The UCYN-A transit peptide: a novel C-terminal targeting system with distinctive biophysical substrate signatures

Author One[1,*], Author Two[2]

[1]Institution One, City, Country
[2]Institution Two, City, Country

[*]Corresponding author: email@example.com

Abstract

UCYN-A is a nitrogen-fixing cyanobacterial endosymbiont that has evolved into an organelle-like structure (nitroplast) in the marine alga *Braarudosphaera bigelowii*. Approximately 900 host-encoded proteins are imported into UCYN-A via a novel C-terminal transit peptide (uTP), unlike the N-terminal signals used by mitochondria and chloroplasts.

1 Introduction

Mitochondria and chloroplasts import the vast majority of their proteins from the host cell cytoplasm. This import is mediated by N-terminal transit peptides that are recognized by dedicated translocon complexes (TOM/TIM in mitochondria, TOC/TIC in chloroplasts) and typically cleaved upon import (**?**). These targeting systems are ancient, highly conserved, and essential for organelle function. The evolution of efficient protein import was a key innovation that enabled the extensive genome reduction characteristic of endosymbiont-to-organelle transitions.

UCYN-A (*Candidatus* Atelocyanobacterium thalassa) is a nitrogen-fixing cyanobacterial endosymbiont that has recently been characterized as an organelle in the marine haptophyte *Braarudosphaera bigelowii* (Coale et al., 2024). The symbiosis originated approximately 90–100 million years ago, and UCYN-A has undergone extreme genome reduction, retaining only ∼22% of genes compared to its free-living relative *Crocosphaera watsonii* (Frail and Others, 2025). This makes UCYN-A—termed the "nitroplast"—the first nitrogen-fixing organelle to be characterized in eukaryotes.

Coale et al. identified approximately 368 host-encoded proteins that are imported into UCYN-A, representing a substantial complement of the nitroplast proteome (Coale et al., 2024). These imported proteins carry a novel C-terminal targeting sequence of ∼100–150 amino acids, termed the UCYN-A transit peptide (uTP). Using a hidden Markov model (HMM) derived from experimentally validated proteins, approximately 900 proteins in the *B. bigelowii* proteome are predicted to carry uTP. The imported proteins fill critical metabolic gaps in UCYN-A, including biosynthesis of threonine, serine, proline, pyrimidines, and tetrahydrofolate.

Despite the identification of uTP as a targeting signal, several fundamental questions remain unanswered. What are the conserved sequence and structural features of uTP? Why is uTP C-terminal, in contrast to the N-terminal signals used by mitochondria and chloroplasts? What is the evolutionary origin of uTP—was it co-opted from an existing cellular function or did it evolve de novo? And what determines which host proteins acquire uTP for nitroplast targeting?

Here, we present a systematic computational characterization of the uTP system. We identify conserved motif architecture with invariant anchor sequences and a conserved U-bend structure. We demonstrate that uTP has no detectable homologs in related haptophyte genomes, supporting a de novo evolutionary origin. Most importantly, we show that uTP-containing proteins share a distinctive biophysical signature—more disordered, acidic, and stable than the general proteome—that enables prediction of uTP status with 90% accuracy from mature domain sequence alone. These findings establish uTP as a novel solution to organellar protein targeting and provide a foundation for understanding the mechanisms underlying nitroplast protein import.

2 Results

2.1 uTP comprises conserved motif architecture with invariant anchor sequences

To systematically characterize the uTP system, we applied the HMM profile from Coale et al. to the full *B. bigelowii* proteome, identifying 933 proteins with putative uTP sequences. The uTP region spans 58–1091 amino acids (median ∼206 aa), located at the C-terminus of each protein.

De novo motif discovery using MEME identified 8 conserved motifs within the uTP region (Table **??**). Two motifs showed near-universal prevalence: motif 2 (93% of sequences) and motif 1 (91%). These "anchor motifs" appear in a stereotyped order, with motif 2 consistently preceding motif 1 at the N-terminal end of the uTP region. The remaining motifs (3–9) appear in variable combinations downstream, creating combinatorial diversity among uTP sequences.

Extended scanning with MAST across all 933 HMM-predicted proteins confirmed this architecture. Of 745 proteins with detectable motif hits, 81.5% showed one of four canonical terminal motif

patterns (ending with motif 4, 5, 7, or 9), with terminal motif 7 being most common (61.7%). The remaining proteins either lacked detectable motifs (20.2%) or showed non-canonical patterns.

2.2 uTP adopts a conserved U-bend structure

To assess structural conservation, we predicted three-dimensional structures for representative uTP sequences using AlphaFold2. Despite sequence variation in the variable motif region, all predicted structures converged on a conserved fold: two α-helices connected by a turn, forming a characteristic "U-bend" configuration.

Structural alignment across predictions revealed low variance, with pairwise RMSD values consistently below 4 Å. This structural conservation suggests that the U-bend fold is functionally constrained, potentially reflecting requirements for recognition by the import machinery.

