Identification and Characterisation of Proteins within the Algal Pyrenoid

Caroline McKenzie

MSc by Research

University of York
Biology

March 2023

Abstract

Eukaryotic algae are responsible for approximately one-third of global carbon fixation. To maximise photosynthetic efficiency, algae have evolved biophysical carbon concentrating mechanisms (CCMs), concentrating inorganic carbon (Ci) to saturate Ribulose-1,5-bisphosphate carboxylase-oxygenase (Rubisco) with CO_2. Central to the algal CCM is the pyrenoid, a dynamic, Rubisco-containing organelle located in the chloroplast. Diatoms, an algal group with CCMs, are found in freshwater and ocean environments. They are important primary producers responsible for up to 20% of global carbon fixation. Current understanding of the diatom CCM is incomplete with key pyrenoid components unknown or uncharacterized. These components include linker proteins which bind Rubisco to drive pyrenoid formation, carbonic anhydrases (CAs) which release CO_2 for fixation by Rubisco and Shell proteins hypothesized to prevent CO_2 loss from the pyrenoid. This study uses bioinformatic techniques to investigate linker proteins, CAs and Shell proteins, particularly focusing on the model diatom *Thalassiosira pseudonana*. A novel combination of structural and phylogenetic analysis gives insight into the complexities of CCM component evolution between ecologically relevant algal lineages.

List of Contents

List of Tables

List of Figures

List of Supplementary Documents and Accompanying Material

Acknowledgements

I would like to thank the members of the Mackinder lab for supporting me during my MSc by research. In particular, I would like to thank Prof Luke Mackinder who has provided excellent guidance, supervision, and encouragement throughout. I have always been supported in carrying out high quality research, in spite of the barriers and challenges I have experienced. I would also like to thank Dr Benjamin Lichman for his contributions on my TAP panel, particularly regarding phylogenetic analysis.

I would like to thank the postdocs, students and technicians in the Mackinder lab, who have not only supported me academically, but throughout the joys and struggles of the past two years. Thank you to Dr Onyou Nam and Sabina Musical for discussions about diatoms, and Dr Charlotte Walker for her guidance, knowledge, and willingness to engage with my research. Dr Walker and Dr Abi Perrin have provided me with more encouraging words, hugs and cups of tea than I can count. For this I am very grateful.

To my friends, parents and church family, thank you for your enduring encouragement, prayers, practical support, listening ears and loving challenges. With the amount I have shared about my research in the past two years, I would not be surprised to find you as co- authors of this thesis!

Finally, thank you to the one who created the beautiful diversity of this earth. Thank you for the gift of asking questions, revealing complexity, and piecing it together. Thank you for research, scientists, and algae! Soli Deo Gloria!

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Identification and Characterisation of Proteins within the Algal Pyrenoid

Chapter 1: Introduction

Carbon Concentrating Mechanisms and the Algal Pyrenoid

Photosynthetic algae, in conjunction with cyanobacteria, are responsible for up to 50% of global carbon fixation(1–3). This conversion of CO_2 into organic carbon is catalysed by the enzyme Rubisco and supports nearly all life on earth(4). Despite being a highly abundant protein Rubisco has a slow catalytic rate and poor selectivity for CO_2 over O_2, resulting in energetically wasteful photorespiration reactions(5–7). To overcome these limitations multiple strategies have evolved including the biochemical C4, and CAM pathways in higher plants and mostly biophysical carbon concentrating mechanisms (CCMs) in algae(8). Structural and temporal differences in leaf anatomy, and vascular physiology and stomatal opening allow C4 and CAM plants to reduce photorespiration and increase the availability of CO_2 at Rubisco. On the other hand, biophysical CCMs operated by prokaryotic cyanobacteria and eukaryotic microalgae actively concentrate inorganic carbon (Ci) via a series of Ci transporters, conversion of the Ci species (CO_2 in HCO_3^-) by carbonic anhydrases (CAs), and aggregation of Rubisco to a distinct micro compartment(9–12). In cyanobacteria Rubisco and CAs are encapsulated by a self-assembling proteinaceous sheath to form a carboxysome (150-200 nm in diameter) (13–16). HCO_3^- is actively transported into the cytosol before diffusing into the carboxysome where it is dehydrated to CO_2 by CA at the site of Rubisco(17).

Eukaryotic algal, and a group of non-vascular plants called hornworts, also aggregate Rubisco but utilise larger microcompartments known as pyrenoids (~1-2 µm)(18). The pyrenoid is a dynamic phase separated organelle, penetrated by thylakoid tubules deriving from the wider thylakoid network(11). Through a series of transporters, HCO_3^- is concentrated across the plasma and plastid membranes, delivered to the acidic thylakoid lumen, and converted to CO_2 by CAs. CO_2 defuses into the surrounding Rubisco-containing matrix and is prevented from dissipating out of the pyrenoid by an outer sheath. This pyrenoid-based CCM is shared between eukaryotic algae, however CCM components and pyrenoid morphology vary greatly between species, reflecting the polyphyletic nature of eukaryotic algae and multiple evolutions of the pyrenoid(6). To date, the majority of research into algal CCMs has been focused on the model green alga *Chlamydomonas reinhardtii* however, algae are polyphyletic and recent years have seen growing interest in the model diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*.

The biogeochemical importance and evolution of diatoms

Diatoms are unicellular algae found in freshwater and marine environments. They play a key role within contemporary oceans, accounting for up to 20% of global photosynthesis (19). As a primary producer diatoms are principal contributors to biological carbon cycling. Through photosynthesis, inorganic atmospheric carbon (CO_2) is transformed into organic carbon, generating organic matter. This is either consumed or becomes dead organic matter, which is broken down and stored in sediments(20). Unlike other algal species, diatoms possess a silicified cell wall (known as a frustule) and are a key component in the biogeochemical cycling of silicon(21,22). Frustule shape can be used to morphologically classify diatoms as either circular centric (e.g. *T. pseudonana*) or oblong shaped

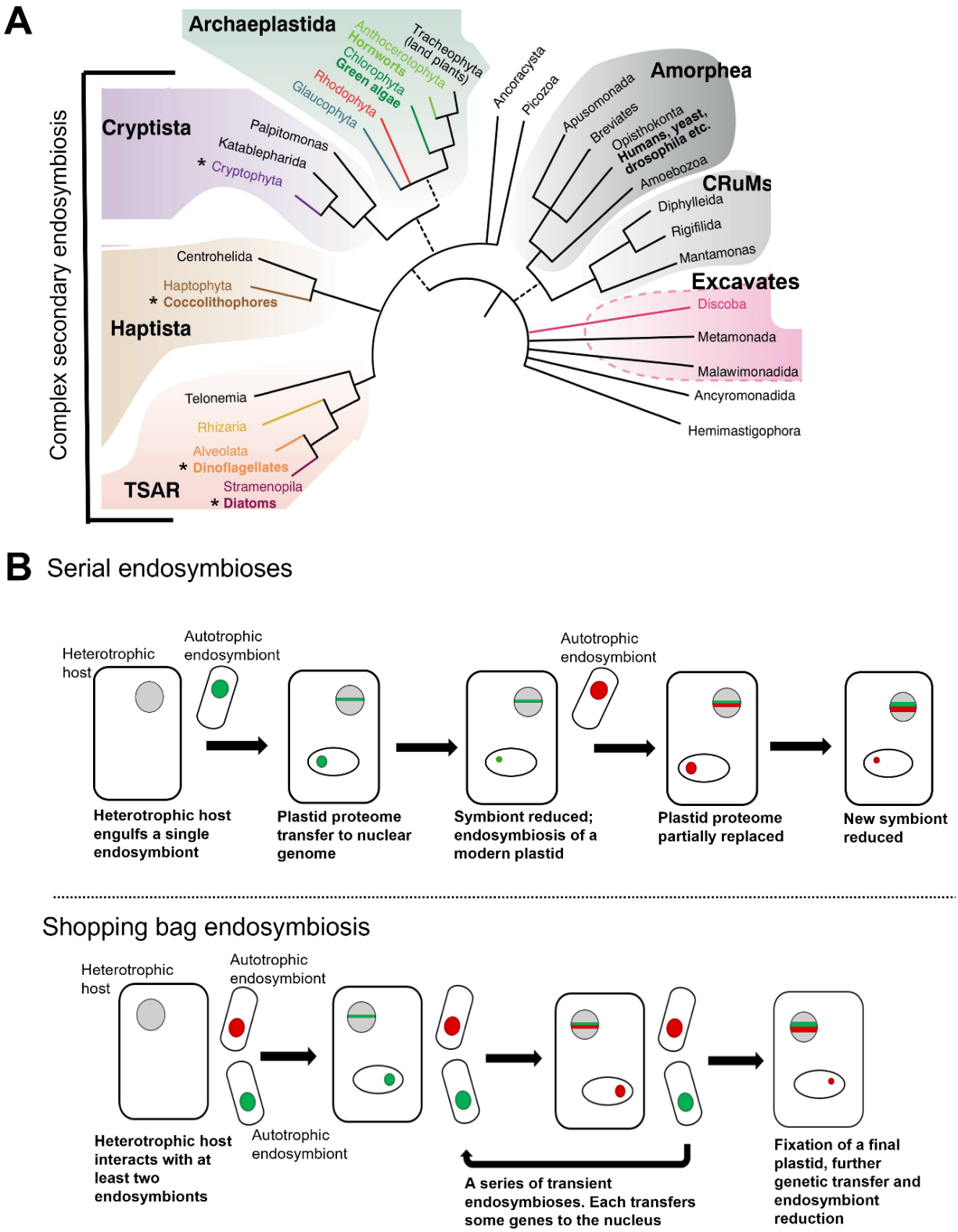


Figure 1 Pyrenoid containing algae are found across the eukaryotic tree of life. A) Tree adapted from Mackinder et al (11) . Names of pyrenoid-containing algae are coloured. Algae with plastid acquisition by complex secondary endosymbiosis are found in the TSAR, Haptista, and Cryptista superfamilies. Complex red plastid origin is indicated by an asterisk and includes diatoms. **B)** Schematic outlining the Serial and Shopping bag endosymbiosis models. Schematic based on Morozov et al (27). Circles represent origin of genes in a given genome (grey: heterotrophic host, red and green: corresponding algae). Smaller circles repeat represent reduced organellar genomes

pennate diatoms (e.g. *P. Tricornutum)*. In recent years, genomic analysis has provided additional understanding of diatom biodiversity and speciation. Wang et al carried out comparative Phylogenetic analysis of 35 diatom mitochondrial DNA genomes(23). They identified three classes within the Bacillariophta and Ochrophyta phyla to which diatoms belong. Centric diatoms are generally found in the classes Coscinodiscophyceae and Mediophycae which diverged from the pennate-containing Bacilariophyceae, around 131 Mya. Details of the diversification event leading to this morphological split are murky. It is thought that a combination of horizontal gene transfer, exchange of transposable elements and gene gain and loss led to the split around 70 Ma(20).

As stramenopiles, diatoms are situated within the TSAR superfamily (Fig. 1A) and were derived by secondary endosymbiosis of a photosynthetic plastid-containing organism. A non-synthetic eukaryote acted as a heterotrophic host, engulfing an autotrophic photosynthetic eukaryote and acquiring a chloroplast in the process(24). A legacy of this is the four-layer chloroplast membrane we see today in cryptophytes, haptophytes, dinoflagellates, and diatoms. It has long been hypothesised that an endosymbiosis event involving a red algal symbiont led to the evolution of diatoms. However, in recent years genomic studies have suggested plastid acquisition by secondary endosymbiosis is far more complex (25,26).

A phylogenomic study by Morozov and Galachyants found approximately equal evidence for red and green algal origins for diatom genes(27). Indeed, both *P. tricornutum* and *T. pseudonana* contain genes of red and green algal lineage, a phenomenon termed red-green mosaicism(28). Considering these findings two models of secondary endosymbiosis have been proposed. Firstly, the 'Serial endosymbiosis' model whereby plastids are acquired and then lost leading to so-called cryptic plastids, and secondly the 'Shopping Bag' endosymbiosis model where several plastid acquisition events and subsequent losses lead to red-green mosaicism of the nuclear genome (Fig 1B)(29–31). How complex secondary endosymbiosis may have influenced the role of plastids in modern diatoms is still uncertain. The subsequent impact on diatom pyrenoids and CCMs is an exciting avenue of research yet to be pursued.

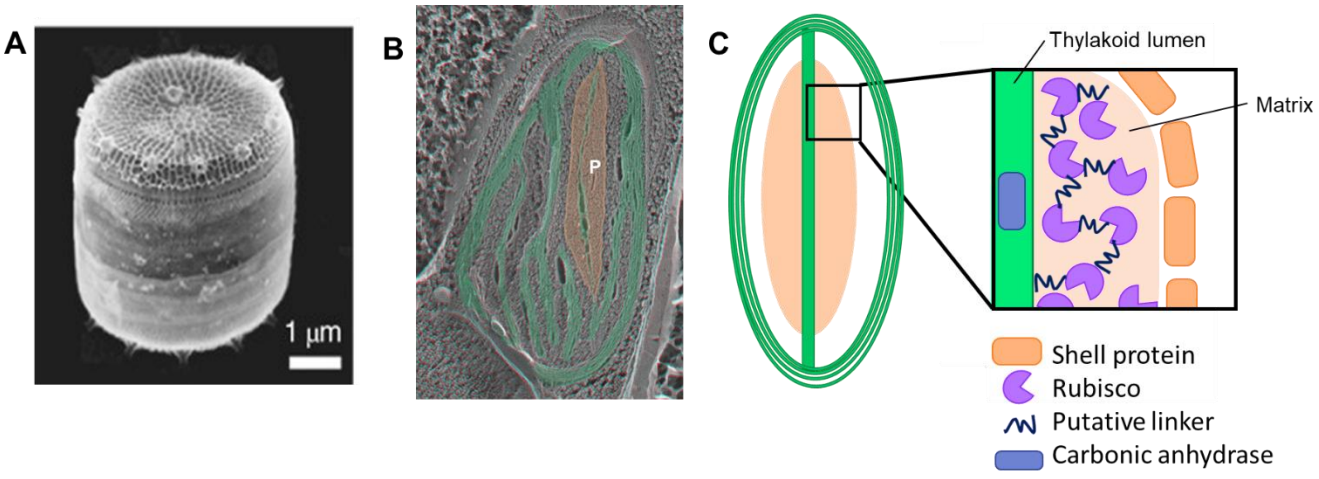


Figure 2 ***T. pseudonana*** **contains a lenticular pyrenoid A)** SEM image, Nils Kröger(110) **B)** TEM image Ursula Goodenough, P = pyrenoid, pyrenoid matrix highlighted in orange, surrounding thylakoids and pyrenoid-penetrating thylakoid highlighted in green. C) Cartoon representation of the T. pseudonana lenticular pyrenoid. Zoom in shows localisation of Thylakoid luminal CA, Rubisco, putative linker and shell proteins

Thalassiosira pseudonana: a model diatom

The genome of the centric model diatom *T. pseudonana* (Fig 2A) was first sequenced in 2004, and since multiple genetic engineering methodshave been developed(32,33). Despite this, little is currently known about the *T. pseudonana* CCM, particularly the pyrenoid and its components. *T. pseudonana* possesses a lenticular pyrenoid with a singular thylakoid tubule traversing the matrix (Fig. 2B). As in *C. reinhardtii*, *T. pseudonana* Rubisco localises to the pyrenoid matrix, however a comprehensive study of T. pseudonana components has yet to be undertaken. To identify CCM-related target proteins a colleague (Onyou Nam) carried out a series of co-immunoprecipitation experiments using *T. pseudonana* Rubisco. Proteins associated with Rubisco were 'pulled down' and separated from cellular debris with the aim of identifying true Rubisco interactors. Mass spectrometry identified protein targets potentially associated with Rubisco, the pyrenoid, or wider *T. pseudonana* CCM. Several novel proteins were identified, including a family of 7 proteins that were later shown to localise to the pyrenoid outer layer (Fig 1C). These 7 proteins, previously identified as unknown proteins by Bowler et al, were termed 'Shell proteins'(34). The pulldown results also identified the novel CA Tp1766. CAs regulate the flux of CO_2 throughout algal CCMs and characterising Tp1766 may give insight into the ways *T. pseudonana* delivers CO_2 to Rubisco.

CO_2 delivery to Rubisco

The model algae *P. tricornutum, T. pseudonana* and *C. reinhardtii* share similar CCM architecture and mechanisms despite containing evolutionary distinct components. Diatoms take-up both CO_2 in $HCO_3^-$from the environment. Uptake of CO_2 is by passive diffusion, driven by the low cytoplasmic CO_2 concentration and permeability of lipid bilayers to CO_2 (Fig. 3)(35). By contrast membranes are impermeable to HCO_3^- and transmembrane transporters are required. In *P. Tricornutum* a study by Nakajima et al showed Solute-Like Carrier 4 (SLC4) SLC4 family transporters directly uptake HCO_3^- from seawater across the plasma membrane(36). Fluorescent tagging localised SLC4 transporters to sites on the plasma and chloroplast membranes, supporting SLC4 as a carbon pump within the *P. tricornutum* CCM (Fig. 2)(36). Transcriptional data suggests the SLC4-2 homologue functions as a major uptake mechanism under low CO_2 (LC) (~0.04% CO_2) conditions.

Although SLC4 homologues have been found in *T. pseudonana* the picture here is less clear. Phylogenetic analysis of SLC4 HCO_3^- transporters in *P. tricornutum, T. pseudonana*, and phytoplankton by Hopkinson et al showed that the clade including SLC4-2 does not include a homologue from *T. pseudonana*(37). There is also little evidence showing SLC4 is responsible for HCO_3^- uptake at the plasma membrane in both *P. tricornutum* and *T. pseudonana*, leaving this area still open for further research.

The freshwater green alga *C. reinhardtii* uses a different array of transporters facilitate C_i flux across intracellular membranes. Under VLC conditions High Light Activated 3 (HLA3) and limiting CO_2 inducible 1 (LCI1) localise to the plasma membrane facilitating Ci transport into the cytoplasm (Fig 3)(12,38,39). At the chloroplast envelope LCIA transports HCO_3^- into the stroma, before delivery to the thylakoid lumen by Bestrophin-like proteins 1-3 (BST1-3) (40). CAs play a key role in this carbon concentration process. Limiting-CO_2-inducible B and C (LCIB and C) complex and localise to the pyrenoid periphery under VLC conditions, and within the thylakoid lumen CAH3 converts HCO_3^- two CO_2 for fixation by Rubisco(38,41).

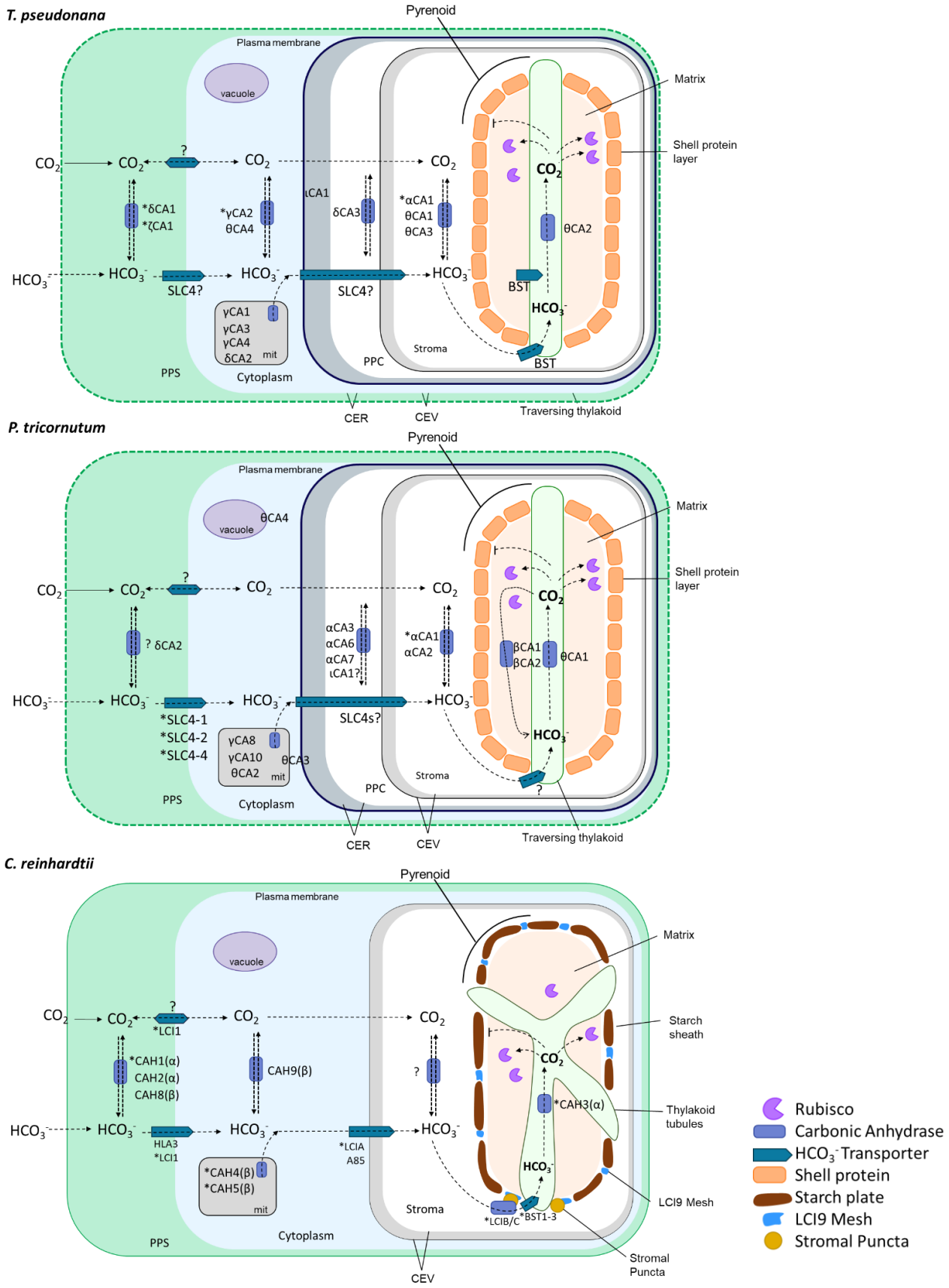


Figure 3. CCM models of *T. pseudonana*, *P. tricornutum* and *C. reinhardtii i*(42,44–46,51,78,111,112)**.** PPC: Periplasmic Space; mit: Mitochondrion; CER: Chloroplast Endoplasmic reticulum; PPC; Periplasmic Compartment; CEV; Chloroplast Envelope.

The Role of CAs within the CCM

To reduce dissipation of CO_2 from the pyrenoid, chloroplast and cell efflux is CO_2 regulated by interconversion is of CO_2 and HCO_3^- (figure 3). Specific localisation of CAs control the magnitude and direction of Ci fluxes, and therefore carbon concentration towards the pyrenoid. The contrasting membrane permeability of CO_2 and HCO_3^- is capitalised on by CAs which catalyse CO_2/HCO_3^- interconversions (37). This is aided by pH changes across membranes, which shift CO_2: HCO_3^- equilibrium (Equation 1) in favour of either $HCO_3^-$for concentration, or CO_2 release for fixation by Rubisco(11). CAs are classified into eight different subtypes (α, β, γ, δ, ζ, η, θ and ι) based on amino acid sequence, active site structure and metal cofactor(42). All eight CA families are present in photosynthetic microorganisms, however distribution and subcellular localisation of CAs vary between algal species.

$$HCO_3^- + H^+ \rightleftharpoons H_2CO_3 \rightleftharpoons CO_2 + H_2O \quad \textbf{[Equation 1]}$$

In *P. tricornutum* 15 putative CAs from five families (α, β, ɣ, θ, and ι) have been identified and localised to the cytosol, mitochondria, periplasmic compartment, stroma, and pyrenoid-penetrating thylakoid tubule (Fig. 4)(43,44). Interestingly, *P. tricornutum* specific localisation appears to reflect CA subtype: gamma in the mitochondria, alpha in stroma and PPC, beta in the pyrenoid matrix and theta in the pyrenoid tubule(45). However, recent localisation by Matsuda et al has placed a theta CAs in several compartments (46). Although CA activity only been confirmed in the two beta CAs (PtβCA1-2) and theta CA (PtθCA1), intact zinc binding sites in most of the other CAs suggests active site functionality(44,47). PtβCA1 and 2 are highly CO_2 responsive at the transcriptional level, potentially playing a key role within the pyrenoid matrix(48–50).

By contrast, in *T. pseudonana* 14 putative CA genes have been identified, however subcellular localisation does not appear related to CA subtype. CAs from six CA subgroups (α, δ, ɣ, ζ, θ and ι) are distributed in the stroma, PPC, CER, mitochondria, cytoplasm, PPS, and pyrenoid thylakoid lumen (Fig. 3). δCA1, ζCA1 and δCA3 appear to play a role in acquiring HCO_3^- and recapturing leaked CO_2, as transcript levels increase under LC conditions(51). The remaining *T. pseudonana* CAs are CO_2 responsive, including cytoplasmic and external CAs, however further research is required to confirm their involved in the CCM.

Recently four *T. pseudonana* theta CAs (TpθCA1-4) have been localised to distinct subcellular compartments(46). Fluorescent localisation showed TpθCA2 localises to the pyrenoid-penetrating thylakoid lumen, and is hypothesized to have an analogous function to *C. reinhardtii* CAH3 (Fig 3)(46). A fifth novel CA (Tp1766) was identified by Onyou Nam, however much is still unknown about this protein including the CA subclass categorisation, subcellular localisation and functionality within the *T. pseudonana* CCM.

Rubisco assembly within the pyrenoid matrix

Within eukaryotic algae tight packing of Rubisco into the pyrenoid matrix is required for an efficient CCM (52,53). The pyrenoid matrix was initially thought to be crystalline, possessing a lattice arrangement(54). However, it is now understood that the pyrenoid is a dynamic organelle formed by a biophysical process known as Liquid-Liquid Phase Separation (LLPS) (55). LLPS is a physical process in which a homogeneous solution spontaneously demixes into two phases: one dense, the biomolecular condensate, where macromolecules are concentrated (e.g. Rubisco), and a second depleted phase (56). This allows compartmentalisation of proteins within a solution, without the formation of a membrane. Banani et al identified LLPS to be driven by the collective protein-protein

interactors of scaffold proteins, which act as linker proteins in the condensate(57). Scaffold proteins are characteristically multivalent, of low sequence complexity and intrinsically disordered proteins (IDPs). Through intra and inter protein interactions, scaffold proteins drive demixing and formation of dense phase liquid droplets. In *C. reinhardtii* this scaffold or linker protein has been identified as Essential Pyrenoid Component 1 (EPYC1). EPYC1 is essential for normal pyrenoid size, number, morphology, Rubisco content, and efficient carbon fixation under LC conditions(52). *In vitro* analysis has shown both EPYC1 and Rubisco are necessary and sufficient for driving LLPS, droplet formation and aggregation of the pyrenoid matrix(58). EPYC1 interacts with Rubisco via several repeated Rubisco binding motifs (RBMs) which directly bind alpha helices on the Rubisco small subunit (SSU), via salt bridges and hydrophobic interactions(52,59). Several further C. reinhardtii pyrenoid proteins, including the starch binding proteins SAGA1 and 2 and the transmembrane proteins RBMP1 and 2, also share this RBM(60). This suggests that RBMs mediate binding between the three pyrenoid sub-compartments (thylakoid tubule, matrix, and outer layer) and facilitate pyrenoid formation.

The organisational principles of RBMs are hypothesised to apply to a broad range of pyrenoid-containing algae, although specific sequences and proteins may differ to *C. reinhardtii* (60). In diatoms, unpublished work by the Mueller-Cajar Lab has identified a putative *P. tricornutum* linker protein, named pyrenoid component one (PYCO1). Early *in silico* and *in vitro* analysis suggests some similarity to EPYC1, however further research is needed to fully understand PYCO1's role in the CCM. An analogous linker protein has yet to be identified in *T. pseudonana*. As such, this study initially aimed to identify a *T. pseudonana* linker protein using a bioinformatic approach to search for EPYC1 and PYCO1 homologs. As discussed in section 2.4, this was unsuccessful, and instead a bioinformatics comparison of EPYC1 and PYCO1 was undertaken to identify conserved linker protein characteristics.

A diffusion barrier surrounding the pyrenoid

To prevent leakage of CO_2 back into the stroma, the third pyrenoid sub-compartment, the outer layer, is proposed to act as a diffusion barrier (61). The structural components and spatial organisation of the pyrenoid outer layer are best understood in *C. reinhardtii*. *In vivo* imaging techniques have identified at least three distinct components to the outer layer surrounding the pyrenoid; 1) a starch sheath, interspersed with 2) a LCI9 containing mesh-like structure, and 3) LCIB/C containing puncta around the periphery (Fig. 3)(62). The starch sheath develops rapidly under LC conditions and comprises of several starch plates organised in a homogenous distribution around the pyrenoid periphery(54,62,63). Fluorescent localisation data suggests that LCI9 has complimentary localisation forming a mesh-like structure between the starch plates, perhaps performing a structural function(62). As previously mentioned, the LCIB/C complex prevents dissipation of CO_2 from the matrix under VLC by recapturing and converting CO_2 to HCO_3^- for uptake into the thylakoid lumen(64).

A different pyrenoid outer compartment is present in diatoms. Within the *T. pseudonana* Rubisco pulldown a Onyou Nam identified a family of proteins, named 'Shell' proteins, which when fluorescently tagged localised to the periphery of the *T. pseudonana* pyrenoid. In parallel, work was presented at the CCM10 conference in August 2022 by Yusuke Matsuda and Ben Engel regarding the identification and characterisation of a proteinaceous pyrenoid Shell within the pennate diatom *P. tricornutum.* Current data indicates that diatoms lack the ability to make starch within the chloroplast and this protein shell presents a possible diffusion barrier surrounding the pyrenoid matrix. Packaging Rubisco within a proteinaceous compartment would not be unique to diatoms as the cyanobacterial carboxysome is also proteinaceous(15).

Aims and Objectives

Many questions remain unanswered regarding the protein components of the *T. pseudonana* CCM. Through bioinformatic methods this study aims to address three categories of CCM protein: linker proteins, CAs, and pyrenoid Shell proteins.

1. To identify pyrenoid linker protein targets in *T. pseudonana* using sequence homology searches of EPYC1 and PYCO1. To briefly compare EPYC1 and PYCO1 through sequence and structural analysis, identifying conserved characteristics.
2. To categorise the CA family of the novel *T. pseudonana* CA Tp1766 through phylogenetic analysis, using this to gain insight into the evolution of algal CCMs.
3. To characterise the seven *T. pseudonana* Shell proteins using phylogenetic, sequence, and structural analysis. The relationships between Shell protein structure, function, localisation and evolutionary origin will be investigated by combining analysis techniques

Pursuing these avenues of research will contribute to understanding of the *T. pseudonana* CCM with evolutionary implications for a wide range of algal species. Combining phylogenetic and structural biology approaches is a novel technique within the field of algal CCM research, and offers an exciting avenue of research.

Chapter 2: Bioinformatic comparison of pyrenoid linker proteins

2.1 Chapter Summary

Over the last 10 years, research in the model green alga, *C. reinhardtii* has identified the pyrenoid as a dynamic structure formed by LLPS (57). LLPS is a dynamic process during which a homogenous liquid separates into distinct phases or compartment (38). The dense compartments are known as biomolecular condensates, the assembly of which can be driven by attaining critical thresholds of specific, multi-valent, intrinsically disordered scaffold proteins (58). In the *C. reinhardtii* pyrenoid matrix, the disordered protein EPYC1 uses multivalency of the five designated RBMs to drive pyrenoid matrix formation (13,57). In the model diatom *P. tricornutum,* an alternate linker protein, Pyrenoid Component 1 (PYCO1) has been proposed by our collaborators in Oliver Mueller-Cajar's group, Nanyang University, Singapore. This study aimed to identify a pyrenoid linker protein in *T. pseudonana* by searching for PYCO1 and EPYC1 homologues. Unfortunately, both BLAST search and an unpublished bioinformatics pipeline in the lab that identifies linker proteins based on physicochemical properties, failed to identify any EPYC1 or PYCO1 *T. pseudonana* homologues/analogues. In lieu of putative *T.pseudonana* linker candidates, this study undertook a brief bioinformatic comparison of EPYC1 and PYCO1, to investigate conserved properties of linker proteins.

2.2 Introduction

The *C. reinhardtii* pyrenoid linker protein EPYC1 is highly abundant, induced under LC conditions, and essential for a functional CCM (35,41). In EPYC1 knockout mutants Rubisco does not phase separate and as a result, pyrenoids do not form (35). During *in vitro* demixing assays both EPYC1 and *C. reinhardtii* Rubisco SSU are required for formation of a biomolecular condensate. The mechanism behind EPYC1's phase separating ability rests largely on its protein structure and physicochemical properties. The EPYC1 sequence consists of five repeat regions each containing a RBM (Fig. 4) (35,37,41). Disordered regions separate RBMs, and span the distance between Rubisco holoenzyme binding sites forming the pyrenoid matrix (37). Using the bioinformatics pipeline Flippr, EPYC1 homologues have been identified in the green algal Ulva and Chlorella species. However, EPYC1 homologues do not seem present in diatoms.

The putative linker PYCO1 phase separates Rubisco in in the model diatom *P. tricornutum*. Unpublished work by Mueller-Cajar et al. has identified a chloroplast transit in PYCO1 sequence suggesting localisation to the chloroplast. Subsequent fluorescent imaging localised PYCO1 to the *P. tricornutum* pyrenoid matrix and demixing assays, and partitions diatom Rubisco into biomolecular condensates. Demixed droplets show slightly different characteristics to those seen in *C. reinhardtii*; droplets appear 'glassy' and phase separation less dynamic. These characteristics suggest PYCO1 is acting a pyrenoid matrix linker. Analysis of the PYCO1 primary sequence revealed a mostly disordered protin with repeated regions. Each repeat contains a 'KWSP' motif predicted to act as a RBM binding the Rubisco SSU to stimulate matrix formation. These findings have led Mueller-Cajar et al to propose PYCO1 as a EPYC1 analog. In this study, bioinformatic analysis techniques were used to compare physiochemical characteristics of these two algal linker proteins.

2.3 Methods

All supplementary files (File) can be found within the online repositary: https://github.com/CGMcKenzie/Supplementary_Files_Thesis_Caroline_McKenzie

Bioinformatic search for EPYC1 and PYCO1 protein homologues. BLAST searches of PYCO1 and EPCYC1 were conducted on the publicly accessible portal Genebank/National Centre for Biotechnology Information(65). Supplementary figures (Sf). 1 and 2. Transcript IDs: EPYC1, Cre10.g436550.t1.2; PYCO1, Phatr3_J49957.t1.

