




In vivo thrombin activity in the diatom *Phaeodactylum tricornutum*: biotechnological insights

Anis Messaabi¹ · Natacha Merindol¹ · Lea Bohnenblust¹ · Elisa Fantino^{1,2} · Fatma Meddeb-Mouelhi^{1,2} · Isabel Desgagné-Penix^{1,2} 

Received: 2 July 2024 / Revised: 24 September 2024 / Accepted: 1 October 2024 / Published online: 8 October 2024
© The Author(s) 2024

Abstract

Diatoms are responsible for 20% of global carbon dioxide fixation and have significant potential in various biotechnological and industrial applications. Recently, the pennate diatom *Phaeodactylum tricornutum* has emerged as a prominent platform organism for metabolic engineering and synthetic biology. The availability of its genome sequence has facilitated the development of new bioengineering tools. In this study, we used in silico analyses to identify sequences potentially encoding thrombin-like proteins, which are involved in recognizing and cleaving the thrombin sequence LVPRGS in *P. tricornutum*. Protein structure prediction and docking studies indicated a similar active site and ligand positioning compared to characterized human and bovine thrombin. The evidence and efficiency of the cleavage were determined in vivo using two fusion-protein constructs that included YFP to measure expression, protein accumulation, and cleavage. Western blot analysis revealed 50–100% cleavage between YFP and N-terminal fusion proteins. Our findings suggest the existence of a novel thrombin-like protease in *P. tricornutum*. This study advances the application of diatoms for the synthesis and production of complex proteins and enhances our understanding of the functional role of these putative thrombin sequences in diatom physiology.

Key points

- Protein structure predictions reveal thrombin-like active sites in *P. tricornutum*.
- Validated cleavage efficiency of thrombin-like protease on fusion proteins in vivo.
- Study advances bioengineering tools for diatom-based biotechnological applications.

Keywords Cleavage peptide · Bioengineering · Synthetic biology · Microalgae · Thrombin-like protease · YFP (yellow fluorescent protein)

Introduction

Synthetic biology relies on efficient heterologous protein production systems, crucial for numerous biotechnology applications involving engineered proteins. Bioengineering combines natural sciences and engineering to address global challenges, such as drug production for disease treatment (Javaid et al. 2023). Current production platforms

include well-characterized hosts like bacteria, yeast, and plants (Demain et al. 2009). The pennate diatom *Phaeodactylum tricornutum* is emerging as a promising chassis for producing high-value compounds. It has been successfully utilized to produce heterologous proteins, such as the SARS-CoV-2 spike receptor-binding domain and plant-specialized metabolites such as geraniol and cannabinoids (Awwad et al. 2023; Butler et al. 2020; Dhaouadi et al. 2020; Fabris et al. 2020; Fantino et al. 2024; Sharma et al. 2021; Slattery et al. 2022). However, the use of this diatom as a bioengineering factory is in its infancy and would benefit from new protein engineering tools.

The production of a metabolite of therapeutical interest often requires the simultaneous expression of multiple genes involved in its biosynthetic pathway. Exogenous episomes delivered to the nucleus via bacterial conjugation have revolutionized metabolic engineering in diatoms (Karas et al.

✉ Isabel Desgagné-Penix
Isabel.Desgagne-Penix@uqtr.ca

¹ Department of Chemistry, Biochemistry and Physics,
Université du Québec À Trois-Rivières, Trois-Rivières, QC,
Canada

² Plant Biology Research Group, Université du Québec À
Trois-Rivières, Trois-Rivières, QC, Canada

2015), enabling the expression of five to six genes from a single episome. Polycistronic constructs allow the simultaneous expression of several genes like a transgene and a reporter gene. Fusion proteins are invaluable for various biotechnological applications, protein purification (Kimple et al. 2013), expression level monitoring (Smale 2010a, 2010b), and cellular localization (Belardinelli and Jackson 2017). These fusion proteins have been utilized in vaccine development and are being evaluated in clinical studies (Ellis et al. 2013; Wang et al. 2023). For example, the extracellular domains of two surface proteins from *Streptococcus pneumoniae* have been tested in a mouse model of pneumonia. Similarly, fusion proteins have been used as antigens in vaccines against *Neisseria meningitidis* serogroup B (Dos Santos et al. 2022; Snape et al. 2010).

However, producing recombinant fusion proteins is challenging due to potential issues like misfolding, instability, or poor expression caused by the proximity of fused proteins (Amet et al. 2009; Zhao et al. 2008). Linkers, or spacers, are short amino acid sequences that separate multiple domains in a single protein, improving structural stability and enhancing biological activity (Bouabe et al. 2008). Linkers are classified into three groups: rigid, flexible, and cleavable. Rigid linkers, rich in Proline residues and adopting α -helical conformations, maintain distance between domains to preserve protein stability and activity (George and Heringa 2002; McCormick et al. 2001). Some other rigid linkers are (EAAAK)_n (Takamatsu et al. 1990). Flexible linkers, composed of small, non-polar (e.g., G) or polar (e.g., S or T) amino acids, allow movement or interaction between proteins (Argos 1990). The most widely used flexible linker is (G-G-G-G-S)_n (Huston et al. 1988). Cleavable linkers, such as those recognized by proteases like thrombin and Factor Xa, are particularly useful in enzyme characterization and allow for tag removal after purification (Araujo et al. 2000). Thrombin, a trypsin-like serine protease, specifically cleaves peptide bonds on the carboxyl side (P'1) of basic amino acid residues, primarily arginine, making it an essential tool in synthetic biology. Proteases play a crucial role in various biological processes, and their application in synthetic biology and metabolic engineering has been well documented in several systems (Rao et al. 1998). The LVPRGS linker, derived from bovine factor XIII and cleaved by thrombin, is frequently used in chemical proteomics workflows and for expressing fusion proteins in *E. coli* without in vivo cleavage (Danckwardt et al. 2013; Jenny et al. 2003; Liu et al. 2017; Lundblad et al. 2004; Takagi and Doolittle 1974).

Despite the advances in using *P. tricornutum* for heterologous protein production, there is limited knowledge regarding the diversity and functionality of endogenous proteases in this diatom. The potential of *P. tricornutum* to harness specific proteolytic activities, such as those resembling thrombin-like serine proteases, remains unexplored

in the literature. Whether the thrombin cleavage sequence can be used in bioengineering strategies of *P. tricornutum* has never been verified. Understanding and leveraging such protease activities could open new avenues for the diatom's application in producing recombinant proteins and associated metabolites.

In this study, we identified candidates of thrombin-like proteases in *P. tricornutum* in silico. We constructed two fusion proteins: mFruits monomeric red fluorescent proteins (mCherry) fused to yellow fluorescent protein (YFP) and cannabidiolic acid synthase (CBDAS) fused to YFP, both separated by the LVPRGS sequence. Both cassettes were expressed under the Highly Abundant Secreted Protein1 promoter (HASP1) and Fucoxanthin chlorophyll a binding protein terminator (FcpA). We introduced the cassettes into *P. tricornutum* by bacterial transkingdom conjugation for extrachromosomal expression. We present evidence of recognition and cleavage of the LVPRGS sequence, confirmed in a bicistronic construct encoding an enzyme fused to the yellow fluorescent protein with the same thrombin cleavage sequence. This study provides a proof of concept that *P. tricornutum* possesses thrombin-like proteases, which could be valuable for projects involving heterologous gene expression in this model diatom.

Materials and methods

Plasmid construction

Plasmids were constructed by Gibson assembly using the NEBuilder® HiFi DNA Assembly Bundle for Large Fragments (New England Biolabs, Canada). Fragments to assemble were amplified by PCR with PrimeSTAR GXL DNA Polymerase (Takara Bio, Japan) following the manufacturer's protocol. Plasmid pPtGE30 was used as a template to construct the expression vectors (Slattery et al. 2018). Genes encoding mCherry, YFP, and cannabidiol acid synthase (CBDAS) were codon optimized for *P. tricornutum* and synthesized by Bio Basic (Markham, Ontario, Canada). To enhance protein production, a minimal Kozak sequence (ACC) was placed directly before each ATG initial codon. Reporter gene *YFP* was tagged with 3 × HA. DNA fragments were expressed under an endogenous promoter *Highly abundant secreted protein 1* (*HASP1*) and *fucoxanthin chlorophyll a/c-binding protein A* (*FcpA*) terminator (Erdene-Ochir et al. 2019) or under the *40S ribosomal protein S8* promoter (40SRPS8) and *fucoxanthin chlorophyll a/c-binding protein A* (*FcpA*) terminator. *N-acetyl transferase* gene (*NAT*) conferring resistance to nourseothricin (NTC) was included as a selection marker under *fucoxanthin chlorophyll a/c-binding protein C* promoter and terminator (*FcpC*) sequences in all constructs.

Construct named here *C-T-YFP* includes *HASPIpro* and *CBDAS* is fused to *YFP* with a thrombin cleavage sequence (LVPRGS). Construct *mCherry-T-YFP* includes *HASPIpro* and *mCherry* is fused to the reporter *YFP* gene with a thrombin cleavage sequence. Constructs *mCherry and YFP* includes *40SRPS8pro* and *mCherry* with a stop codon and a separated expression cassette with *HASPIpro* controlling the expression of *YFP*. A modified version of pPtGE30 vector, harboring *NAT* gene cassette instead of *Sh ble* gene, conferring zeocin resistance, was used as negative control (EV). All DNA sequences, oligonucleotides, and plasmids used to generate these constructs are listed in supplementary data (Tables S1, S2, and S3).

