Analysis of Protein Conservation and Divergence in Four Diatom Genomes

Anni Wang[1], Matthew Parks[2], Matthew Johnson[2], and Norm Wickett[2]
[1]Florida State University, Tallahassee, FL, 32306; [2]Chicago Botanic Garden, Glencoe, IL 60022

Introduction

Diatoms are an unicellular group of unicellular algae responsible for one fifth of the world's primary productivity [1]. While it has been estimated that there are over 100,000 diatom species, they are still a relatively understudied group of organisms with only four complete genomes published. In order to get a better understanding of their evolutionary history, we can use each species' genome sequences to perform large scale comparative analyses.

Proteins are a class of biological macromolecules consisting of amino acids. Their sequence and structure determine organisms' physical and metabolic characteristics. The scope of this project is primarily looking at proteins to see levels of relatedness across four diatom species. By comparing proteins shared between diatom species and assessing their functions, we can infer how diatom species have diverged in an evolutionary context.

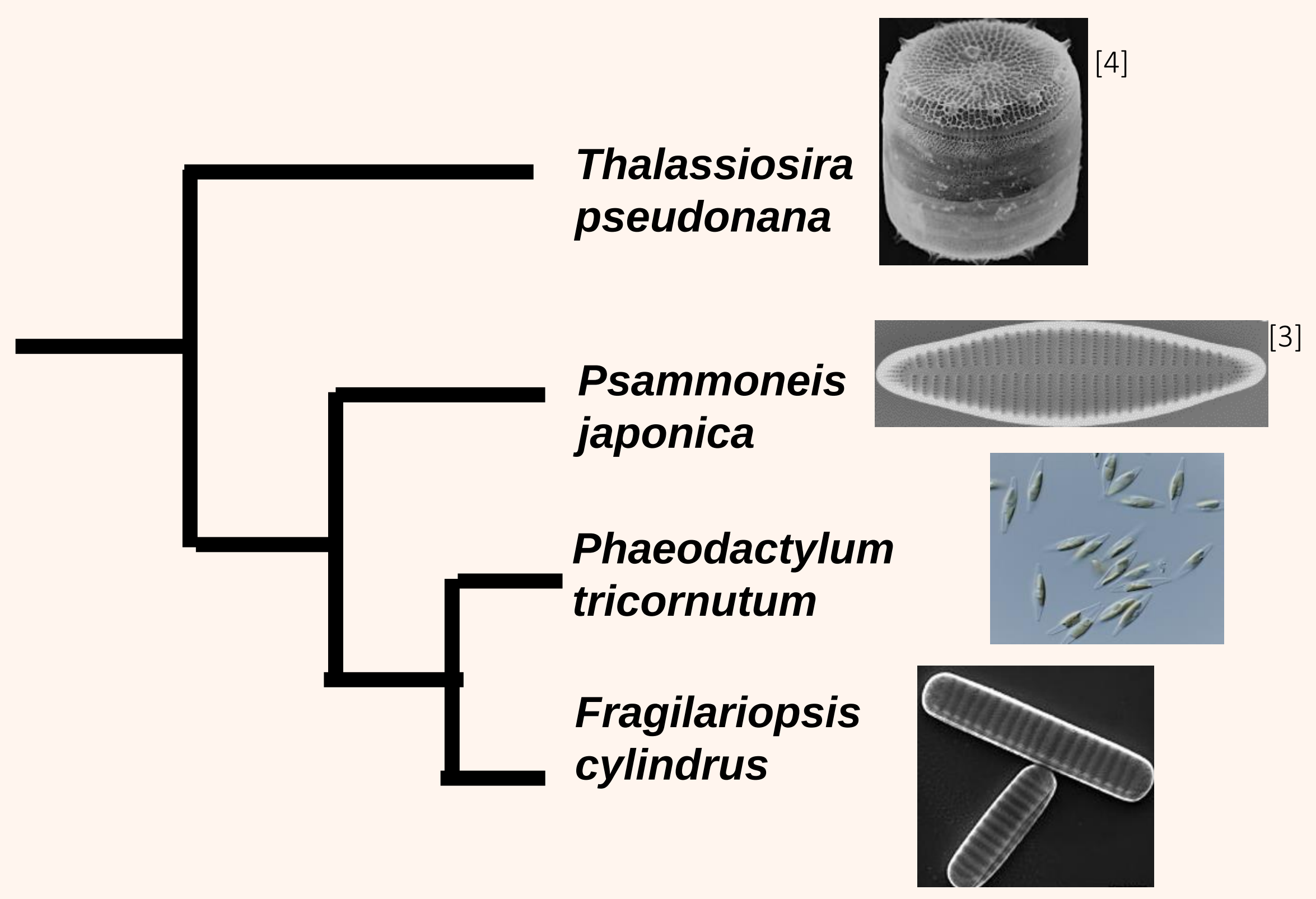
Objective

Look at protein sequences across four species of diatoms and assess their levels of relatedness to better understand their evolutionary history.