Bioinformatic characterisation of EPYC1 and PYCO1. Rapid automatic detection and alignment of repeat (RADAR) regions were generated using the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL- EBI) web service, selecting defaults(66). Protein Disorder Profiles were created using the open access Predictor of Natural Disordered Regions (PONDR) web service selecting the VSL2, VL3, VL-XT, XL1-XT algorithms(67).

Protein modelling of PYCO1.PYCO1 peptide repeat region structural model was generated by entering the sequence 'PGQAYAGSGPRKNYSMVKWSPRGG' PEP-FOLD3.5 selecting no reference model, 5 structures, and all other default settings (File 1 *20211026_PYCO1_repeat region_model1-5*) (68) using PyMOL2.0 (Molecular Graphics System, Version 2.0 Schrödinger, LLC) Models 2 and 3 were aligned to model 1 using the following alignment protocol (Sf. 3, File. 2 *20211026_PYCO1_5 models*)

PyMOL Align protocol: Choose selection > Align function > to molecule/selection.

Protein modelling of Rubisco SSUs The *C. reinhardtii* Rubisco SSU holoenzyme crystal structure (PDB ID 7JFO) was refined in PyMOL2.0 by manually removing residues (File 3 *230328_EPYC1_SSU*) (62). The *P. tricornutum* Rubisco SSU was modelled using the open access web service AlphaFold2 Jupyter Notebook for Google Co-laboratory (referred to as AlphaFold Co-lab)(69). The full FASTA sequence (Accession number: AAV69748) was queried, selecting a model = 1 and all other default settings (File 4 *20211120_PtSSU_1_7f305_unrelaxed_model_1_rank_1*). To model the *T. pseudonana* Rubisco SSU the full length FASTA sequence (Accession number YP_874497) was threaded into *Thalassiosira antarctica* Rubisco SSU crystal structure (PDB ID: 5MZ2/I) using the Protein Homology/analogY Recognition Engine V 2.0 (Phyre2). On 'Expert mode'' One-one threading ' was selected with defaults (File 5 *20211025_Tpthreaded_SSUintoTA_1*) (70). Files 4 and 5 were imported into PyMOL2.0 and were aligned to File. 3 using the PyMOL alignment protocol (Files 6*20211126_TP_PT_CR.2*). Linker binding residues were manually selected and visualised by selecting Show > liquorice sticks

2.4 Results and discussion

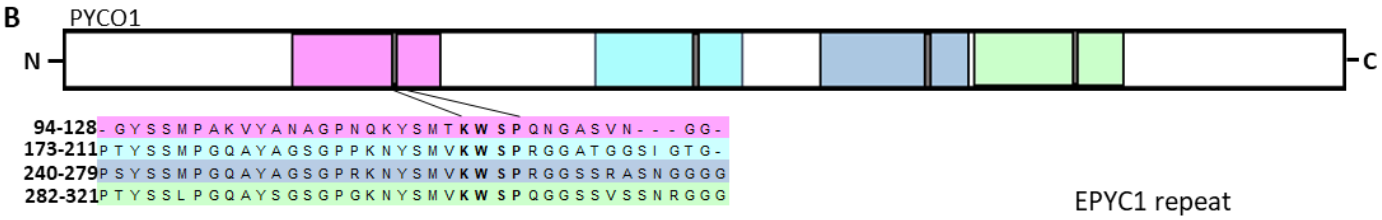
Bioinformatic analysis of EPYC1 and PYCO1 was unsuccessful in identifying a *T. pseudonana* homolog. An EPYC1 BLAST search identified 14 proteins of significant alignment all from green algal species (Sf. 2) suggesting EPYC1 homologs are found within the Chlorphyceae clade. This finding was consistent with research by Meyer et al. who identified homologs of *C. reinhardtii* RBM-containing pyrenoid proteins in several green algal species(60). A lack of hits from the stramenopiles clade suggests EPYC1 homologues are not present in the pyrenoids of diatom species. Interestingly, when PYCO1 was subjected to BLAST search, only a single protein hit (PYCO1 itself) was identified. The complete absence of homologues suggests PYCO1 is unique to *P. tricornutum,* and not conserved between diatom species, including *T. pseudonana*.

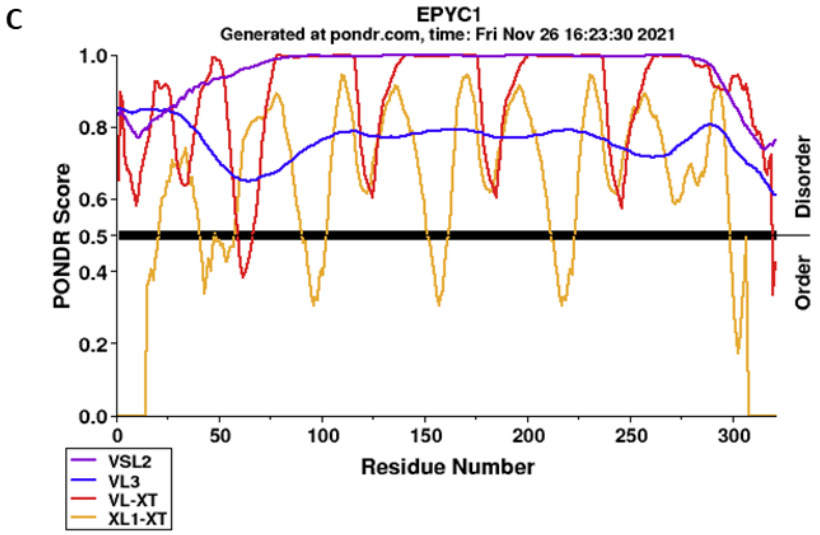
Both PYCO1 and EPYC1 have repeat regions containing RBMs. Sequence analysis of EPYC1 validated the presence of 4 tandem repeats ~60 amino acids in length, and a fifth much shorter repeat at the C-terminus (Fig. 4A). Each repeat contains an 11 residue RBM followed by an additional 6 residue RBM at the C- terminus. Bioinformatic analysis of PYCO1 detected four repeat regions each containing the KWSP RBM (Fig. 4B). In contrast to EPYC1, the PYCO1 repeat regions were much shorter (~37 amino acids), dispersed unevenly throughout the sequence, with no shortened C-terminus RBM. This comparison suggests RBM-containing repeats are a characteristic of linker proteins, however repeat and RBM properties are not conserved. The length of the linker (non-RBM) section of each repeat is much shorter in PYCO1 compared to EPCY1. On the one hand, this is surprising as environmental selection pressures mean that physicochemical properties such as repeat length are often conserved between disordered proteins (60). However, the shorter linker region in PYCO1 may be a consequence of binding a different region of the SSU to EPYC1. The flexible linker region spans the distance between SSUs and binding to the solvent channel may reduce this distance in *P. tricornutum.*

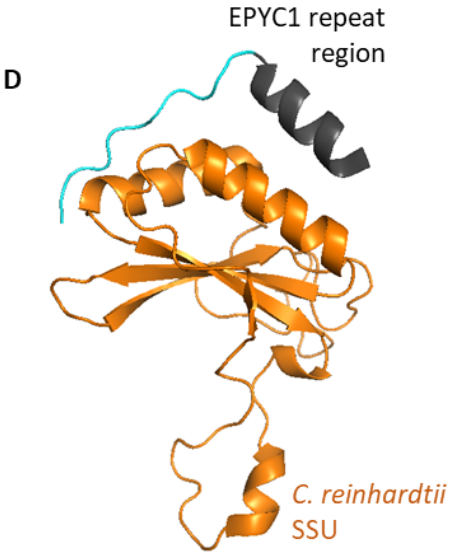
EYPC1 and PYCO1 are both highly disordered but differ in repeat characteristics, disorder distribution, structure and binding mechanism. Comparing the disorder profiles of EPYC1 and PYCO1 reveals these IDPs differ in disorder pattern. The regularly spaced troughs within the EPYC1 disorder profile indicate a repeating section of secondary structure dispersed by disordered regions of a similar length (Fig.4C). The length of these disordered repeats (distance between troughs) reflects the repeat length of the flexible linker region within the primary sequence (Fig. 4A). This is consistent with the Rubisco-bound EPYC1 crystal structure in which each repeat region comprises of an alpha helix (RBM) and flexible domain (linker motif)(Fig.4D) (68). The combination of order and disorder in EPYC1 concurs with research into IDPs which suggests inherently unstructured proteins can retain some preferred structures, such as alpha helices (60). Functionally, this repeating pattern of structure and disorder allows EPYC1 to both bind the Rubisco SSU alpha helix whilst dynamically linking to additional Rubisco holoenzymes enabling pyrenoid matrix formation(55,60).

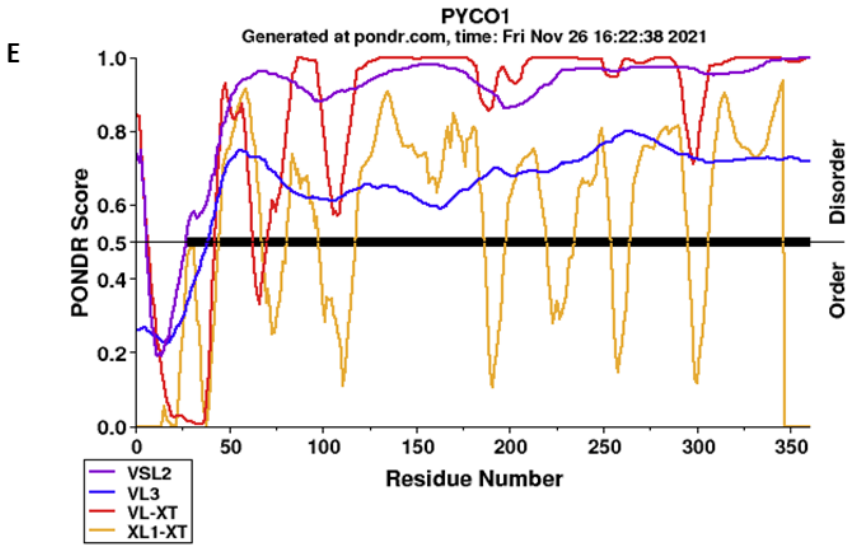
Figure 4. Comparison of the linker proteins EPYC1 and PYCO1. **A)** EPYC1 has 4 repeat regions (highlighted in blue, cyan, green, and magenta), each containing a RBM (black bars and bold type) SSU. **B)** PYCO1 has 4 repeat regions (highlighted in magenta, cyan, blue, and green), each containing the KWSP motif (black bars and bold type) There is a fifth RBM at the C-terminus. **C)** The EPYC1 disorder profile with a repeating pattern, reflecting the alpha helical (grey) and disordered domains (blue) of each repeat (D). **D)** EPYC1 binds the RuBisCO small subunit via a surface alpha helix. **E)** The disorder profile of PYCO1, which lacks a clear pattern of disorder. **F)** Alignment of the 3 best models of PYCO1 repeat region

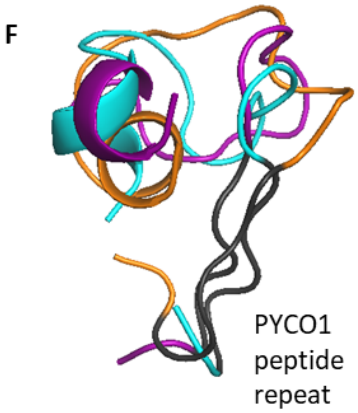

A
EPYC1
N
C
52-114 VSPTRSVLPANWRQELESLRNGNGSSSAASSAPAPARSSSASWRDAAPASSAPARSSSASKKA
115-174 VTPSRSALPSNWKQELESLRS--SSPAPASSAPAPARSSSASWRDAAPASSAPARSSS-SKKA
175-235 VTPSRSALPSNWKQELESLRS--SSPAPASSAPAPARSSSASWRDAAPASSAPARSSSASKKA
236-291 VTPSRSALPSNWKQELESLRS--NSPAPASSAPAPARSSSASWRD-APASS----SSSSADKA
302-317 PEIKRTALPADWRKGL


B
PYCO1
N
C
94-128 -GYSSMPAKVYANAGPNQKYSMTKWSPQNGASVN---GG-
173-211 PTYSSMPGQAYAGSGPPKNYSMVKWSPRGGATGGSIGTG-
240-279 PSYSSMPGQAYAGSGPRKNYSMVKWSPRGGSSRASNGGGG
282-321 PTYSSLPGQAYSGSGPGKNYSMVKWSPQGGSSVSSNRGGG


C
EPYC1
Generated at pondr.com, time: Fri Nov 26 16:23:30 2021
PONDR Score
Disorder
Order
Residue Number
VSL2
VL3
VL-XT
XL1-XT


D
EPYC1 repeat region
C. reinhardtii SSU


E
PYCO1
Generated at pondr.com, time: Fri Nov 26 16:22:38 2021
PONDR Score
Disorder
Order
Residue Number
VSL2
VL3
VL-XT
XL1-XT


F
PYCO1 peptide repeat

In contrast to the regular repeating disorder profile of EPCY1, four out of five neural networks analysed predicted PYCO1 to be almost completely disordered (Fig 1G). Only the long read network XL1-XT predicted multiple ordered regions, however these are not in a regular repeating pattern as would be expected if reflecting the KWSP motif. Indeed, such high disorder prevented confident peptide modelling of the PYCO1 repeat region. Alignment of the three best PYCO1 repeat region structural models generated in Phyre2 (figure 5H) gave Root Mean Square Deviation (RMSD) values of 4.109 Å between models 1 and 2; 5.647 Å, models 1 and 3; 5.738 Å models 2 and 3). A single model could not be confidently assigned as the PYCO1 repeat structure as RMSD values of 0-2 Å are considered to reflect structural similarity.

The *C. reinhardtii*, *T. pseudonana* and *P. tricornutum* Rubisco SSU structures suggest different mechanisms of linker action. *In silico* alignments of the *T. pseudonana*, *P. tricornutum* and *C. reinhardtii* Rubisco SSU showed a similar overall structure, with low RMSD values 0.717 (*T. pseudonana*), and 0.734 (*P. tricornutum*) (Fig. 5). However, analysis of linker binding residues revealed distinctions between *T. pseudonana*, *P. tricornutum* and *C. reinhardtii.* CryoEM imaging has previously shown the EPYC1 RBM binds specific residues on both Rubisco SSU alpha helices (Figs. 4C, 5) (59). A brief comparison of the *T. pseudonana* and *P. tricornutum* Rubisco SSUs revealed not all residues involved in hydrophobic and electrostatic interactions with EPYC1 are conserved. Although all three Rubisco SSUs possessed an aspartate residue aligning to *C. reinhardtii* D23, differently charged amino acids aligned to position V94. In the *C. reinhardtii* SSU at position M67 smaller alanine residues are present in diatom species. These differences could prevent a EPYC1 homolog binding to the SSU helices and may explain why PYCO1 is proposed to bind an alternative region of the Rubisco SSU.

Unpublished CryoEM data from Mueller-Cajar et al has suggested PYCO1 binds to the solvent channel of Rubisco using the repeating KWSP motif. The individual residues predicted to bind PYCO1 are located at the *P. tricornutum* SSU C-terminus, a disordered region lacking secondary structure (Fig. 5). The structural alignment of Rubisco SSUs revealed that, the *C. reinhardtii* SSU does not possess a flexible C-terminus aligning to those of *T. pseudonana* and *P. tricornutum.* Furthermore, the pair of beta strands immediately preceding the C-terminus in the diatom species is absent in *C. reinhardtii*. These finding suggest that although both EPYC1 And PYCO1 phase separate Rubisco, they are employing different mechanisms to do so.

Chapter Conclusion. The comparison of EPYC1 and PYCO1 carried out in this study suggests that algal linker proteins are intrinsically disordered, a feature key to LLPS. However, differences in repeat length, RBM, and Rubisco SSU binding residues show linker mechanisms are not conserved between algal species. Linker proteins may have evolved convergently to facilitate pyrenoid phase separation and matrix formation. If so, this would reflect the multiple evolutionary origins of the algal pyrenoid(11,53).

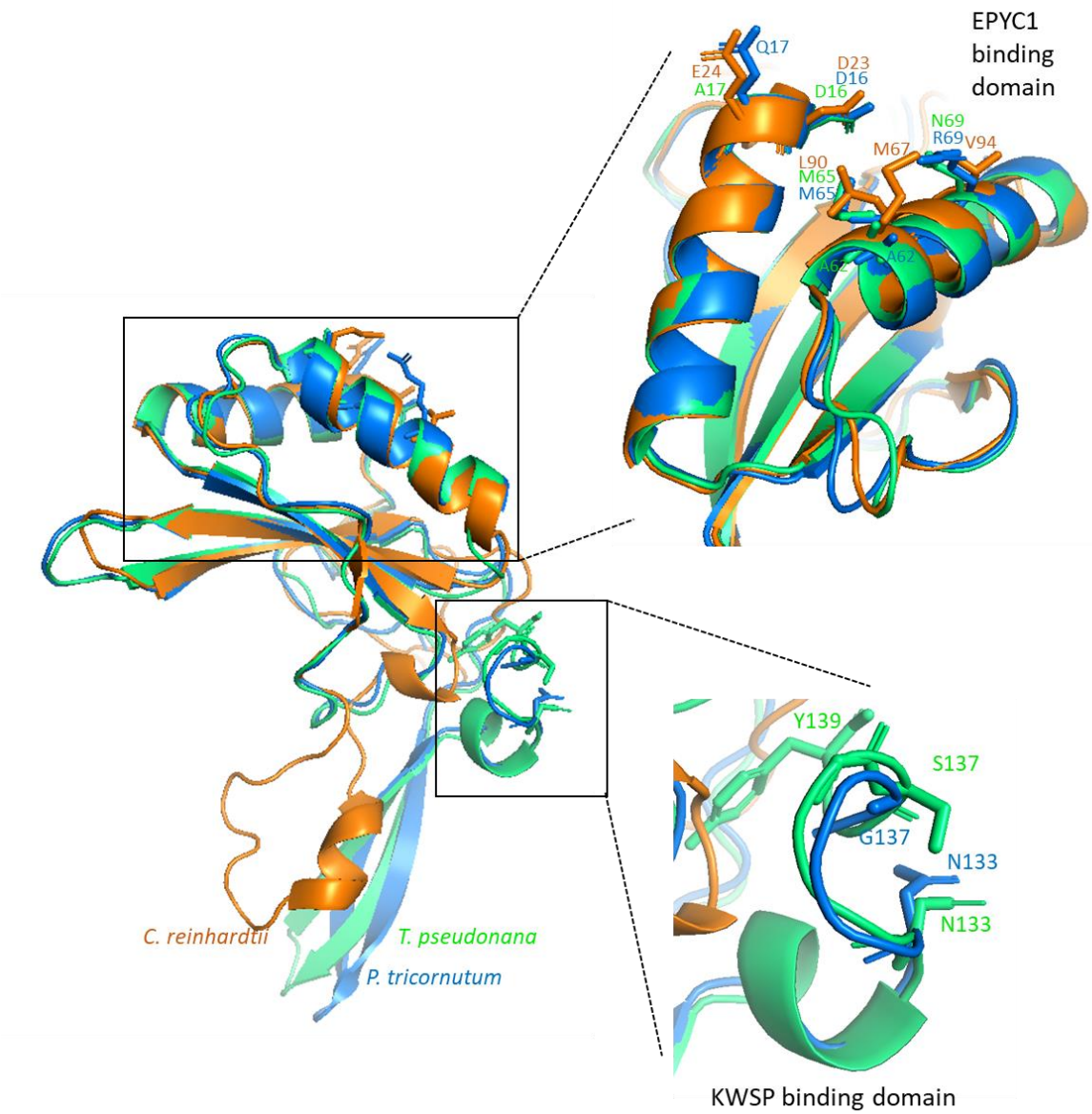


Figure 5. Alignment of the *C. reinhardtii, P. tricornutum*, and *T. pseudonana* Rubisco SSUs. *P. tricornutum* (blue) and *T. pseudonana* (green) SSU models and were aligned to the *C. reinhardtii* R SSU (orange). EPYC1 binding residues are located within the two external facing alpha helices (shown top right hand corner). The KWSP domain is location at the C-terminus of the *P. tricornutum* SSU. Corresponding RBM- binding residues are shown in the aligned Rubiscos.

Chapter 3: Carbonic anhydrases and the CCM

3.1 Chapter Summary

Carbonic anhydrases are crucial within algal CCMs actively concentrating carbon through the interconversion of CO_2 and HCO_3^-. Divided into 8 subclasses (α, β, γ, δ, ζ, η, θ and ι) recent work has characterised four *T. pseudonana* CAS, TpθCA1-4, determining them theta CAs. It is not yet known into which CA family the novel *T. pseudonana* CA Tp1766 is grouped. To investigate Tp1766 using bioinformatic methods a dataset of eukaryotic CAs was created and analysed phylogenetically. Protein sequences of known CA families were identified across stramenopiles, alveolates, haptophytes and green algal lineages. After several rounds of refinement a phylogenetic tree representing the eight distinct CA families was used to categorise Tp1766. Comparing previously determined subcellular localisations of TpθCA1-4 with the phylogenetic tree gave further insight into the relationship between CA localisation and evolutionary distance. Further analysis of the tree provided insight into the categorisation of the *C. reinhardtii* CAs LCIB and LCIC. This phylogenetic analysis of broad range of CAs has provided specific family identification for Tp1766 and has implications for pyrenoid evolution.

3.2 Introduction

The algal CCM relies upon the active uptake of carbon dioxide, transport of HCO_3^- across membranes, and recapturing of CO_2 to prevent leakage(62). Central to these processes are CAs which catalyse the interconversion of CO_2 and HCO_3^- (Fig 3) (71)CAs regulate the flux of Ci across membranes within algal cells, ensuring saturation of CO_2 at the active site of Rubisco(72). Within photosynthetic organisms there are 8 distinct CA protein families (α, β, γ, δ, ζ, η, θ, and ι) (73,74). Phylogenetic analysis can be used to distinguish between CA families and has previously been used to identify the new iota subclass (73). Whilst structural traits are often conserved between CA families, for example delta and alpha CAs have similar active site structures, there is often little primary sequence similarity (74). This is characteristic of convergent evolution, by which CAs are thought to have evolved(75).

Different algal lineages appear to utilise different familial CAs within the CCM. In *C. reinhardtii*, the CAs LCIB and LCIC form a complex. The LCIB/C complex is localised to the stroma, associated with the pyrenoid at below atmospheric CO_2 levels (VLC) and critical for survival under atmospheric (LC) levels of CO_2. This suggests LCIB/C plays key role in the *C. reinhardtii* CCM (76). Categorising LCIB appears less straightforward. Structural analysis reveals resemblance of a beta CA (77)whereas phylogenetic analysis suggests LCIB is a member of the theta CA family (73,74). In *P. tricornutum* several CAs contribute to the CCM. The theta CA PtθCA1 is located within the singular pyrenoid penetrating thylakoid, and directly supplies Rubisco with CO_2 (44). It contains a cys-gly-his-rich (CGHR) domain thought to be a hallmark of theta CA.

Within *T. pseudonana*, CAs from 5 different protein families localise to different cellular sub-compartments(Fig 4) (80). Several of these including extracellular δCA1 and ζCA1, cytoplasmic ɣCA2 and stromal αCA1 are shown to be LC inducible suggesting involvement in the *T. pseudonana* CCM (72,78,79). Localisation of *T. pseudonana* CAs has recently been updated by recent work done by the Matsuda Lab on theta type CAs(46). Four putative theta CAs candidates (TpθCA1-4) (Table 1) have been identified(30). Sequence analysis, GFP fluorescent tagging and immunogold labelling of TpθCA1-4 identified subcellular localisation. TpθCA1 and TpθCA3 join TpαCA1 in the stroma, and

TpθCA4 reside in the cytosol. Perhaps the most interesting finding was TpθCA2 localises to the thylakoid lumen in a way similar to PtθCA1 (Fig 2). This has contributed greatly to our understanding of the *T. pseudonana* CCM however unpublished work from the Mackinder lab suggests this is not the complete picture. In the Rubisco pulldown by Onyou Nam identified a fifth putative theta CA, Tp1766. Containing the chloroplast transit peptide this novel protein may play involvement in the *T. pseudonana* CCM and is a novel target for research.

Table 1 Protein IDs and localisation of T. pseudonana theta CAs

Name	Accession	Protein ID	Localisation	Other names
TpθCA1	XP_002297285	672	Stroma	B8LE19
TpθCA2	XP_002286051	1093	Thylakoid lumen	B8BPY6
TpθCA3	XP_002297284	1765	Stroma	B8LE17
TpθCA4	XP_002290372	5647	Cytosol	
TpθCA5	XP_002297283	1766	Unknown	B8LE18

This study aimed to investigate Tp1766 localisation using molecular cloning techniques to fluorescently tag TP 1766. However, in the early stages lack of laboratory accessibility prevented this from being possible. In the waiting period the project shifted to an accessible phylogeny-based approach. Over time it became clear that molecular biology would not be accessible, however phylogenetic analysis and other computational approaches proved of value and interest such that they became the focal point of this project.

3.3 Methods

The protocol used for phylogenetic analysis of CAs can be broken down into five stages, 1: Sequence Identification; 2: Alignment; 3: Creation of a preliminary phylogenetic tree; 4: Multiple refinement phases; and 5: Formation of a robust phylogenetic tree (Fig. 6).

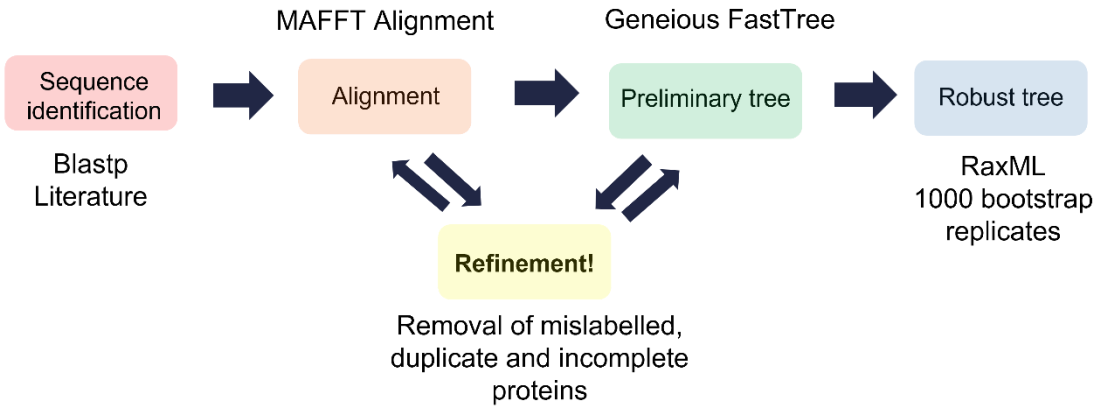


Figure 6. Flowchart illustrating stages in phylogenetic analysis

Sequence identification. The full length FASTA sequences of TpθCA1-5 (Tab.1) were individually subjected to BLASTp search of NCBI database, selecting defaults. All hits were collated in Supplementary Table (St). 1. (File 7 *211125_Table_CA*). Zeta CAs were identified by BLAST search (defaults) of the *T. weissflogii* zeta CA (Accession Q50LE4) using the open access data base UniProtKB(80). Additional proteins from iota and eta CA subclasses were identified in the study by Jensen et al and Del Prete et al (73,81). Finally, the known *C. reinhardtii* CCM CAs, CAH1-6 and CAG1 were added to St. 1 totalling 53 CAs.

Alignment. Full amino acid sequences from CAs File 7 were imported into Geneious2.0 (File 8 *220111_Table_CA_geneious*). Sf aligned by the MAFFT Alignment protocol (See section 2.3) (File 9 *220111 _CA_Alignment_*)

Preliminary tree. A preliminary tree was created selecting File 9 following the FastTree protocol (section 2.3). (Sf. 4, File 10 *220111_CA_Alignment_1 FastTree Tree*). This showed mostly correct grouping of CA families.

FastTree protocol: Select MSA file > Tree > FastTree (with defaults)

Refinement. Analysis of preliminary trees and sequence alignments led to rounds of refinement. The duplicate protein THAPSDRAFT_bd1766 was removed (Files 11-13 *220118_CA_Alignment_2, 220118_CA_Alignment_2 FastTree Tree, 220118_CA_Alignment_3*), LCIB and LCIC were then added *220314_CA_Alignment_4_MAFFT, 220314_CA_Alignment_4_MAFFT FastTree Tree,* followed by removal of Tp_XP_002290372_theta/calmodulin as uncertainty it was a CA (Note this sequence was later readded as TP_5674.) PtBAV00143 was also mistakenly removed, this is PtθCA4. The resulting dataset contained 53 sequences was realigned (File 14 *220314_Protein_Alignment_5_MAFFT)* and FastTree created (Sf. 5, File 15 *220314_Protein_Alignment_5_MAFFT FastTree Tree*. CA protein families mostly clustered together slight crossover between the alpha and eta families.

Robust tree. An initial robust phylogenetic tree with 100 bootstrap replicates was created using following RaXML and Consensus tree protocols (File 16 *220314_Protein_Alignment_5_MAFFT RAxML Tree RAxML Bootstrapping Trees consensus.* This tree was successful in distinguishing between CA protein families and identified FsCA2_alpha as a mislabelled alpha CA; phylogenic analysis suggests gamma.

RaXML protocol: Select MAFFT file > Tree> RAxML > select following the settings: Protein Model: GAMMABLOSSOM62, Algorithm: Rapid Bootstrapping, Number of Starting Trees or Bootstrap Replicates = (100), Parsimony Random Seed =1.

Consensus Tree protocol: Select RAxML file >Tree > Consensus Tree Builder > select defaults and changing only Support Threshold = 0.

Final tree. In a final refinement of the MSA (File 14) the mislabelled FsCA2_alpha was removed, TpθCA4/ TpCA_5674 added and TpCA_theta relabelled as TpCA_Tp1093. (St. 2, File 17 *220809_CA_table_53_*) This dataset (File 17) was aligned (File 18 *220809_MAFFTALIGN_53*) and FastTree created (File 19 *220809_MAFFTALIGN_53 FastTree Tree)*. A final robust tree was created by selecting the MSA (File 18) using the RaXML (selecting 1000 bootstrap replicates) and Consensus Tree protocols (Sf. 6, Files 20-21 *220809_MAFFTALIGN_53 RAxML Bootstrapping Trees, 220809_MAFFTALIGN_53 RAxML Bootstrapping Trees consensus*).

3.4 Results and Discussion

Theta CAs are widespread in aquatic photosynthetic organisms. Phylogenetic analysis of eukaryotic CAs showed proteins cluster in eight distinct clades reflecting the CA families (Fig. 7). The tree is unrooted due to the convergent evolutionary nature of CAs. Consequently, it is not appropriate to draw conclusions relating to the evolutionary relationships between CA families, however phylogenetic analysis can be used to categorise individual CAs and discuss relationships within clades. Recently, Jensen et al identified the novel *T. pseudonana* protein LCIP63 as an iota CA through phylogenetic analysis (73,81). In this study theta CAs were identified in *C. reinhardtii* and diverse diatom species (Fig.7). This is consistent with an extensive worldwide metatranscriptomic study by Karlusic et al. which identified chloroplast-targeted theta CAs in stramenopiles, cryptomonads and haptophyte genomes (82). The largest proportion of theta CAs were found in diatom species suggesting a predominance of this specific innovation (chloroplast theta CAs) within diatom chloroplasts (83). Indeed, chloroplast-targeted theta CAs were identified in the early diverging radial centric genius *Leptocylindrus*, implying theta CA presence in the diatom common ancestor (83). Nonoyma et al suggested theta CAs were acquired through horizontal gene transfer from haptophytes into a common ancestor, rather than via plastid secondary endosymbiosis (Fig. 1B) (83). Further meta studies analysing theta CAs in red and green algal lineages could give further insight into the evolution of theta CAs. Indeed the categorisation of *C. reinhardtii* LCIB and LCIC as theta CAs (Fig 7) shows this CA family is widespread throughout eukaryotes. The

Phylogenetic analysis suggests Tp1766 is a theta CA. All five *T. pseudonana* CAs including Tp1766 fall within the theta CA clade with 99.5 bootstrap confidence. Within the theta clade they are grouped into two distinct clusters. TpθCA2 and TpθCA4 form a sub-clade (bootstrap value with 89.9) separate from TpθCA1, TpθCA3, and Tp1766 which cluster with FcθCA1(bootstrap value with 99.9). The close evolutionary relationships of TpθCA1, TpθCA3, and Tp1766 is unsurprising as these three proteins are next to each other on the genome. The Tp1766 gene immediately follows TpθCA3 suggesting a gene duplication and adding confidence to the categorisation of Tp1766 as TpθCA5.

Both TpθCA1 and TpθCA3 localise to the stroma (Fig. 3), and as a duplication of TpθCA3, it could be inferred that TpθCA5 will have similar subcellular localisation stroma (34). However, the different localisations of closely

Figure 7 Theta CAs appear widespread in aquatic photosynthetic organisms Unrooted phylogenetic tree of Tp1766 and CAs from photosynthetic eukaryotes constructed using Maximum Likelihood (ML). Node labels represent species and CA family. Node values represent 1000 bootstrap replicates. The 8 CA families (α, β, γ, δ, ζ η, θ and ι) are represented. Back arrow indicates Tp1766 within the theta clade. Bspp, *Blastocystis sp.*; Cr, *Chlamydomonas reinhartii*; Eh, *Emiliania huxleyi*; Fc, *Fragilariopsis cylindrus*; Fs, *Fistulifera solaris*; Mc, *Micromonas commoda*; Mp, *Micromons pusilla*; Pt, *Phaeodactylum tricornutum*; Pf, *Plasmodium falciparum*; Pr, *Plasmodium reichenowi*; Pvp, Plasmodium vinckei petteri; Sj, *Saccharina japonica*; To, *Thalassiosira oceanica*; Tp, *Thalassiosira pseudonana*; Tw *Thalassiosira weissflogii.*

related TpθCA2 and TpθCA4 suggests evolutionary relationships between theta CAs do not necessarily reflect localisation. Instead, but presence of sequence features such as chloroplast transit peptide and ER transit peptide are more accurate predictors of subcellular localisation (34). The presence of signal sequences and fluorescent localisation of in TpθCA5 should be investigated to further current understanding of the role theta CA undertake in *T. pseudonana*.