Microbial strains, growth conditions, and transformation

Escherichia coli (NEB® 10-beta, New England Biolabs, Canada) was grown in Luria Broth (LB, ThermoFisher, Canada) supplemented with appropriate antibiotics (chloramphenicol, 30 mg.L⁻¹, ThermoFisher, Canada). Plasmid DNA constructs were extracted using a miniprep kit allowing the extraction of large vectors (Biobasic EZ10 miniprep kit, NY, USA), and the integrity of these plasmids DNA was verified by next-generation sequencing in CCIB DNA Core (Massachusetts General Hospital, United States of America). Plasmid constructs were transformed to *E. coli* Epi 300 strain containing pTA-MOB plasmid to allow conjugation with diatoms, as described previously (Karas et al. 2015). Epi 300 strain was grown in LB supplemented with chloramphenicol (30 mg. L⁻¹) and gentamicin (40 mg. L⁻¹, ThermoFisher, Canada). Microalgae *P. tricornutum* (Culture Collection of Algae and Protozoa CCAP 1055/1) was grown in L1 media, as described previously (Karas et al. 2015) at 18 °C under white fluorescent lights (75 µE m⁻² s⁻¹) and a photoperiod of 16 h light:8 h dark with an agitation of 130 rpm for liquid cultures. *P. tricornutum* conjugation was performed as in Awwad et al. (2023). Transconjugant colonies appeared after 2 weeks of incubation where 36 colonies per construct were plated on new selective media supplemented with nourseothricin (200 µg mL⁻¹). After 7 days, transconjugants were transferred from the agar plates to 96-well plates with to 200 µL of L1 supplemented with 200 µg mL⁻¹ in each well and incubated at 18 °C under gentle agitation. After 14 days of liquid growth, putative transconjugants were screened for fluorescence (YFP, $I_{ex}/\lambda_{em} = 513/527$ nm and mCherry, $I_{ex}/\lambda_{em} = 587/610$ nm) using a plate reader (Synergy H1 BioTek microplate reader (Agilent, Santa Clara, CA, USA)). The selected nourseothricin-resistant colonies were transferred to a 25-mL L1 liquid medium containing nourseothricin in 125-mL flasks for kinetic growth, measurement of the OD at 680 nm, and analysis.

Protein extraction

Bioengineered 10-day-old *P. tricornutum* liquid cultures of 25 mL were centrifuged at 3500×g for 15 min at 4 °C. Pellets were weighed and resuspended in lysis buffer (50 mM Tris pH 7.4, 500 mM NaCl, 0.1% Tween20, 1×protease inhibitor cocktail) and kept for 20 min on ice. Sonication was performed 6 times at 35% amplitude, 30 s on, 30 s off (3 min total). Cell debris was removed by centrifugation at 12,000×g for 60 min at 4 °C. Total protein samples were quantified using RC DCTM Protein Assay Kit I (Bio-Rad, Hercules, CA, USA). Protein extracts were stored at −80 °C until further use.

Western blot

For each clone, a quantity of 60 µg of total protein extract was loaded in a 10% SDS-PAGE. Protein separation was performed at 80 V until passage through to the stacking gel, followed by 120 V for the rest of the run. Wet electroblotting was performed at 120 V and was performed for 2 h to transfer the separated proteins to a 0.2-µm PVDF membrane using the Bio-Rad Mini-PROTEAN Trans-Blot Turbo Transfer System. The blot was equilibrated in 1×TBS solution and blocked with 5% milk in 1×TBS for 1 h at room temperature followed by three washes. Primary antibodies were added for an overnight incubation at 4 °C. Primary anti-YFP (cat no# CLH106AP), anti-HA (cat no# A01244-100), and anti-mCherry (cat no# MA5-32977) were purchased from Cedarlane (Canada) and ThermoFisher Scientific (USA), respectively, and used at a 1:1000 dilution in 3% BSA. After three washes with TBST solution, the blots were incubated for 1 h with a secondary antibody at a 1:20,000 dilution of Immun-Star Goat Anti-Mouse (GAM)-HRP conjugate from Bio-Rad (Canada), in 5% milk. A quantity of 10 ng of multiple Tag (GenScript, USA) and 10 ng of purified recombinant GFP protein (10 ng) from *E. coli* were used as a positive control for HA Tag and YFP detection, respectively (Diamond et al. 2023). Two confirmed clones were used as extra controls for the detection of CBDAS, one of them expressing YFP-tagged with a single HA-tag (YFP-1×HAtag) and the other one expressing YFP-tagged with three HA-tag (YFP-3×HAtag). After three washes with TBST solution, protein detection was performed by using Clarity Max Western ECL Substrate-Luminol solution from Bio-Rad. Chemiluminescence detection and Ponceau red staining (Glacial acetic acid 5% v/v, Ponceau red dye 0.1% m/v) of the blots were visualized using the ChemiDoc Imaging System with Image Lab Touch Software (Bio-Rad) and Image Lab™ Software (Bio-Rad).

Flow cytometry (FC)

At least 10,000 events per transconjugant strain were analyzed on a CytoFLEX S flow cytometer (Beckman) equipped with violet (405 nm), blue (488 nm), yellow/green (561 nm),

and red (638 nm) lasers. A volume of 200 μ L from each culture was filtered and transferred to a clear 96-well plate as described in Awwad et al. (2023). Autofluorescence was detected and gated using the PerCP channel (690/50 nm). Cells with high non-specific autofluorescence were excluded based on emission in the PB450 channel (450/45 nm). The fluorescence of YFP and mCherry was analyzed in the FITC channel (525/40 nm) and the ECD/mCherry channel (610/20 nm), respectively. Empty vector (with *NAT*) was used as a negative control to set the gates.

Bioinformatics analysis and phylogenetic tree

To identify *P. tricornutum* thrombin-like proteins, the amino acid sequences of bovine (OX = 9913), human (OX = 9606), mouse (OX = 10,090), pig (OX = 9823), orangutan (OX = 9601), and rat (OX = 10,116) prothrombins from UniProtKB/Swiss-Prot were blast searched in the *P. tricornutum* protein database version 3. The selected amino acid coding sequences were analyzed using different web-site applications: PROSITE (Sigrist et al. 2013), PESTfind (Rechsteiner and Rogers 1996), UniProt, ClustalO, and Unipro UGENE (Okonechnikov et al. 2012). Solvent accessibility and secondary structure analysis of putative *P. tricornutum* thrombin-like candidates were performed using Predict protein (<https://www.predictprotein.org/>). Phylogenetic analyses were carried out after multiple amino acid sequence alignments of 55 prothrombin candidates (Table S4) and 100 non-thrombin serine proteases sequences (Fig. S2) using MEGA11 software (Tamura et al. 2021). Phylogeny was inferred using the neighbor-joining method, and the evolutionary distances were computed using the maximum composite likelihood method for both trees in this study.

Protein structure prediction and docking study

The models of the 7 putative thrombins of *P. tricornutum* were predicted using ColabFold v1.5.3 (AlphaFold2 using MMseqs2) (Mirdita et al. 2022). Structures were visualized and overlaid using Pymol (Shrödinger). MOE2022.09 software (Chemical Computing Group) was used to analyze the conformation of resulting models and prepare the receptors for docking, as described in Majhi et al. (2023). The pocket residues detected by *Site Finder* were superimposed on the crystal structure of *H. sapiens* α -thrombin (PDB:1PPB and 1NRS Bode et al. 1989; Mathews et al. 1994) in MOE and the active site was predicted based on residues interacting with ligands from orthologous crystal structures. Receptors were then prepared using structure preparation, correcting issues, capping, charging termini, selecting appropriate alternatives, and calculating optimal hydrogen positions and charges using Protonate 3D. Structures were energy

minimized with 1PPB ligand (D-Phe-Pro-Arg chloromethylketone) inside the pocket.

The folding of the thrombin cleavage sequence LVPRGS was predicted using ColabFold v1.5.3 (AlphaFold2 using MMseqs2) and further prepared in MOE. Two predicted protomers, at pH = 7, were included as possible ligands. The MMFF94 \times force field was applied. Triangle Matcher was used as the placement method for 200 poses and tethered induced fit as the refinement to perform flexible docking, yielding 10 best poses. Docking poses were analyzed by comparison with crystallized human α -thrombin ligand complex (1PPB with $_D$ FPR chloromethylketone, 1NRS with hirugen), and the first coherent pose (with the best docking score) consistent with the catalysis of cleavage of LVPR↓GS was selected for each ligand. H-bonds were predicted, and images were further processed using PyMOL (Shrödinger).