Hypothesis

Levels of conservation and divergence within orthologous proteins should reflect species relationships across four species of diatoms with varying levels of relatedness.

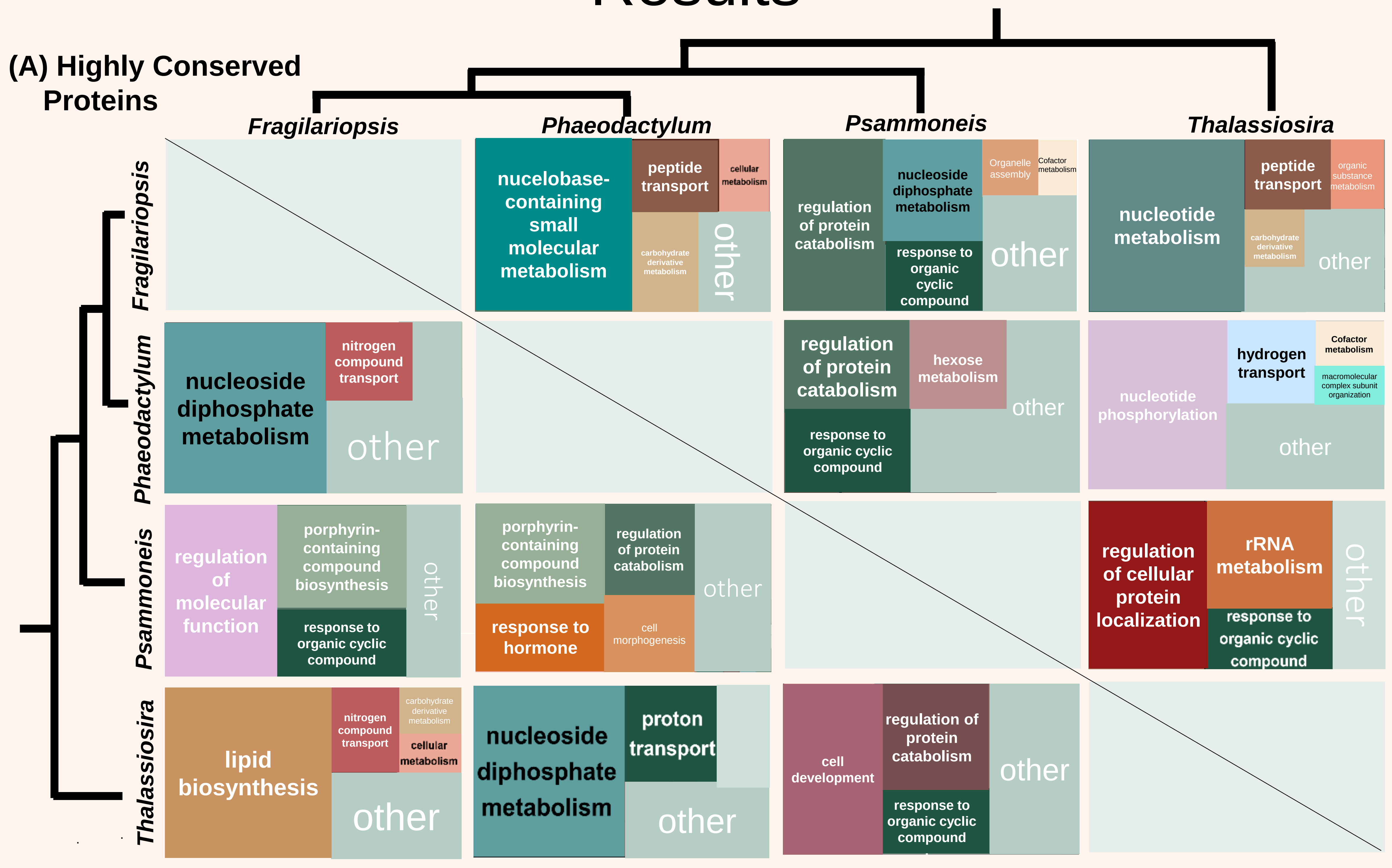
Study System


[4]
Thalassiosira pseudonana
[3]
Psammoneis japonica
Phaeodactylum tricornutum
Fragilariopsis cylindrus