LCIB and LCIC are theta CAs. The *C. reinhardtii* CAs LCIB and LCIC localised to the stroma and trap incoming CO_2 as HCO_3^- under conditions LC. Under VLC conditions these proteins form a complex (LCIB/C), localised to the vicinity of the pyrenoid and prevent CO_2 dissipating from the matrix(64,76). Despite the importance of the LCIB/C complex, characterisation of these proteins into a CA family has been less than straightforward. Jin et al. compared the crystal structures of LCIB the *P. tricornutum* beta CA1 (PtβCA1) identifying both a general structural resemblance and conservation of residues within the beta catalytic core(77). Conflicting characteristics of LCIB were observed, (oligomerisation, disorder of important conserved residues, and apparent mutations within key CA activity residues) however thought to be insignificant enough that LCIB was categorised as a beta CA. A subsequent study by Kikutani et al. determined LCIB and LCIC contain a CGHR motif within their genetic sequences, suggesting LCIB and LCIC are theta CAs(44). This finding is supported by phylogenetic analysis (Fig. 7) which places LCIB and LCIC in the theta clade. The presence of these green algal CAs within the theta clade suggests this CA family is not unique within stramenopiles, cryptophytes and haptopytes. As the pyrenoid is thought to have evolved multiple times it is possible that a pyrenoid localised theta CA may have convergently evolved to be a widespread feature of algal CCMs.

Iota CAs are not unique to diatom species. Analysis of the algal species present within clades suggests evolutionary restricted distributions of certain CA families. Within the iota clade there are four proteins from the diatom species *T. pseudonana, P. tricornutum, T. oceanica* and *F. solaris* (Fig. 7). The iota family are a group of recently identified Low CO_2 Inducible Proteins (LCIPs). The *T. pseudonana* homologue LCIP63 or TpιCA is rapidly and substantially upregulated in cells grown under LC conditions (82). Immunogold labelling localises TpιCA to the chloroplast, a feature consistent with the presence of a chloroplast transit peptide (73). Collectively these characteristics suggest involvement of TpιCA with in the *T. pseudonana* CCM. As all sequenced diatoms, including centric and pennate diatoms, possess a TpιCA homologue it may seem that diatoms have evolved iota CAs as an additional string in their CCM bows (73). Indeed, phylogenetic analysis (Fig. 7) seems to suggest iota CAs are unique to diatoms. This conclusion may be misplaced as searches made against the NCBI database for functional CA homologues may have unintentionally missed members of the iota family not yet recognised as CAs (73). When Karlusich et al searched the TARA oceans database for iota homologues, without restricting results to functional cas, a wide geographic and taxonomic distribution of iota homologues was revealed(83). Species present included in archaea, bacteria and the eukaryotic green algal species *C. reinhardtii*, *Chlorella variabilis* and *Tetrabaena socialis*. These findings suggest that the iota subclass is in fact not unique to diatoms, and iota CAs may play a role in green algal CCMs. The precise role of iota CAs in algal CCMs is yet to be established and presents an exciting avenue of further research.

Chapter conclusion. Phylogenetic analysis and comparison of photosynthetic microalgal CAs has identified the novel protein TP 1766 as a theta CA and revealed the widespread distribution of several CA subclasses including theta and iota. The acquisition and evolution of CAs within algal appears complex, reflecting the convergent evolution of both CAs and pyrenoids.

Chapter 4: Characterising putative pyrenoid Shell proteins

4.1 Chapter Summary

Recent unpublished *T. pseudonana* Rubisco pulldown assays from the lab have identified seven putative Shell proteins (Shell 1-7) which localise to specific regions around the pyrenoid periphery. These novel proteins are hypothesised to act as a CO_2 diffusion barrier and/or maintain the structural integrity of the pyrenoid. Little is currently known about similarities and differences between Shell 1-7, particularly their protein structures, evolutionary relationships and algal homologues. This study characterises Shell 1-7 using a threefold approach. Firstly, Shell protein homologues were identified and a consensus sequence determined by MSA. Secondly, phylogenetic analysis revealed evolutionary relationships between Shell 1-7 and suggested mechanisms of Shell protein evolution. Finally, modelling of the predicted Shell protein structures revealed distinctions in beta fold positioning. These three approaches were then combined with localisation data to hypothesise about the distinct structural roles of Shell proteins in *T. pseudonana.*

4.2 Introduction

Diffusion barriers surrounding the algal paranoid prevent CO_2 leakage from the matrix. An extensive review of pyrenoids, by Barrett et al, revealed pyrenoid outer-layer morphology varies greatly between algal species(11). Analysis of TEM images categorised pyrenoid starch sheath morphologies into five groups (with example species): a single cytosolic plate (for example, dinoflagellates *A. carterae* and *S. tridacnidorum*), two starch plates (for example, dinoflagellate *C. aff. operosum* and chlorophyte *C. glomerata*), a loosely associated globular sheath (rhodophyte *R. Violacea*), multiple starch plates (chlorophytes *C. geminata* and *C. reinhardtii*) and no sheath (diatoms *P. tricornutum* and *A baicalensis*). The starch sheath has been proposed to function as a structural barrier preventing CO2 leaking from the paranoid matrix. Evidence for this is most strongly supported by research in *C. reinhardtii*. Mutatnt models have shown the starch sheath required for LCIB operation, and starch sheath formation correlates with the induction of the CCM under LC conditions(61,84). A study by Itakura et al showed increasing starch sheath surface area was favourable for pyrenoid number and CCM function(85).

As algae derived by complex secondary endosymbiosis are unable to synthesise starch within the chloroplast it has been previously assumed that diatoms do not possess a pyrenoid outer-layer. Recent unpublished studies now present an alternative solution. Instead of the starch outer sheath *P. tricornutum* and *T. pseudonana* are posited to possess a protein layer surrounding the matrix. Due to their localisation, these proteins have been termed ‘Shell proteins’ are are yet to be fully characterised.

Localisation of pyrenoid Shell proteins in *T. pseudonana*. Unpublished microscopy images from fluorescent localisation experiments by Onyou Nam, show two distinct Shell protein sub-pyrenoid localisation patterns. Shell1,2, and 3 form a girdle-like band surrounding the longest regions of the lenticular pyrenoid, whereas Shell4 and 5 appear to cover a larger curved region. These findings have been combined to propose a model of Shell protein placement within the *T. pseudonana* pyrenoid outer layer (Fig. 7).*In vivo* cryET images presented at CCM10, August 2022 by Matsuda and Engel, suggested *P. tricornutum* Shell protein homologues form a macromolecular structure a single protein in diameter. They predicted Shell proteins dock at 90° angles, although this has not been confirmed *in vivo*. The current *T. pseudonana* model proposes a similar one protein thick pyrenoid outer layer (Fig. 7). Localisation experiments for Shell 6 and 7 are currently being undertaken, with

the aim of incorporating these proteins into the *T. pseudonana* pyrenoid model. The structural and functional implications of distinct sub-pyrenoid Shell protein localisations are currently unknown. By modelling Shell protein 3D structures, this study aims to suggest a structural basis underlying localisation.

AlphaFold computationally predicts Shell protein structures. Knowledge of protein structure provides insight into mechanisms of protein formation, molecular interactions and guides biochemical experimentation(86). At present all experimentally resolved macromolecular structures are available in the Protein Data Bank (PDB), an open access database containing thousands of protein structures (87). Typically, these structures have been solved using experimental techniques such as cryoEM and X-ray crystallography, which are both technically challenging and time-consuming(69). Recent computational developments in protein structure prediction have revolutionised structural biology determination of novel proteins. AlphaFold is a machine learning algorithm which predicts protein structure from an amino acid query sequence(69). Physical and geometric constraints of protein structures are learnt from data within the PDB and combined with understanding of evolutionary conservation derived from MSAs (88). Using this multilevel approach, AlphaFold can predict 3D protein structure with near experimental accuracy (69). Despite the obvious benefits, the algorithm is not perfect, and limitations include an inability to accurately fold IDPs, protein multimers, and difficulty predicting docking interactions. Ideal candidates for analysis by AlphaFold include novel single-subunit proteins of low disorder. Excitingly, the *T. pseudonana* Shell proteins fall within this category, enabling analysis by AlphaFold.

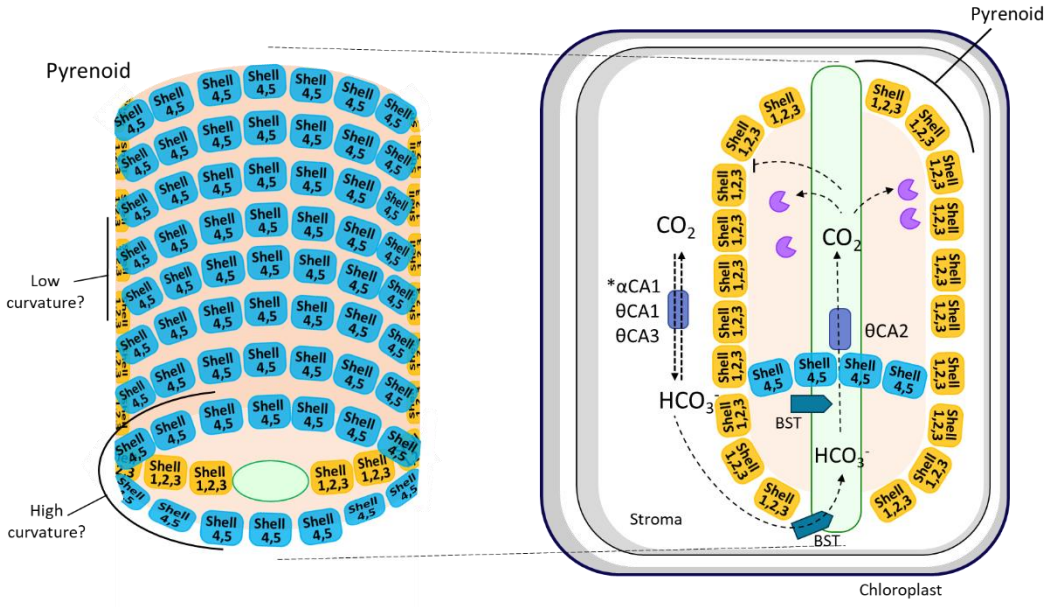


Figure 8 The proposed model of Shell protein localisation around the T. pseudonana pyrenoid. Unpublished fluorescent localisation images suggest Shell1,2 and 3 localise to a different region of the pyrenoid periphery to Shell4 and 5. Putative regions of high curvature (Shell4 and 5) and low curvature (Shell 1,2 and 3) are indicated.

Molecular phylogeny provides insight into the evolutionary relationships between proteins. Phylogenetic analysis can use amino acid MSAs to illustrate how proteins have evolved. They can identify protein origin, the presence of a common ancestor, give insight into evolutionary mechanisms and act as a 'molecular clock' measuring the rate protein evolution (89,90). Molecular phylogenetics using protein MSAs has several benefits over their genetic counterparts. Variations in protein, rather than genetic, sequence directly impact structure and function and it is proteins which interact with the environment and are the object on which natural selection acts. (89,91). The availability of open access proteomic databases such as NCBI provide rich source of molecular data. Proteins from distantly related species can be easily identified and used to construct phylogenetic trees (65,89). As algae are polyphyletic this ability to mine databases for diverse algae Shell protein homologues is an exciting research prospect. The dataset generated can be analysed phylogenetically to characterise the evolutionary relationships between the *T. pseudonana* Shell proteins and their homologues.

4.3 Methods

Shell1-7 proteins in *T. pseudonana* were identified by Rubisco pulldown experiments by Onyou Nam. Protein sequence, Accession code and Genbank ID are detailed in supplementary table 3. All MSAs in this section (4.3) were undertaken using the MAFFT Alignment Protocol, detailed in section 2.3.All phylogenetic analysis used the FastTree, RaXLML, and Consensus protocols detailed in section 3.3. Any deviation in method is detailed.

Identification of Shell protein homologues preliminary MSAs and phylogenetic analysis. The Shell1-6 full protein sequences were individually submitted for BLASTp search (with defaults) selecting an E value = 1. Hits were sorted by highest to lowest alignment length, and a cut-off length of 100 amino acids was employed. (File 22 *220526_Collated shell csv files_Shell1-6).* Proteins were categorised into diatoms, haptophytes, coccolithophores, pelagophytes, gram-positive bacteria, proteobacteria, Gram-negative bacteria, and Archaea.

Protein hits were imported into GeneiousR11 and a preliminary MSA undertaken of the top 100 sequences (File 23 *220530_MAFFTalignment_100.1).* (Sf. 7)

As this alignment revealed highly conserved consensus regions a full MSA was undertaken with the 144 Shell 1-BLASTprotein hits(File 24 *220606_MAFFTAlignment_ 144* sequences). From this MSA an initial FastTree was created (File 25 *220606_MAFFTAlignment_ 144 sequences FastTree Tree*). Phylogeny revealed a lack of clear clade groupings between Shell proteins. At this point in the project it became apparent that analysis of Shell7 should be included within this study.

Initial Shell7 BLASTp analysis was undertaken with the full amino acid sequence selecting defaults and E = 1. (File 26 *220801_Collated shell csv files_Shell1-7*). None of the hits overlapped with hits from the Shell1-6 BLAST search (St. 4) and MSA revealed protein hits from Shell1-7 lacked the Shell1-6 consensus region (File 27 *220607_MAFFT_Alignment_154)* (Sf. 8). To ensure the Shell7 hits contained the characteristic Shell beta fold the structure of the top hit *Chrysochromulina tobinii* calmodulin KOO34179.1 was modelled. The full KOO34179.1 amino acid sequence was entered into AlphaFold Co-lab selecting defaults, expect cycles =5, and rank = 5. The top ranked model showed lack of characteristic Shell protein beta fold (Sf. 9, File 28 *220608_KOO34179_c5ceb_unrelaxed_rank_1_model_1*). This suggested the extended C-terminal region of Shell7 was skewing the search results. To identify true Shell7 homologues a new BLAST search (defaults, e = 1) was undertaken querying only the Shell7 beta sheet region.

QGGRIKSFTFGEEIESVEVLLVTKHRNLKAMLEILQGPNDDNEIIEVETEDGRVHPFYTVIQTPGGANTLRVVNRSPV EFPFEAFVRPFV.

The protein hits unique to Shell7 were selected (KAH8046553, KAH8064147, KAH8066543, KAH8085487, KAH8094245, KOO28333, KOO34466, KOO34832, VEU33671) imported into GeneiousR11 and aligned with combined with Shell1-6 homologues(Files 29-30 *220801_Collated shell csv files_Shell1-7*; *220629_MAFFTShell protein144).* The consensus region now appears present in all species.

Consensus sequence identification and visualisation To identify Shell protein consensus sequence, residues >95%, >90%, >85%, and >75% conserved were identified in GeneiouR11 by selecting Display > Consesnsus > more options > threshold > 95% in the File 26 MSA. Residues were manually highlighted onto the consensus sequence (Sf 10). The selection process was repeated with threshold values of 90%, 85%, and 75%. The File 26 MSA was exported into Jalview2(92) and visualized as a

histogram: Consensus > show histogram. The histogram was copied and pasted into PowerPoint, residue single letter codes added beneath each corresponding bar and the consensus residues coloured. This identified small regions of low percentage conservation within the consensus region, suggesting insertions within certain Shell protein homologues (Sf. 11).

MSA refinement. Several refinement phases removed proteins with duplicates or insertions. Returning to the MSA (File 26) in GeneiousR11, four proteins with insertions were identified and removed from the dataset (Aa_KAH8043819; Aa_KAH8097243; Aa_KAH8046392; Aa_KAH8070514). The dataset was realigned (File 31 *220704_MAFFTAlignShellprot138*). The duplicate proteins Aa_KAH8072502, Aa_KAH8075838 were identified, removed and dataset realigned (File 32 *220704_MAFFTAlignShellprot136*). One further protein Aa_KAH8059281 removed from the dataset before alignment(Sf. 12, File 33 *220704_MAFFTALIGN_Shell135*. The final dataset contained 135 proteins

The final MSA (File 33) was exported into Jalview2(92) and trimmed to the consensus region (File 34 *220407_MAFFTALIGN_Shell135 Copy_ TRIMMED*). A screenshot of the consensus sequence histogram was copied into Microsoft PowerPoint. The consensus sequence was typed out (single letter codes), aligned to the corresponding histogram bar, and manually coloured to reflect percentage conservation identified in GeneiousR11 (File 33). Hydrophobic residues were manually identified underlined in the sequence.

Visualisation of consensus residue positioning using *T. pseudonana* Shell proteins structural models. To visualise consensus residue structural positioning the AlphaFold Co-lab structures of Shell1-7 were imported into PyMol2.0. The following structures (Files 35-41) were created using AlphaFold Co-lab by Onyou Nam by entering the full protein sequence, selecting defaults.

Tp7881_unrelaxed_model_1_Shell1,
Tp23918_unrelaxed_model_1_Shell2,
Tp7883_unrelaxed_model_1_Shell3,
THAPSDRAFT_24512_56be9_unrelaxed_model_1_Shell4
THAPSDRAFT_3883_df672_unrelaxed_model_1_Shell
Tp8449_f7947_unrelaxed_model_1_Shell6
220613_Shell7_ef9c4_unrelaxed_rank_1_model_3_Shell7

Shell protein beta folds were manually selected and renamed ShellX_betafold where X = Shell number. Beta folds were aligned by selecting Align > ShellX_betafold > Shell1_betafold (Sf. 13). Consensus residues at each percentage threshold were manually selected for to create four selections (95, 90, 85, 75). These selections were visualised using Show > Liquorice > sticks. The selections were differentially coloured by sequentially selecting colour > red (95); colour > orange (90); colour> limegreen (85); colour > marine (75) (Sf. 14). Amino acids were manually removed either side of the consensus region to enable clear beta fold (Sf. 15, File 42 *220726_Shell1-7consensus*).

To determine the presence of consensus residues at the Shell protein surface the alignment in File 42 was visualised as a surface model by selecting each Shell protein then Show > Surface (File 43 *220726_Shellsurfaceconsensus).* Surface charge of each Shell proteins was visualised by selecting APBS electrostatics > main > polymer > Shell protein (with defaults) > run. The model 'prepared01' was selected (File 44 *220726_Shellsurfacechargeconsensus*) The red to blue sliding scale represents positive to negative charge.

Docking simulations of Shell protein dimers using AlphaFold Co-lab.
An initial homodimer interaction between two Shell1 protein beta folds was modelled in AlphaFold Co-lab (Sf. 16) selecting defaults, cycles = 3, rank = 3, 3D structure (File 45 *220725_Shell1homo1.1_2*). Analysis of the highest ranking model showed low confidence the relative positions of the two proteins as the predicted pLDDT has low rank at the strands interfacing the proposed interaction (Sf. 17). The predicted aligned error (PAE plots) have low confidence in relative domain position. To check if this was a consequence of querying a trimmed Shell1 structure, a query with the full Shell1 homodimer was queried in AlphaFold Co-lab (Sf.18), selecting defaults (File 46 *220726_Shell1fullhomo_1).*The full sequence Shell1 homodimer query also showed no obvious protein interaction (Sf. 19) with low pLDDT at the point of interface, and PAE plots giving no confidence in relative domain position.

Phylogenetic analysis
Phylogenetic analysis of Shell protein homologues was undertaken using the stages illustrated in Figure 6. MAFFT alignment, FastTree, RAxML tree and Consensus tree GeneiousR11 protocols were consistent with those outlined in section 3.3. Any deviation from these protocols as detailed below. All alignments undertaken used the MAFFT protocol.

Preliminary phylogenetic analysis. The full amino acid sequence of Shell3 was subjected to BLAST search (defaults)(93). The top 100 hits were collated, full amino acid sequences inputted into GeneiousR11, and aligned (File 47 *220530_MAFFTalignment_100.1*, Sf. 20). This MSA was selected and FastTree created (File 48 *220530_MAFFTalignment_100.1 FastTree Tree)*. Although this tree was unrooted, *T. pseudonana* Shell proteins occupy distinct clades (Sf. 21). This gave confidence to pursue this methodology of research.

Full Shell protein homolog MSA and phylogenetic analysis. The MSA of 144 Shell homologs (File 26) was used to create a FastTree (File 49 *220606_MAFFTAlignment_ 144 sequences FastTree Tree*, Sf. 22). This MSA (File 49) was duplicated creating File 50 *220606_MAFFTAlignment_ 144 sequences Copy extraction 2* and File 51 *220629_MAFFT142Shell* (note these alignments/trees contain 142 sequences despite the filenames). A FastTree was created in which phylogeny is unclear and unrooted (File 52 *220606_MAFFTAlignment_ 144 sequences Copy extraction 2 FastTree* Tree, Sf. 23)

Initial FastTree rooting. The *P. tricornutum* CA PTCABETA1 (Accession AAL07493) was added to the File 51 dataset and realigned (File 53 *220622_MAFFT_164+CA)*. This file was wrongly labelled, there are 143 sequences including the CA. *T. pseudonana* Shell proteins were relabelled to ShellX where X is the Shell protein number. The addition of PTCABETA1 successfully rooted tree (File 54 *220622_MAFFT_164+CA FastTree Tree*, Sf 24).

MSA edited to contain the consensus region only. A copy of the File 51 dataset was made and trimmed to the previously identified consensus region (File 55 *220801_Shell142_beta)*. This was realigned (File 56 *220801_Shell142_beta – realigned*, Sf. 25) and preliminary FastTree created (File 57 *220801_Shell142_beta - realigned FastTree Tree*, Sf. 26).

Rooting the FastTree with CCM proteins. Sequences of the *C. reinhardtii* CCM proteins Bestrophin1-3 (BST-1: A0A2K3CTN0, BST-2: A0A2K3CTQ2, BST-3: A0A2K3CTP3) were added to the dataset (File 58 *220801_Shell142_beta - realigned (modified)* dataset). This dataset was realigned (File 59 *220802_MAFFTShellwithBST1-3)* and FastTree created (File 60 *220802_MAFFTShellwithBST1-3 FastTree Tree*, Sf. 27). Addition of BST1-3 successfully rooted the tree.

Further dataset refinement. Analysis of the MSA consensus region (File 59) s and FastTree (File 60) led to the removal of the following proteins from the dataset: Aa_KAH8066543 containing an insertion in consensus region, Ctob_KOO28333 a protein fragment, and 4 proteins from non-eukaryotic species (Ob_HBO57542.1, Ea_ MBA45229.1, Ab_NCY17740.1, and Pb_ NDD31680.1). The resulting dataset containing 137 proteins was realigned (File 61 *221007_MAFFT_137)* and FastTree created (File 62 *221007_MAFFT_137 FastTree Tree*, Sf. 28). Branches were coloured manually in GeneiousR11 to reflect the algal lineage.

Determining a robust phylogenetic tree. A robust phylogenetic tree with bootstrapping was created by selecting *the MSA* (File 61) and following the RaXML protocol in section 3.3 (Bootstrap Replicates = 1000.) A consensus tree was created following the protocol in section 3.3 (File 62 *221007_MAFFT_137 RAxML Bootstrapping Trees consensus Copy*, Sf. 29). Branches were coloured manually to distinguish proteins from haptophyte, pelagophyte, and centric and pennate diatom groups. The *T. pseudonana* Shell proteins Shell1-7 were manually highlighted in bold. Multiple clade A homologues from *A. anophagefferens* and *C. tobinii* were hidden manually for clearer visualisation.

Analysis of Shell1-7 predicted AlphaFold structures. The AlphaFold structures of Shell1-7 (Files 35-41) were imported into PyMOL2.0, refined, and aligned to compare beta fold positioning.

Tp7881_unrelaxed_model_1_Shell1, renamed Shell1_Tp7881_
Tp23918_unrelaxed_model_1_Shell2, renamed Shell2_Tp23918
Tp7883_unrelaxed_model_1_Shell3, , renamed Shell3_Tp7883_
THAPSDRAFT_24512_56be9_unrelaxed_model_1_Shell4, renamed Shell4_Tp24512
THAPSDRAFT_3883_df672_unrelaxed_model_1_Shell, renamed Shell5_Tp3883
Tp8449_f7947_unrelaxed_model_1_Shell6, renamed Shell6_Tp8449
220613_Shell7_ef9c4_unrelaxed_rank_1_model_3_Shell7, renamed Shell7_Tp1762

To align Shell1-7 a 'selection' for each Shell protein beta fold consensus region was created (File 63 *221014_7shells.1*.) Residues aligning with consensus residues QGGS---MTAAVVP were highlighted in sequence pane. Selection were renamed to ShellX_betafold where X is the shell number (Sf. 30). ShellX_betafold alignments were made between several Shell protein combinations selecting > alignment > action > align > to selection > select alignment. Nine initial alignments were created, RMSD values given (St. 5). The similarity between alignments one, two, and three led to final decision to align Shell2-7 to 'Shell1_betafold'. Alignments to Shell 1 were chosen (1,2,4,7,8 and 9 in St. 5) and refined by manual deletion of residues outside the consensus region *(*File 63, Sf. 31).

Analysis of beta fold strand positioning between Shell1-5. In PyMOL2.0 Selections were made for strand 1A of Shells1-5 (St. 6) in the File 63 structural alignment. Strand residues were highlighted in the selection pane and renamed by Action > Rename selection > 1AShellX (Sf. 32). Shell2-5 strand 1A selection (1AShell2, 1AShell3, 1AShell4, and 1AShell5) were individually aligned to 1AShell1 using Alignment> action > align > to selection > 1AShell1. Alignments were refined by manual selection and removal of residues to visualise strand combinations(Files 64-67 *221018_7shells.7_1A2A; 221017_7shell.3_6A7A*; *221017_7shells.4_13B_14B*; *221018_7shells.6_8B9B*).

4.4 Results

The Shell1-7 protein structures share a characteristic beta fold. Initial AlphaFold Co-lab models of Shell1-7 were generated by Onyou Nam. Viewing the structures side-by-side reveals a characteristic beta fold comprising of eight beta strand pairs and two short additional outer strands (Fig. 9A). The beta fold consists of two halves (designated A and B); strands are curved and there is an overall twist to the fold (Fig. 9B). Shell 1-5 show additional secondary structure in the form of two short alpha helices directly above the beta fold. There are subtle distinctions between the seven Shell protein structures. Beta fold strands appear at slight different angles and orientations, and alpha helices differ in length and position. Each protein possesses a disordered N-terminal region containing transit peptide and chloroplast targeting domains. In addition, Shell7 contains a bulky alpha helical C-terminal domain. It still requires experimental validation if the C-terminal extension is part of the Shell7 gene model or codes for a separate protein.

Broad parameter BLAST search of Shell1-7 created a dataset of Shell protein homologues from diverse algal species. After refinement, the dataset comprised of 135 proteins from haptophytes, diatoms (centric and pennate) and pelagophytes. Performing a MSA analysis of the Shell homologues identified a two-part consensus sequence, each part around 100 amino acids in length (Fig. 10A). Residues >75%, >85%, >90%, and >95% conserved (56 total) were highlighted and termed consensus residues. Short stretches (~3-5 residues long) of consensus residues were observed in both consensus regions however, the overall distribution of consensus residues lack a clear pattern. Characteristically hydrophobic consensus residues were identified and accounted for a high proportion of the total consensus residues (38 out of the 56 residues).

To investigate the structural positioning of consensus residues, *T. pseudonana* Shell protein AlphaFold models were aligned and consensus residues visualised in correspondence to their percentage conservation (Fig. 10B). Consensus residue location was analysed in 3D space by rotating the Shell protein alignment and capturing images of different orientations. Several observations regarding consensus residue placement were made. Firstly, consensus residues, particularly with >85%, >90%, and >95% conservation, are external facing and located to the 'edges' of the beta fold. Secondly, a high proportion of these residues are hydrophobic, located in the loop regions between beta strands, with outward facing side chains (Fig. 11). Thirdly, the most highly conserved residues (>95%) appear to localise in the vicinity of the 'corner' regions of the beta fold (see Fig. 10B underside orientation).

To further investigate consensus residue structural positioning the Shell protein alignment in Figure 10B was shown as surface model (Fig 11). From the right hand side (RHS) orientation consensus residues mapped to similar surface positions in Shell1-7, with clusters of consensus residues facing the external environment. To see if these similarities extended to surface charge patterning, hydrophobicity of the corresponding protein models was analysed. This shows visible differences in surface hydrophobicity between the seven Shell proteins. Distinctively, Shell7 has a large negatively charged pocket within the centre region of the beta fold (RHS orientation), a feature absent in Shell 1-6. There are some similarities in charge patterning between Shell1-6, however these were not quantified.

Similarities between Shell1-7 consensus residue localisations suggested a Shell protein docking mechanism. Shell1 homodimer docking was simulated using AlphaFold Co-lab software. Unfortunately, when 'docked' homodimers did not form a confident replicable interaction (supplementary figure X). These findings leading to a halt and this avenue of research.

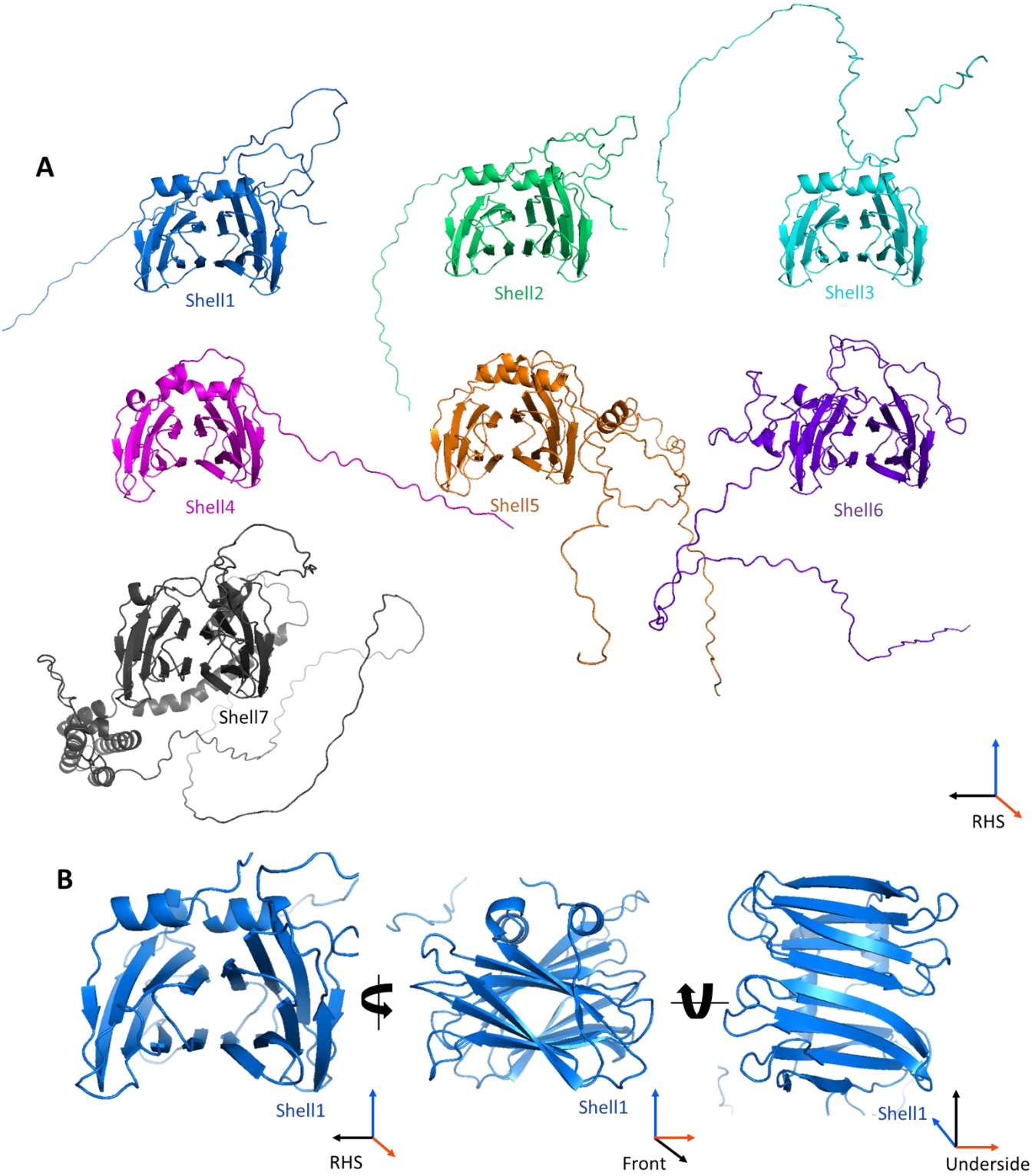


Figure 9) Predicted AlphaFold structures of Shell1 (blue), Shell2 (green), Shell3 (cyan), Shell 4 (magenta), Shell5 (orange), Shell6 (purple), Shell7 (black), viewed from a RHS orientation **B)** The Shell1 beta fold structure in three orientations

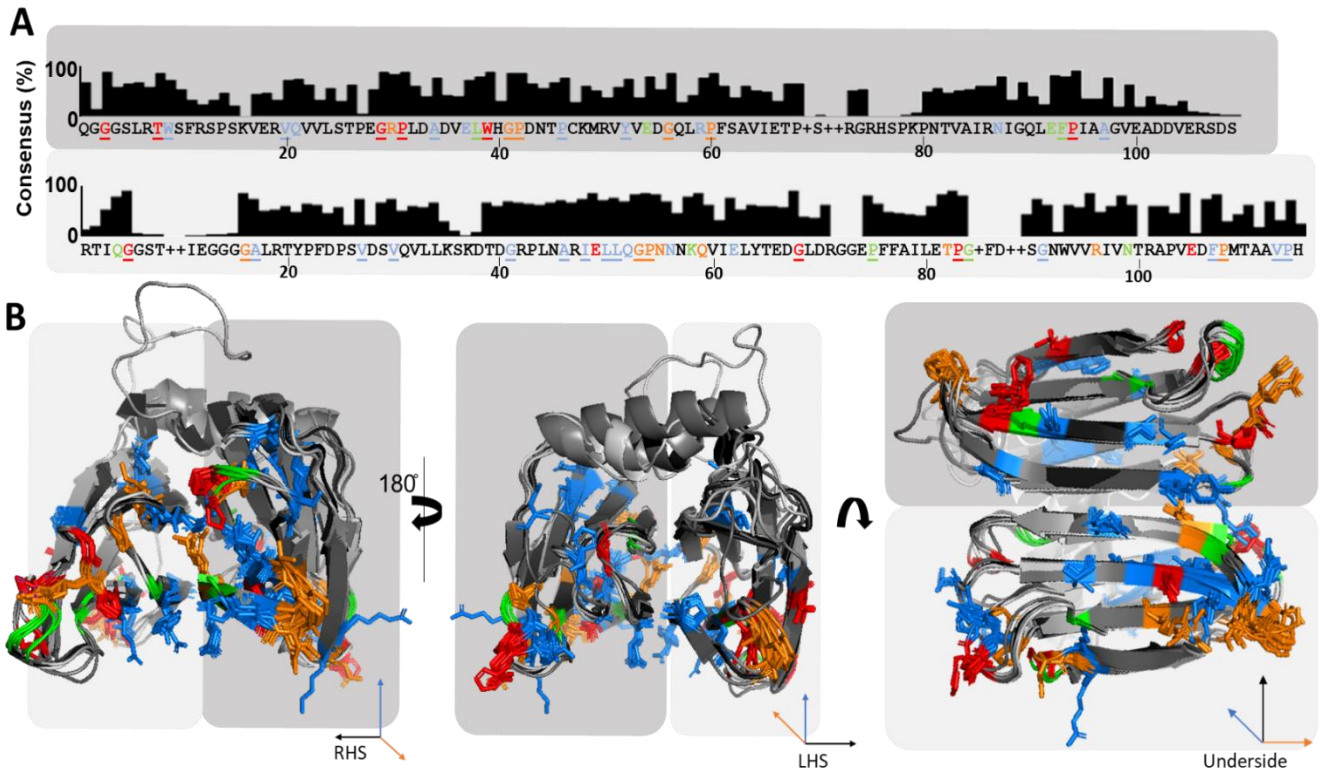


Figure 10 A) Two part Shell proteins consensus sequences constructed using Geneious and Jalview. Consensus residues >95% (red), >90% (orange),> 85% (green), and >75% (blue) are highlighted consensus sequence constructed using Geneious and Jalview. Hydrophobic residues are underlined. **B)** Aligned beta fold structure of TP Shell 1-7. Consensus residues highlighted corresponding to (A). Protein structures determined by AlphaFold, visualised in PyMOL2.0.