Results

Serine protease prothrombin candidates in *P. tricornutum* proteome and phylogenetic analysis

A comprehensive analysis of the *P. tricornutum* proteome using characterized sequences as baits identified seven potential amino acid sequences with features of prothrombin and thrombin-like proteins (Table S4). Interestingly, over the Bacillariophyta phylum (taxid:2836), 113 hits were obtained using these sequences as baits (Table S5), including 35 in the order Bacillariales, 31 in Thallassiosirales, 14 in Chaetocerotales, 24 in Naviculales, and 9 in Fragilariales (Schoch et al. 2020), suggesting that orthologs of thrombin-like serine proteases are present in diverse diatom species. The seven *P. tricornutum* candidate sequences were aligned with prothrombin and thrombin-like sequences from human, cow, mouse, pig, orangutan, rat, salmon, and snake species. Phylogenetic analysis revealed that the selected *P. tricornutum* sequences formed a distinct cluster (Fig. S1), with only one of the *P. tricornutum* sequences (Phatr3 J45961) clustering closely with the complete and partial prothrombin sequences of human, cow, pig, orangutan, mouse, rat, and salmon. Further phylogenetic analysis, which included the candidates, thrombin baits, and non-thrombin serine proteases, revealed that even though *P. tricornutum* sequences cluster with mammalian prothrombins, the node bootstrap values were low and insufficient to support their characterization as thrombin serine proteases (Fig. S2). *Homo sapiens* and *Bos taurus* sequences were selected for a detailed comparison with the *P. tricornutum* thrombin candidates as they are well characterized due to their pharmacological interest (Bhandari et al. 2011; Cheng et al. 2009). The identity and similarity between *P. tricornutum* candidates and human thrombin ranged from 25 to 32% and 39 to 47%,

respectively. Phatr3 J49772 and J54319 exhibited the highest identity (32% and 30%) and similarity (47% and 43%) to the human sequence. Protein domain predictions using the PROSITE database confirmed that all *P. tricornutum* candidates contained a trypsin domain with the catalytic triad (Histidine (H), aspartate (D), and serine (S)), similar to human and bovine thrombin sequences (Fig. 1).

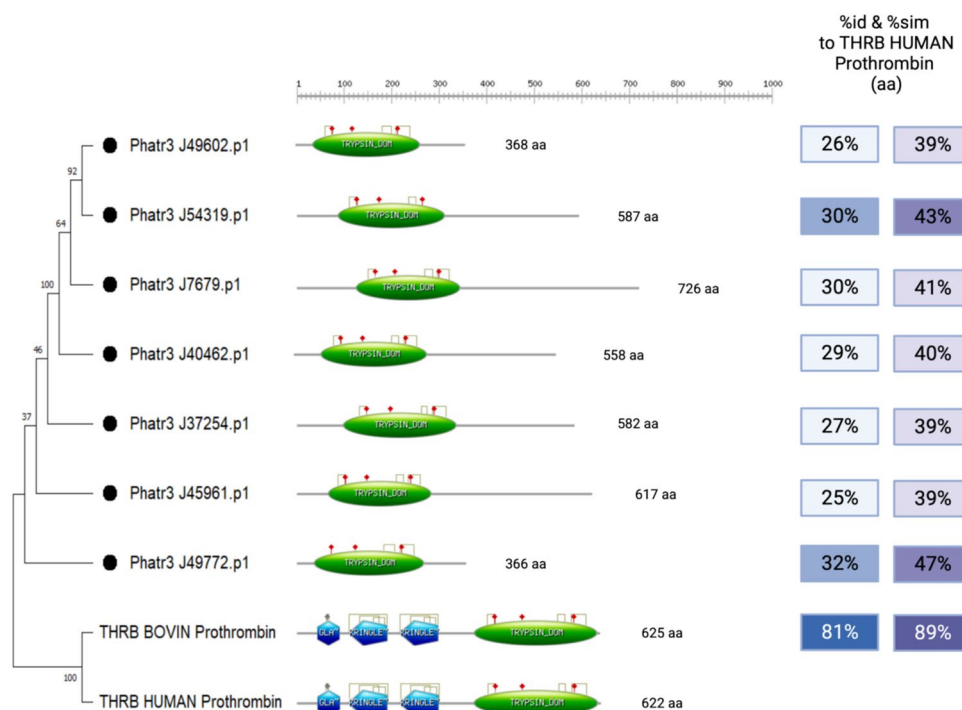
Lengths, domains, and percentages of identity (% id) and similarity (% sim) to the Human Thrombin sequence (amino acids) are shown. Potential Thrombin-like sequences contain a Trypsin domain (TRYPSIN_DOM) shown in green like the Human and Bovine thrombin sequence. Active sites are represented in red arrow and disulfide bonds in gray. Gamma-carboxyglutamic acid-rich (Gla) and Kringle domains are represented in blue.

An alignment of the trypsin domains showed further similarity, with 26 strictly conserved residues including the catalytic triad (Fig. 2). The gamma-carboxyglutamic acid-rich (Gla) domain and the two Kringle domains of human and bovine sequences were absent in *P. tricornutum* candidates. The Gla domain is known to be a membrane-binding motif that interacts with phospholipid membranes in the presence of calcium ions (Freedman et al. 1996). Kringles are disulfide cross-linked domains found in some serine proteases (Castellino and McCance 1997). They play a role in the regulation of proteolytic activity, but also in binding to membranes and other proteins (Wei et al. 1997). Their absence in *P. tricornutum* may be due to their function in blood coagulation by platelet activation, which are not relevant to diatom.

In silico evidence for thrombin-like function

The folding of the seven putative thrombins with trypsin domain was predicted using ColabFold v1.5.3 (AlphaFold2 using MMseqs2) (Mirdita et al. 2022) and compared with crystallized human and bovine thrombin (Fig. S3a and S3b). All predicted models displayed a trypsin-like domain with two interacting six-stranded β -sheet barrel domains surrounded by helical and loop regions (Fig. S3). The active site residues were predicted and are shown in Table S6. Superposition of the predicted *P. tricornutum* trypsin domains with human β -thrombin showed similar overall folding and orientation of the active site residues (Fig. S4). The conserved catalytic triad (H, D, and S) is located at the rims of the two interacting barrels, as expected. The oxyanion hole, composed of the catalytic serine and a near-adjacent glycine, which stabilizes the oxyanion intermediate by hydrogen bonding to the oxygen of the P1 residue, was also observed in all structures (Schechter and Berger 2012). Additionally, equivalent structures to the sodium binding site loop, the 180 loop, the autolysis γ -loop, and a covalent disulfide bond between two cysteines connecting the catalytic serine with the oxyanion hole were identified in the diatom candidates, as seen in human thrombin structures, but not in other serine proteases. However, the exosites and the 60-loop of human β -thrombin were not detected (Johnson et al. 2005). Outside of the trypsin domain, *P. tricornutum* predicted structures contained β -barrels (Fig. S3f, g, h, i) or α -helices (Fig. S3c, d, g, h, i), suggesting that they might be transmembrane proteins with additional functions.

Fig. 1 A phylogenetic tree of seven potential serine protease thrombin-like sequences in *P. tricornutum*