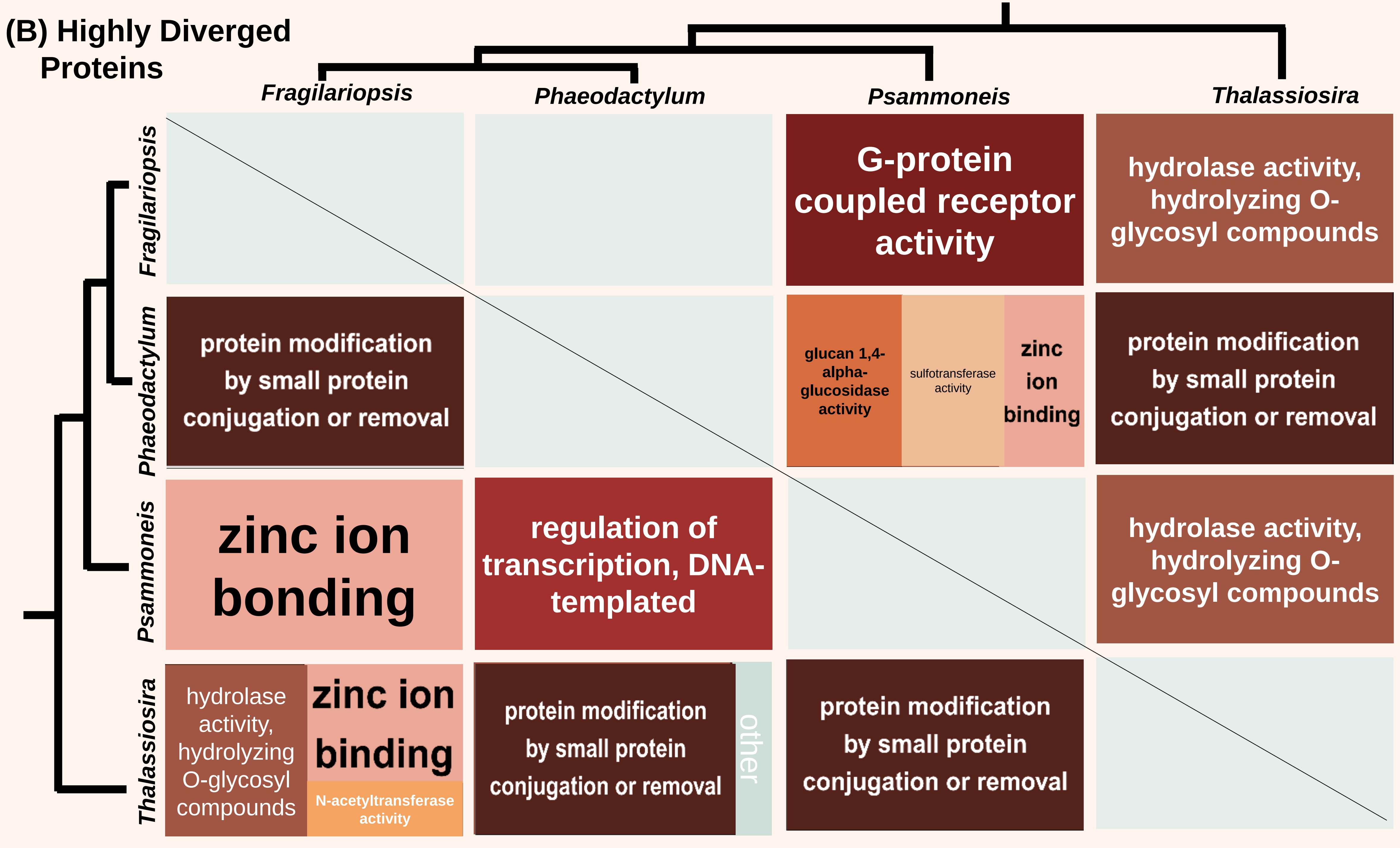
Methods

1. Download protein FASTA files for *Thalassiosira* and *Phaeodactylum* from NCBI genome database. FASTA file for *Psammoneis* was produced in-house and *Fragilariopsis* was downloaded from JGI.
2. Use BLAST to identify orthologous proteins based on best pair-wise similarity.
3. Use MAFFT to align all positions in orthologous proteins.
4. Estimate levels of divergence in orthologous protein pairs.
5. Compare highly conserved & diverged orthologs to assess patterns in biological processes across 4 species.