Phylogenetic analysis of Shell proteins showed homologues from 11 species are distributed throughout the six clades, termed A, B, C, D, E and X (Fig. 12). Species are categorised as haptophytes (*C. tobinii* and *E. huxleyi*), centric diatoms (*T. pseudonana, T. oceanica,* and *C. tenuissimus*), pennate diatoms (*N. inconspicua*, *F. cylindrus*, *P. tricornutum*, and *P. multistriata*) and pelagophytes (*A. anophagefferens*). Many of these species contain multiple Shell protein homologues, often closely related within a clade. For example, in clade A alone there are 27 *A. anophagefferens* and 12 *C. tobinii* homologs.

Shell1-7 are located within five distinct clades termed A-E. *T. pseudonana* Shell proteins fall into the following clades: A) Shell1, Shell2, Shell3; B) Shell6; C) Shell4; D) Shell7; E) Shell5. Shell 1-6 appear to have at least one closely related *T. oceanica* homologue, often a sister branch. Interestingly, a sixth clade (X) contains near exclusively pennate diatom homologues and lacks a *T. pseudonana* Shell protein. In clade X there are only two Shell protein homologues from the centric diatom *C. tobinii*.

To place phylogenetic analysis in a structural context the Shell 1-7 beta fold structures were aligned. Shell2-7 were individually aligned to Shell1 giving RMSD values indicating structural similarity (Fig. 13). These values show Shell1-3 (0.194Å, 0.208 Å) possess the highest structural similarity, followed by Shell6 (0.413Å), Shell4 (0.583Å), Shell7 (0.627Å) and Shell5 (0.893Å). Mapping each Shell protein RMSD value to the corresponding phylogenetic position (Fig. 12) showed differences in Shell protein beta fold structure are reflected by evolutionary distance. In clade A close evolutionary relationships between Shell1, 2, and 3 are consistent with the close structural similarity and low RMSD value observed.

To investigate nuances in beta fold positioning, alignments between Shell1-5 beta strands were made. The 14 beta strands were assigned the identities 1A-7A and 8B-14B (Fig. 14) with A and B representing the two halves of the beta fold consensus regions (Fig. 10A). Alignments of the selections 'strand1AShell2' and '1AShell3' to '1AShell1' gave RMSD values of 0.04Å and 0.09Å respectively, indicating near identical structural positioning. The structural similarity of strand positioning between Shell1,2, and 3 continues throughout the beta fold, with strands 6A, 7A, 13B, 14B, 8B and 9B appearing interchangeable (Fig. 14). Following alignment of 1AShell4 to 1AShell1, there is a slight shift in the central strands (6A, 7A, 14B, and 13B) position. The Shell4 outer strands 8B, 9B remain closely aligned to Shell1 8B and 9B. Conversely, there is

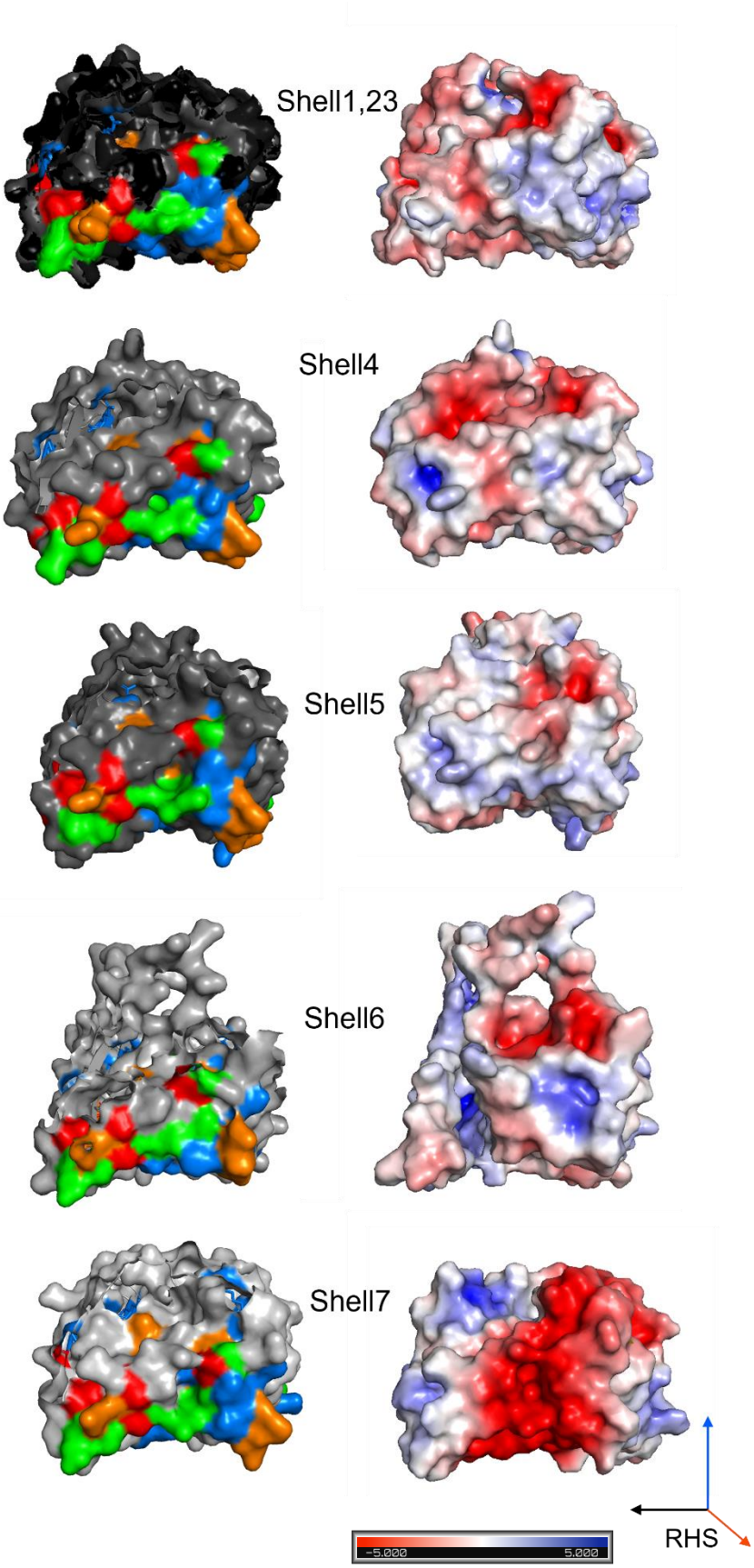


Figure 11 Surface representations Shell 1-7 beta fold structures visualising Consensus residues and surface charge. From the RHS orientation**.** Consensus residues >95% (red), >90% (orange), >85% (green), and >75% (blue) are highlighted consensus sequence. Surface charge represented as a sliding scale from red (negative) to blue (positive.) Structures created in PyMOL2.0. Note Shell6 structure is missing strand 1A and 2A.

distinct shift in beta strand positioning throughout the fold following alignment of ‘1AShell5’ to ‘1AShell1’. In Shell5 strands 6A, 7A, 13B and 14B curve away from their Shell1 beta strand counterparts, indicating increased curvature within the beta fold.

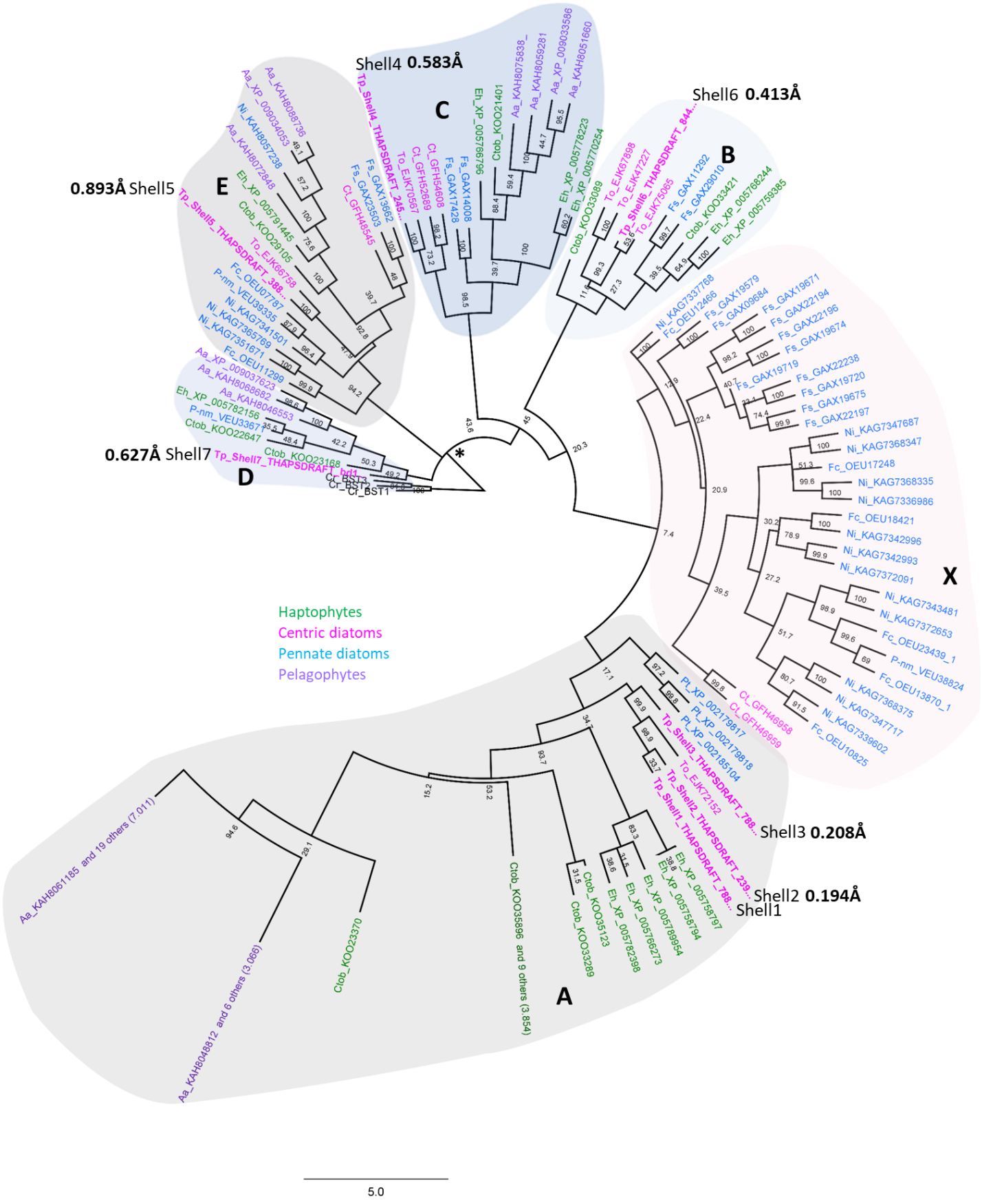

Shell4 0.583Å
Shell6 0.413Å
0.893Å Shell5
0.627Å Shell7
Shell3 0.208Å
Shell2 0.194Å
Shell1
A
B
C
D
E
X
Haptophytes
Centric diatoms
Pennate diatoms
Pelagophytes
5.0

Figure 12 Phylogenetic tree of Shell protein homologues proteins constructed using Maximum Likelihood (ML). Node values represent 1000 bootstrap replicates. Protein colour reflects algal grouping. *T. pseudonana* Shell proteins are highlighted in bold. RMSD values are given for alignment of Shell2-6 structures to Shell1. Shell1-7 are group into clades **A)** Shell1, Shell2, Shell3 **B)** Shell6 **C)** Shell4 **D)** Shell7 **E)** Shell5. Asterisk indicates common ancestor. *C. reinhartii* BST1-3 root the tree.

Aa*, A. anophagefferens*; Cr, *Chlamydomonas reinhartii*; Ct, *C. tenuissimus*; Ct, *C. tobinii* ; Eh, *Emiliania huxleyi*; Fc, *Fragilariopsis cylindrus*; Fs, *Fistulifera solaris*; Nt, *N. inconspicua*; Pt, *P.tricornutum*; P-mn, *P. multistriata*; To, *T. oceanica*; Tp, *T. pseudonana.*

4.4 Discussion

Shell protein homologues are found in geographically and morphologically diverse algae sharing similar pyrenoid structure and and evolutionary origin. Mining proteomic databases for Shell1-7 homologues identified 135 proteins from species grouped within the haptophyte and stramenopiles clades. These species are geographically widespread, found in coastal (*C. tenuissimus* and *T. pseudonana*), open ocean (*T. oceanica*), widespread marine (*E. huxleyi*, *C. tobinii*, *N. inconspicua*, *P.tricornutum*, and *A. anophagefferens*), freshwater (*N. inconspicua*, and *C. tobinii*) and polar oceanic (*F. cylindrus*) environments(94–102). They also exhibit diverse cellular morphologies varying in shape, size, pigments, and cellular outer layer. These findings suggest Shell proteins are not unique to algal species found within the same environment, or with similar cellular composition, to *T. pseudonana*.

Despite these differences the Shell protein containing algal species have two characteristics in common: pyrenoid morphology and the mechanism of evolutionary origin. The 11 algal species all possess pyrenoids, consistent with the putative role of Shell proteins as a pyrenoid diffusion barrier. Except for the golden alga *A. anophagefferens,* these pyrenoids are similar in morphology being elongated and containing a single penetrating thylakoid tubule, in contrast to the green alga *C. reinhardtii* which possesses multiple thylakoid tubules branching from a central thylakoid knot(11). Haptophytes, diatoms and pelagophytes fall into the haptista and stramenopile algal clades (Fig.1). As deatiled in section 1, these algae are derived by complex secondary endosymbiosis. The clear absence of Shell protein homologues in red and green algal species suggests an relationship between Shell proteins and evolutionary origin. This will be discussed further in conjunction with phylogenetic analysis (Fig.12).

An unexpected finding was the presence of multiple shell homologues in the pelagophyte species *A. anophagefferens*. SEM images of *A. anophagefferens* have shown the pyrenoid is rarely traversed by thylakoids(100). The sheer number (37) of Shell proteins identified in database mining. Raises the possibility *A. anophagefferens* is employing Shell protein homologues for completely different function to *T. pseudonana*. Fluorescent localisation of *A. anophagefferens* Shell homologues from different phylogenetic clades (Fig. 12) may reveal an alternative function. As *A. anophagefferens* is the only pelagophyte species to have its genome sequenced, it is possible more algal species of the same subclass possess multiple Shell protein homologues.

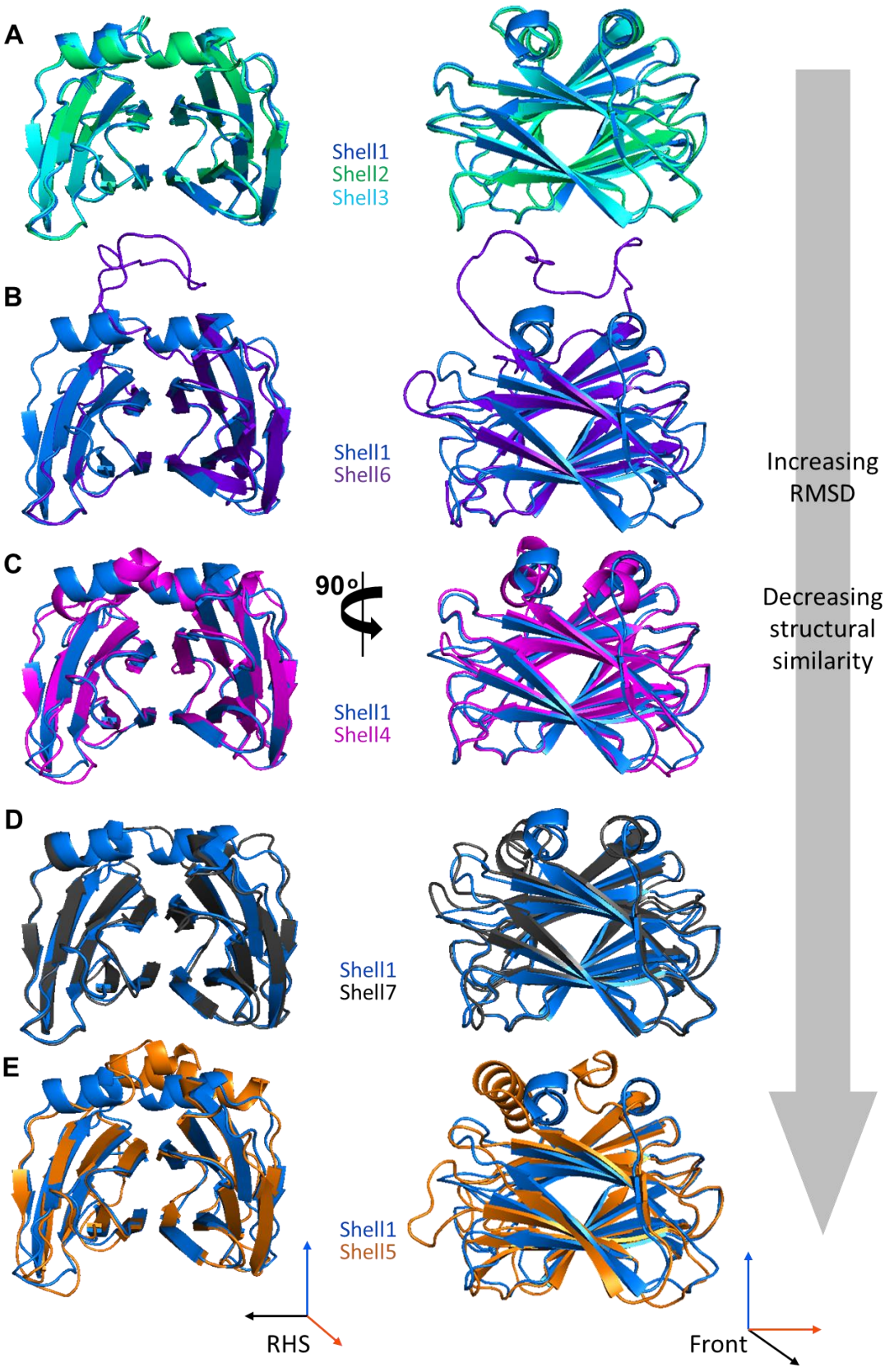


Figure 13 Alignment of predicted show beta fold after fold structures into orientations, aligned to shell one (marine blue). **A** shell to (limegreen) and Shell three (cyan). **B** Shell6 (purple) **C** Shell4 (magenta), **D** Shell7 (black) **E** Shell5(orange),. Arrow represents increasing difference in structure and increasing RMSD value

MSA of the Shell protein homologues revealed two highly conserved regions within primary protein sequence (Fig. 10). Analysis of the Shell protein consensus sequence revealed residues >75% conserved are distributed throughout the consensus sequence, often clustering as short stretches of consecutive residues ~2-8 amino acids in length. Visualisation of the consensus sequence as a histogram also shows many of the residues <75% are often >50% conserved (Fig. 10A). Further analysis of the consensus sequence identified highly conserved hydrophobic residues across the 137 Shell protein homologues. Such high conservation of consensus residues across diverse species suggests possible structural and/or functional involvement of these residues within algal Shell proteins.

Shell protein consensus sequence residue localised to distinct regions in Shell protein 3D structure. Mapping consensus regions onto the Shell1-7 structural alignment (Fig 10A) reveals the two consensus sequence sections (Fig. 10A) are located to the two halves of the Shell protein beta fold. Consensus residues tend to cluster at the edges, underside, and corners of the beta fold with the more highly conserved residues (>95%) appearing closer to the edges. Viewing the Shell protein alignment from the underside shows highly conserved residues (95%) in the corner regions with exposed side chains. This residue structural patterning may explain why the primary consensus sequence contains seemingly sporadic short clusters of consensus residues. In particular, the flexible loop regions appear to be populated with highly conserved hydrophobic G, P, T and W residues. This suggests these residues are critical within Shell protein structure and/or function. However, within beta sheets there are a limited number of residues which can occupy loop regions in a given secondary structure confirmation. Furthermore, the presence of hydrophobic residues at the corners of the beta fold may be due to the increased frequency of hydrophobic residues in beta rich

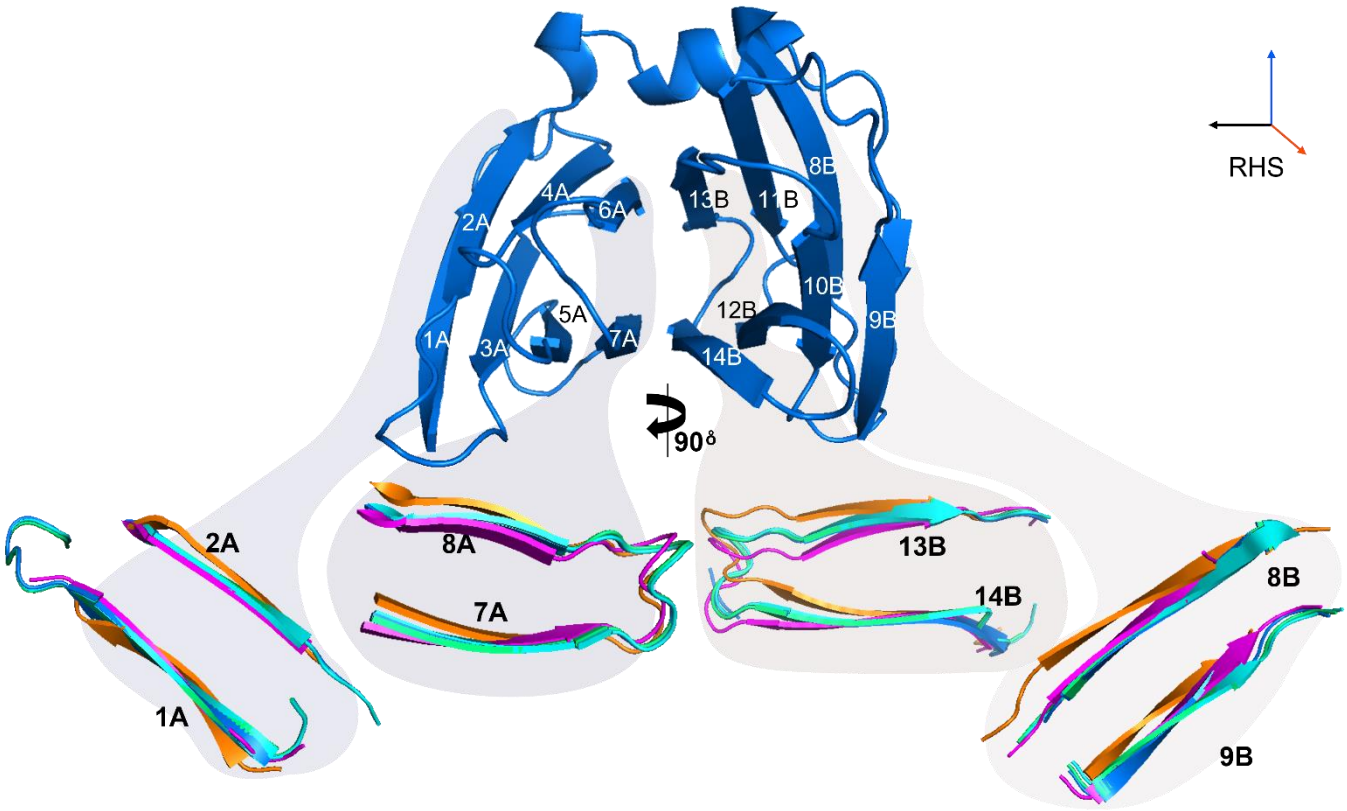


Figure 14 Shell1 beta fold with strands numbered. Cut through images depict Shell2-5 aligned to Shell1 strand 1A, rotated by 90°

structures(103). The current understanding of diatom Shell proteins in *P. tricornutum* (proposed by Ben Engel and Yusuke Matsuda) suggests that Shell proteins form a sheath with a single protein layer surrounding the pyrenoid. Within this layer proteins are predicted to interact at 90 degree angles along the edges/front/back of the beta fold. If this is replicated within *T. pseudonana*, the structural positioning of these highly conserved residues into loop regions and corners may play a role in Shell protein docking.

Modelling Shell protein surface. The clustering of consensus residues to the edges and corners of the secondary structural model (Fig. 10B) is reflected in the surface models. The surface models of Shell1-7 show consensus residue side chains at the protein surface (Fig. 11). Although not identical, specific localisation of side chains appears conserved between Shell1-7. This suggests involvement consensus residues in Shell protein-Shell protein interactions. Modelling surface hydrophobicity showed consensus residue positioning does translate into neutral patches, although these are reduced in size (Fig. 11). The most consistent feature between Shell1-6 is a region of negative charge (red) in the cleft where an alpha helix sometimes sits. From the right hand side (RHS) orientation the beta fold surface exhibits patches of high, low and uncharged regions. Conversely, Shell7 appears strongly negatively charged in the central beta fold region. Whether this charge pattern affects Shell7 localisation and protein interactions is yet unknown.

Shell protein docking simulations do not suggest an obvious method of Shell protein interaction Computational simulations in AlphaFold Co-lab failed to confidently predict Shell protein homodimer interactions (supp fig). As modelling simulations improve it may be possible to predict Shell protein docking *in silico*, however shared localisation of multiple Shell proteins (Fig. 8) presents that multiple Shell proteins interact in a complex pattern to form the pyrenoid outer layer. There is also the possibility that an additional protein is involved in macromolecular structure. Future research, using *in vivo* imaging techniques, may reveal Shell protein associations and pyrenoid outer layer structure in *T. pseudonana*.

Phylogenetic analysis suggests divergent evolution of Shell proteins from a common ancestor. Shell protein phylogenetic analysis forms six distinct clades (A, B, C, D,E and X) with branching patterns suggesting multiple rounds of evolution(Fig. 12). The *C. reinhardtii* proteins BST1-3 are an outgroup, rooting the tree and suggesting the node position of a Shell protein common ancestor. This ancestor underwent four divergence events (Fig. 14) first separating into clades D, E and the remaining 4 clades. Two subsequent events led to the divergence of clades C and B, followed by a final divergence event forming clades A and X. Despite these multiple rounds of evolution Shell proteins retain high sequence similarity, as demonstrated through presence of a consensus sequence (Fig. 10A). It is plausible that a Shell protein common ancestor also possessed such a consensus sequence, probably most similar to the consensus region of proteins in clade D.

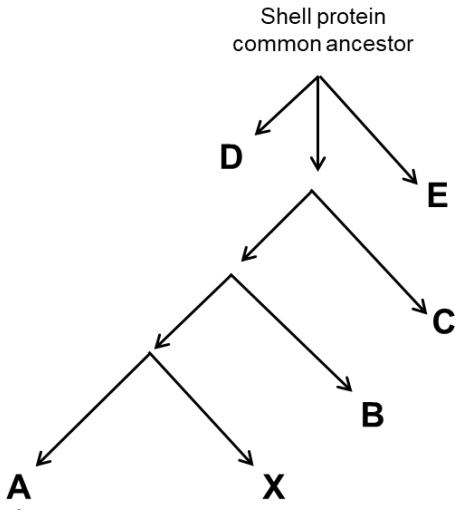


Figure 15 Diagram outlining Shell protein divergence events identified by phylogenetic analysis (Fig. 12). Asterisks represent divergence events. Line length are not quantitative representations of evolutionary time

The clear absence of Shell protein homologues in green and red algal lineages suggests Shell proteins

were not present in the algal endosymbionts forming haptophytes and stramenopiles (Fig.1). If they were, it is unlikely that Shell proteins have subsequently disappeared from all sequenced red and green algal species. Instead, it is more probable the Shell protein common ancestor was present in the heterotrophic host pre-endosymbiosis. If so, Shell proteins would predate the origin of both centric and pennate diatoms originating around 141Mya (23).

Pennate diatoms have evolved a Shell protein distinct from *T. Pseudonana.* Shell protein homologs from diverse species are distributed throughout the phylogenetic tree (Fig 12). Proteins from diatoms, pelagophytes and haptopytes are present throughout clades D,E,C and A, and clade B lacks only *A. anophagefferens* proteins .In contrast, clade X lacks *T. pseudonana* and comprises nearly entirely of Shell protein homologues from pennate diatom species of the class Bacillariophycae. If clade X possessed exclusively pennate diatoms, the date of the split of Bacillariophycae from the centric diatoms class Mediophycae could be used to date the divergence of this clade X (23). However, two of the 33 clade X proteins are from the centric diatom, *C. tenuissimus*. This makes the picture within Clade X less clear and suggests a greater complexity to the evolution of Shell protein homologs.

Combining sequence-based phylogeny with structural analysis can answer questions relating to protein evolution and biological function. In an attempt to reveal the deep evolutionary relationships among proteins, computational biologists have combined structural data with phylogenetic analysis. This is desirable as protein structures are often better conserved than the sequences from which MSAs are derived (104). Generating phylogeny from structural data alone is challenging as there is a need to create a quantifiable metric for structural similarity. When there is evidence of common descent, RMSD values between protein structures can be used with distance-based methods to reconstruct structural phylogeny(105). However, to determine the statistical significance of molecular relationships, a metric termed the 'Q score' combining RMSD value, alignment length, and molecular dynamics can be used. Using the 'Q score' metric Malik et al. were able to add statistical support to qualitative structural phylogeny of the ferritin-like protein superfamily(105,106). Such structural-phylogeny techniques can provide insight into the evolutionary relationships of structurally similar proteins. Inspired by this combined approach, Shell protein structural alignments were created and RMSD values determined. Unfortunately, the 'Q score' metric could not be employed as query structures need to be registered in the PDB, not simulated in AlphaFold. If the structures of Shell1-7 are solved experimentally, analysis via the 'Q score' metric presents and interesting future direction of study.

Structural similarity between of *T. Pseudonana* Shell proteins is reflected by phylogenetic positioning. Aligning the structures of Shell2-6 to Shell1 revealed a pattern of decreasing structural similarity (Fig. 13). Shell2 and 3 were the most structurally similar to Shell1, giving RMSD values of 0.208 and 0.194 respectively. The RMSD values then increased in the following order: Shell6, 4, 7, and 5 which is the least structurally similar to Shell1. Interestingly, when these RMSD values were mapped onto the corresponding phylogenetic branches, the trend in structural similarity was reflected with increased with phylogenetic distance from Shell1 (Fig. 12). This is perhaps unsurprising as protein sequence influences both phylogenetic positioning and protein structure, however it does suggest a consistency and evolutionary conservation of both Shell protein sequence and structure in *T. pseudonana*. Combining structural and phylogenetic analysis in this way is a novel approach within the study of pyrenoid-associated proteins.

Analysis of beta strand positioning suggests a structural rationale behind shell protein localisation patterns. Alignments of individual beta strands in Shell 1-5 showed differences in strand positioning and fold curvature (Fig. 14). Shell 1-3 show near identical strand positioning throughout the beta fold. In comparison the slight shifts in shell4 strands 6A, 7A, 13B, and 14B suggest an overall increased curvature towards the centre of the beta fold. The Shell5 central strands show a more marked separation in positioning from Shell1,2, and 3. Furthermore, at the opposite end to the aligned strand (1A) strands 8B and 9B are twisted away from the other four Shell proteins analysed. This suggests not only increased curvature towards the centre of the beta fold but an overall twisting in the fold structure. These distinct twists in strand position may explain why the Shell 5 RMSD value is the greatest (0.893Å) (Fig. 13). The apparent increased curvature of Shell4 and 5 could explain their distinct sub-pyrenoid localisation. Fluorescent microscopy images have shown Shell4 and 5 localise to the cylindrical surface of the *T. pseudonana* lenticular pyrenoid (Fig. 8). This circular region may possess higher curvature than the rectangular localisation pattern of Shell1,2, and 3. If so, the increased beta fold curvature observed in Shell4 and 5 may influence the overall curvature of the Shell protein layer. Using this hypothesis of distinct high/low curvature, Shell6 and 7 would be predicted to localise with Shell4 and 5 in the more curved regions surrounding the pyrenoid. As localisations experiments for Shell6 and 7 are currently being undertaken it will be interesting to see how they fit into the *T. pseudonana* Shell protein model.

Chapter conclusions. The *T. pseudonana* Shell proteins share a conserved sequence with protein homologs from algal species derived by secondary endosymbiosis. It is likely that the diatom common ancestor possessed a Shell protein which then diverged into the homologs present today. Structural analysis of Shell1-7 has highlighted a characteristic beta fold and conserved positioning of consensus residues. Subtle differences in strand positioning between the *T. pseudonana* Shell proteins has been used form a high/low curvature hypothesis which suggests a structural basis for Shell protein localisation patterns.

Chapter 5: Final Conclusions and Future Perspectives

Despite the great diversity in the pyrenoid structures, three key parts of the algal pyrenoid-based CCMs appear to be common between different algal and hornworts species. These are: the formation of a pyrenoid with a densely aggregated Rubisco matrix, the release of CO_2 by CAs increasing CO_2 concentration at the site of Rubisco, and thirdly, the formation of a diffusion barrier surrounding the matrix which prevents CO_2 leakage from the pyrenoid(11). Little is currently known about the presence presence of traversing thylakoid membranes that appear to contain a CA to release CO_2, particularly in the ecologically relevant diatom species. Through bioinformatics analysis this study has highlighted the underlying evolutionary complexity of pyrenoid components, with implications to the origins of algal species derived by secondary endosymbiosis.