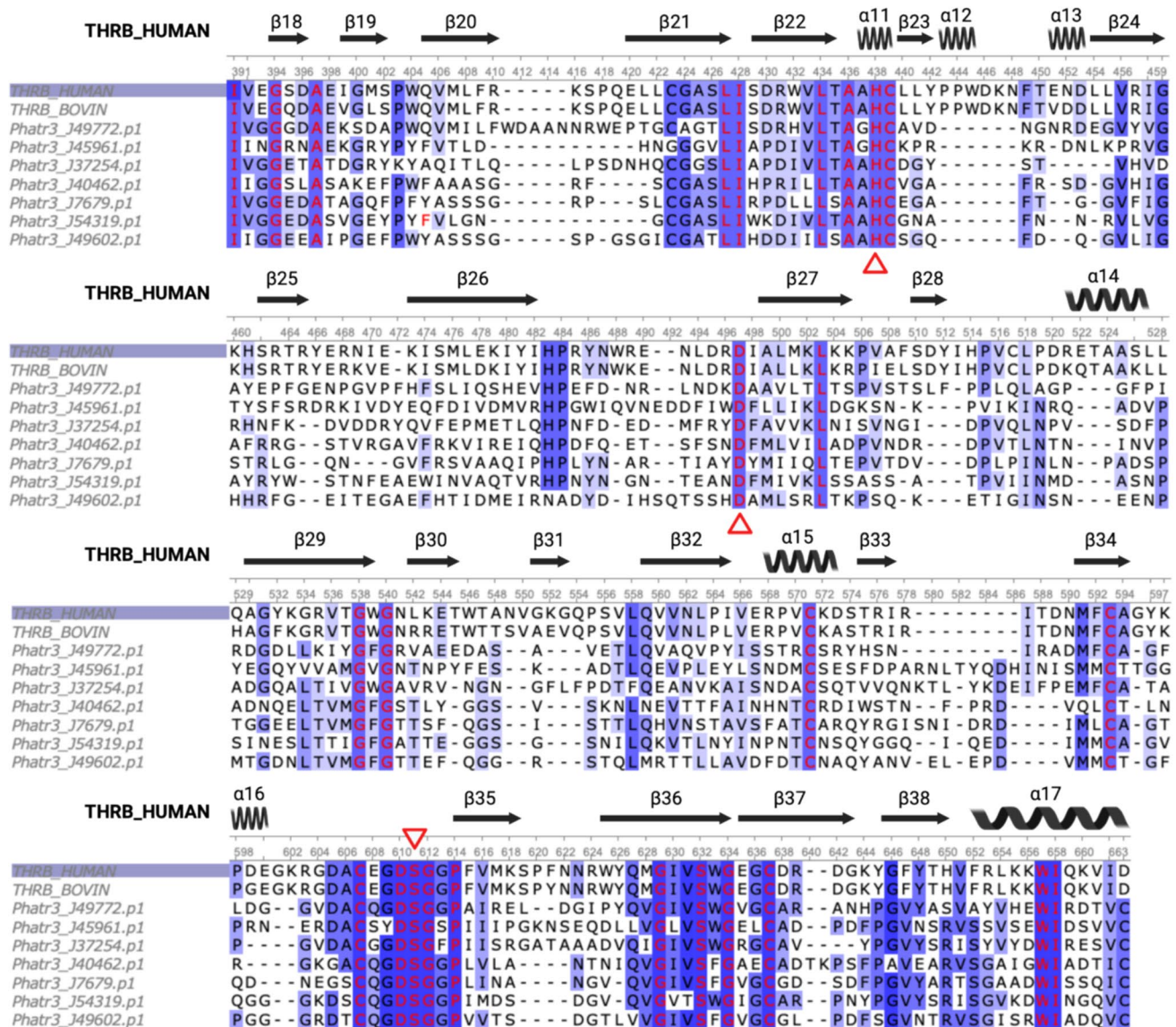


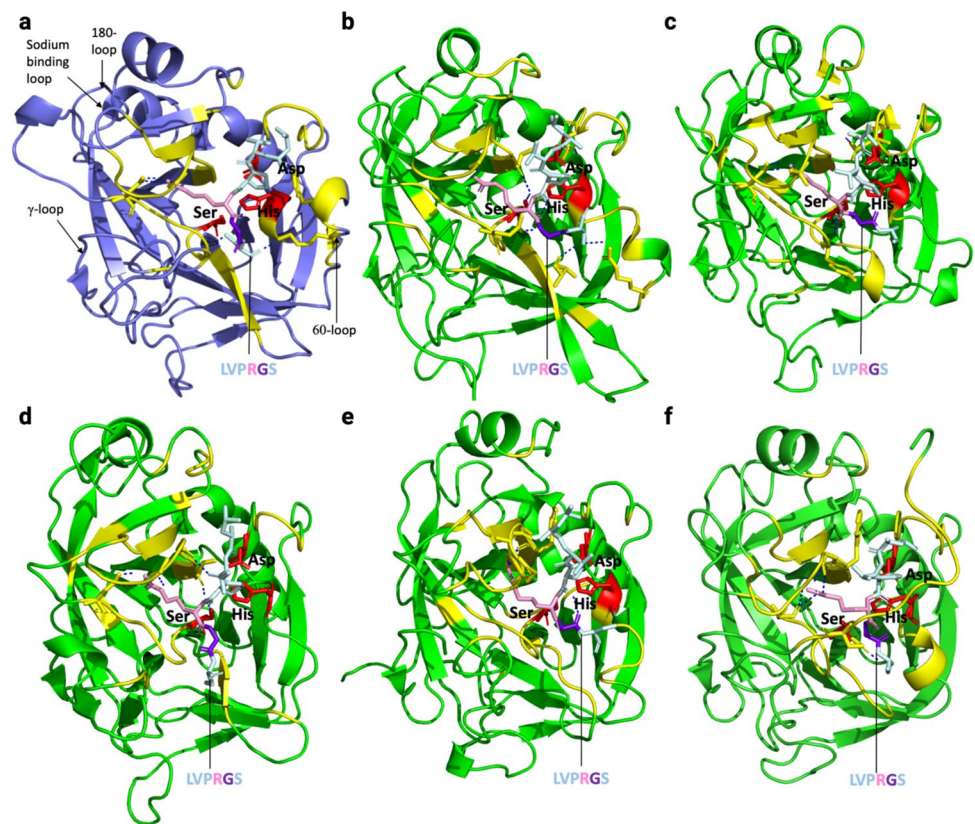
Fig. 2 Amino acid sequence alignment of predicted Trypsin domains from *P. tricornutum* with two known Human (Accession number: P00734) and Bovin (Accession number: P00735) structures. Identical

residues are highlighted in blue, and strictly conserved residues are highlighted in blue with red characters. Red triangles represent the active sites

As the diatom's candidate displayed a predicted trypsin domain that closely matched with the human's, the thrombin cleavage peptide was docked into their active site and the interactions were analyzed (Table S6). LVPRGS docked with a score of -10.87 kcal/mol in human β -thrombin active site. In the case of Phatr3_J7679 and Phatr3_J49602 candidates, the ligand docked further from the catalytic residues than 1PPB and 1NRS crystallized ligands. Therefore, the docking poses of these structures were rejected. In five of the putative predicted structures, the ligand docked in a position consistent with a cleavage between R and G (Fig. 3), with scores ranging from -8.92 to -10.6 kcal/mol (Table S6).

Overall, LVPRGS docked very similarly into human and trypsin domains of selected putative *P. tricornutum* thrombin, with the cleavage site positioned in between the catalytic serine and histidine. Several H-bonds were detected between the docked peptide and the active site, including with catalytic residues in the case of 1PPB, J49772, J45961, J37254, and J40462. J54319 showed the highest docking score with a position most similar to the docking configuration in the human protein (Fig. 3a–f, and Table S6). These results suggest that *P. tricornutum* produces endogenous proteins containing a trypsin domain that could efficiently cleave the thrombin cleavage sequence in vivo.

Fig. 3 Cartoon representation of docked thrombin-cleavage peptide in the active site of human (a 1PPB in purple) and 5 putative thrombins from *P. tricornutum* (b Phatr49772, c Phatr45961, d Phatr37254, e Phatr40462, and f Phatr54319 in green). The cleavage peptide (LVPRGS) is shown as light blue sticks with pink arginine and violet glycine. The active site is shown in yellow, and catalytic residues (His, Asp, and Ser) are displayed as red sticks. Only the trypsin domain of each protein is shown. H-bonds are shown as dashed blue lines



P. tricornutum transformation and screening of transconjugant strains

To analyze potential thrombin activity in *P. tricornutum* cells, we constructed and introduced four expression cassettes into *P. tricornutum* wild type (Fig. 4). An empty vector (EV) encoding antibiotic resistance and a plasmid harboring two genes (*mCherry* and *YFP*) expressed under two distinct promoters (40SRPS8 and *HASPI*, respectively) were used as controls (Fig. 4b). Additionally, two other plasmids were constructed containing bicistronic cassettes encoding two recombinant proteins (CBDAS or *mCherry*) fused to YFP by a thrombin cleavable sequence hereafter referred to as C-T-YFP and *mCherry*-T-YFP (Fig. 4a, b). Highly Abundant Secreted Protein1 (*HASPI*) promoter drove the expression of *mCherry*-T-YFP and C-T-YFP recombinant proteins. *P. tricornutum* wild type cells grown to stationary phase were transformed by conjugation (Karas et al. 2015). Transconjugants were selected with the appropriate antibiotic and screened for *mCherry* and YFP fluorescence using a microplate reader and flow cytometer. A total of 108 colonies were screened, with 36 transconjugants per construct. The percentage of positive clones was higher in the *mCherry* and YFP lines (69% of clones were *mCherry*⁺) compared to the

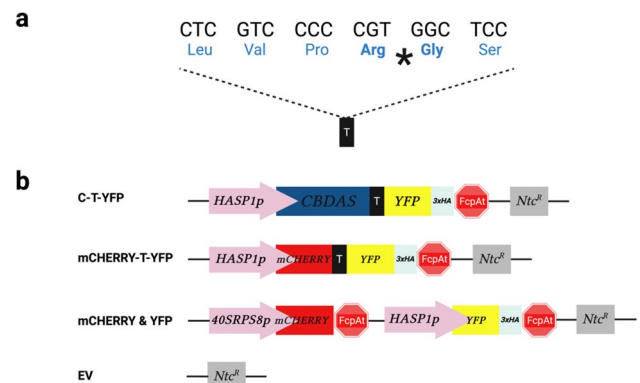


Fig. 4 Expression cassette of recombinant constructs with a thrombin cleavable sequence. **a** Thrombin cleavable sequence (LVPRGS) derived from the sequence in bovine factor XIII, cleavage occurs between Arginine (Arg) and Glycine (Gly). **b** Scheme of the expression cassettes, *mCherry*-Thrombin-YFP-3×HAtag (*mCherry*-T-YFP) and CBDAS-Thrombin-YFP-3×HAtag (C-T-YFP). The expression of *mCherry* and CBDAS genes is driven by the endogenous highly abundant secreted protein 1 promoter (*HASPI*), fused to YFP reporter gene with a thrombin cleavable sequence in between. Construct *mCherry* and YFP containing *mCherry* and YFP genes are driven by two different promoters, *HASPI* and 40SRPS8, respectively. EV containing only *Nourseothricin* resistance gene (*Ntc^R*) *mCherry* and YFP were cloned in two different vectors and used as controls

mCherry-T-YFP lines (39%) (Table 1). Additionally, 25% of the C-T-YFP clones were YFP⁺. Three to six independent transgenic lines from each construct were selected for further analysis (Table 1).

n.a. not applicable, *n.t.* not tested. Percentage of mCherry⁺ and YFP⁺ clones, *n* = 36; fluorescence was detected with a microplate reader (PR) or with a flow cytometer (FC).