Results


(A) Highly Conserved Proteins

Fragilariopsis
Phaeodactylum
Psammoneis
Thalassiosira
nucelobase-containing small molecular metabolism
peptide transport
cellular metabolism
carbohydrate derivative metabolism
other
regulation of protein catabolism
nucleoside diphosphate metabolism
Organelle assembly
Cofactor metabolism
response to organic cyclic compound
other
nucleotide metabolism
peptide transport
organic substance metabolism
carbohydrate derivative metabolism
other
nucleoside diphosphate metabolism
nitrogen compound transport
other
regulation of protein catabolism
hexose metabolism
response to organic cyclic compound
other
nucleotide phosphorylation
hydrogen transport
Cofactor metabolism
macromolecular complex subunit organization
other
regulation of molecular function
porphyrin-containing compound biosynthesis
response to organic cyclic compound
other
porphyrin-containing compound biosynthesis
regulation of protein catabolism
response to hormone
cell morphogenesis
other
regulation of cellular protein localization
rRNA metabolism
response to organic cyclic compound
other
lipid biosynthesis
nitrogen compound transport
carbohydrate derivative metabolism
cellular metabolism
other
nucleoside diphosphate metabolism
proton transport
other
cell development
regulation of protein catabolism
response to organic cyclic compound
other


(B) Highly Diverged Proteins

Fragilariopsis
Phaeodactylum
Psammoneis
Thalassiosira
G-protein coupled receptor activity
hydrolase activity, hydrolyzing O-glycosyl compounds
protein modification by small protein conjugation or removal
glucan 1,4-alpha-glucosidase activity
sulfotransferase activity
zinc ion binding
protein modification by small protein conjugation or removal
zinc ion bonding
regulation of transcription, DNA-templated
hydrolase activity, hydrolyzing O-glycosyl compounds
hydrolase activity, hydrolyzing O-glycosyl compounds
zinc ion binding
N-acetyltransferase activity
protein modification by small protein conjugation or removal
other
protein modification by small protein conjugation or removal

Figure 1. Biological processes shared between orthologous pairings of proteins. For (A) and (B) processes above and below diagonal lines generated using orthologous protein pairs and all proteins as background populations, respectively.
(A) Treemap of highly conserved proteins' biological processes.
(B) Treemap of highly diverged proteins' biological processes.