There appears to be convergent evolution of linker proteins between algal clades. Whilst *C. reinhardtii* EPYC1 homologues have been found in in green algal species, this study was unable to identify homologues in other lineages, including diatoms. As the pyrenoid containing algae are polyphyletic alternative mechanisms of matrix formation, perhaps analogous to EPYC1, must occur. In *P. tricornutum*, PYCO1 partitions RuBisCO by LLPS forming the pyrenoid matrix. The differences in, RBM, RuBisCO binding site, and secondary structural features highlighted in this study suggest convergent evolution of algal linker function but not necessarily linker mechanism. Convergent evolution of CCM components is consistent with the multiple evolutionary origins of pyrenoids(53). Linker analysis suggests the *P. tricornutum* and *C. reinhardtii* evolved independently developing distinct mechanisms of linker action within the pyrenoid.

The specific localisation of CAs within *T. pseudonana*, *P. tricornutum* and *C. reinhardtii* sub compartments also suggests convergent evolution of pyrenoid components. In *T. pseudonana* and *P. tricornutum*, theta CAs localise to pyrenoid-penetrating thylakoid lumen, whereas in *C. reinhardtii* it is the alpha CA (CAH3) which is present (107). CAH3 releases CO_2 for Rubisco and is essential for CCM function (108). Although functional studies of luminal PtθCA1 and TpθCA2 are yet to be undertaken, they are proposed to fulfil the same function. If so, it would appear diatoms have convergently evolved a thylakoid-localised CA, distinct from the green algal lineage. This study's classification of the green algal LCIB and C as theta CAs show that this CA family is not unique to diatom species. It is possible future studies will reveal additional theta CAs across increasingly diverse algae, perhaps with implications for algal CCMs.

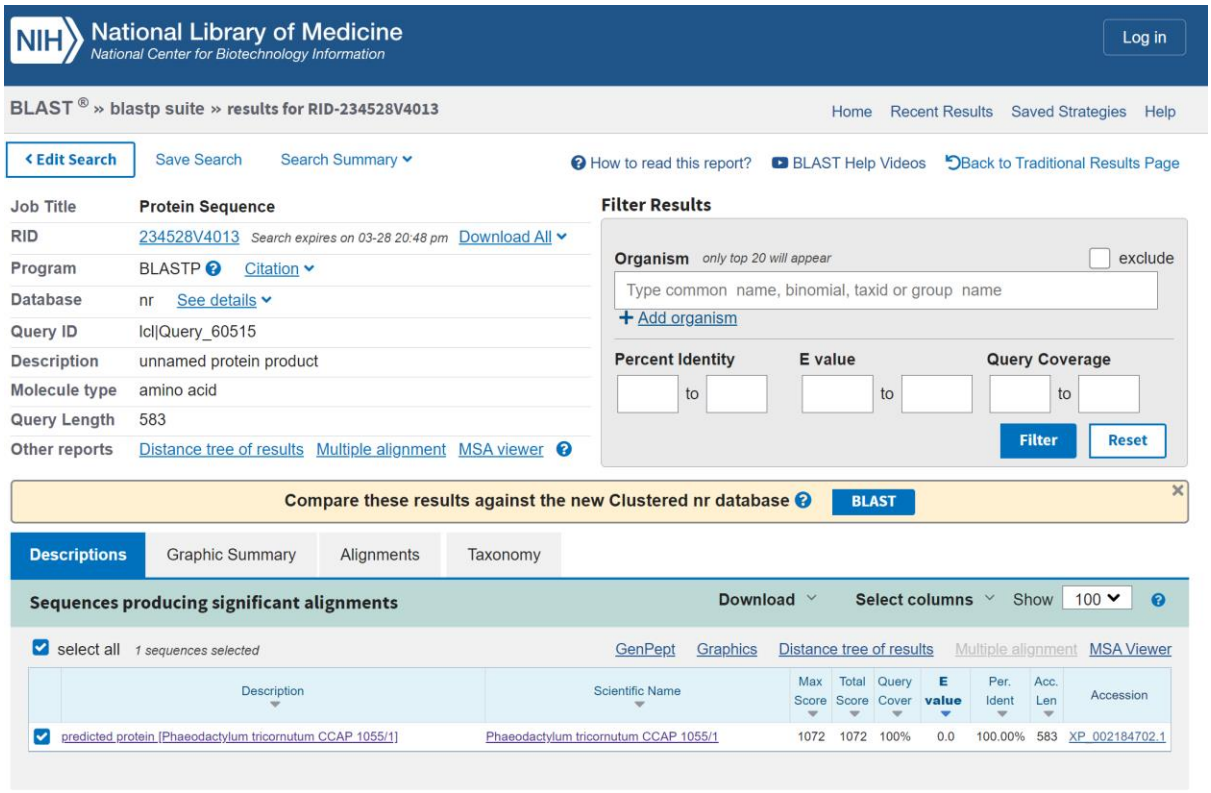
The putative role of Shell proteins as a CO_2 barrier, suggests an analogous function to the starch sheath observed in diverse algal species. If so, this suggests convergent evolution of pyranoid substructure and protein components. Despite this, phylogenetic analysis carried out in this study suggests within the shell protein family there is divergent evolution deriving from a common ancestor. The absence of Shell protein homologues from red and green algal lineages implies Shell proteins are unique to algae derived by complex secondary endosymbiosis. If so, Shell proteins would predate diatoms, originating before 141Mya and possibly even earlier if present in the haptophyte common ancestor (23,109). It is interesting to speculate on the role of Shell proteins within the heterotrophic host. It is currently debated whether the diatom common ancestor possess the peer annoyed. If so, did Shell proteins act as a CO_2 diffusion barrier and if not, what was the alternative function of Shell proteins? Until further studies shed light on diatom pyrenoid evolution, this question may remain unanswered.

Through combining bioinformatic analysis techniques, this study has highlighted the convergent evolution of algal pyrenoids and CCM components. Characterisation of the novel *T. pseudonana* Shell proteins was used to hypothesise the role differences in Shell protein structure relate to the

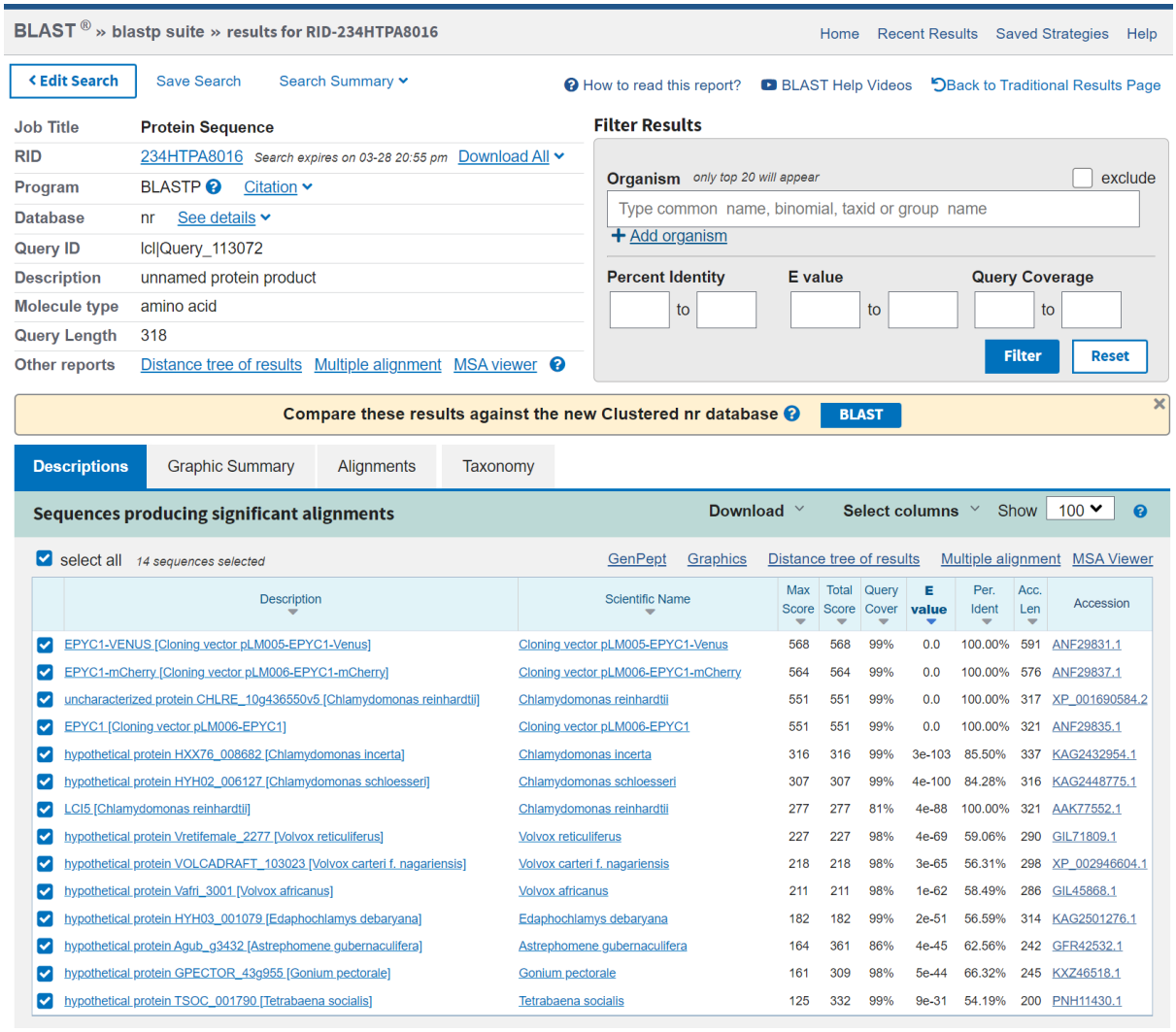
Shell protein localisation model. Structural phylogenetics, a novel technique in investigating algal CCMs, provided insight into the structural and sequence conservation between Shell proteins, relationships to Shell protein homologues and evolutionary origin. These findings, contribute to our understanding of the algal CCM within diverse algal lineages. The insight gained into diatom evolution, CCMs and relationship to other algal species is vital for understanding how diatoms shape global carbon flows, both now and in the future.

Supplementary documents

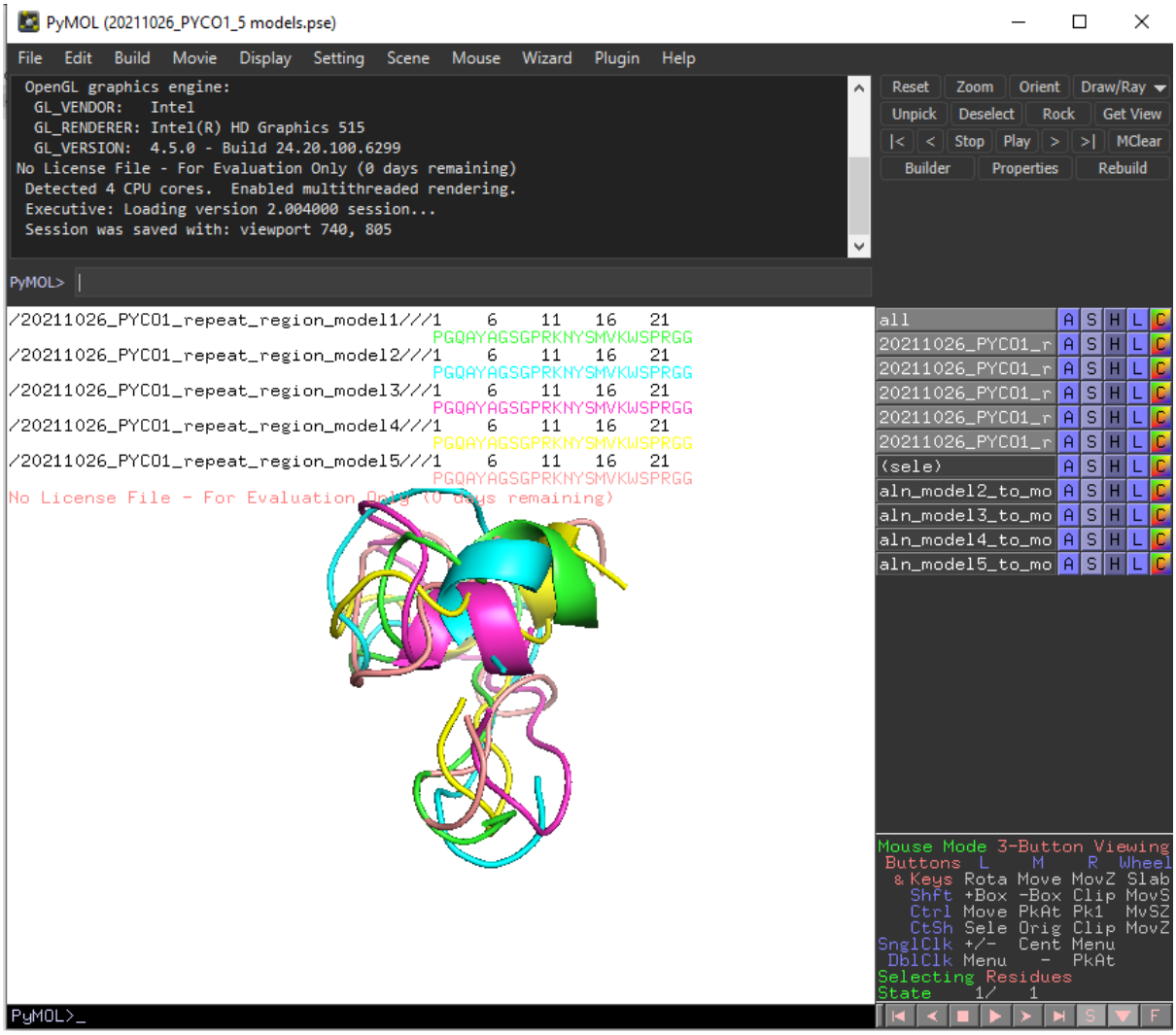
Supplementary Figures



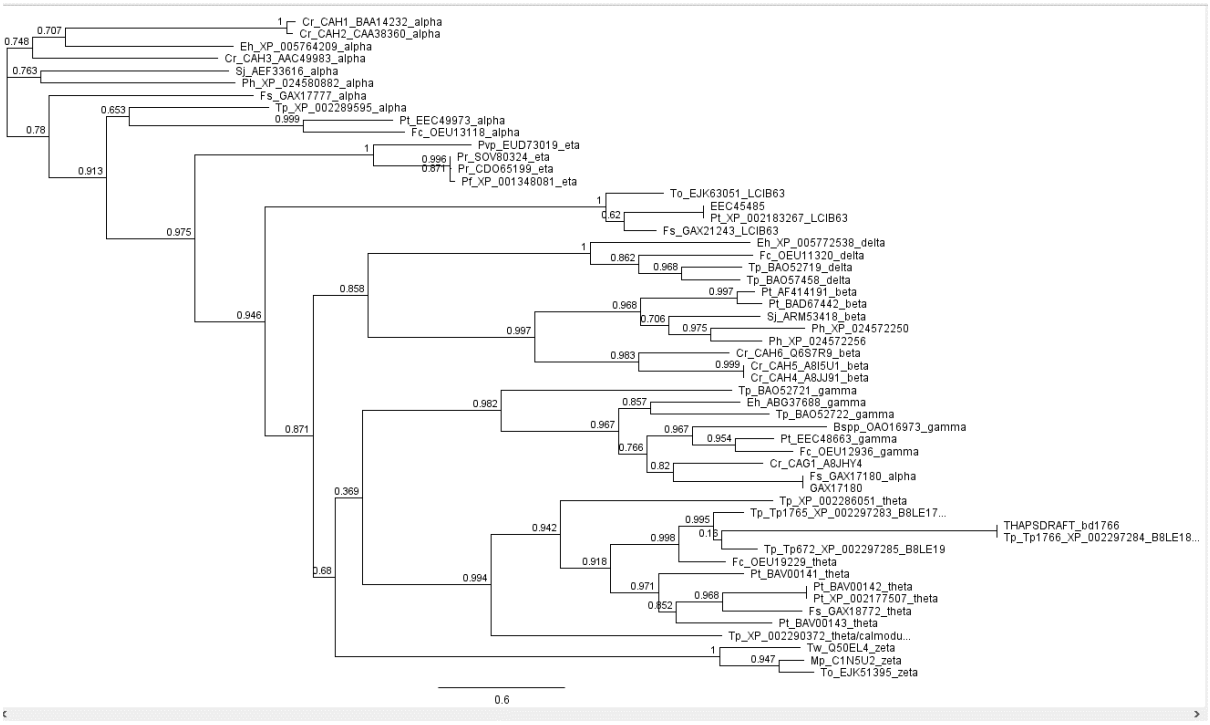
Supplementary Figure 1:PYCO1 BLAST search results



Supplementary Figure 2: EPYC1 BLAST search results

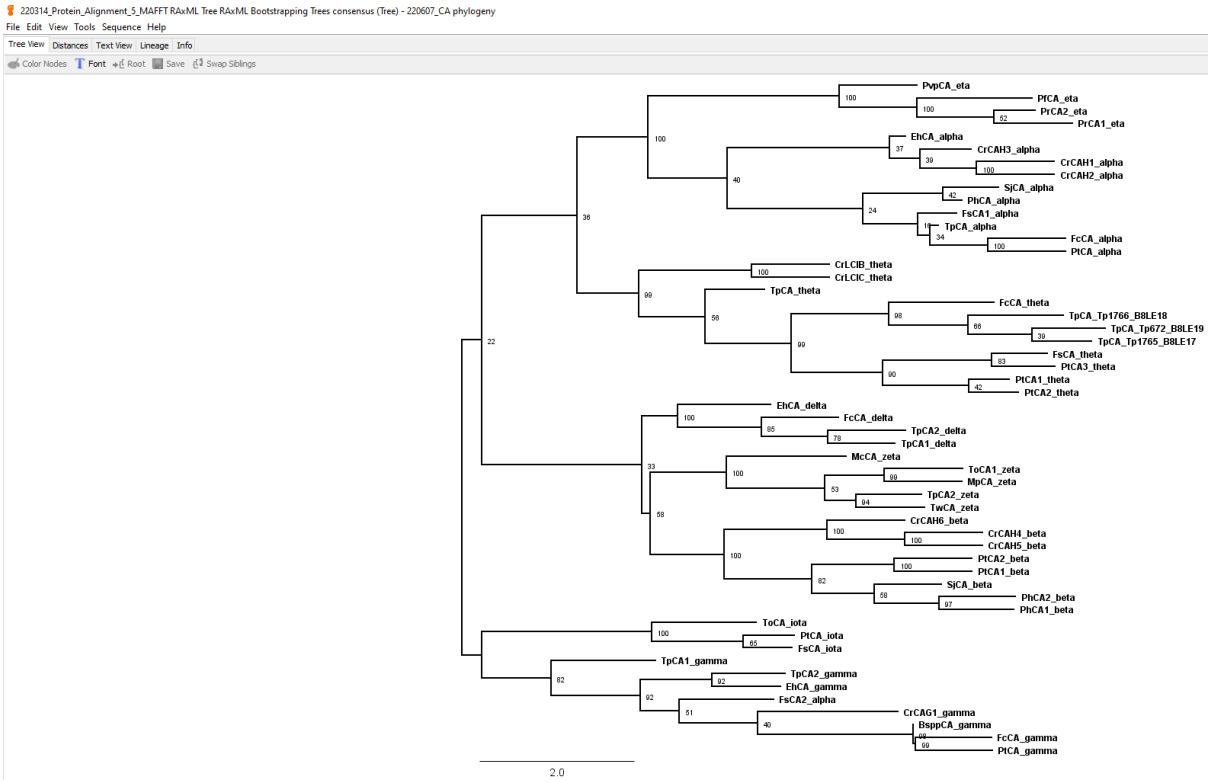


Supplementary Figure 3: Alignment of PYCO1 repeat region peptide models

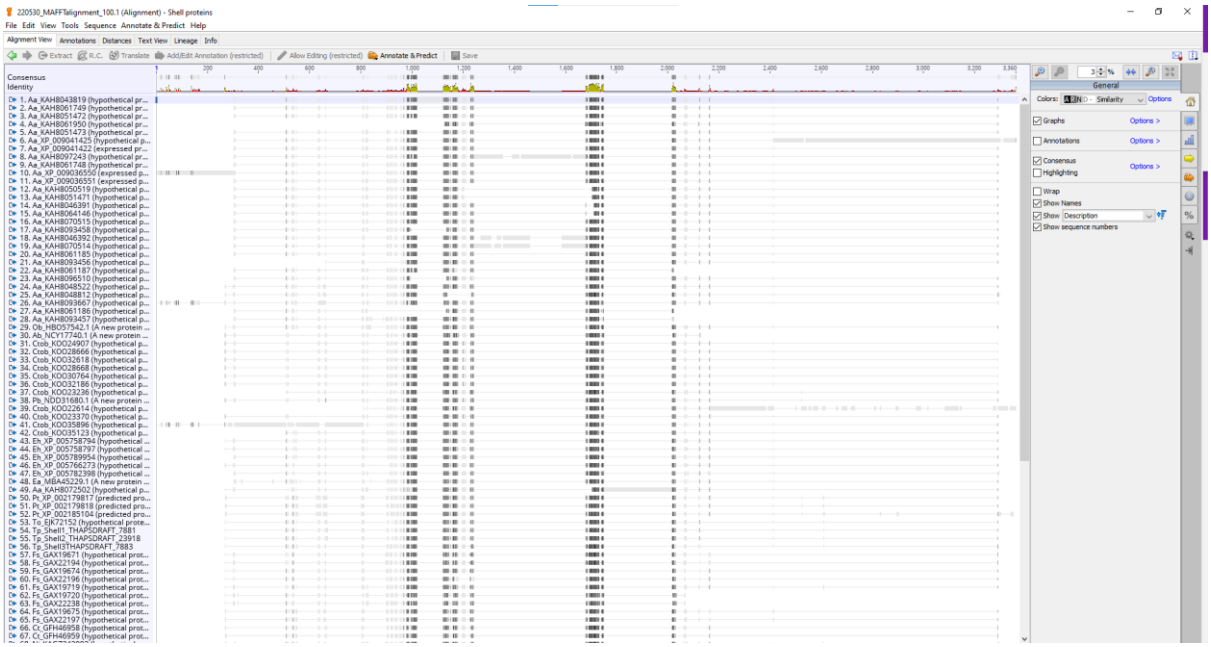


Supplementary Figure 4. *220111_CA_Alignment_1 FastTree Tree*

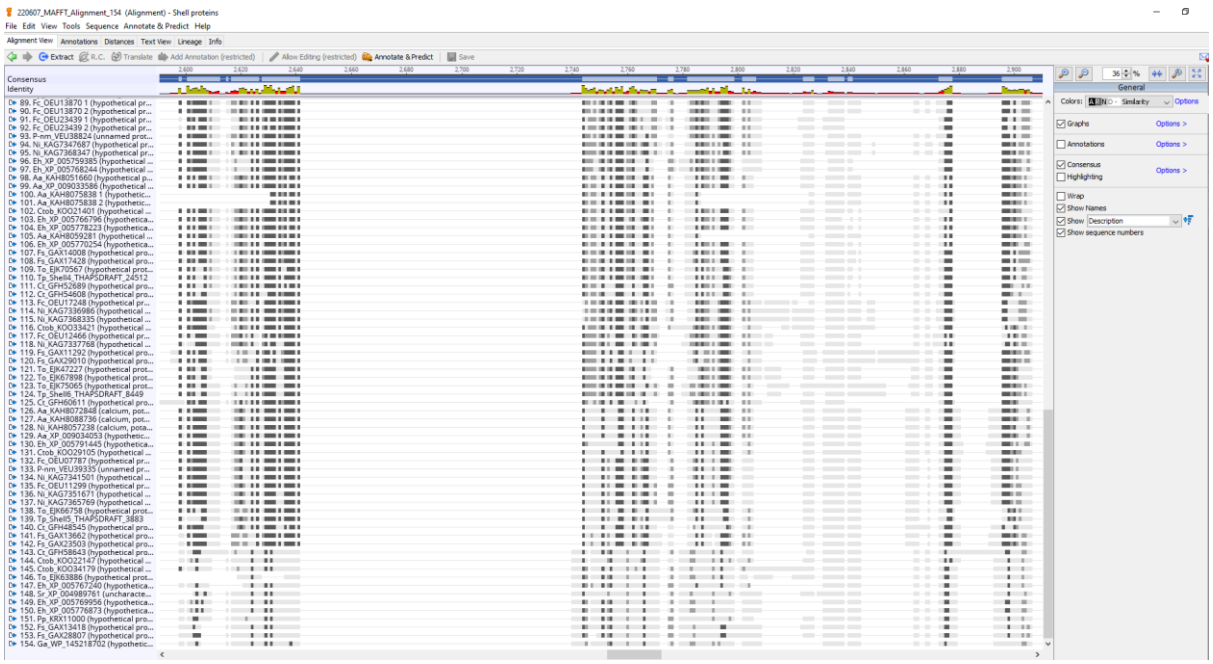
Supplementary Figure 5. *220314_Protein_Alignment_5_MAFFT FastTree Tree*



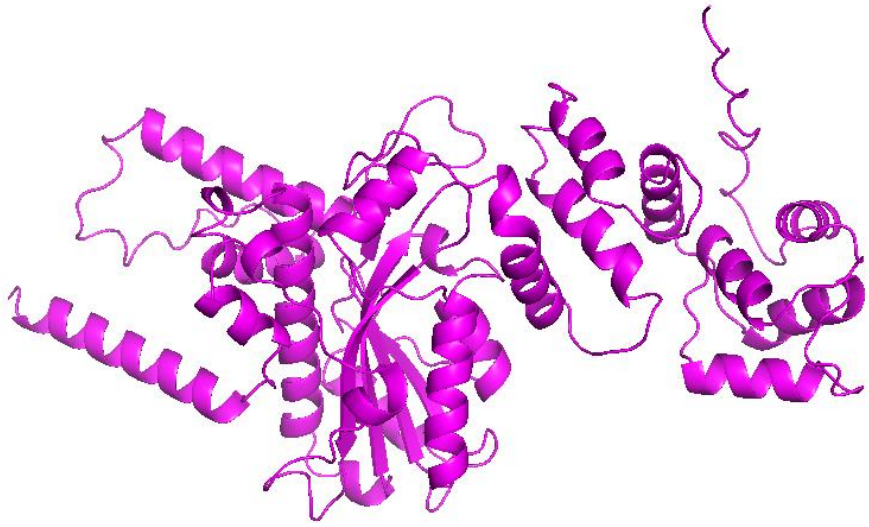
Supplementary Figure 6. *220314_Protein_Alignment_5_MAFFT RAxML Tree RAxML Bootstrapping Trees consensus*



Supplementary Figure 7. MAFFT alignment of 100 Shell protein homologues (including 6 TP Shell's) from early blast search. Consensus sequence regions are found at~1000- 1800 aa in this alignment *220530_MAFFTalignment_100.1*

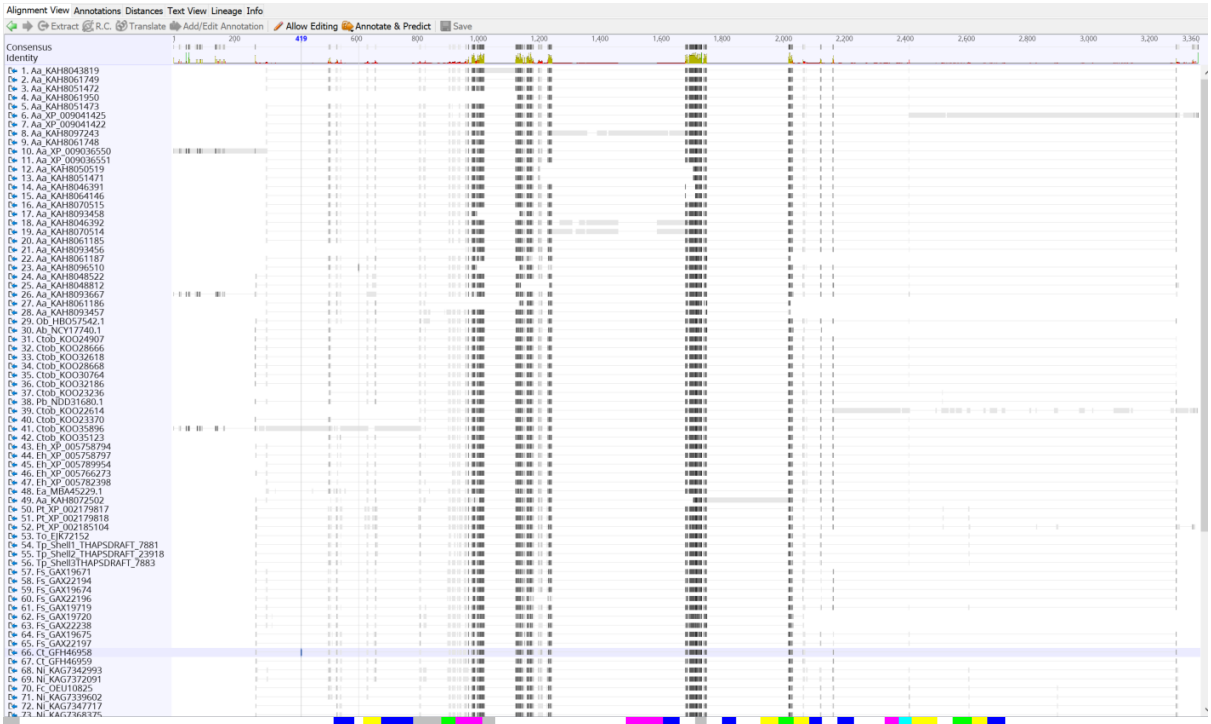


Supplementary Figure 8. MAFFT alignment of 154 Shell protein homologues (including 7 TP Shell's). Consensus sequence regions differ in hits from shell7 blasy 143-154 *220607_MAFFT_Alignment_154*

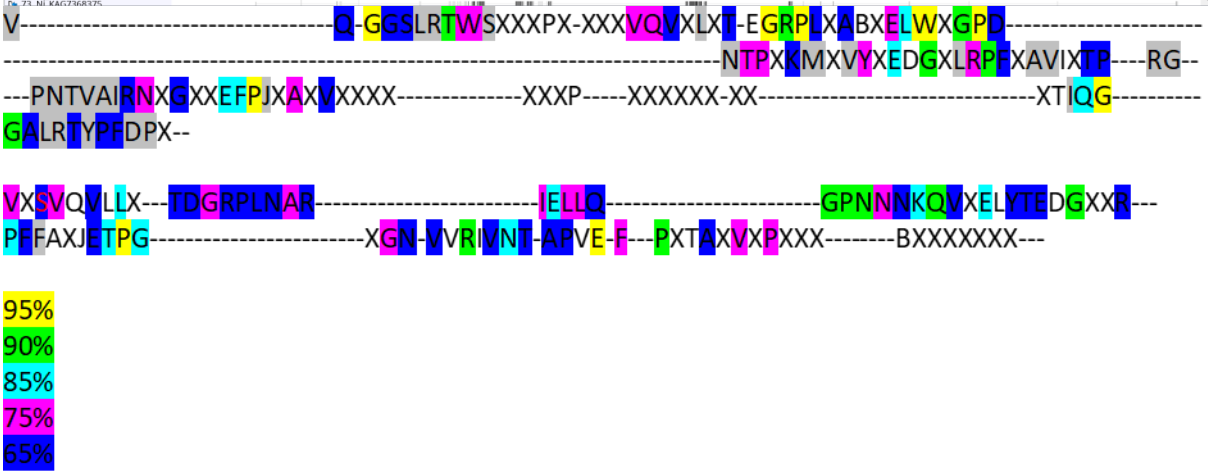


Supplementary Figure 9. AlphaFold structure of K0034179.1 C. tobinii show lack of characteristic shell protein bet fold *220608_KOO34179_c5ceb_unrelaxed_rank_1_model_1*