Different levels of transgene fluorescence in *P. tricornutum* transconjugants

Transconjugant growth was monitored for 10 days, and all selected transgenic lines exhibited similar results with no significant differences compared to the negative control (EV) ($*p > 0.05$, one-way ANOVA, Fig. 5a, upper panel), indicating that transgenes did not affect the growth of *P. tricornutum* cells. The percentages of mCherry⁺ and YFP⁺ cells and relative fluorescence were measured on day 10 by flow cytometry and microplate reader, respectively, and normalized to the autofluorescence of the negative control (EV) (Fig. 5a, middle and lower panels). The percentage of mCherry⁺ cells varied significantly between constructs ($*p < 0.05$, Student's *t*-test), with a mean of 55% for mCherry and YFP transconjugants compared to 7.5% for mCherry-T-YFP (Fig. 5a, middle panel). However, since mCherry was expressed under different promoters in these constructs, the differences in fluorescence may be due to changes in the expression levels of the 40SRPS8 and HASP1 promoters rather than the construction design. Nonetheless, the percentage of YFP⁺ cells was also significantly higher with mCherry and YFP ($*p < 0.05$, *t*-test), although the difference was smaller, with a mean of 3.4% for mCherry and YFP transconjugants compared to 1.2% for mCherry-T-YFP transconjugants. C-T-YFP transconjugants did not show a significant difference in YFP⁺ percentages compared to mCherry and YFP (1.8% of YFP⁺, $*p > 0.05$, *t*-test). A similar trend was observed for relative fluorescence measured with the microplate reader (Fig. 5a, lower panel). Comparison of mCherry and YFP clones with mCherry-T-YFP clones revealed a significantly higher

relative mCherry fluorescence in mCherry and YFP clones (mean of 830 RFU) ($*p < 0.05$, Student's *t*-test). In contrast, mCherry-T-YFP clones exhibited a mean relative fluorescence unit (RFU) of 11. Moreover, the relative fluorescence emitted by YFP in mCherry and YFP clones (152 RFU) was significantly higher than in mCherry-T-YFP (29 RFU) ($*p < 0.05$, Student's *t*-test), while C-T-YFP clones showed no significant changes (mean of 75 RFU). Taken together, these results demonstrate that the fluorescence levels of the transgenes varied, with higher mCherry emission under the 40SRPS8 promoter than the HASP1 promoter. Furthermore, YFP expression, driven by the HASP1 promoter, slightly decreased when fused to the C-terminal of another protein.

In vivo evidence of recognition and cleavage of the thrombin cleavable sequence in *P. tricornutum*

To verify the possible cleavage of the thrombin sequence in *P. tricornutum* transconjugants, we analyzed the transgenes by western blotting using anti-HA antibody (Fig. 5b). No bands were detected in the total protein extract of *P. tricornutum* transconjugants 4, 5 and 6 harboring the *mCherry-T-YFP* cassette consistent with the low fluorescence detected by both flow cytometry and microplate reader. In contrast, two bands corresponding to the cleaved (32 kDa) and uncleaved forms (58 kDa) of mCherry-T-YFP protein from transconjugants 1, 2, and 3 were observed (Fig. 5b). The presence of the cleaved protein form was confirmed by using an anti-mCherry antibody (Fig. S5). Cleavage efficiency was determined by relative quantification of band intensity using Image Lab software. In transconjugants (mCherry-T-YFP) 1, 2 and 3, the fusion proteins were partially cleaved with a mean ratio of uncleaved to cleaved forms of approximately 53/47%. Heterogeneity between clones was observed, with transconjugant 1 showing the highest cleaved form (57%) compared to transconjugant 2 and 3, which showed 37% and 46%, respectively. The recognition and cleavage of the thrombin cleavage site in vivo by *P. tricornutum* were further confirmed using another construct C-T-YFP, which contained the gene of interest (*C: CBDAS*) linked to the YFP gene through the thrombin (*T*) sequence (Fig. 4b). This construct showed a total cleaved fraction (100%) with a size of 28.9 kDa (Fig. 5c). All C-T-YFP transconjugants displayed a corresponding size of YFP with a single HA tag instead of three, explained by the deletion of two HA tags, confirmed by DNA sequencing analysis (Table S7). Additionally, several substitutions were detected in the plasmid DNA sequence introduced into the diatom cells. However, all these substitutions were outside the coding sequence (CDS) and were not expected to affect the results. Altogether, these results confirm the in vivo cleavage of the thrombin cleavage site in *P. tricornutum*.

Table 1 Transconjugant positive strains screened for mCherry or YFP fluorescence

Transconjugants	mCherry ⁺ (FC)	YFP ⁺ (PR)	No. of selected strains
mCherry and YFP	69% (<i>n</i> = 36)	<i>n.t.</i>	6
mCherry-T-YFP	39% (<i>n</i> = 36)	<i>n.t.</i>	6
C-T-YFP	<i>n.a.</i>	25% (<i>n</i> = 36)	3

n.a. not applicable, *n.t.* not tested. Percentage of mCherry⁺ and YFP⁺ clones, *n* = 36; fluorescence was detected with a microplate reader (PR) or with a flow cytometer (FC)

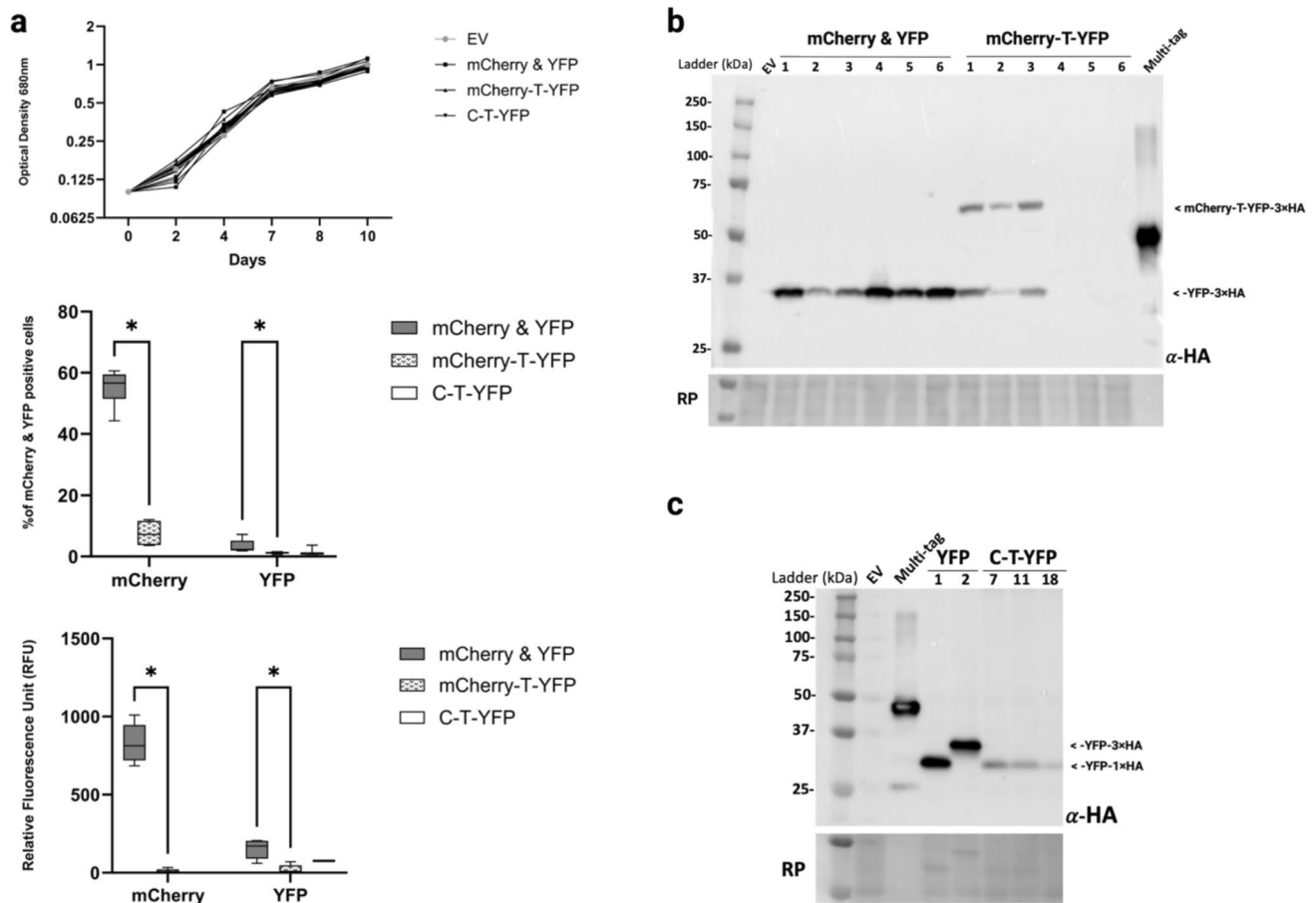


Fig. 5 Characterization of *P. tricornutum* clones and production of protein with specific cleavage of the thrombin cleavable sequence. **a** Upper panel, cell growth curves for each *P. tricornutum* strain of each construct, the optical density (OD) at 680 nm was monitored for 10 days. Middle panel, the mean percentage of mCherry and YFP positive cells were monitored for each construct by flow cytometer. Lower panel, mean relative fluorescence intensity of mCherry and YFP for each construct measured by plate reader. All strains were grown in L1 culture media supplemented with the appropriate antibiotic and incubated at 18 °C under gentle agitation of 130 rpm. **b**