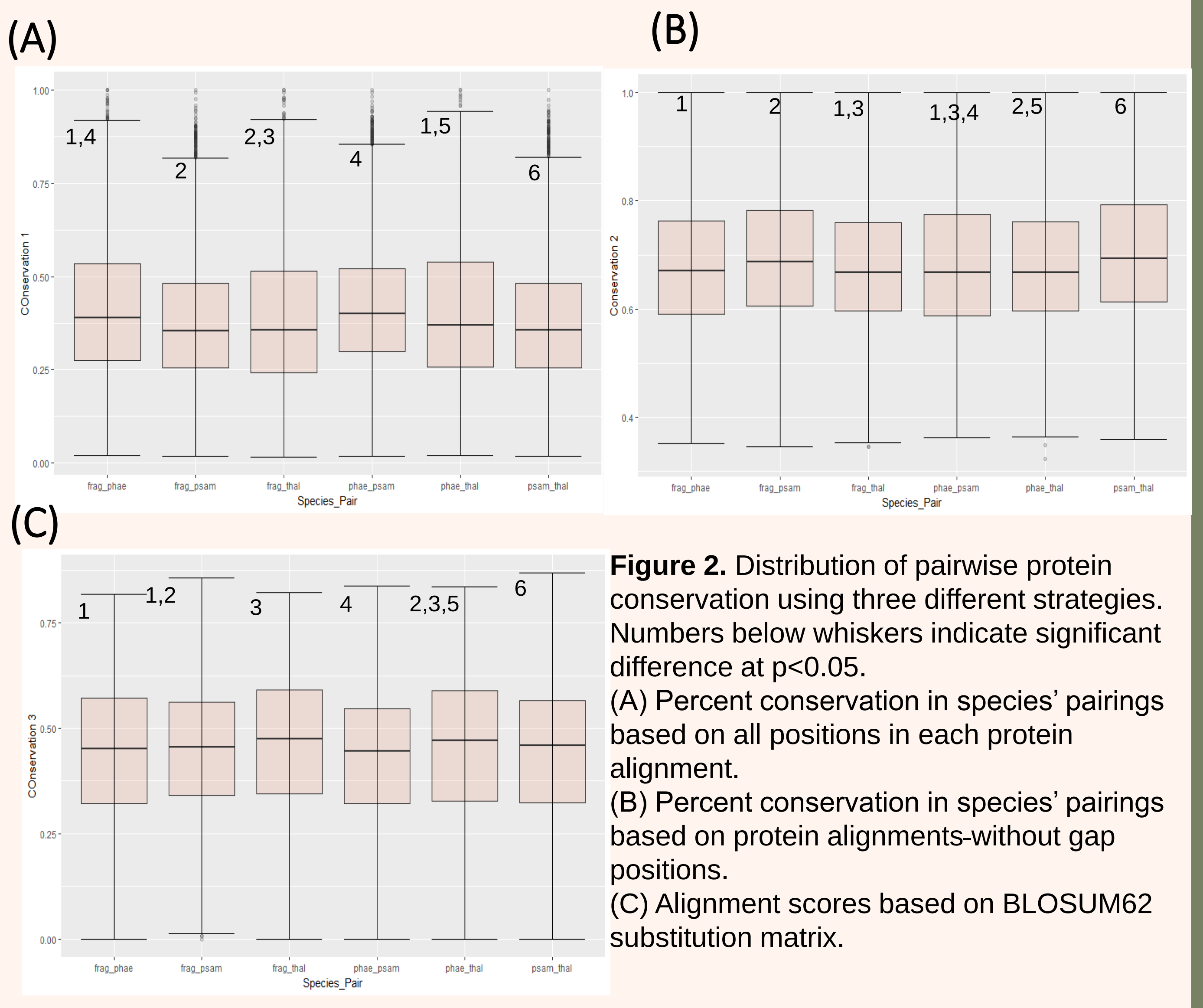


Figure 2. Distribution of pairwise protein conservation using three different strategies. Numbers below whiskers indicate significant difference at p<0.05.
(A) Percent conservation in species' pairings based on all positions in each protein alignment.
(B) Percent conservation in species' pairings based on protein alignments-without gap positions.
(C) Alignment scores based on BLOSUM62 substitution matrix.

Conclusion

Amongst both highly diverged and conserved orthologous proteins, some biological processes are shared across species comparisons. Conservation of protein function was more common across species comparisons of highly conserved proteins. This indicates that species' relatedness does not have a direct correlation to protein function or sequence conservation. Further, highly diverged proteins may contribute to species-specific diatom characteristics. Future research should include more species of diatoms to study as more genomes are assembled.

Highly Conserved Proteins

Cofactor metabolism and **response to organic cyclic compound** were commonly enriched functions with species-pair proteins as background population.

Protein folding and **generation of precursor metabolites and energy** were enriched in all species pairs with orthologous proteins as background population.

Highly Diverged Proteins

Protein modification by small protein conjugation or removal was enriched in all species pairs with species-pair proteins as background population.

No functions were enriched across all species pairs with orthologous proteins as background population. **Hydrolase activity and hydrolyzing O-glycosyl compounds** were shared between *Thalassiosira-Fragilariopsis* and *Thalassiosira-Psammoneis* comparisons.

Acknowledgments

We would like to thank NSF-REU grant DBI-1461007 for support. I would also like to thank my mentors, Matt Parks, Matt Johnson, and Norm Wickett, and the Chicago Botanic Garden for their assistance with this research project. Also special thanks to Jacky Vargas.

References

1. Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244 (2008)
2. Sato, S. *et al.* A new araphid diatom genus *Psammoneis* gen. nov. (Plagiogrammaceae, Bacillariophyta) with three new species based on SSI and LSU rDNA sequence data and morphology. *Phycologia* **47**, 510-528 (2008)
3. Sato, Shinya, *et al*. "A new araphid diatom genus Psammoneis gen. nov.(Plagiogrammaceae, Bacillariophyta) with three new species based on SSU and LSU rDNA sequence data and morphology." *Phycologia* **47.5** (2008): 510-528.
4. Demers, Serge, *et al*. "Rapid light-induced changes in cell fluorescence and in xanthophyll-cycle pigments of *Alexandrium excavatum* (Dinophyceae) and *Thalassiosira pseudonana* (Bacillariophyceae): a photo-protection mechanism." *Marine Ecology Progress Series* (1991): 185-193.