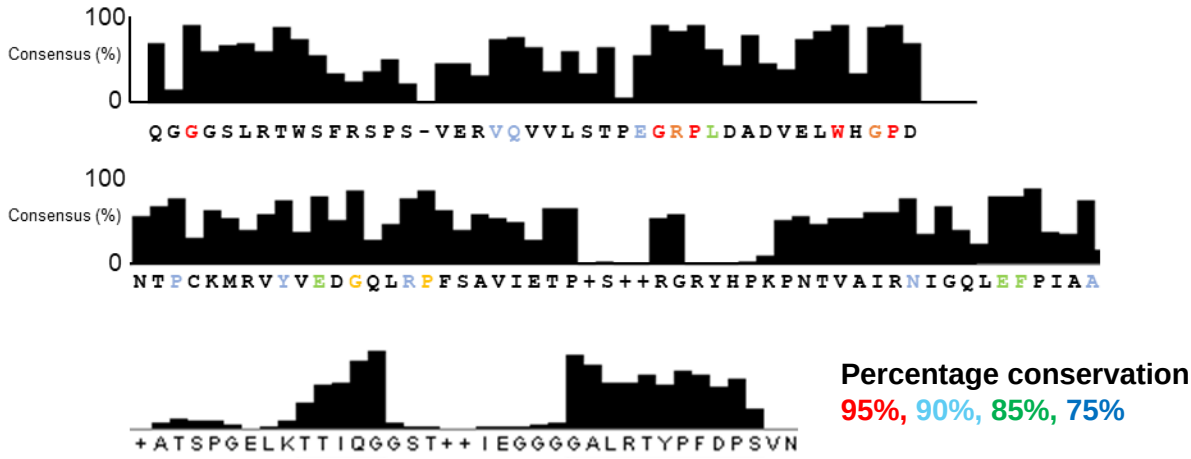
A



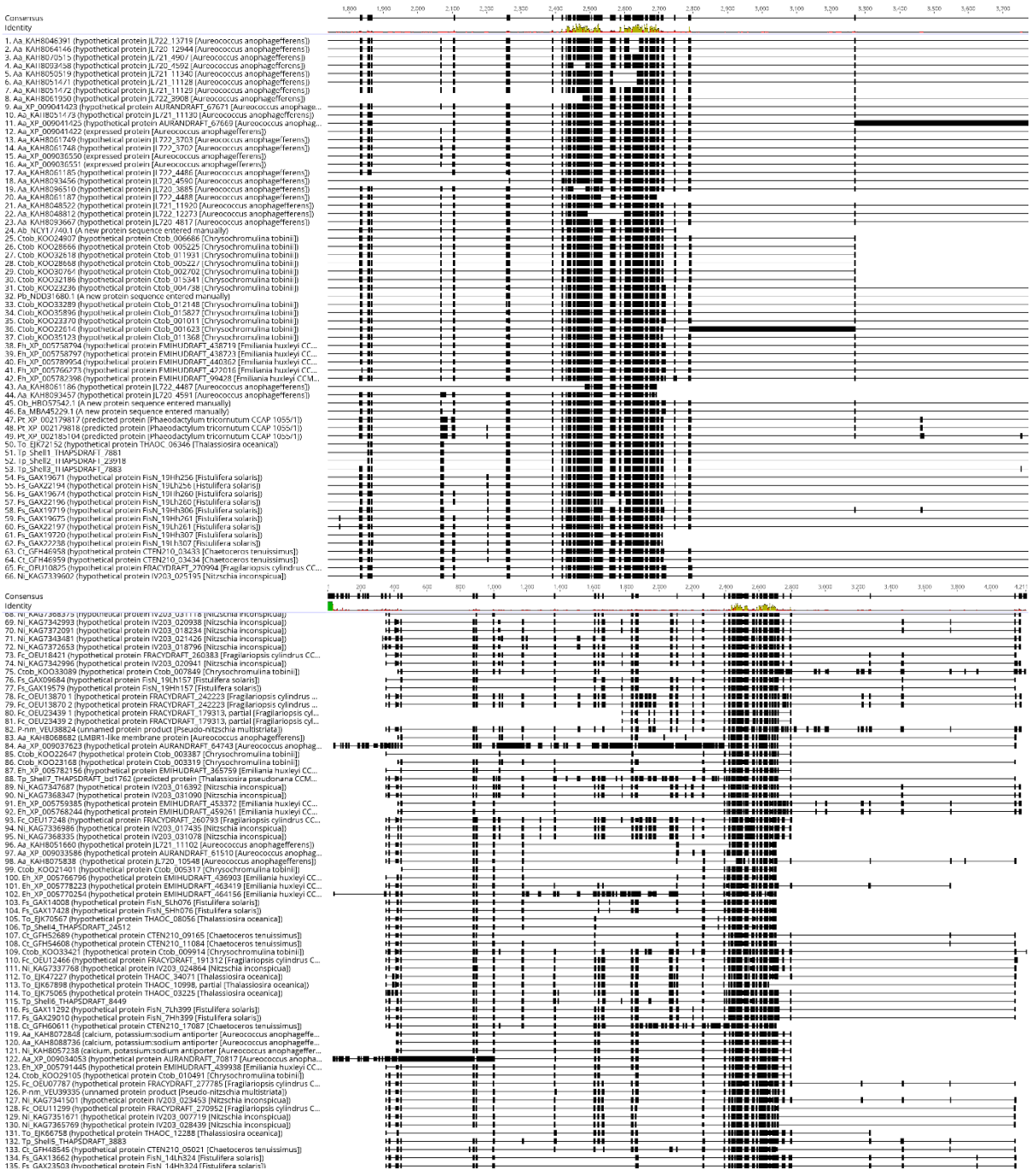
B



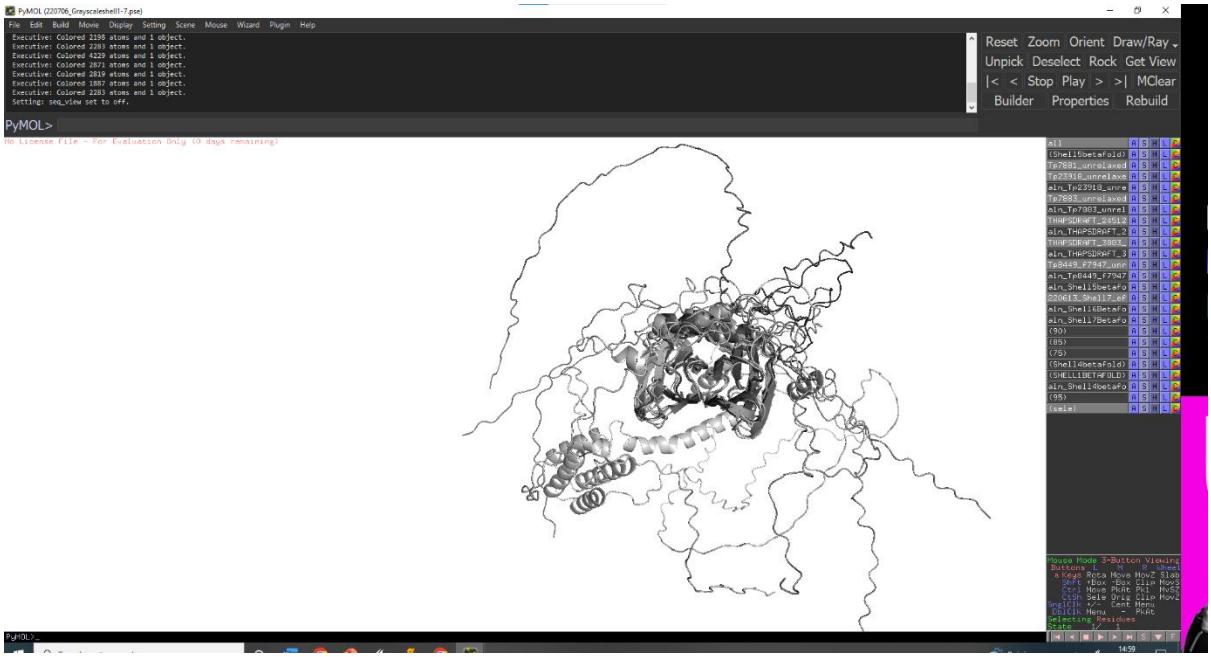
Supplementary Figure 10 a)MAFFT alignment of 144 Shell protein homologues showing consensus region at the top. *220606_MAFFTAlignment_ 144 sequences* B) manual highlighting of consensus residues with different percent conserved within the consensus sequence



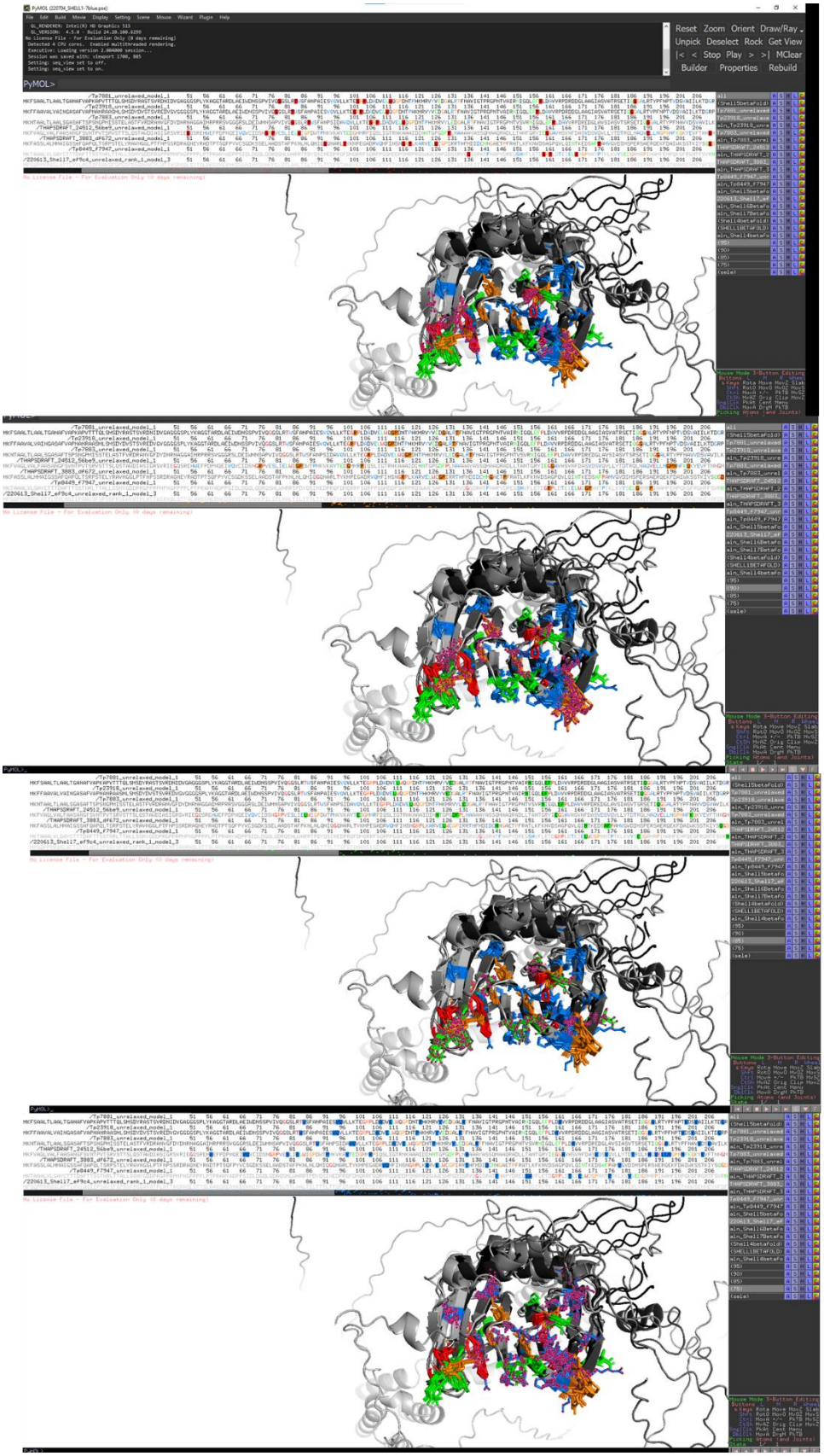
Supplementary Figure 11. Sections of consensus histogram from *220629_MAFFT144Shell.* Gaps indicate proteins contain insertions within the consensus sequence



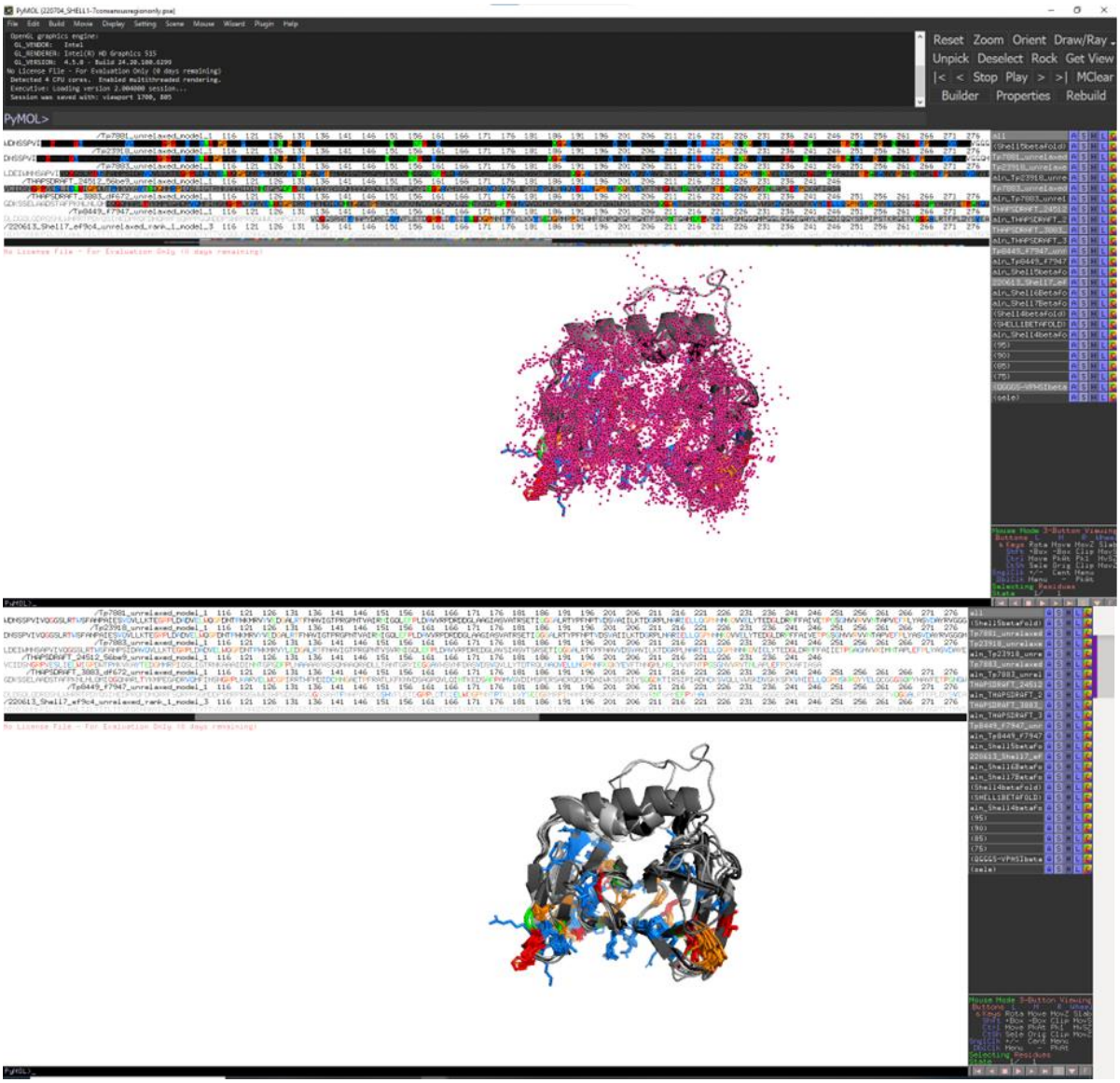
Supplementary Figure 13. *220704_MAFFTALIGN_Shell135*



Supplementary Figure 13. Alignment of consensus regions in the Shell1-7 models. Full protein structures shown



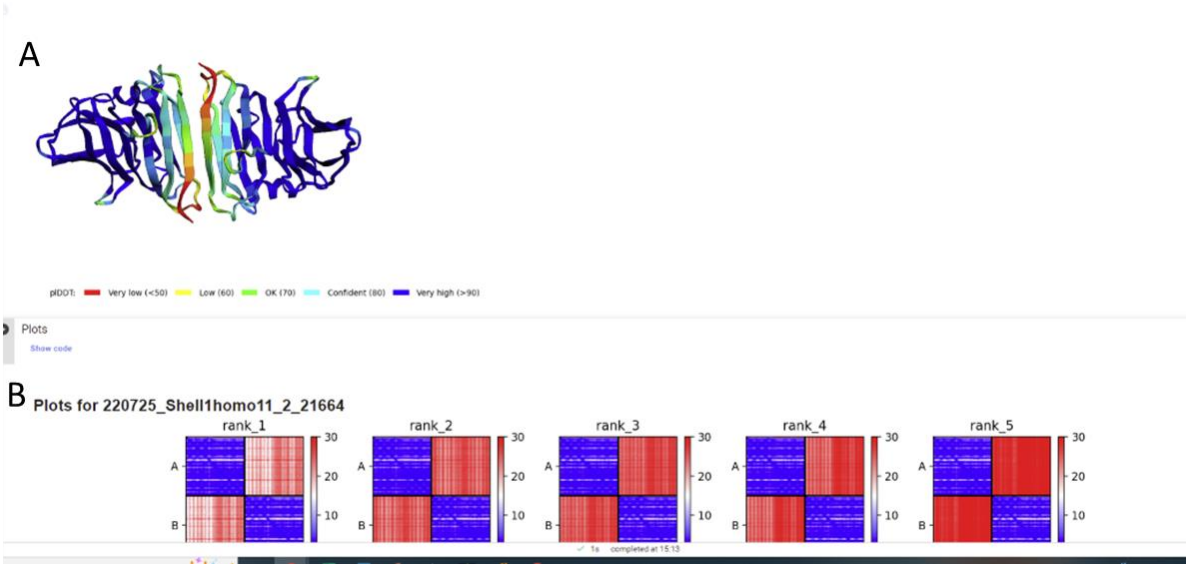
Supplementary Ffigure 14. Visualisation of consensus residues in Shell1-7. A) 95, red B) 90, orange C) 85, green D) 75, blue



Supplementary Figure 15. Removal of amino acids either side of the consensus region

220725_Shell1homo1.1_2
QGGSLRTWSFANPAIESVQVLLKTEGRPLDADVELWQGPDNTPHKMRVYVEDGALRTFNAVIGTPRGPNTVAIR
NIGQLEFPLDAVVRPDRDDGLAAGIASVATRSETIQGGALRTYPFNPTVDSVAIILKTDGRPLNARIELLQGPN
NNKQVVELYTEDGLDRPFFAIVETPGSGNVVRVVNTAPVEFPLYASVDAYRV:QGGSLRTWSFANPAIESVQVL
LKTEGRPLDADVELWQGPDNTPHKMRVYVEDGALRTFNAVIGTPRGPNTVAIRNIGQLEFPLDAVVRPDRDDGL
AAGIASVATRSETIQGGALRTYPFNPTVDSVAIILKTDGRPLNARIELLQGPNNNKQVVELYTEDGLDRPFFAI
VETPGSGNVVRVVNTAPVEFPLYASVDAYRV

Supplementary figure 16. Query sequence *220725_Shell1homo1.1_2*

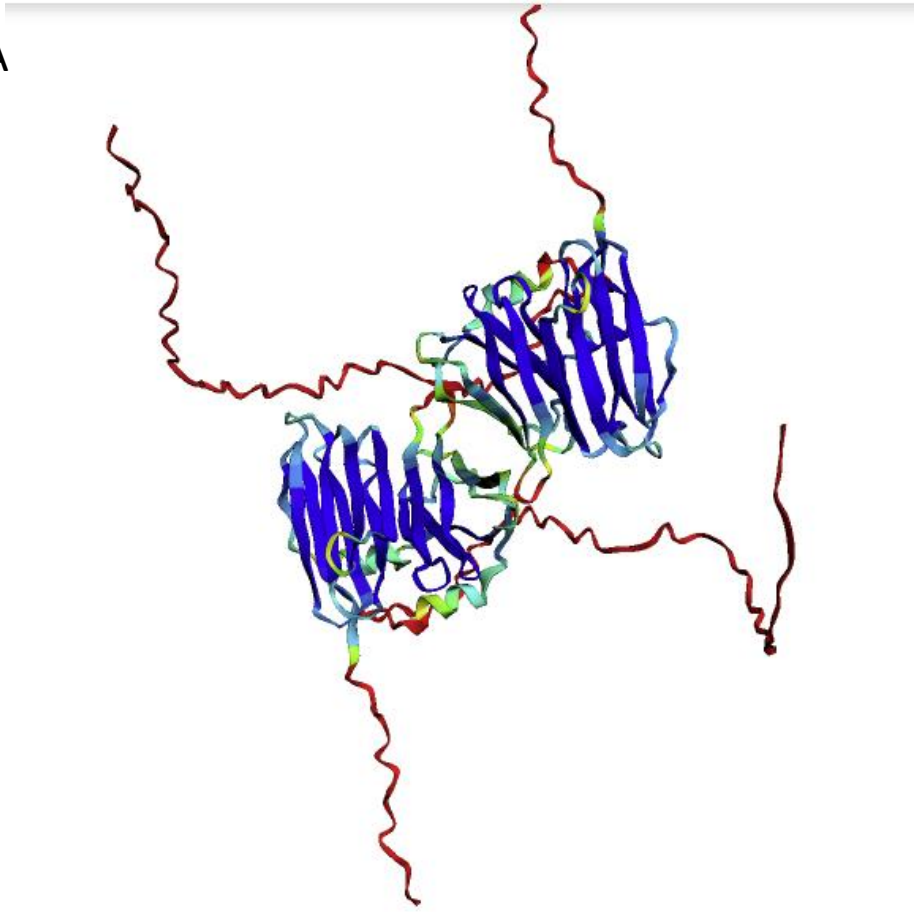


Supplementary Figure . 17 *220725_Shell1homo1.1_2*A) Predicted structural interaction B)PAE and pLDDT plots

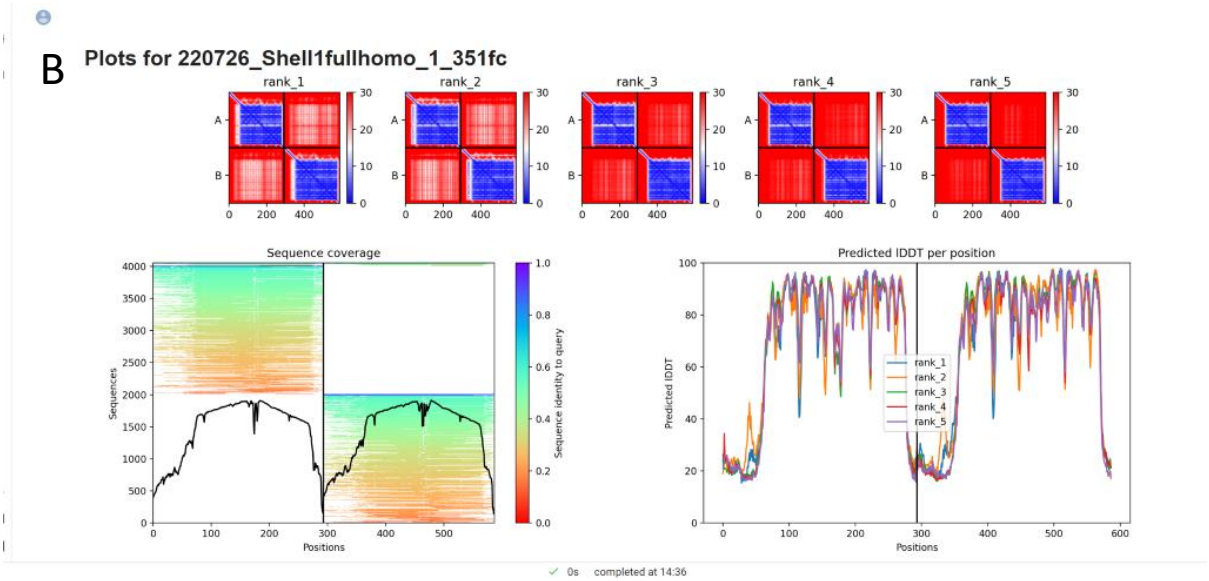
220726_Shell1fullhomo_1
MKFSAALTLAALTGANAFVAPKAPVTTTQLSMSDYRASTSVRDNIDVGAGGGSPLYKAGGTARDLAEIWDNSSP
VIVQGGSLRTWSFANPAIESVQVLLKTEGRPLDADVELWQGPDNTPHKMRVYVEDGALRTFNAVIGTPRGPNTV
AIRNIGQLEFPLDAVVRPDRDDGLAAGIASVARSETIQGGALRTYPFNPTVDSVAIILKTDGRPLNARIELLQG
PNNNKQVVELYTEDGLDRPFFAIVETPGSGNVVRVVNTAPVEFPLYASVDAYRVGGGGDWADDGLMIGRAF:MK
FSAALTLAALTGANAFVAPKAPVTTTQLSMSDYRASTSVRDNIDVGAGGGSPLYKAGGTARDLAEIWDNSSPVI
VQGGSLRTWSFANPAIESVQVLLKTEGRPLDADVELWQGPDNTPHKMRVYVEDGALRTFNAVIGTPRGPNTVAI
RNIGQLEFPLDAVVRPDRDDGLAAGIASVATRSETIQGGALRTYPFNPTVDSVAIILKTDGRPLNARIELLQGP
NNNKQVVELYTEDGLDRPFFAIVETPGSGNVVRVVNTAPVEFPLYASVDAYRVGGGGDWADDGLMIGRAF

Supplementary Figure 18. Query sequence *220725_Shell1homo1.1_2*

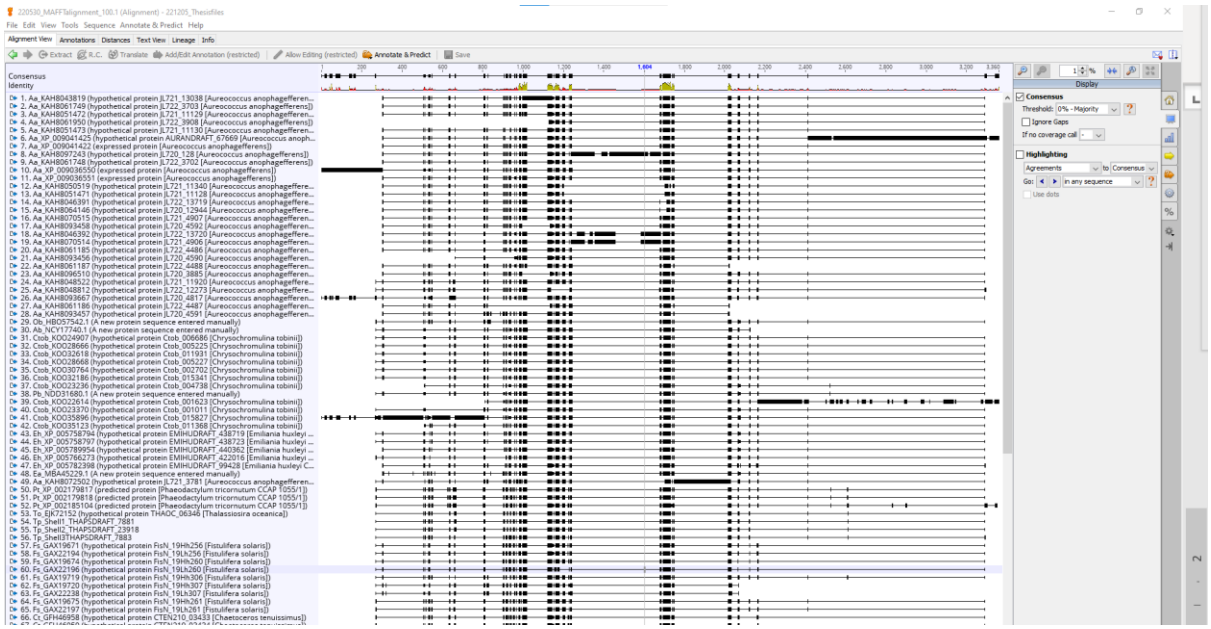
A



B

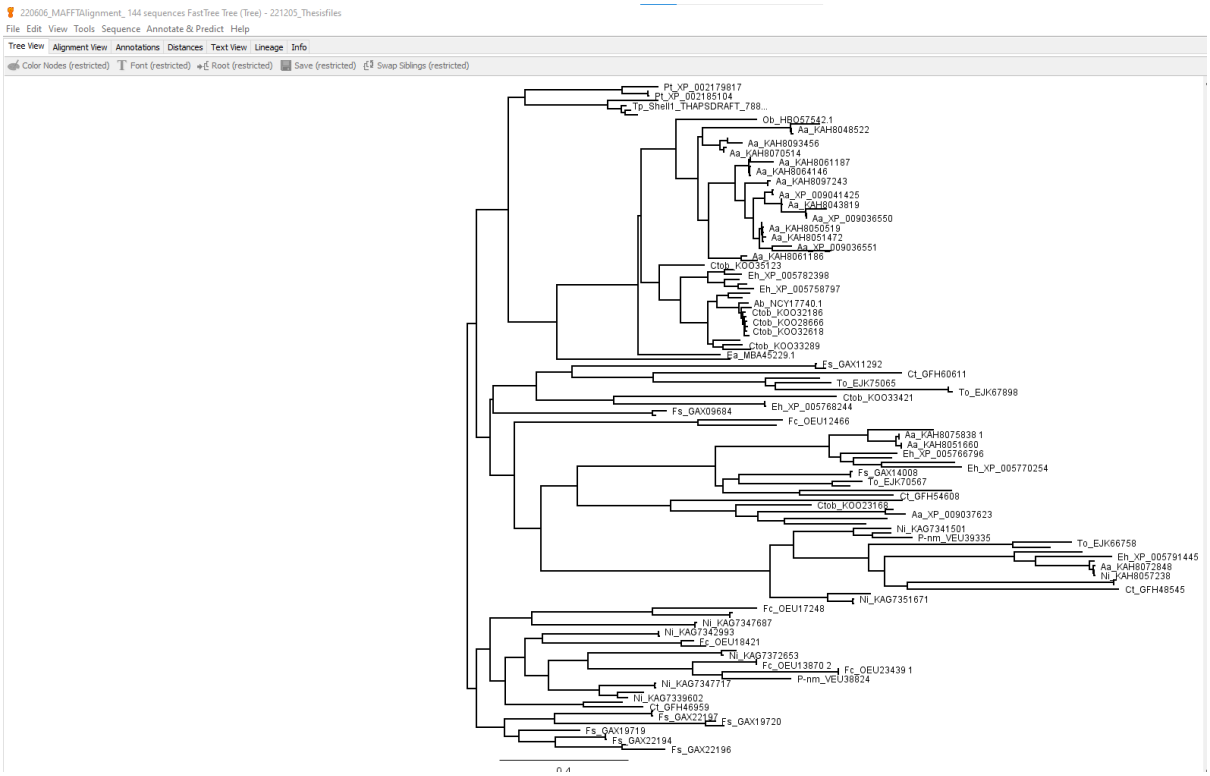


Supplementary Figure. 19 *220725_Shell1homo1.1_2* A) Predicted structural interaction B)PAE and pLDDT plots

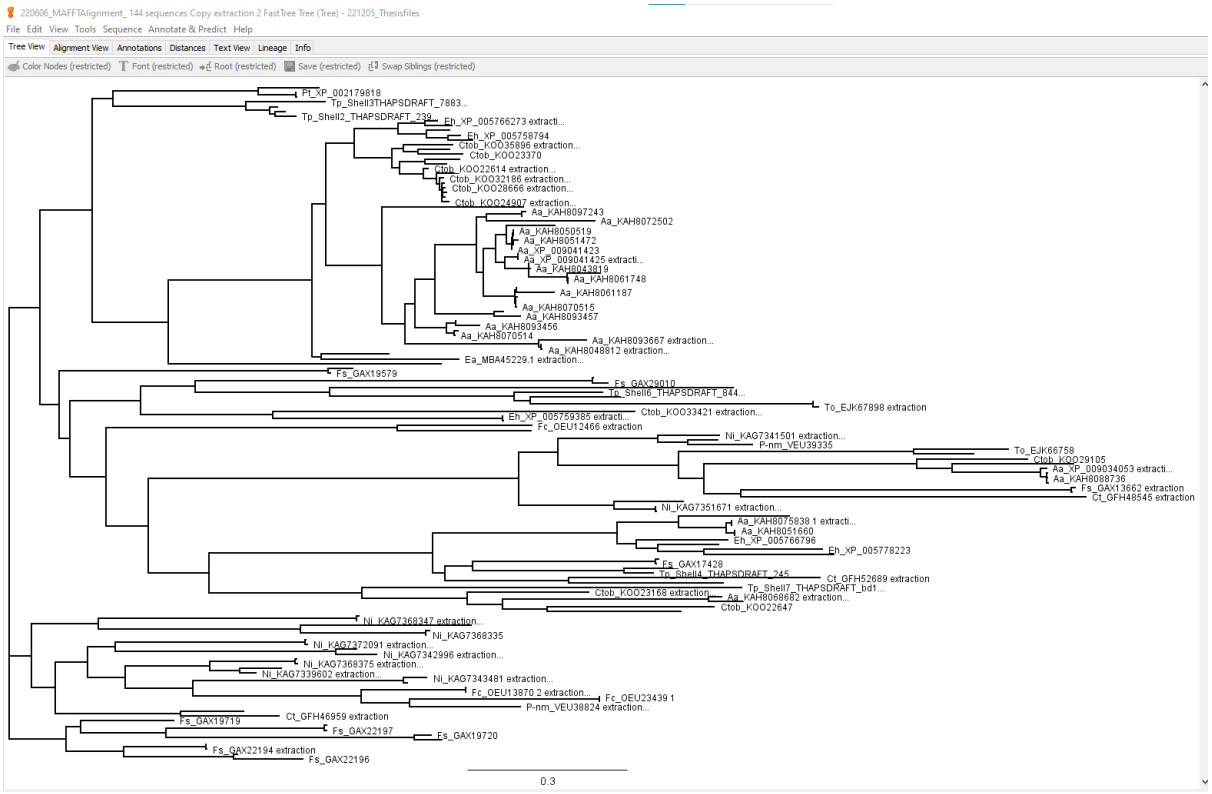


Supplementary Figure 20 220530_MAFFTalignment_100.1

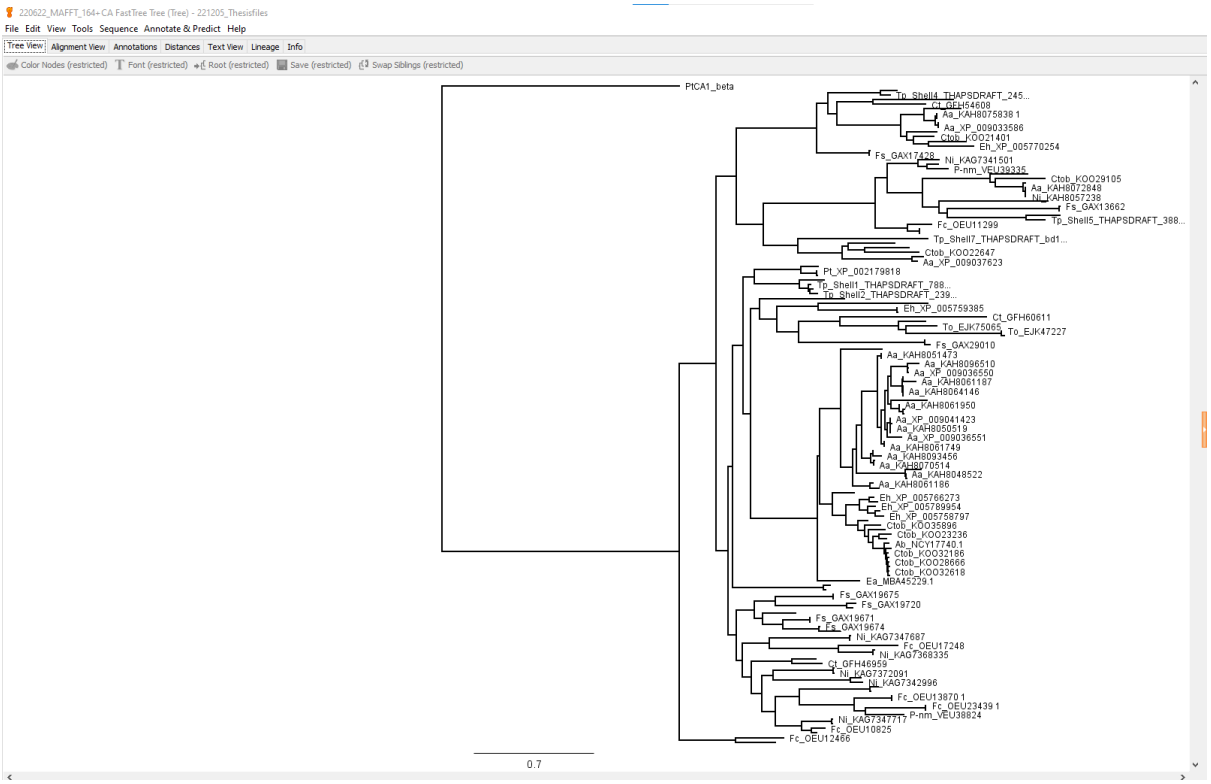
Supplementary Figure .21 220530_MAFFTalignment_100.1 FastTree Tree\