Western blot analysis with an anti-HA tag antibody of total protein extracts from mCherry and YFP clones (32 kDa), mCherry-T-YFP clones (58 kDa), and Multi-tag (40 kDa). **c** Western blot with an anti-HA antibody of total protein extracts from C-T-YFP clones (28.9 kDa), YFP clones used as controls (1: 28.9 kDa and 2: 32 kDa) and Multi-tag (40 kDa). a-HA: anti-Hemagglutinin antibody, RP: Ponceau staining of the membrane is shown as a loading control. $n=6$ or $n=3$ number of biologically independent samples for each strain was plotted, t -test ($*p < 0.05$) was performed

Discussion

Cloning multiple genes to produce multiple proteins from a single vector has been challenging in the heterologous host *P. tricornutum*. Previous applied studies have developed a multi-gene expression system based on the *Thossea asigna* virus 2A self-cleaving peptide (2A), replacing older constructs reliant on the presence of internal ribosomal entry sites (IRES) (Sonenberg et al. 2016; Tang et al. 2016; Wang et al. 2015). The 2A-based multigene expression method relies on ribosomal skipping during translation, though varying cleavage efficiencies among different 2A sequences remain under investigation for future research (Liu et al. 2017). Hence, alternative methods are being explored.

In this study, we used several known thrombin and thrombin-like sequences from various species to identify putative thrombin sequences from *P. tricornutum* proteome database. Initially, seven sequences were identified with identities ranging from 25 to 32% and similarities from 39 to 47% compared to human thrombin. All sequences exhibited the characteristic trypsin domain of serine proteases family but the evidence supporting their classification as thrombin-like proteins remains insufficient, based on the current phylogenetic tree analysis. Structural modeling predicted their folding similarity to human and bovine thrombin domains, with conserved catalytic residues (H, D, and S). The predicted structures exhibited key features typical of serine proteases, along with specific elements characteristic of thrombin

proteases, such as a sodium binding site, an autolysis γ -loop, a 180 loop, and a covalent disulfide bond linking the catalytic serine to the oxyanion hole. The sodium binding site enhances catalytic efficiency, the autolysis γ -loop facilitates substrate, and regulator interactions essential for coagulation, the 180 loop is crucial for substrate recognition and specificity, and the disulfide bond stabilizes the active site structure (Johnson et al. 2005; Schechter and Berger 2012). Docking analysis indicated that the active site conformation was compatible for interaction with the thrombin-cleavage peptide and with potential cleavage between R and G of LVPRGS sequence in five putative *P. tricornutum* thrombin candidates.

While the domains of the *P. tricornutum* candidates displayed structural similarities to characterized thrombin, these findings did not preclude the possibility that these candidates are serine proteases without thrombin-like peptide cleavage activity. To confirm the cleavage capability of the thrombin sequence in *P. tricornutum*, we designed three constructs using two promoters (40SRPS8 and HASP1), with and without the LVPRGS cleavage site between two fluorescent proteins (mCherry and YFP), or between an enzyme (CBDAS) and YFP. Initially, we evaluated fluorescence levels associated with transgene expression. Notably, mCherry fluorescence was brighter under 40SRPS8 promoter (mCherry and YFP) than under HASP1 (mCherry-T-YFP) at day 10, possibly due to higher 40SRPS8 promoter activity in this growth phase rather than protein fusion or cleavage site addition. Although reports suggest detectable HASP1-dependent expression in the exponential phase, our findings indicated specificity to the late stationary phase (Erdene-Ochir et al. 2019). Furthermore, we assessed differences in YFP fluorescence (expressed under HASP1 in all constructs). YFP fluorescence levels were slightly lower in mCherry-T-YFP transconjugants compared to mCherry and YFP transconjugants, suggesting that fusion with mCherry reduced fluorescence intensity. Of note, the sensitivity of the method to measure YFP fluorescence was suboptimal due to significant overlap between YFP's excitation and emission spectra, limiting precise excitation and emission measurements at their peak wavelengths. The flow cytometer is equipped with a 488-nm laser, which excites only 41% of YFP (Inc. 2024), further constrain YFP emission strength.

Western blot analysis revealed that the constructs with the thrombin cleavage sequences in between two genes were cleaved, at a mean ratio of uncleaved/cleaved of ~53/47% for mCherry-T-YFP *P. tricornutum* transconjugants. This finding is consistent with the analysis of YFP fluorescence emission levels in mCherry-T-YFP clones. The fluorescence emitted by the uncleaved fraction may be lower than normal when two fluorescent proteins are fused and separated by a small distance (Zhuang et al. 2000). Interestingly, 100% of CBDAS-T-YFP clones showed cleaved proteins.

This homogeneity in CBDAS-T-YFP clones compared to mCherry-T-YFP clones may be due to lower efficiency in detecting larger proteins by Western blot. Further investigations must be done. Further functional studies are required to confirm the thrombin-like activity of these proteins, also to elucidate their cleavage efficiency and uncover the elements that could affect the cleavage in *P. tricornutum*.

The physiological role of these thrombin candidates in *P. tricornutum* is unknown. Intriguingly, candidates were also observed in various diatom orders. Mammals' thrombin proteases are involved in the blood clotting process. Aggregation is possible via activation of protease-activated receptors (PARs) (Coughlin 2000; Di Cera 2008). PARs are a subfamily of G protein-coupled receptors (GPCRs). These transmembrane proteins are activated by proteolytic cleavage such as performed by thrombin (Lindahl et al. 2017). GPCRs are involved in signal transduction pathways in response to external stimuli in mammalian cells (Tuteja 2009). Interestingly, the availability of complete genome sequence and comparative genomics studies have revealed the presence and the conservation of these receptors in four diatoms, including *P. tricornutum* (Bowler et al. 2008). While the function of GPCRs in diatoms has not been elucidated, previous findings suggest a role in intracellular signaling pathways in response to environmental changes. For example, gamma-aminobutyric acid (GABA) in plants plays a role in protecting against abiotic and biotic stress (Li et al. 2021). In diatoms, it may act in a similar way protecting against salinity or temperature changes (Allen et al. 2006). Other GPCRs could also contribute to biofilm formation and surface colonization in *P. tricornutum* (Fu et al. 2020). Interestingly, we observed that several thrombin candidates also encode for transmembrane domains. The presence of transmembrane domains in the predicted thrombin candidates in *P. tricornutum* suggests that they could co-localize with PARs and be involved in intercellular communication.

Future studies should investigate the diatom candidates and uncover their relative efficiency, as well as their specificity for the cleavage peptide to confirm the thrombin-like activity of these proteins. Overall, our results suggest that thrombin cleavage site linkers can be added to the molecular toolkit, allowing the expression of multi-gene cassettes in an episomal system in *P. tricornutum*. In comparison to 2A self-cleavage peptides, the use of thrombin strategy is based on proteolytic activity. Moreover, to enhance thrombin proteolytic activity, kinetic studies must be conducted to determine the best consensus substrate recognition sequence of thrombin in *P. tricornutum* (Gallwitz et al. 2012). This tool could be adapted to develop a new approach for assessing simultaneous in vivo functional and subcellular localization studies of endogenous or heterologous enzymes linked to a fluorescent protein in *P. tricornutum*.

To conclude, our results highlight for the first time in the microalgae diatom *P. tricornutum* the outcome and the potential of adding a thrombin cleavage sequence for the simultaneous expression of more than one gene under a single promoter. In studies, it will be interesting to unravel the functional role of these putative thrombin sequences in the diatom *P. tricornutum*. Our findings may contribute to the understanding of diatom's biology, which is important for both marine ecology and biotechnology. *P. tricornutum* is a promising bioengineering platform that will benefit from further studies of its endogenous processes.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00253-024-13322-z>.

Acknowledgements The authors wish to thank all the staff and students of the lab for their cooperation and kind support. Their contributions have improved the quality of this work.

Author contribution A.M.: conceptualization, investigation, laboratory work, data curation, formal analysis, methodology, and led manuscript drafting and editing; N.M.: conceptualization, investigation, laboratory work, data curation, project administration, and writing—review and editing; L.B.: laboratory work, formal analysis, and writing—review and editing; E.F.: laboratory work, formal analysis, and writing—review and editing; F.M.-M.: project administration and writing—review and editing; I.D.-P.: conceptualization, funding acquisition, resources, supervision, project administration, writing—review and editing, and resources.

Funding The authors acknowledge that financial support for this research work was funded by the Canada Research Chair on plant specialized metabolism Award No 950–232164 to I.D.-P. Many thanks are extended to the Canadian taxpayers and to the Canadian government for supporting the Canada Research Chairs Program.