2.3 No uTP homologs detected in related haptophyte genomes

To investigate the evolutionary origin of uTP, we searched for homologous sequences in related haptophyte species. We compiled proteomes from seven haptophyte species spanning estimated divergence times of 330–920 million years from *B. bigelowii* (Table **??**).

Profile HMM search with the uTP model yielded 84 total hits across all haptophyte proteomes. However, detailed analysis of hit coordinates revealed that all 33 significant hits (E-value < 0.01) aligned exclusively to the mature domain region of the HMM (positions 100–713), not the uTP-specific region (positions 1–100). These hits represent orthologs of the mature protein domains, not uTP homologs.

We also searched for individual uTP motifs using position weight matrices (MAST). While 530 individual motif hits were detected across haptophyte proteomes, the defining feature of uTP—co-occurrence of the two anchor motifs (MEME-1 and MEME-2) at the C-terminus—was completely absent. Only one protein in any haptophyte genome contained both anchor motifs, and these were located internally (not C-terminal), in a protein 1269 amino acids long.

As a control, we performed the same motif search against the *Arabidopsis thaliana* proteome. Similar hit rates confirmed that individual motif matches represent generic sequence patterns, not uTP-specific features.

The complete absence of proteins with the uTP motif architecture in any haptophyte genome strongly supports de novo evolution of uTP in the *B. bigelowii* lineage, likely after establishment of the UCYN-A symbiosis.

2.4 uTP-containing proteins have distinctive biophysical properties

To identify features distinguishing uTP-containing proteins from the general proteome, we constructed a carefully controlled comparison. From each uTP protein, we extracted the mature domain by removing the uTP region identified by HMM. We then selected length-matched control proteins from the *B. bigelowii* proteome, filtering by subcellular localization prediction (CELLO) to exclude proteins with other targeting signals (signal peptides, mitochondrial/chloroplast transit peptides). This yielded 605 uTP mature domains and 773 cytoplasmic/nuclear controls.

We trained binary classifiers using ProtT5 protein language model embeddings as features. Logistic regression achieved 92.8% accuracy (F1 = 0.92), with significance confirmed by permutation testing ($p = 0.002$, 1000 permutations). The ROC AUC was 0.92, indicating strong discriminative power.

To understand what drives this classification, we computed biophysical properties for all proteins. Three properties showed large effect sizes distinguishing uTP from control proteins (Table **??**):

- **Intrinsic disorder:** uTP proteins have significantly higher predicted coil/disordered content (Cohen's $d = +1.05$, $p < 0.006$)
- **Isoelectric point:** uTP proteins are more acidic (mean pI = 6.5 vs 8.5; $d = -0.89$, $p < 0.006$)
- **Stability:** uTP proteins have lower instability index, indicating greater stability ($d = -0.81$, $p < 0.006$)

All p-values were Bonferroni-corrected for 8 tests.

To validate these findings at scale, we applied the trained classifier to the entire *B. bigelowii* proteome (933 uTP vs 43,430 non-uTP proteins). Despite the extreme class imbalance, the classifier maintained strong performance (ROC AUC = 0.920, recall = 80.3

These biophysical signatures have plausible mechanistic interpretations. Increased disorder may facilitate protein unfolding during translocation across the nitroplast membrane. The acidic character could mediate electrostatic interactions with positively charged components of the import machinery. The greater stability may reflect selection for function in the nitroplast environment.

3 Discussion

4 Methods

4.1 Data sources

Proteomics data, genome annotations, and the uTP HMM profile were obtained from Coale et al. (Coale et al., 2024). The *B. bigelowii* transcriptome contained approximately 44,000 predicted proteins.

Haptophyte proteomes were downloaded from NCBI GenBank: *Emiliania huxleyi* CCMP1516 (GCA_000372725.1, ∼38,500 proteins), *Chrysochromulina tobinii* CCMP291 (GCA_001275005.1, ∼16,700 proteins), *Diacronema lutheri* (GCA_019448385.1, ∼14,400 proteins), *Pavlovales* sp. CCMP2436 (GCA_026770615.1, ∼26,100 proteins), *Prymnesium parvum* 12B1 (GCA_041296205.1, ∼23,700 proteins), and *Prymnesium* sp. SGEUK-05 (GCA_046255225.1, ∼15,000 proteins).

Acknowledgements

[To be added]

Author Contributions

[To be added]

Data Availability

[To be added]

References

T. H. Coale, V. Loconte, K. A. Turk-Kubo, J. P. Zehr, and D. Bhattacharya. Nitrogen-fixing organelle in a marine alga. *Science*, 384(6692):217–222, 2024. doi: 10.1126/science.adk1075.

S. Frail and Others. Genomes of nitrogen-fixing eukaryotes reveal an alternate path for organellogenesis. *Nature*, 2025. In press.