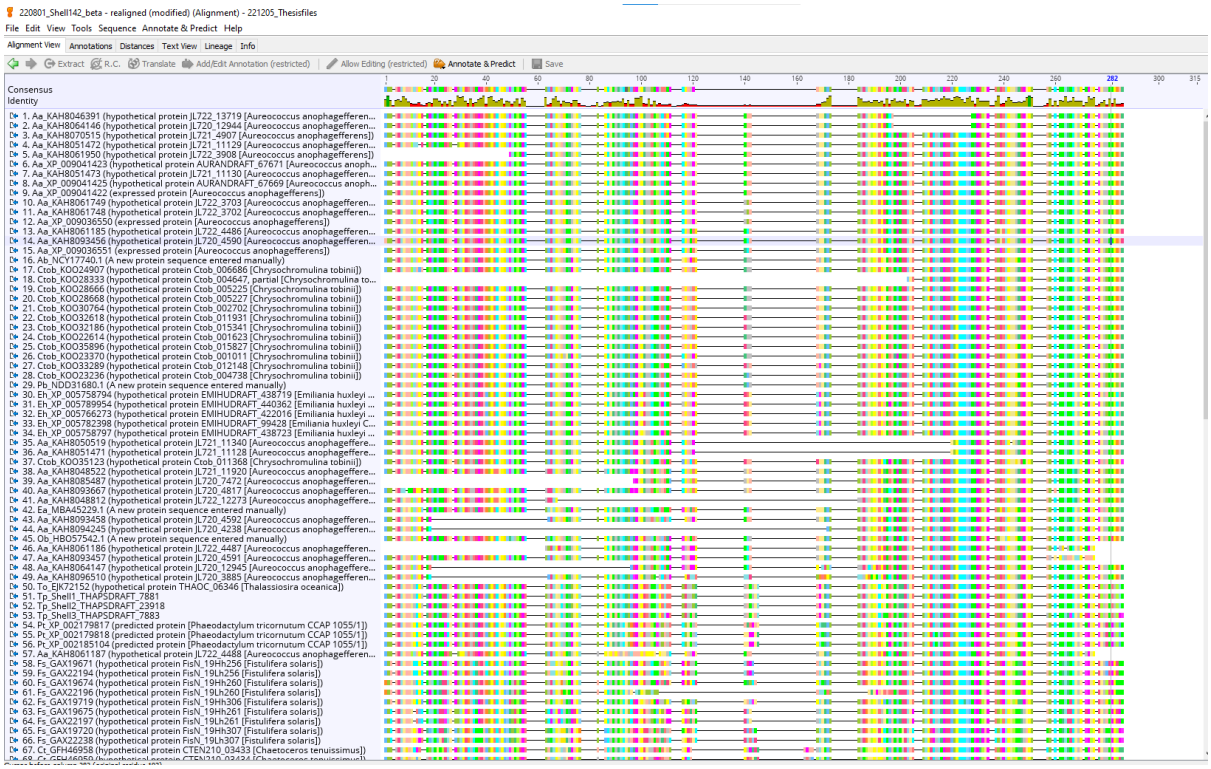
Supplementary Figure 22 220606_MAFFTAlignment_ 144 sequences FastTree Tree



Supplementary Figure 23 *220606_MAFFTAlignment_ 144 sequences Copy extraction 2 FastTree Tree*



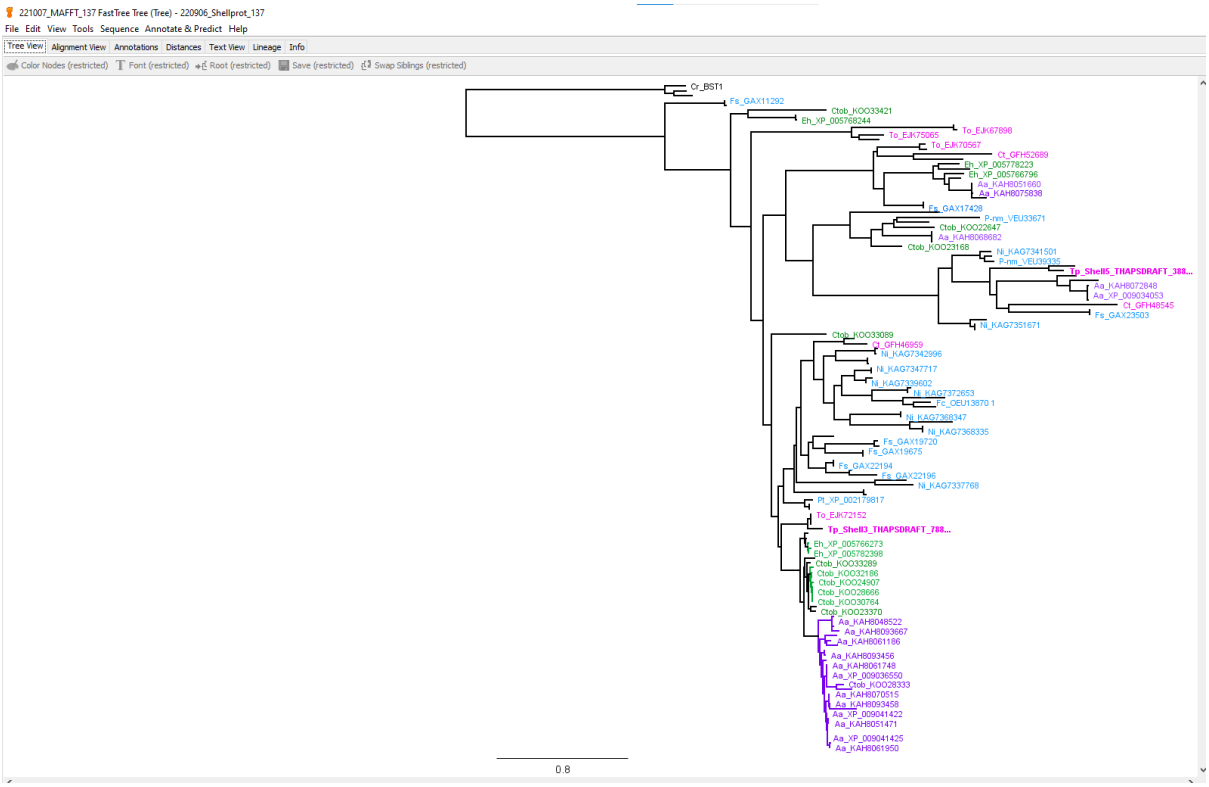
Supplementary Figure 24 *220622_MAFFT_164+CA FastTree Tree*



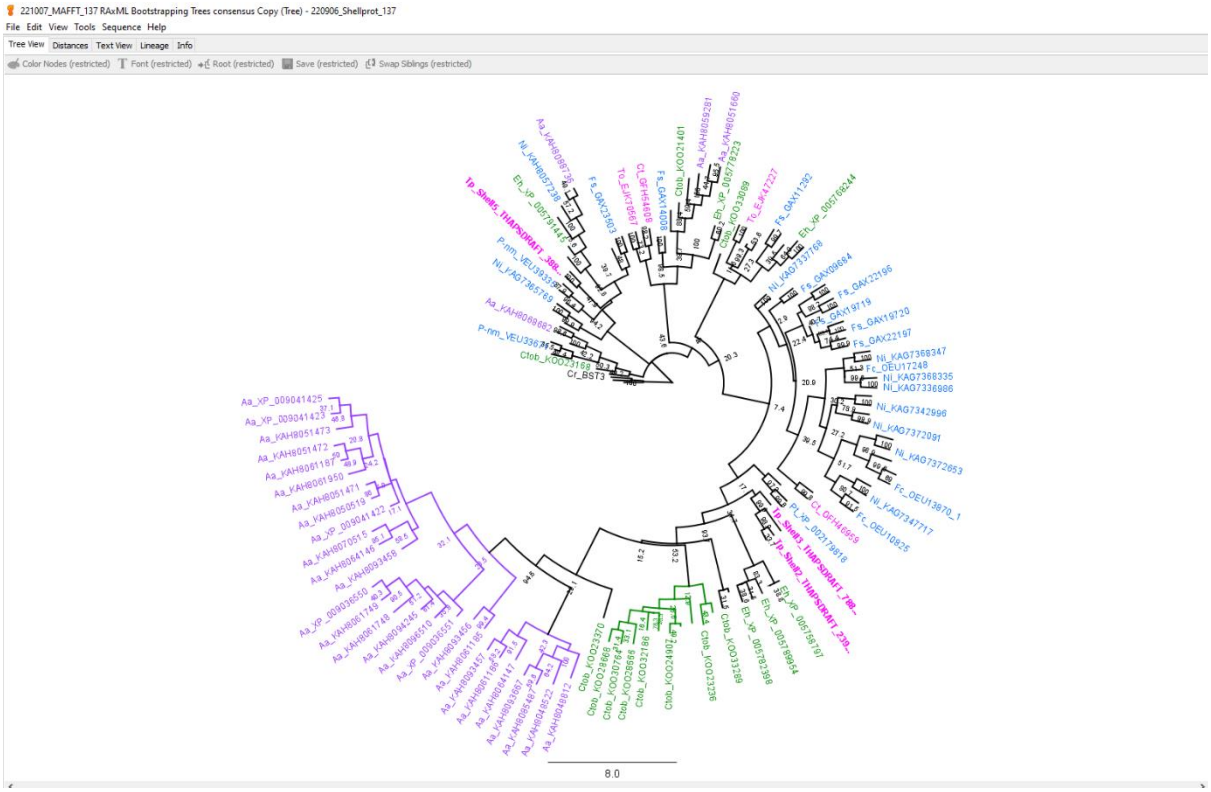
Supplementary Figure .25 *220801_Shell142_beta - realigned (modified)*

Supplementary Figure .26 *220801_Shell142_beta - realigned FastTree Tree*

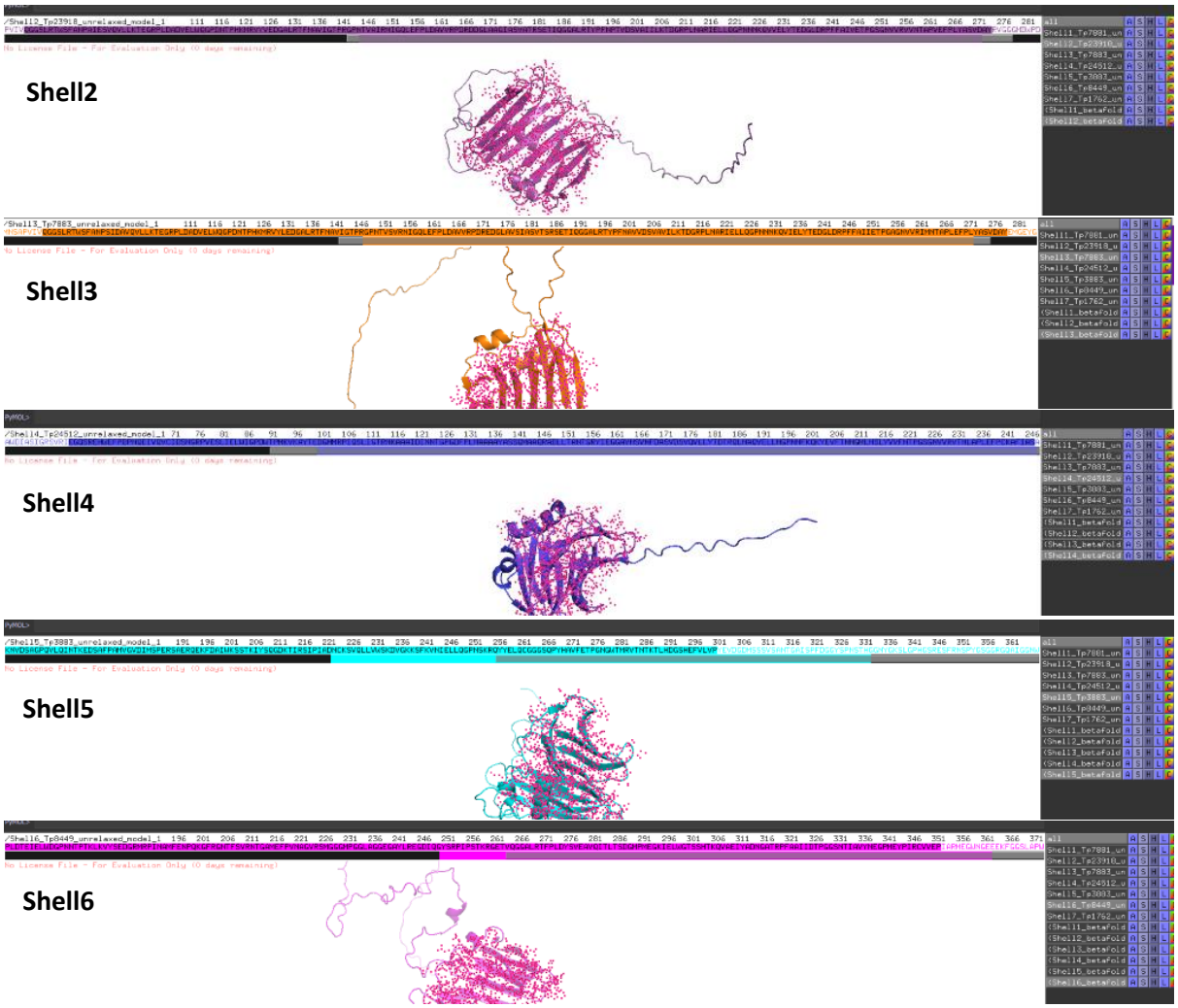
Supplementary Figure .27 *220802_MAFFTShellwithBST1-3 FastTree Tree*



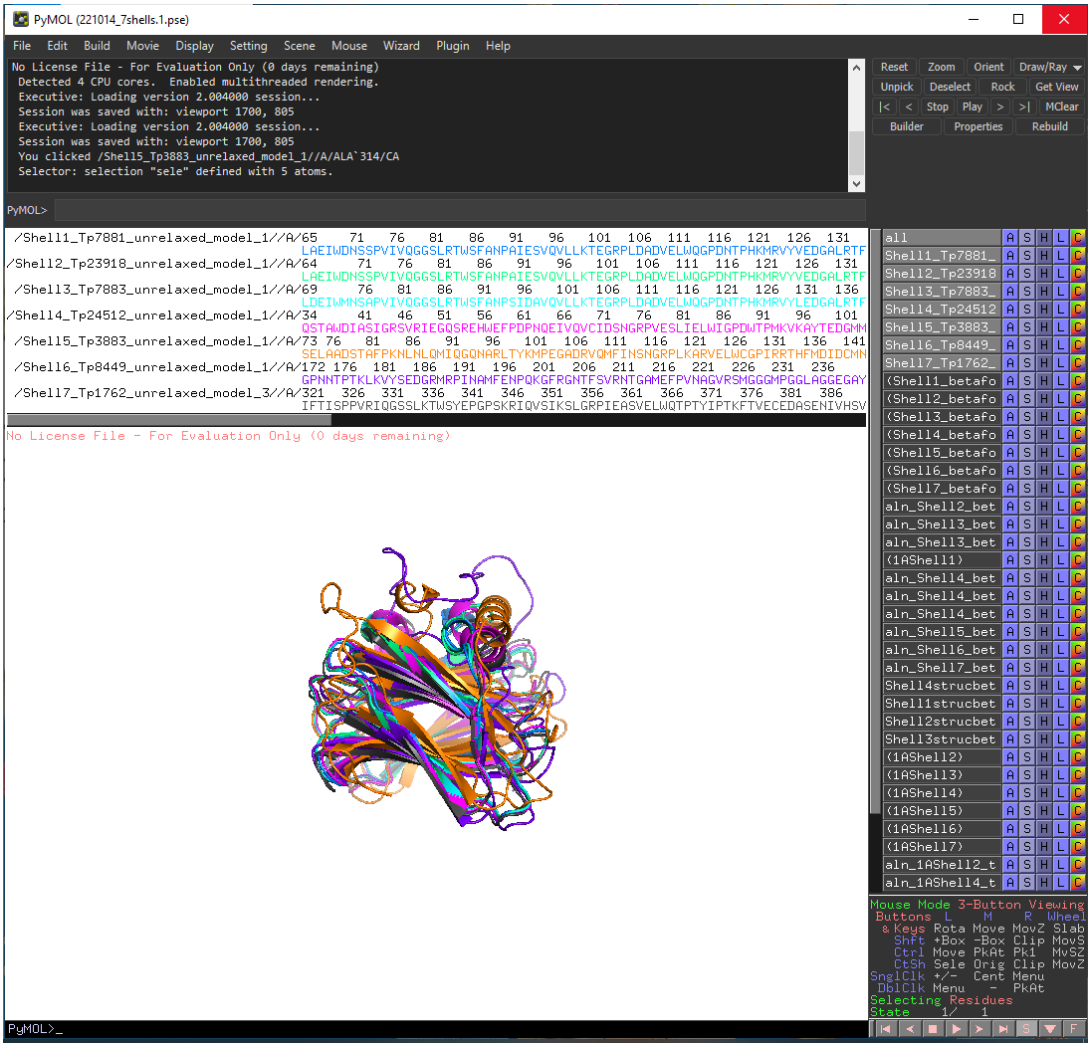
Supplementary Figure .28 *221007_MAFFT_137 FastTree Tree*



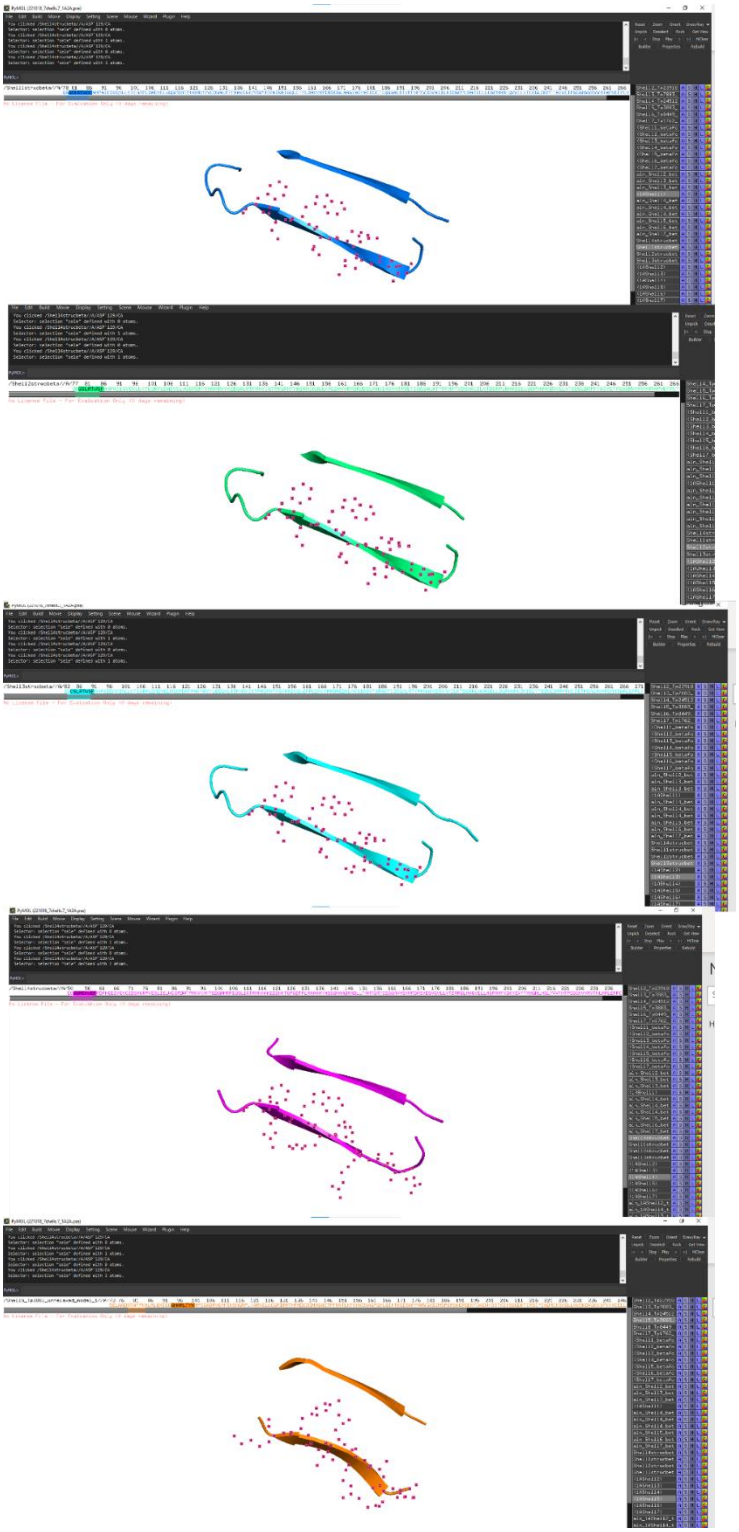
Supplementary Figure .29 *221007_MAFFT_137 RAxML Bootstrapping Trees consensus Copy*



Supplementary Figure .30 Selection of consensus regions *221014_7shells.1*



Supplementary Figure .31 Shell1-7 aligned by consensus region to Shell1_beta fold. Deletion of nonconsensus residues



Supplementary Figure .32 Shell1-5 Aligned by strand 1A, showing Strand 1A and 2A

Supplementary Tables

Supplementary Table 1. *211125_Table_CA*

Accession Number	Species	Predicted carbonic anhydrase
XP_005764209.1	Emiliania huxleyi	alpha
A8JHY4	Chlamydomonas reinhartii	Gamma
AAC49983.1	Chlamydomonas reinhartii	Alpha
AAL07493	Phaeodactylum tricornitum	Beta
ABG37688.1	Emiliania huxleyi	Gamma
AEF33616	Saccharina japonica	alpha
ARM53418.1	Saccharina japonica	Beta
BAA14232	Chlamydomonas reinhartii	Alpha
BAD67442.1	Phaeodactylum tricornitum	Beta
BAO52718.1	Thalassiosira pseudonana	Delta
BAO52721.1	Thalassiosira pseudonana	Gamma
BAO52722.1	Thalassiosira pseudonana	Gamma
BAO57458	Thalassiosira pseudonana	Delta
BAV00141	Phaeodactylum tricornitum	theta putative
BAV00142.1	Phaeodactylum tricornitum	theta
BAV00143	Phaeodactylum tricornitum	Theta
CAA38360	Chlamydomonas reinhartii	Alpha
CAH4	Chlamydomonas reinhartii	Beta
CAH5	Chlamydomonas reinhartii	Beta
cah6	Chlamydomonas reinhartii	Beta
CDO65199	Plasmodium reichenowi	Eta
EEC48663.1	Phaeodactylum tricornitum	Gamma
EEC49973.1	Phaeodactylum tricornitum	Alpha
EJK5395	Thalassiosira oceanica	zeta
EJK63051	Thalassiosira oceanica	LCIB63
EUD73019	Plasmodium vpPlasmodium v	Eta
GAX17180	Fistulifera solaris	Gamma
GAX17777	Fistulifera solaris	alpha
gax18772	Fistulifera solaris	theta
GAX21243	Fistulifera solaris	LCIB63
OAO16973	Blastocystis sp	Gamma
OEU11320.1	Fragilariopsis cylindrus	Delta
OEU12936.1	Fragilariopsis cylindrus	Gamma
OEU13118	Fragilariopsis cylindrus	alpha
OEU19229	Fragilariopsis cylindrus	theta
Q50LE4	Thalassiosira weissflogii	zeta
SOV80324	Plasmodium reichenowi	Eta
XP 002183267	Phaeodactylum tricornitum	LCIB63
XP 002295227	Thalassiosira pseudonana	zeta
XP 002297283	Thalassiosira pseudonana	
XP 002297284	Thalassiosira pseudonana	unk
XP 002297285	Thalassiosira pseudonana	unk
XP 002860	Thalassiosira pseudonana	unk
XP 00363214	Micromonas pusilla	zeta
XP_001348081.2	Plasmodium falciparum	Eta
XP_002177507.1	Phaeodactylum tricornitum	Theta
XP_002286051	Thalassiosira pseudonana	theta putative
XP_002290372	Thalassiosira pseudonana	theta putative
XP_002290372.1	Thalassiosira pseudonana	theta putative
XP_005772538.	Emiliania huxleyi	Delta
XP_024580882	Plasmopara halstedii	alpha
XP002504722	Micromonas commoda	zeta
	Thalassiosira pseudonana	theta putative

Supplementary Table 2 *220809_CA_MAFFTALIGN_53*

Geneious Name	Accession number	Species	Known CA Family	Phylogeny predicted CA Family
BsppCA_gamma	OAO16973	Blastocystis sp	Gamma	Gamma
CrCAG1_gamma	A8JHY4_CHLRE	Chlamydomonas reinhartii	Gamma	Gamma
CrCAH1_alpha	BAA14232	Chlamydomonas reinhartii	Alpha	Alpha
CrCAH2_alpha	CAA38360	Chlamydomonas reinhartii	Alpha	Alpha
CrCAH3_alpha	AAC49983.1	Chlamydomonas reinhartii	Alpha	Alpha
CrCAH4_beta	A8JJ91_CHLRE	Chlamydomonas reinhartii	Beta	Beta
CrCAH5_beta	A8I5U1_CHLRE	Chlamydomonas reinhartii	Beta	Beta
CrCAH6_beta	Q6S7R9_CHLRE	Chlamydomonas reinhartii	Beta	Beta
CrLCIB_theta	BAD16682	Chlamydomonas reinhartii	Theta/Beta	Theta
CrLCIC_theta	BAD16683	Chlamydomonas reinhartii	Theta/Beta	Theta
EhCA_alpha	XP_005764209.1	Emiliania huxleyi	Alpha	Alpha
EhCA_delta	XP_005772538.	Emiliania huxleyi	Delta	Delta
EhCA_gamma	ABG37688.1	Emiliania huxleyi	Gamma	Gamma
FcCA_alpha	OEU13118	Fragilariopsis cylindrus	Alpha	Alpha
FcCA_delta	OEU11320.1	Fragilariopsis cylindrus	Delta	Delta
FcCA_gamma	OEU12936.1	Fragilariopsis cylindrus	Gamma	Gamma
FcCA_theta	OEU19229	Fragilariopsis cylindrus	Theta	Theta
FsCA_alpha	GAX17180	Fistulifera solaris	Alpha	Alpha
FsCA_iota	A0A1Z5K4T2	Fistulifera solaris	Iota	Iota
FsCA_theta	GAX18772	Fistulifera solaris	Theta	Theta
McCA_zeta	XP002504722	Micromonas commoda	Zeta	Zeta
MpCA_zeta	XP 00363214	Micromonas pusilla	Eta	Eta
PfCA_eta	XP_001348081.2	Plasmodium falciparum	Eta	Eta
PhCA1_beta	XP_024572250	Plasmopara halstedii	Beta	Beta
PhCA2_beta	XP_024572256	Plasmopara halstedii	Beta	Beta
PhCA_alpha	XP_024580882	Plasmopara halstedii	Alpha	Alpha
PrCA1_eta	CDO65199	Plasmodium reichenowi	Eta	Eta
PrCA2_eta	SOV80324	Plasmodium reichenowi	Eta	Eta
PtCA1_beta	AAL07493	Phaeodactylum tricornitum	Beta	Beta
PtCA1_theta	BAV00141	Phaeodactylum tricornitum	Theta	Theta
PtCA2_beta	BAD67442.1	Phaeodactylum tricornitum	Beta	Beta
PtCA2_theta	BAV00143	Phaeodactylum tricornitum	Theta	Theta
PtCA3_theta	XP_002177507.1	Phaeodactylum tricornitum	Theta	Theta
PtCA_alpha	EEC49973.1	Phaeodactylum tricornitum	Alpha	Alpha
PtCA_gamma	EEC48663.1	Phaeodactylum tricornitum	Gamma	Gamma
PtCA_iota	XP 002183267	Phaeodactylum tricornitum	Iota	Iota
PvpCA_eta	EUD73019	Plasmodium vinckei petteri	Eta	Eta
SjCA_alpha	AEF33616	Saccharina japonica	Alpha	Alpha
SjCA_beta	ARM53418.1	Saccharina japonica	Beta	Beta
ToCA1_zeta	EJK5395	Thalassiosira oceanica	Zeta	Zeta
ToCA_iota	EJK63051	Thalassiosira oceanica	Iota	Iota
TpCA1_delta	BAO52719	Thalassiosira pseudonana	Delta	Delta
TpCA2_delta	BAO57458	Thalassiosira pseudonana	Delta	Delta
TpCA2_gamma	BAO52722.1	Thalassiosira pseudonana	Gamma	Gamma
TpCA2_zeta	XP 002295227	Thalassiosira pseudonana	Zeta	Zeta
TpCA_alpha	XP_002289595	Thalassiosira pseudonana	Alpha	Alpha
TpCA_iota	EED8870	Thalassiosira pseudonana	Iota	Iota
TpCA_Tp672_B8LE19	XP_002297285	Thalassiosira pseudonana	Theta	Theta
TpCA_Tp1765_B8LE17	XP_002297284	Thalassiosira pseudonana	Theta	Theta
TpCA_Tp1766_B8LE18	XP_002297283	Thalassiosira pseudonana	Theta/unkown	Theta
TpCA_Tp5647		Thalassiosira pseudonana	Theta	Theta
TpCA_ Tp1093_B8BPY6	XP_002286051	Thalassiosira pseudonana	Theta	Theta
TwCA_zeta	Q50LE4	Thalassiosira weissflogii	Zeta	Zeta

Supplementary Table 3. *T. pseudonana* Shell1-7 with NCIB Accession, Gene name and primary amino acid sequence

Shell name	Genbank ID	Accession	Primary sequence
Shell1	THAPSDRAFT_8449	XP_002292853	MKTAAALVLGAYCTTTDAFTTSSTNRLTTALLSSNYGDDPRGPP RPMPNGPPPPLPTPRANYADPPSIDLDGQLQDRQSNLWHRRTPD YQSINEDPRQFDMQRRFSQRPPGGMDDPSMRPRQSWWESHPGDS SRVQGGSRATFNAPYDREGSHVFLETDGRPLDTEIELWDGPNNT PTKLKVYSEDGRMRPINAMFENPQKGFRGNTFSVRNTGAMEFPV NAGVRSMGGGMPGGLAGGEGAYLREGDIQGYSRPIPSTKRGETV QGGALRTFPLDYSVEAVQITLTSDGMPMEGKIELWGTSSHTKQV AEIYADNGATRPFAAIIDTPGGSNTIAVYNEGPMEYPIRCVV EPIARMEGWNGEEEKFGGSLAPW
Shell2	THAPSDRAFT_23918	XP_002292183	MKFFAAVALVAINGASAFVAPNANRAASMLSMSDYDVSTSVRED VGVGGGGSPLYKAGGTARDLAEIWDNSSPVIVQGGSLRTWSFAN PAIESVQVLLKTEGRPLDADVELWQGPDNTPHKMRVYVEDGALR TFNAVIGTPRGPNTVAIRNIGQLEFPLDAVVRPDRDDGLAAGIA SVATRSETIQGGALRTYPFNPTVDSVAIILKTDGRPLNARIELL QGPNNNKQVVELYTEDGLDRPFFAIVETPGSGNVVRVVNTAPVE FPLYASVDAYRVGGGMDWPDDGLMIGRSF
Shell3	THAPSDRAFT_7883	XP_002292184	MKNTAALTLAALSGASAFTSPSNGPMISSTELASTFVRDRANVG FDYDNRNAGGAIMRPRRSVGGGRSLDEIWMNSAPVIVQGGSLRT WSFANPSIDAVQVLLKTEGRPLDADVELWQGPDNTPHKMRVYLE DGALRTFNAVIGTPRGPNTVSVRNIGQLEFPLDAVVRPDREDGL AVSIASVTSRSETIQGGALRTYPFNAVVDSVAVILKTDGRPLNA RIELLQGPNNNKQVIELYTEDGLDRPFFAIIETPGAGNVVRIMN TAPLEFPLYASVDAYEMGEYGSWNDEGYQLGAFPRL
Shell4	THAPSDRAFT_24512	XP_002293045	MKFVAGLVALFAASANGFSVNTPVTSRVSTTSLQSTAWDIASIG RSVRIEGQSREHWEFPDPNQEIVQVCIDSNGRPVESLIELWIGP DWTPMKVKAYTEDGMMRPIQSLIGTRNKAAAIDINNTGPGDFPL NAAAAYASSQMAAQRADLLTANTGRYIEGGAVHSVNFDASVDSV QVLLYTDTRQLNAQVELLNGPNNFKQKYEVFTNNGMLNSLYVVF NTPGSGNVVRVTNLAPLEFPCKAFIASA
Shell5	THAPSDRAFT_3883	XP_002289037	MKFASSLALMMAIGSSAFQAPQLTSRPSTELYRAVNGGLPTFNP SSRDRAQNEYRADTPTSQFPYVCSGDKSSELAADSTAFPKNLNL QMIQGQNARLTYKMPEGADRVQMFINSNGRPLKARVELWCGPIR RTHFMDIDCMNGAETPFRATLKFKNVDSAGPQVLQINTKEDSAF PAMVGVDIMSPERSAERQEKFDAIWKSSTKIYSQGDKTIRSIPI ADNCKSVQLLVWSKDVGKKSFKVNIELLQGPNSKRQYYELQCGG GSQPYHAVFETPGNGWTMRVTNTKTLHDGSHEFVLVPYEVDGDM SSSVSANTGAISPFDGGYSPNSTHGGNYGKSLGPHGSRESFRNS PYGSGGRGQAIGGNW
Shell6	THAPSDRAFT_8449	XP_002292853	MKTAAALVLGAYCTTTDAFTTSSTNRLTTALLSSNYGDDPRGPP RPMPNGPPPPLPTPRANYADPPSIDLDGQLQDRQSNLWHRRTPD YQSINEDPRQFDMQRRFSQRPPGGMDDPSMRPRQSWWESHPGDS SRVQGGSRATFNAPYDREGSHVFLETDGRPLDTEIELWDGPNNT PTKLKVYSEDGRMRPINAMFENPQKGFRGNTFSVRNTGAMEFPV NAGVRSMGGGMPGGLAGGEGAYLREGDIQGYSRPIPSTKRGETV QGGALRTFPLDYSVEAVQITLTSDGMPMEGKIELWGTSSHTKQV AEIYADNGATRPFAAIIDTPGGSNTIAVYNEGPMEYPIRCVVEP IARMEGWNGEEEKFGGSLAPW

Shell7	THAPSDRAFT_bd1762	XP_002297261	MPSQPPRRQSLATTLVVLASIIPSAVSFSPSSFRSGAVATPLHTTASRAQLSTTSLAASSNERNKDSVNINYSNNKPILSTLAALTILTTTLTTSFQTANAYEESDYASETVTTVVQQLKENAGDVDKTFGTLEEIAKIITEGKGVGGSLSYDGIRLTEGYVADEDTTIYNPGLSLLTNSEKERLVSAIISNRKTGLSTNHWSENNEYAFDFLKTKLDPLHMYELEGYLSILPYYGAVLYLGALFVQKNFVTMGSTDRLPPPTFHHNAKPFASPTLTMVSQLQTYWLLSSFLFGSADSFSYNVRANQSATQVRSSPINEDIFTISPPVRIQGSSLKTWSYEPGPSKRIQVSIKSLGRPIEASVELWQTPTYIPTKFTVECEDASENIVHSVFEVPENTPVTIAIYNTENVQFPLEVSVSDTGLESAIDSFEGEQSEHIQGGRIKSFTFGEEIESVEVLLVTKHRNLKAMLEILQGPNDDNEIIEVETEDGRVHPFYTVIQTPGGANTLRVVNRSPVEFPFEAFVRPFVTVEDGNTQYNRGGPYF

Supplementary Table 4. Tom 14 hits from blast P of Shell 7 full amino acid sequence. Species key: diatom's pink, haptophytes , coccolithophores cyan, alveolate gold, opisthokonta umber. _*Collated shell csv files_Shell1-7,* sheet 220607_bd1762_ Shell7_all

Species	Accession	Found in Shell1-6 Blastp	% identity	alignment length	mismatches	gap opens	q.start,	q.end	s.start	s.end	e value	bit score	% positives
Thalassiosira pseudonana	XP_002290372.1	x	84.058	207	7	11	1	181	610	816	7.17E-89	288	84.06
Fragilariopsis cylindrus	OEU14939.1	x	45.868	242	69	11	1	180	2	243	1.23E-44	159	56.2
Chrysochromulina tobinii	KOO34179.1	x	44.49	245	70	10	2	180	177	421	5.88E-40	152	53.06
Thalassiosira oceanica	EJK63886.1	x	55.814	172	45	10	28	176	5	168	8.26E-35	132	65.7
Emiliania huxleyi	XP_005767240.1	x	44.24	217	80	9	2	179	220	434	3.21E-33	134	57.6
Chrysochromulina tobinii	KOO22147.1	x	39.08	174	89	7	24	180	72	245	3.98E-23	103	55.17
Chaetoceros tenuissimus	GFH58643.1	x	34.043	188	86	9	29	180	149	334	1.86E-13	78.6	47.34
Fistulifera solaris	GAX13418.1	x	36.416	173	86	7	28	178	126	296	1.04E-12	76.3	49.71
Fistulifera solaris	GAX28807.1	x	34.911	169	90	6	28	178	84	250	7.48E-11	70.9	48.52
Pseudocohnilembus persal	KRX11000.1	x	34.783	161	96	5	29	180	100	260	7.55E-11	70.9	50.93
Gimesia alba	WP_145218702.1	x	32.911	158	93	7	29	174	67	223	1.62E-10	69.3	48.73
Salpingoeca rosetta	XP_004989761.1	x	29.798	198	118	9	4	181	54	250	2.31E-10	69.3	45.45
Emiliania huxleyi	XP_005769956.1	x	31.72	186	116	6	5	180	80	264	2.88E-10	69.3	48.39
Emiliania huxleyi	XP_005776873.1	x	31.818	176	111	5	14	180	101	276	5.12E-10	69.3	49.43

Supplementary Table 5. Shell protein structural alignments and corresponding RMSD values

Alignment number	Shell numbers aligned	RMSD	Command line
1	2 to 1	0.208 (169 to 169 atoms)	Executive: object "aln_Shell2_betafold_to_Shell1_betafold" created.
2	3 to 1	0.194 (175 to 175 atoms)	Executive: object "aln_Shell3_betafold_to_Shell1_betafold" created.
3	3 to 2	0.172 (173 to 173 atoms)	Executive: object "aln_Shell3_betafold_to_Shell2_betafold" created.
4	4 to 1	0.583 (158 to 158 atoms)	Executive: object "aln_Shell4_betafold_to_Shell1_betafold" created
5	4 to 2	0.625 (160 to 160 atoms)	Executive: object "aln_Shell4_betafold_to_Shell2_betafold" created.
6	4 to 3	0.620 (159 to 159 atoms)	Executive: object "aln_Shell4_betafold_to_Shell3_betafold" created.
7	5 to 1	0.893 (131 to 131 atoms)	Executive: object "aln_Shell5_betafold_to_Shell1_betafold" created
8	6 to 1	0.413 (155 to 155 atoms)	Executive: object "aln_Shell6_betafold_to_Shell1_betafold" created.
9	7 to 1	0.627 (165 to 165 atoms)	Executive: object "aln_Shell7_betafold_to_Shell1_betafold" created.