Data availability The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files. Should any raw data files be needed in another format, they are available from the corresponding author upon reasonable request.

Declarations

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Allen AE, Vardi A, Bowler C (2006) An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr Opin Plant Biol* 9(3):264–273. <https://doi.org/10.1016/j.pbi.2006.03.013>
- Amet N, Lee HF, Shen WC (2009) Insertion of the designed helical linker led to increased expression of tf-based fusion proteins. *Pharm Res* 26(3):523–528. <https://doi.org/10.1007/s11095-008-9767-0>
- Araujo AP, Oliva G, Henrique-Silva F, Garratt RC, Caceres O, Beltrami LM (2000) Influence of the histidine tail on the structure and activity of recombinant chlorocatechol 1,2-dioxygenase. *Biochem Biophys Res Commun* 272(2):480–484. <https://doi.org/10.1006/bbrc.2000.2802>
- Argos P (1990) An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *J Mol Biol* 211(4):943–958. [https://doi.org/10.1016/0022-2836\(90\)90085-Z](https://doi.org/10.1016/0022-2836(90)90085-Z)
- Awwad F, Fantino EI, Héneault M, Diaz-Garza AM, Merindol N, Custeau A, Gélinas S-E, Meddeb-Mouelhi F, Li J, Lemay J-F, Karas BJ, Desgagne-Penix I (2023) Bioengineering of the marine diatom *Phaeodactylum tricornutum* with Cannabis genes enables the production of the cannabinoid precursor, olivetolic acid. *Int J Mol Sci*. <https://doi.org/10.3390/ijms242316624>
- Belardinelli JM, Jackson M (2017) Green fluorescent protein as a protein localization and topological reporter in mycobacteria. *Tuberculosis (Edinb)* 105:13–17. <https://doi.org/10.1016/j.tube.2017.04.001>
- Bhandari M, Ofosu FA, Mackman N, Jackson C, Doria C, Humphries JE, Babu SC, Ortel TL, Hoffman Van Thiel D, Walenga JM, Wahi R, Teoh KH, Fareed J (2011) Safety and efficacy of thrombin-JMI: a multidisciplinary expert group consensus. *Clin Appl Thromb Hemost* 17(1):39–45. <https://doi.org/10.1177/1076029610385674>
- Bode W, Mayr I, Baumann U, Huber R, Stone SR, Hofsteenge J (1989) The refined 1.9 Å crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J* 8(11):3467–3475. <https://doi.org/10.1002/j.1460-2075.1989.tb08511.x>
- Bouabe H, Fassler R, Heesemann J (2008) Improvement of reporter activity by IRES-mediated polycistronic reporter system. *Nucleic Acids Res* 36(5):e28. <https://doi.org/10.1093/nar/gkm1119>
- Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Dettler JC, Durkin C, Falcatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y, Grigoriev IV (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456(7219):239–244. <https://doi.org/10.1038/nature07410>
- Butler T, Kapoore RV, Vaidyanathan S (2020) *Phaeodactylum tricornutum*: a diatom cell factory. *Trends Biotechnol* 38(6):606–622. <https://doi.org/10.1016/j.tibtech.2019.12.023>
- Castellino FJ, McCance SG (1997) The kringle domains of human plasminogen. *Ciba Found Symp* 212:46–60. <https://doi.org/10.1002/9780470515457.ch4>

- Cheng CM, Meyer-Massetti C, Kayser SR (2009) A review of three stand-alone topical thrombins for surgical hemostasis. *Clin Ther* 31(1): <https://doi.org/10.1016/j.clinthera.2009.01.005>
- Coughlin SR (2000) Thrombin signalling and protease-activated receptors. *Nature* 407:258–264. <https://doi.org/10.1038/35025229>
- Danckwardt S, Hentze MW, Kulozik AE (2013) Pathologies at the nexus of blood coagulation and inflammation: thrombin in hemostasis, cancer, and beyond. *J Mol Med (Berl)* 91(11):1257–1271. <https://doi.org/10.1007/s00109-013-1074-5>
- Demain AL, Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. *Biotechnol Adv* 27(3):297–306. <https://doi.org/10.1016/j.biotechadv.2009.01.008>
- Dhaouadi F, Awwad F, Diamond A, Desgagné-Penix I (2020) Diatoms' breakthroughs in biotechnology: *Phaeodactylum tricornutum*; as a model for producing high-added value molecules. *Am J Plant Sci* 11(10):1632–1670. <https://doi.org/10.4236/ajps.2020.1110118>
- Di Cera E (2008) Thrombin. *Mol Aspects Med* 29(4):203–254. <https://doi.org/10.1016/j.mam.2008.01.001>
- Diamond A, Diaz-Garza AM, Li J, Slattery SS, Merindol N, Fantino E, Meddeb-Mouelhi F, Karas BJ, Barnabé S, Desgagné-Penix I (2023) Instability of extrachromosomal DNA transformed into the diatom *Phaeodactylum tricornutum*. *Algal Research* 70. <https://doi.org/10.1016/j.algal.2023.102998>
- Dos Santos TW, Goncalves PA, Rodriguez D, Pereira JA, Martinez CAR, Leite LCC, Ferraz LFC, Converso TR, Darrieux M (2022) A fusion protein comprising pneumococcal surface protein A and a pneumolysin derivative confers protection in a murine model of pneumococcal pneumonia. *PLoS ONE* 17(12):e0277304. <https://doi.org/10.1371/journal.pone.0277304>
- Ellis RW, Rappuoli R, Ahmed S (2013) Technologies for making new vaccines. In: *Vaccines* (sixth edition), pp 1182–1199. <https://doi.org/10.1016/B978-1-4557-0090-5.00013-6>
- Erdene-Ochir E, Shin BK, Kwon B, Jung C, Pan CH (2019) Identification and characterisation of the novel endogenous promoter HASP1 and its signal peptide from *Phaeodactylum tricornutum*. *Sci Rep* 9(1):9941. <https://doi.org/10.1038/s41598-019-45786-9>
- Fabris M, George J, Kuzhiumparambil U, Lawson CA, Jaramillo-Madrid AC, Abbriano RM, Vickers CE, Ralph P (2020) Extrachromosomal genetic engineering of the marine diatom *Phaeodactylum tricornutum* enables the heterologous production of monoterpenoids. *ACS Synth Biol* 9(3):598–612. <https://doi.org/10.1021/acssynbio.9b00455>
- Fantino E, Awwad F, Merindol N, Diaz Garza AM, Gélinas S-E, Gajón Robles GC, Custeau A, Meddeb-Mouelhi F, Desgagné-Penix I (2024) Bioengineering *Phaeodactylum tricornutum*, a marine diatom, for cannabinoid biosynthesis. *Algal Res* 77:103379. <https://doi.org/10.1016/j.algal.2023.103379>
- Freedman SJ, Blostein MD, Baleja JD, Jacobs M, Furie BC, Furie B (1996) Identification of the phospholipid binding site in the vitamin K-dependent blood coagulation protein factor IX. *J Biol Chem* 271(27):16227–16236. <https://doi.org/10.1074/jbc.271.27.16227>
- Fu W, Chaiboonchoe A, Dohai B, Sultana M, Baffour K, Alzahmi A, Weston J, Al Khairy D, Daakour S, Jaiswal A, Nelson DR, Mystikou A, Brynjolfsson S, Salehi-Ashtiani K (2020) GPCR genes as activators of surface colonization pathways in a model marine diatom. *iScience* 23(8):101424. <https://doi.org/10.1016/j.isci.2020.101424>
- Gallwitz M, Enoksson M, Thorpe M, Hellman L (2012) The extended cleavage specificity of human thrombin. *PLoS ONE* 7(2):e31756. <https://doi.org/10.1371/journal.pone.0031756>
- George RA, Heringa J (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng* 15(11):871–879. <https://doi.org/10.1093/protein/15.11.871>
- Huston JS, Levinson D, Mudgett-Hunter M, Tai MS, Novotny J, Margolies MN, Ridge RJ, Brucoleri RE, Haber E, Crea R et al (1988) Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Escherichia coli*. *Proc Natl Acad Sci U S A* 85(16):5879–5883. <https://doi.org/10.1073/pnas.85.16.5879>
- Inc. AB (2024) Spectrum [YFP]. Publisher. <https://www.aatbio.com/fluorescence-excitation-emission-spectrum-graph-viewer/yfp>. Accessed 14 Sept 2024
- Javaid M, Haleem A, Singh RP, Suman R (2023) Sustaining the healthcare systems through the conceptual of biomedical engineering: a study with recent and future potentials. *Biomed Technol* 1:39–47. <https://doi.org/10.1016/j.bmt.2022.11.004>
- Jenny RJ, Mann KG, Lundblad RL (2003) A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expr Purif* 31(1):1–11. [https://doi.org/10.1016/S1046-5928\(03\)00168-2](https://doi.org/10.1016/S1046-5928(03)00168-2)
- Johnson DJ, Adams TE, Li W, Huntington JA (2005) Crystal structure of wild-type human thrombin in the Na⁺-free state. *Biochem J* 392(Pt 1):21–28. <https://doi.org/10.1042/BJ20051217>
- Karas BJ, Diner RE, Lefebvre SC, McQuaid J, Phillips AP, Noddings CM, Brunson JK, Valas RE, Deerinck TJ, Jablanovic J, Gillard JT, Beeri K, Ellisman MH, Glass JJ, Hutchison CA 3rd, Smith HO, Venter JC, Allen AE, Dupont CL, Weyman PD (2015) Designer diatom episomes delivered by bacterial conjugation. *Nat Commun* 6:6925. <https://doi.org/10.1038/ncomms7925>
- Kimple ME, Brill AL, Pasker RL (2013) Overview of affinity tags for protein purification. *Curr Protoc Protein Sci* 73:9. <https://doi.org/10.1002/0471140864.ps0909s73>
- Li L, Dou N, Zhang H, Wu C (2021) The versatile GABA in plants. *Plant Signal Behav* 16(3):1862565. <https://doi.org/10.1080/15592324.2020.1862565>
- Lindahl T, Bjerke M, Ramström S, Vretenbrant K (2017) Platelet activation via PAR4 is involved in the initiation of thrombin generation and in clot elasticity development. *Thromb Haemost* 97(03):417–424. <https://doi.org/10.1160/th06-07-0397>
- Liu Z, Chen O, Wall JBJ, Zheng M, Zhou Y, Wang L, Vaseghi HR, Qian L, Liu J (2017) Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci Rep* 7(1):2193. <https://doi.org/10.1038/s41598-017-02460-2>
- Lundblad RL, Bradshaw RA, Gabriel D, Ortel TL, Lawson J, Mann KG (2004) A review of the therapeutic uses of thrombin. *Thromb Haemost* 91(5):851–860. <https://doi.org/10.1160/TH03-12-0792>
- Majhi BB, Gélinas S-E, Méridol N, Desgagné-Penix I (2023) Characterization of norbelladine synthase and noroxomaritidine/nor-craugosidine reductase reveals a novel catalytic route for the biosynthesis of *Amaryllidaceae* alkaloids including the Alzheimer's drug galanthamine. *Front Plant Sci* 14:1231809. <https://doi.org/10.3389/fpls.2023.1231809>
- Mathews II, Padmanabhan KP, Ganesh V, Tulinsky A, Ishii M, Chen J, Turck CW, Coughlin SR, Fenton JW 2nd (1994) Crystallographic structures of thrombin complexed with thrombin receptor peptides: existence of expected and novel binding modes. *Biochemistry* 33(11):3266–3279. <https://doi.org/10.1021/bi00177a018>
- McCormick AL, Thomas MS, Heath AW (2001) Immunization with an interferon- γ -gp120 fusion protein induces enhanced immune responses to human immunodeficiency virus gp120. *J Infect Dis*. <https://doi.org/10.1086/324371>
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold: making protein folding accessible to all. *Nat Methods* 19(6):679–682. <https://doi.org/10.1038/s41592-022-01488-1>
- Okonechnikov K, Golosova O, Fursov M, team U (2012) UniProt KEGG: a unified bioinformatics toolkit. *Bioinformatics* 28(8):1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>
- Rao MB, Tanksale AM, Ghatge MS, Deshpande VV (1998) Molecular and biotechnological aspects of microbial proteases. *Microbiol*

