYLQGRCSAMTSDRS QLAAARSGFNNPEQHVILPEVLSKEPLAPLAAGGDQRLADAMRWVIYALIAAEELGITQ QNIGDKLEEAQRRPELTQLRRFLGVEGELGEKLGLNNDFIVKVIQAVGNYGEIYNRHLG PDSAVPIPRGLNNLYRNGGVLTSPPFQ
CK_Syn_CC9902_00847	SDGTSSRLDLVQSRGELLCGVSGKIPGFSFLSPEGRYAGFDVDICKAMAAAFVGDSAKV QYRPLTAPERFTALRSGEIDLLSRNTTHTLSRDAAGGNGLSFAPTVFHDGQGLMVSVGS GVRSIADLSGKAICVGSGTTTEQNLNDAFAAQGLPYTPIKYQDLNQVVGGYLQGRCAAM TSDRSQLASARSGFKDPQQHVILEDRLSKEPLAPAVVGGDQRMADATNWVVYALIEAEE RGITQANLDDVLERAVADPSQAALRRFLGVDGGLGAKLGLPDDFVVQVIRATGNYGEIY NRHLGPGSAVTIPRGPNRLGVEGGLMIAPPFT
CK_Syn_MITS9220_01218	TLGDGDASRVDLIRKRGELRCGVSGKIPGFSFLQRDGSYAGLDVDLCKAFAAAIIDDPS KVQYRPLTAPERFTALKTGEIDLLSRNTTFNLSRDAAGGNGVSFAPVVFHDGQGLLVQR NSGIRSLKDLKGQAICVGSGTTTEQNLNDAFQSLGLPYKPIKYQDLNQVVAGYLQGRCQ AMTSDRSQLASARSGFERPEQHVILGDILSKEPLAPLSSGGDQQLSDAMRWVVYAVITA EEMGITQANVDNKLQEAQNNPDLTKLRRFLGVEGDLGDKLDLSNDFVVNVIRATGNYGE IYNRHLGPDSAVPIPRGLNQLHRDGGVITAPPFQ
CK_Syn_WH8102_00840	DEAGRSRLDLVKQRDELLCGVSGKIPGFSFLTPKGTYTGLDVDICRAMAAAFLGDADKV QYRPLTAPERFTALRSGEIDLLSRNTTHTLSRDAEGGNGLGFAPVVFHDGQGLIVPVNS GVTSLADLSGQAICVGSGTTTEQNLNDAFASKGIPYTPIKYQDLNQVVGGYLQGRCQAM TSDRSQLAAARSGFIEPEGHVILEDRLSKEPLAPAVVGGDQRLVDAMTWVVYALIEAEE RGITQANLDRQLRAAEADPSQAAMRRFLGVDGGLGRKLGLPDDFVVQAIRATGNYGEIY DRHLGPESPVAIPRGANRLAEDGGLMIAPPFT
CK_Syn_WH8109_00877	AGSRLDLVKARGELLCGVSGKIPGFSFLSSDGRYTGLDVDICRAMAAAFVGDAEKVQYR PLTAPERFTALRSGEIDLLSRNTTHTLSRDAVGGNGLRFGPVVFHDGQGLMVNAASGVR SLADLSGKSICVGSGTTTEQNLNDAFASKGLPYIPIKYQDLNQVVGGYLQGRCAAMTSD RSQLAAARSGFSDPEAHQILDDRISKEPLSPAVVGGDQPMGDAMTWVINGLIEAEERGI TQANVDAVVKQAAADSSQTALRRFLGVDPGLGRKLGLADDFVVQVIRATGNYGEIYNRH LGPDSAVAIPRGANRLAGEGGLMISPPFT
CK_Syn_CC9311_00634	AQQKDTQITSDSKNTILLGQSLPLSGPSAQIGKKYQAGAQAWFNEVNRQGGINGKKIRL ISLDDQYEPELTISNTKTLLEKPNLLALFGYVGTPTTKEILPFIEERKVPLIAPLTGAS ILRDEKLKMVVNLRASYQMEIDKIVDSLVRNARQKIAIIYQDDAFGKDGLKSAESALKK HGLKPIATATVQRNSAKIQSALQVLISSQPNAIIIISTYVSSAALSRELLKRDVNAQIM NVSFVGTRALEQSLPVGLANGIGVSQVVPFPWDRWIPVVADYQRLMRVNNSSARFGFTS LEGFMAARLITEGIKKVQGPLTKESLIKSLKSIKKVDLGGFQLDLSSNNKQASDYVELT FFGAQQWEP
CK_Syn_CC9311_02631	MTGALFAIWGGVNQTTKQDQFETAQTTTIASNDANRSVASENLDSLLDEEPRPWLLAQS LPLKGPSGLIGERFALGIDTVLRELNNRGGIAGRPVKIWRIDDGYEPENTLRNTRWFAA EPDVLALFGFFGTPTSKAALPIAKKAGLTLVAPLTGASALRAQGQTGVLHFRASYGEEA RRIVNHLVNDGFVRIAVAYQNDAYGKDVLASTEKALQSHDLRAVSTAALPRNSTETTQA ADTIVKSKPDALIVISLSKTMASLVNNVHRSGSRPQLMTISPTGTKALFRDLPQAAAFG VGVTQVVPFPWDARHPDVASYQRLLRQQQPDGDVDFDFYSLEGFMAAQWLVKAMEAIAP NVTRERLMKELKRTTPGLHRSIDLVFLGSDPWEP