Supplementary Table 6. Shell1-7 beta strand sequences

	Strand 1A	Strand 2A	Strand 6A	Strand 7A	Strand 13B	Strand 14B	Strand 9B	Strand 8B
Shell1:	GSLRTWSF	LDAVVRP	FNAVIG	HKMRV	FFAIVE	NNKQVVELTYT	LYASVDA	GALRTY
Shell2:	GSLRTWSF	LDAVVRP	FNAVIG	HKMRV	FFAIVE	NNKQVVELTYT	LYASVDA	GALRTY
Shell3:	GSLRTWSF	LDAVVRP	FNAVIG	HKMRV	FFAIIE	NNKQVVELTYT	LYASVDA	GALRTY
Shell4:	QSREHWEF	LNAAAAY	IQSLIG	MKVKA	LYVVFN	NFKQKYEVFT	VKAFIAS	GAVHSV
Shell5:	QNARLTYK	AMVGVDI	FRATLK	HFMDI	YHAVFE	SKRQYYELQC	SHEFVGA	DKTIRSI
Shell6:	GSRATFN-	VNAGVRS	INAMFE	TKLKV	FAAIID	HTKQVAEIYA	IRCVVEP	GALRTF
Shell7:	SSLKTWSY	LEVSVSD	VHSVFE	TKFTV	FYTVIQ	DDNEIIEVET	FEAFVRP	GRIKSF

References

1. Mackinder LCM, Chen C, Leib RD, Patena W, Blum SR, Rodman M, et al. A Spatial Interactome Reveals the Protein Organization of the Algal CO2-Concentrating Mechanism. Cell [Internet]. 2017 Sep 21 [cited 2022 Aug 15];171(1):133-147.e14. Available from: http://www.cell.com/article/S0092867417310024/fulltext

2. Reinfelder JR. Carbon Concentrating Mechanisms in Eukaryotic Marine Phytoplankton. http://dx.doi.org/101146/annurev-marine-120709-142720 [Internet]. 2010 Dec 15 [cited 2022 Aug 15];3:291–315. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev-marine-120709-142720

3. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science (1979) [Internet]. 1998 Jul 10 [cited 2022 Aug 15];281(5374):237–40. Available from: https://www.science.org/doi/10.1126/science.281.5374.237

4. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science [Internet]. 1998 Jul 1 [cited 2023 Feb 24];281(5374):237–40. Available from: https://escholarship.org/uc/item/9gm7074q

5. Whitney SM, Houtz RL, Alonso H. Advancing Our Understanding and Capacity to Engineer Nature's CO2-Sequestering Enzyme, Rubisco. Plant Physiol [Internet]. 2011 Jan 3 [cited 2023 Feb 24];155(1):27–35. Available from: https://academic.oup.com/plphys/article/155/1/27/6111403

6. Barrett J, Girr P, Mackinder LCM. Pyrenoids: CO2-fixing phase separated liquid organelles. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2021 Apr 1;1868(5):118949.

7. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, et al. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011 May 31;50(21):4402–10.

8. Clapero V, Arrivault S, Stitt M. Natural variation in metabolism of the Calvin-Benson cycle. Semin Cell Dev Biol [Internet]. 2023 Mar 21 [cited 2023 Mar 29]; Available from: https://linkinghub.elsevier.com/retrieve/pii/S1084952123000587

9. Clement R, Dimnet L, Maberly SC, Gontero B. The nature of the CO2-concentrating mechanisms in a marine diatom, Thalassiosira pseudonana. New Phytologist [Internet]. 2016 Mar 1 [cited 2023 Mar 29];209(4):1417–27. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/nph.13728

10. Meyer MT, Whittaker C, Griffiths H. The algal pyrenoid: Key unanswered questions. J Exp Bot. 2017 Jun 22;68(14):3739–49.

11. Barrett J, Girr P, Mackinder LCM. Pyrenoids: CO2-fixing phase separated liquid organelles. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2021 Apr 1;1868(5):118949.

12. Wang Y, Stessman DJ, Spalding MH. The CO2 concentrating mechanism and photosynthetic carbon assimilation in limiting CO2: How Chlamydomonas works against the gradient. Plant Journal. 2015 May 1;82(3):429–48.

13. Kerfeld CA, Melnicki MR. Assembly, function and evolution of cyanobacterial carboxysomes. Curr Opin Plant Biol. 2016 Jun 1;31:66–75.

14. Melnicki MR, Sutter M, Kerfeld CA. Evolutionary relationships among shell proteins of carboxysomes and metabolosomes. Curr Opin Microbiol. 2021 Oct 1;63:1–9.

15. MacCready JS, Basalla JL, Vecchiarelli AG, Echave J. Origin and Evolution of Carboxysome Positioning Systems in Cyanobacteria. Mol Biol Evol. 2020 May 1;37(5):1434–51.

16. Rae BD, Long BM, Badger MR, Price GD. Functions, Compositions, and Evolution of the Two Types of Carboxysomes: Polyhedral Microcompartments That Facilitate CO 2 Fixation in Cyanobacteria and Some Proteobacteria . Microbiology and Molecular Biology Reviews [Internet]. 2013 Sep [cited 2023 Mar 11];77(3):357–79. Available from: https://journals.asm.org/journal/mmbr

17. Kaplan A, Reinhold L. CO2 CONCENTRATING MECHANISMS IN PHOTOSYNTHETIC MICROORGANISMS. https://doi.org/101146/annurev.arplant501539 [Internet]. 2003 Nov 28 [cited 2023 Mar 29];50:539–70. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev.arplant.50.1.539

18. Meyer MT, Goudet MMM, Griffiths H. The Algal Pyrenoid. 2020;179–203.

19. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science [Internet]. 1998 Jul 10 [cited 2023 Mar 29];281(5374):237–40. Available from: https://pubmed.ncbi.nlm.nih.gov/9657713/

20. Benoiston AS, Ibarbalz FM, Bittner L, Guidi L, Jahn O, Dutkiewicz S, et al. The evolution of diatoms and their biogeochemical functions. Philos Trans R Soc Lond B Biol Sci [Internet]. 2017 Sep 5 [cited 2023 Mar 15];372(1728). Available from: https://pubmed.ncbi.nlm.nih.gov/28717023/

21. Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles [Internet]. 1995 Sep 1 [cited 2022 Aug 15];9(3):359–72. Available from: https://onlinelibrary.wiley.com/doi/full/10.1029/95GB01070

22. Tréguer PJ, de La Rocha CL. The World Ocean Silica Cycle. http://dx.doi.org/101146/annurev-marine-121211-172346 [Internet]. 2013 Jan 2 [cited 2022 Aug 15];5:477–501. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev-marine-121211-172346

23. Wang Y, Liu S, Wang J, Yao Y, Chen Y, Xu Q, et al. Diatom Biodiversity and Speciation Revealed by Comparative Analysis of Mitochondrial Genomes. Front Plant Sci. 2022 Mar 24;13:680.

24. Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, et al. The evolution of modern eukaryotic phytoplankton. Science (1979) [Internet]. 2004 Jul 16 [cited 2023 Mar 15];305(5682):354–60. Available from: https://www.science.org/doi/10.1126/science.1095964

25. Sibbald SJ, Archibald JM. Genomic Insights into Plastid Evolution. Genome Biol Evol [Internet]. 2020 Jul 1 [cited 2022 Aug 15];12(7):978–90. Available from: https://academic.oup.com/gbe/article/12/7/978/5836826

26. Morozov AA, Galachyants YP. Diatom genes originating from red and green algae: Implications for the secondary endosymbiosis models. Mar Genomics. 2019 Jun 1;45:72–8.

27. Morozov AA, Galachyants YP. Diatom genes originating from red and green algae: Implications for the secondary endosymbiosis models. Mar Genomics. 2019 Jun 1;45:72–8.

28. Sibbald SJ, Archibald JM. Genomic Insights into Plastid Evolution. Genome Biol Evol [Internet]. 2020 Jul 1 [cited 2023 Mar 29];12(7):978–90. Available from: https://academic.oup.com/gbe/article/12/7/978/5836826

29. Keeling PJ. The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. http://dx.doi.org/101146/annurev-arplant-050312-120144 [Internet]. 2013 May 2 [cited 2022 Aug 15];64:583–607. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev-arplant-050312-120144

30. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. The origin of plastids. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. 2008 Aug 27 [cited 2022 Aug 15];363(1504):2675–85. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2008.0050

31. Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science (1979) [Internet]. 2009 Jun 26 [cited 2022 Aug 15];324(5935):1724–6. Available from: https://www.science.org/doi/10.1126/science.1172983

32. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, et al. The genome of the diatom Thalassiosira Pseudonana: Ecology, evolution, and metabolism. Science (1979) [Internet]. 2004 Oct 1 [cited 2023 Mar 15];306(5693):79–86. Available from: https://www.science.org/doi/10.1126/science.1101156

33. Hopes A, Nekrasov V, Belshaw N, Grouneva I, Kamoun S, Mock T. Genome Editing in Diatoms Using CRISPR-Cas to Induce Precise Bi-allelic Deletions. Bio Protoc [Internet]. 2017 [cited 2023 Mar 15];7(23). Available from: https://pubmed.ncbi.nlm.nih.gov/34595293/

34. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 2008 456:7219 [Internet]. 2008 Oct 15 [cited 2023 Mar 20];456(7219):239–44. Available from: https://www.nature.com/articles/nature07410

35. Tsuji Y, Nakajima K, Matsuda Y. Molecular aspects of the biophysical CO2-concentrating mechanism and its regulation in marine diatoms. J Exp Bot [Internet]. 2017 Sep 8 [cited 2022 Aug 15];68(14):3763–72. Available from: https://academic.oup.com/jxb/article/68/14/3763/3867358

36. Nakajima K, Tanaka A, Matsuda Y. SLC4 family transporters in a marine diatom directly pump bicarbonate from seawater. Proc Natl Acad Sci U S A [Internet]. 2013 Jan 29 [cited 2022 Aug 15];110(5):1767–72. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1216234110

37. Hopkinson BM, Dupont CL, Matsuda Y. The physiology and genetics of CO2 concentrating mechanisms in model diatoms. Curr Opin Plant Biol. 2016 Jun 1;31:51–7.

38. Wang Y, Duanmu D, Spalding MH. Carbon dioxide concentrating mechanism in Chlamydomonas reinhardtii: inorganic carbon transport and CO2 recapture. Photosynth Res [Internet]. 2011 Sep [cited 2023 Mar 30];109(1–3):115–22. Available from: https://pubmed.ncbi.nlm.nih.gov/21409558/

39. Kono A, Chou TH, Radhakrishnan A, Bolla JR, Sankar K, Shome S, et al. Structure and function of LCI1: a plasma membrane CO2 channel in the Chlamydomonas CO2 concentrating mechanism. The Plant Journal [Internet]. 2020 Jun 1 [cited 2023 Mar 30];102(6):1107–26. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.14745

40. Mukherjee A, Lau CS, Walker CE, Rai AK, Prejean CI, Yates G, et al. Thylakoid localized bestrophin-like proteins are essential for the CO2 concentrating mechanism of Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A [Internet]. 2019 Aug 20 [cited 2023 Mar 15];116(34):16915–20. Available from: https://pubmed.ncbi.nlm.nih.gov/31391312/

41. Duanmu D, Wang Y, Spalding MH. Thylakoid lumen carbonic anhydrase (CAH3) mutation suppresses air-dier phenotype of LCIB mutant in Chlamydomonas reinhardtii. Plant Physiol. 2009;149(2):929–37.

42. Tsuji Y, Mahardika A, Matsuda Y, Griffiths H. Evolutionarily distinct strategies for the acquisition of inorganic carbon from seawater in marine diatoms. J Exp Bot [Internet]. 2017 [cited 2023 Mar 1];68(14):3949–58. Available from: http://jxb.oxfordjournals.org/open_access.html

43. Tachibana M, Allen AE, Kikutani S, Endo Y, Bowler C, Matsuda Y. Localization of putative carbonic anhydrases in two marine diatoms, Phaeodactylum tricornutum and Thalassiosira pseudonana. Photosynth Res. 2011 Sep;109(1–3):205–21.

44. Kikutani S, Nakajima K, Nagasato C, Tsuji Y, Miyatake A, Matsuda Y. Thylakoid luminal Θ-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. Proc Natl Acad Sci U S A. 2016 Aug 30;113(35):9828–33.

45. Matsuda Y, Hopkinson BM, Nakajima K, Dupont CL, Tsuji Y. Mechanisms of carbon dioxide acquisition and CO2 sensing in marine diatoms: a gateway to carbon metabolism. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. 2017 Sep 5 [cited 2023 Mar 2];372(1728). Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0403

46. Nawaly H, Tanaka A, Toyoshima Y, Tsuji Y, Matsuda Y. Localization and characterization θ carbonic anhydrases in Thalassiosira pseudonana. Photosynth Res [Internet]. 2023 Mar 2 [cited 2023 Mar 7]; Available from: https://pubmed.ncbi.nlm.nih.gov/36862281/

47. Kikutani S, Tanaka R, Yamazaki Y, Hara S, Hisabori T, Kroth PG, et al. Redox regulation of carbonic anhydrases via thioredoxin in chloroplast of the marine diatom Phaeodactylum tricornutum. Journal of Biological Chemistry. 2012 Jun 8;287(24):20689–700.

48. Harada H, Nakajima K, Sakaue K, Matsuda Y. CO2 sensing at ocean surface mediated by cAMP in a marine diatom. Plant Physiol. 2006;142(3):1318–28.

49. Hennon GMM, Ashworth J, Groussman RD, Berthiaume C, Morales RL, Baliga NS, et al. Diatom acclimation to elevated CO 2 via cAMP signalling and coordinated gene expression. Nat Clim Chang. 2015 Jul 24;5(8):761–5.

50. Ohno N, Inoue T, Yamashiki R, Nakajima K, Kitahara Y, Ishibashi M, et al. CO2-cAMP-responsive cis-elements targeted by a transcription factor with CREB/ATF-like basic zipper domain in the marine diatom Phaeodactylum tricornutum. Plant Physiol. 2012;158(1):499–513.

51. Samukawa M, Shen C, Hopkinson BM, Matsuda Y. Localization of putative carbonic anhydrases in the marine diatom, Thalassiosira pseudonana. Photosynth Res. 2014;121(2–3):235–49.

52. Mackinder LCM, Meyer MT, Mettler-Altmann T, Chen VK, Mitchell MC, Caspari O, et al. A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. Proc Natl Acad Sci U S A. 2016 May 24;113(21):5958–63.

53. Badger MR, Andrews TJ, Whitney SM, Ludwig M, Yellowlees DC, Leggat W, et al. The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae. Canadian Journal of Botany [Internet]. 1998 [cited 2023 Mar 11];76(6):1052–71. Available from: https://research-repository.uwa.edu.au/en/publications/the-diversity-and-coevolution-of-rubisco-plastids-pyrenoids-and-c

54. Engel BD, Schaffer M, Cuellar LK, Villa E, Plitzko JM, Baumeister W. Native architecture of the chlamydomonas chloroplast revealed by in situ cryo-electron tomography. Elife. 2015 Jan 13;2015(4).

55. He S, Chou HT, Matthies D, Wunder T, Meyer MT, Atkinson N, et al. The structural basis of Rubisco phase separation in the pyrenoid. Nature Plants 2020 6:12 [Internet]. 2020 Nov 23 [cited 2023 Mar 11];6(12):1480–90. Available from: https://www.nature.com/articles/s41477-020-00811-y

56. Alberti S. Phase separation in biology. Current Biology [Internet]. 2017 Oct 23 [cited 2022 Aug 15];27(20):R1097–102. Available from: http://www.cell.com/article/S0960982217311090/fulltext

57. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: Organizers of cellular biochemistry. Nat Rev Mol Cell Biol. 2017 May 1;18(5):285–98.

58. Wunder T, Cheng SLH, Lai SK, Li HY, Mueller-Cajar O. The phase separation underlying the pyrenoid-based microalgal Rubisco supercharger. Nature Communications 2018 9:1 [Internet]. 2018 Nov 29 [cited 2023 Mar 11];9(1):1–10. Available from: https://www.nature.com/articles/s41467-018-07624-w

59. He S, Chou HT, Matthies D, Wunder T, Meyer MT, Atkinson N, et al. The structural basis of Rubisco phase separation in the pyrenoid. Nature Plants 2020 6:12 [Internet]. 2020 Nov 23 [cited 2023 Mar 14];6(12):1480–90. Available from: https://www.nature.com/articles/s41477-020-00811-y

60. Meyer MT, Itakura AK, Patena W, Wang L, He S, Emrich-Mills T, et al. Assembly of the algal CO2-fixing organelle, the pyrenoid, is guided by a Rubisco-binding motif. Sci Adv [Internet]. 2020 Nov 11 [cited 2023 Mar 14];6(46). Available from: https://www.science.org/doi/10.1126/sciadv.abd2408

61. Villarejo A, Martínez F, Plumed MDP, Ramazanov Z. The induction of the CO2 concentrating mechanism in a starch-less mutant of Chlamydomonas reinhardtii. Physiol Plant. 1996;98(4):798–802.

62. Mackinder LCM, Chen C, Leib RD, Patena W, Blum SR, Rodman M, et al. A Spatial Interactome Reveals the Protein Organization of the Algal CO2-Concentrating Mechanism. Cell. 2017 Sep 21;171(1):133-147.e14.

63. Kuchitsu K, Tsuzuki M, Miyachi S. Changes of Starch Localization within the Chloroplast Induced by Changes in CO2 Concentration during Growth of Chlamydomonas reinhardtii: Independent Regulation of Pyrenoid Starch and Stroma Starch. Plant Cell Physiol [Internet]. 1988 Dec 1 [cited 2023 Mar 3];29(8):1269–78. Available from: https://academic.oup.com/pcp/article/29/8/1269/1828285

64. Yamano T, Tsujikawa T, Hatano K, Ozawa SI, Takahashi Y, Fukuzawa H. Light and low-CO2-dependent LCIB-LCIC complex localization in the chloroplast supports the carbon-concentrating mechanism in Chlamydomonas reinhardtii. Plant Cell Physiol [Internet]. 2010 Sep [cited 2023 Mar 8];51(9):1453–68. Available from: https://pubmed.ncbi.nlm.nih.gov/20660228/

65. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res [Internet]. 2022 Jan 7 [cited 2023 Mar 25];50(D1):D20–6. Available from: https://pubmed.ncbi.nlm.nih.gov/34850941/

66. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. Proteins [Internet]. 2000 Nov 1 [cited 2023 Mar 27];41(2):224–37. Available from: http://europepmc.org/article/MED/10966575

67. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Research 2009 19:8 [Internet]. 2009 Jul 14 [cited 2023 Mar 27];19(8):929–49. Available from: https://www.nature.com/articles/cr200987

68. Lamiable A, Thevenet P, Rey J, Vavrusa M, Derreumaux P, Tuffery P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Res [Internet]. 2016 Jul 8 [cited 2023 Mar 27];44(W1):W449–54. Available from: https://pubmed.ncbi.nlm.nih.gov/27131374/

69. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021 [cited 2023 Mar 18];596:583. Available from: https://doi.org/10.1038/s41586-021-03819-2

70. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc [Internet]. 2015 Jun 30 [cited 2023 Mar 27];10(6):845–58. Available from: https://pubmed.ncbi.nlm.nih.gov/25950237/

71. Supuran CT. Structure and function of carbonic anhydrases. Biochem J [Internet]. 2016 [cited 2023 Mar 29];473(14):2023–32. Available from: https://pubmed.ncbi.nlm.nih.gov/27407171/

72. Tsuji Y, Nakajima K, Matsuda Y. Molecular aspects of the biophysical CO2-concentrating mechanism and its regulation in marine diatoms. J Exp Bot [Internet]. 2017 Sep 8 [cited 2023 Mar 29];68(14):3763–72. Available from: https://academic.oup.com/jxb/article/68/14/3763/3867358

73. Jensen EL, Clement R, Kosta A, Maberly SC, Gontero B. A new widespread subclass of carbonic anhydrase in marine phytoplankton. ISME Journal. 2019 Aug 1;13(8):2094–106.

74. DiMario RJ, Machingura MC, Waldrop GL, Moroney J V. The many types of carbonic anhydrases in photosynthetic organisms. Plant Sci [Internet]. 2018 Mar 1 [cited 2023 Mar 29];268:11–7. Available from: https://pubmed.ncbi.nlm.nih.gov/29362079/

75. Hewett-Emmett D, Tashian RE. Functional Diversity, Conservation, and Convergence in the Evolution of the α-, β-, and γ-Carbonic Anhydrase Gene Families. Mol Phylogenet Evol. 1996 Feb 1;5(1):50–77.

76. Yamano T, Toyokawa C, Shimamura D, Matsuoka T, Fukuzawa H. CO2-dependent migration and relocation of LCIB, a pyrenoid-peripheral protein in Chlamydomonas reinhardtii. Plant Physiol [Internet]. 2022 Feb 1 [cited 2023 Mar 29];188(2):1081–4. Available from: https://pubmed.ncbi.nlm.nih.gov/34791500/

77. Jin S, Sun J, Wunder T, Tang D, Cousins AB, Sze SK, et al. Structural insights into the LCIB protein family reveals a new group of β-carbonic anhydrases. Proc Natl Acad Sci U S A. 2016 Dec 20;113(51):14716–21.

78. Matsuda Y, Hopkinson BM, Nakajima K, Dupont CL, Tsuji Y. Mechanisms of carbon dioxide acquisition and CO2 sensing in marine diatoms: a gateway to carbon metabolism. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. 2017 Sep 5 [cited 2023 Mar 29];372(1728). Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0403

79. Hopkinson BM, Dupont CL, Matsuda Y. The physiology and genetics of CO2 concentrating mechanisms in model diatoms. Curr Opin Plant Biol. 2016 Jun 1;31:51–7.

80. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res [Internet]. 2021 Jan 8 [cited 2023 Mar 29];49(D1):D480–9. Available from: https://academic.oup.com/nar/article/49/D1/D480/6006196

81. Del Prete S, Vullo D, Fisher GM, Andrews KT, Poulsen SA, Capasso C, et al. Discovery of a new family of carbonic anhydrases in the malaria pathogen Plasmodium falciparum—The η-carbonic anhydrases. Bioorg Med Chem Lett. 2014 Sep 15;24(18):4389–96.

82. Clement R, Lignon S, Mansuelle P, Jensen E, Pophillat M, Lebrun R, et al. Responses of the marine diatom Thalassiosira pseudonana to changes in CO_2 concentration: a proteomic approach. Sci Rep [Internet]. 2017 Feb 9 [cited 2023 Mar 6];7:42333–42333. Available from: https://europepmc.org/articles/PMC5299434

83. Pierella Karlusich JJ, Bowler C, Biswas H. Carbon Dioxide Concentration Mechanisms in Natural Populations of Marine Diatoms: Insights From Tara Oceans. Front Plant Sci. 2021 Apr 30;12:659.

84. Toyokawa C, Yamano T, Fukuzawa H. Pyrenoid starch sheath is required for LCIB localization and the CO2-concentrating mechanism in green algae. Plant Physiol. 2020 Apr 1;182(4):1883–93.

85. Itakura AK, Chan KX, Atkinson N, Pallesen L, Wang L, Reeves G, et al. A Rubisco-binding protein is required for normal pyrenoid number and starch sheath morphology in Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A. 2019 Sep 10;116(37):18445–54.

86. Jisna VA, Jayaraj PB. Protein Structure Prediction: Conventional and Deep Learning Perspectives. Protein Journal [Internet]. 2021 Aug 1 [cited 2023 Mar 26];40(4):522–44. Available from: https://link.springer.com/article/10.1007/s10930-021-10003-y

87. Berman HM. Synergies between the Protein Data Bank and the community. Nature Structural & Molecular Biology 2021 28:5 [Internet]. 2021 May 7 [cited 2023 Mar 18];28(5):400–1. Available from: https://www.nature.com/articles/s41594-021-00586-6

88. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics [Internet]. 2019 Sep 14 [cited 2023 Mar 18];20(1):1–15. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3019-7

89. Downard KM. Protein phylogenetics with mass spectrometry. A comparison of methods. Analytical Methods [Internet]. 2021 Apr 1 [cited 2023 Mar 18];13(12):1442–54. Available from: https://pubs.rsc.org/en/content/articlehtml/2021/ay/d1ay00153a

90. Harms MJ, Thornton JW. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet [Internet]. 2013 Aug [cited 2023 Mar 18];14(8):559–71. Available from: https://pubmed.ncbi.nlm.nih.gov/23864121/

91. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface [Internet]. 2014 Nov 6 [cited 2023 Mar 18];11(100). Available from: https://royalsocietypublishing.org/doi/10.1098/rsif.2014.0419

92. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics [Internet]. 2009 May 1 [cited 2023 Mar 25];25(9):1189–91. Available from: https://academic.oup.com/bioinformatics/article/25/9/1189/203460

93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990 [cited 2023 Mar 25];215(3):403–10. Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

94. Maeda Y, Kobayashi R, Watanabe K, Yoshino T, Bowler C, Matsumoto M, et al. Chromosome-Scale Genome Assembly of the Marine Oleaginous Diatom Fistulifera solaris. Marine Biotechnology [Internet]. 2022 Aug 1 [cited 2023 Mar 27];24(4):788–800. Available from: https://link.springer.com/article/10.1007/s10126-022-10147-7

95. Garrido JL, Brunet C, Rodríguez F. Pigment variations in Emiliania huxleyi (CCMP370) as a response to changes in light intensity or quality. Environ Microbiol [Internet]. 2016 Dec 1 [cited 2023 Mar 27];18(12):4412–25. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.13373

96. Deodato CR, Barlow SB, Hovde BT, Cattolico RA. Naked Chrysochromulina (Haptophyta) isolates from lake and river ecosystems: An electron microscopic comparison including new observations on the type species of this taxon. Algal Res. 2019 Jun 1;40:101492.

97. Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, et al. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature [Internet]. 2017 Jan 26 [cited 2023 Mar 27];541(7638):536–40. Available from: https://pubmed.ncbi.nlm.nih.gov/28092920/

98. Basu S, Patil S, Mapleson D, Russo MT, Vitale L, Fevola C, et al. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. New Phytologist [Internet]. 2017 Jul 1 [cited 2023 Mar 27];215(1):140–56. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/nph.14557

99. Hongo Y, Kimura K, Takaki Y, Yoshida Y, Baba S, Kobayashi G, et al. The genome of the diatom Chaetoceros tenuissimus carries an ancient integrated fragment of an extant virus. Scientific Reports 2021 11:1 [Internet]. 2021 Nov 24 [cited 2023 Mar 27];11(1):1–13. Available from: https://www.nature.com/articles/s41598-021-00565-3

100. Sieburth JMcN, Johnson PW. Picoplankton Ultrastructure: A Decade of Preparation for the Brown Tide Alga, Aureococcus Anophagefferen. 2012 Jul 24 [cited 2023 Mar 26];1–21. Available from: https://onlinelibrary.wiley.com/doi/full/10.1029/CE035p0001

101. Round F, Crawford R, Mann D. Diatoms: biology and morphology of the genera [Internet]. 1990 [cited 2023 Mar 24]. Available from: https://books.google.com/books?hl=en&lr=&id=xhLJvNa3hw0C&oi=fnd&pg=PP13&ots=qRkiSgV60p&sig=BXSzmLS3qRatdC01QJh5w3-SMiU

102. Haunost M, Riebesell U, Bach LT. The Calcium Carbonate Shell of Emiliania huxleyi Provides Limited Protection Against Viral Infection. Front Mar Sci. 2020 Sep 11;7:735.

103. Thomas A, Meurisse R, Charloteaux B, Brasseur R. Aromatic side-chain interactions in proteins. I. Main structural features. Proteins [Internet]. 2002 Sep 1 [cited 2023 Mar 22];48(4):628–34. Available from: https://pubmed.ncbi.nlm.nih.gov/12211030/

104. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins [Internet]. 2009 Nov 15 [cited 2023 Mar 22];77(3):499–508. Available from: https://pubmed.ncbi.nlm.nih.gov/19507241/

105. Malik AJ, Poole AM, Allison JR. Structural Phylogenetics with Confidence. Mol Biol Evol [Internet]. 2020 Sep 1 [cited 2023 Mar 25];37(9):2711–26. Available from: https://pubmed.ncbi.nlm.nih.gov/32302382/

106. Lundin D, Poole AM, Sjöberg BM, Högbom M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. J Biol Chem [Internet]. 2012 Jun 8 [cited 2023 Mar 25];287(24):20565–75. Available from: https://pubmed.ncbi.nlm.nih.gov/22535960/

107. Mitra M, Lato SM, Ynalvez RA, Xiao Y, Moroney J V. Identification of a New Chloroplast Carbonic Anhydrase in Chlamydomonas reinhardtii. Plant Physiol [Internet]. 2004 May 1 [cited 2023 Mar 31];135(1):173–82. Available from: https://academic.oup.com/plphys/article/135/1/173/6111960

108. Duanmu D, Wang Y, Spalding MH. Thylakoid Lumen Carbonic Anhydrase (CAH3) Mutation Suppresses Air-Dier Phenotype of LCIB Mutant in Chlamydomonas reinhardtii. Plant Physiol [Internet]. 2009 Feb 6 [cited 2023 Mar 31];149(2):929–37. Available from: https://academic.oup.com/plphys/article/149/2/929/6108026

109. Burki F, Okamoto N, Pombert JF, Keeling PJ. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc Biol Sci [Internet]. 2012 Jun 7 [cited 2023 Mar 31];279(1736):2246–54. Available from: https://pubmed.ncbi.nlm.nih.gov/22298847/

110. Poulsen N, Kröger N. Silica morphogenesis by alternative processing of silaffins in the diatom Thalassiosira pseudonana. J Biol Chem [Internet]. 2004 Oct 8 [cited 2023 Mar 29];279(41):42993–9. Available from: https://pubmed.ncbi.nlm.nih.gov/15304518/

111. Satoh D, Hiraoka Y, Colman B, Matsuda Y. Physiological and molecular biological characterization of intracellular carbonic anhydrase from the marine diatom Phaeodactylum tricornutum. Plant Physiol. 2001;126(4):1459–70.

112. Tanaka Y, Nakatsuma D, Harada H, Ishida M, Matsuda Y. Localization of soluble β-carbonic anhydrase in the marine diatom Phaeodactylum tricornutum. Sorting to the chloroplast and cluster formation on the girdle lamellae. Plant Physiol. 2005;138(1):207–17.