- Mol Biol Rev 62(3):597–635. <https://doi.org/10.1128/MMBR.62.3.597-635.1998>
- Rechsteiner M, Rogers SW (1996) PEST sequences and regulation by proteolysis. Trends Biochem Sci 21(7):267–271. [https://doi.org/10.1016/S0968-0004\(96\)10031-1](https://doi.org/10.1016/S0968-0004(96)10031-1)
- Schechter I, Berger A (2012) On the size of the active site in proteases. I. Papain. 1967. Biochem Biophys Res Commun 425(3):497–502. <https://doi.org/10.1016/j.bbrc.2012.08.015>
- Schoch CL, Ciufo S, Domrachev M, Hottot CL, Kannan S, Khovan-skaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi M (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020:62
- Sharma N, Simon DP, Diaz-Garza AM, Fantino E, Messaabi A, Meddeb-Mouelhi F, Germain H, Desgagné-Penix I (2021) Diatoms biotechnology: various industrial applications for a greener tomorrow. Front Marine Sci 8. <https://doi.org/10.3389/fmars.2021.636613>
- Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41(Database issue):D344–D347. <https://doi.org/10.1093/nar/gks1067>
- Slattery SS, Diamond A, Wang H, Therrien JA, Lant JT, Jazey T, Lee K, Klassen Z, Desgagne-Penix I, Karas BJ, Edgell DR (2018) An expanded plasmid-based genetic toolbox enables Cas9 genome editing and stable maintenance of synthetic pathways in *Phaeodactylum tricornutum*. ACS Synth Biol 7(2):328–338. <https://doi.org/10.1021/acssynbio.7b00191>
- Slattery SS, Giguere DJ, Stuckless EE, Shrestha A, Briere LK, Galbraith A, Reaume S, Boyko X, Say HH, Browne TS, Frederick MI, Lant JT, Heinemann IU, O'Donoghue P, Dsouza L, Martin S, Howard P, Jedeszko C, Ali K, Styba G, Flatley M, Karas BJ, Gloor GB, Edgell DR (2022) Phosphate-regulated expression of the SARS-CoV-2 receptor-binding domain in the diatom *Phaeodactylum tricornutum* for pandemic diagnostics. Sci Rep 12(1):7010. <https://doi.org/10.1038/s41598-022-11053-7>
- Smale ST (2010a) Beta-galactosidase assay. Cold Spring Harb Protoc 2010(5):pdb.prot5423. <https://doi.org/10.1101/pdb.prot5423>
- Smale ST (2010b) Luciferase assay. Cold Spring Harb Protoc 2010(5):5421. <https://doi.org/10.1101/pdb.prot5421>
- Snape MD, Dawson T, Oster P, Evans A, John TM, Ohene-Kena B, Findlow J, Yu L-M, Borrow R, Ypma E, Toneatto D, Pollard AJ (2010) Immunogenicity of two investigational serogroup B meningococcal vaccines in the first year of life. Pediatr Infect Dis J 29(11):e71–e79. <https://doi.org/10.1097/inf.0b013e3181f59f6d>
- Sonenberg N, Svitkin Y, Siddiqui N (2016) Internal ribosome entry site-mediated translation encyclopedia of cell biology. In: Bradsha RA, Stahl PD (eds) Encyclopedia of Cell Biology, Academic Press, Waltham, pp 307–316
- Takagi T, Doolittle RF (1974) Amino acid sequence studies on factor XIII and the peptide released during its activation by thrombin. Biochemistry 13(4):750–756. <https://doi.org/10.1021/bi00701a018>
- Takamatsu N, Watanabe Y, Yanagi H, Meshi T, Shiba T, Okada Y (1990) Production of enkephalin in tobacco protoplasts using tobacco mosaic virus RNA vector. FEBS Lett 269(1):73–76. [https://doi.org/10.1016/0014-5793\(90\)81121-4](https://doi.org/10.1016/0014-5793(90)81121-4)
- Tamura K, Stecher G, Kumar S, Battistuzzi FU (2021) MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38(7):3022–3027. <https://doi.org/10.1093/molbev/msab120>
- Tang X, Liu X, Tao G, Qin M, Yin G, Suo J, Suo X (2016) “Self-cleaving” 2A peptide from porcine teschovirus-1 mediates cleavage of dual fluorescent proteins in transgenic *Eimeria tenella*. Vet Res 47(1):68. <https://doi.org/10.1186/s13567-016-0351-z>
- Tuteja N (2009) Signaling through G protein coupled receptors. Plant Signal Behav 4(10):942–947. <https://doi.org/10.4161/psb.4.10.9530>
- Wang Y, Wang F, Wang R, Zhao P, Xia Q (2015) 2A self-cleaving peptide-based multi-gene expression system in the silkworm *Bombyx mori*. Sci Rep 5:16273. <https://doi.org/10.1038/srep16273>
- Wang S, Ju X, Heuler J, Zhang K, Duan Z, Patabendige HMLW, Zhao S, Sun X (2023) Recombinant fusion protein vaccine containing *Clostridioides difficile* FliC and FliD protects mice against *C. difficile* infection. Am Soc Microbiol. <https://doi.org/10.1128/iai.00169-22>
- Wei Wu J, Bancroft D, Suttie JW (1997) Structural features of the kringle domain determine the intracellular degradation of under-glycosylated prothrombin: studies of chimeric rat/human prothrombin. Cell Biology 94. <https://doi.org/10.1073/pnas.94.25.1365>
- Zhao HL, Yao XQ, Xue C, Wang Y, Xiong XH, Liu ZM (2008) Increasing the homogeneity, stability and activity of human serum albumin and interferon-alpha2b fusion protein by linker engineering. Protein Expr Purif 61(1):73–77. <https://doi.org/10.1016/j.pep.2008.04.013>
- Zhuang X, Ha T, Kim HD, Centner T, Labeit S, Chu S (2000) Fluorescence quenching a tool for single-molecule protein-folding study. Biophys Comput Biol. <https://doi.org/10.1073/pnas.97.26.14241>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.