Supplementary Information

Post-translational Claisen Condensation and Decarboxylation en Route to the Bicyclic Core of Pantocin A.

Swapnil V. Ghodge[1], Kristen A. Biernat[2], Sarah Jane Bassett[1], Matthew R. Redinbo[2], and Albert A. Bowers[1]

[1]Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Table of contents

Materials and methods

Figures and tables

Materials and methods

Protected amino acids were purchased from Chempep Inc., buffers and solvents were purchased from Fischer Scientific, and the remaining chemicals and reagents were purchased from Sigma-Aldrich, unless mentioned otherwise. Preparatory HPLC was performed in a Shimadzu UFLC CBM-20A with a dual channel wavelength detector at 220 nm and 260 nm with a LUNA 10 u C18(2) 100 A, AXIA (Phenomenex®) semi-preparatory column with a 15 mL/min flow rate. Purification was carried out with a two solvent system (solvent A = 0.1% trifluoroacetic acid in water; solvent B = 0.1% trifluoroacetic acid in acetonitrile) using gradient flow. LC-MS analysis was done using Kinetex 2.6u C18 column, and mass spectrometry measurements were recorded using an Agilent 6520 Accurate-Mass Q-TOF ESI positive in high-resolution mode. Predicted masses were extracted to ±5 ppm.

Cloning PaaA. PaaA gene was amplified from the cosmid template pCPP702 using Phusion® High-fidelity mastermix with HF buffer (NEB) following manufacturer's instructions. The primer sequences used for PCR reactions were as follows:
Forward: 5'-GATCGATCCCATGGGCATGTCACTAACGAATGTTAAACC-3'
Reverse: 5'-CTCGAGGATCGATCTTTAAAAATCTCT-3'

The PCR product was digested using NcoI and XhoI restriction enzymes (NEB) according to standard procedures. The vector pET28b was digested with the same two restriction enzymes as above, followed by treatment with Antarctic Phosphatase. The digested vector and insert were incubated together along with T4-DNA ligase in 1x-T4 ligase buffer, and the reaction was transformed into One-Shot® Top10 *E.coli* competent cells (Agilent). A single colony was picked and transferred to LB medium containing 50 µg/mL kanamycin and incubated overnight at 37 °C. The PaaA-pET28b cloned plasmid was purified using Purelink™ plasmid purification kit (Invitrogen) and stored at -20 °C.

Expression and Purification of PaaA. PaaA was expressed in BL21-Gold (DE3) *E. coli* cells (Invitrogen) in LB medium containing 40 µg/mL kanamycin. Cells were grown at 37°C until an OD_{600} of 0.5 was attained, at which point the temperature was lowered to 18°C. At an OD_{600} of 0.8, protein expression was induced with the addition of 0.1 mM isopropyl-1-thio-D-galactopyranoside (IPTG) and incubated overnight at 18°C. Cells were collected by centrifugation at 4500xg for 20 min at 4°C in a Sorvall (model RC-3B) swinging bucket centrifuge. Cell pellets were resuspended in Buffer A (20 mM Tris, pH 7.5, 30 mM imidazole, 300 mM NaCl, 5% Glycerol, 1 mM DTT) and a Roche complete-EDTA free protease inhibitor tablet per 50 mL buffer. Resuspended cells were sonicated and centrifuged at 14,500xg for 60 min in a Sorvall (model RC-5B) centrifuge to clarify the lysate. The lysate was passed through a Ni-NTA HP column (GE Healthcare), loaded

onto the Äktaxpress FPLC system (Amersham Biosciences) and washed with Buffer A. Protein was eluted with Buffer B (20 mM Tris, pH 7.5, 500 mM imidazole, 300 mM NaCl, 5% glycerol, 1 mM DTT). Purity of eluted fractions was assessed by SDS-PAGE. Purest fractions were combined and passed over a HiLoad™ 16/60 Superdex™ 200 gel filtration column. The protein was eluted into 50 mM HEPES, pH 7.5, 150 mM NaCl, 5% glycerol, and 1 mM DTT. Fractions (1.5 mL each) were collected based on highest ultraviolet absorbance at 280 nm. Fractions were analyzed by SDS-PAGE (which indicated >95% purity), combined, and concentrated using a Milllipore centrifugal filter at 3000xg to 15 mg/mL.

To express selenomethionine-substituted PaaA, SelenoMet^{TM} Medium and Nutrient Mix (AthenaES) was prepared for growth, with 60 mg of selenomethionine added for each liter of medium. Cells were grown at 37°C until an OD_{600} of 0.7 was attained and then induced with 0.3 mM IPTG. The cultures were then grown overnight at 18°C. Purification was performed as for the wild-type enzyme (see above).

Crystal Structure Determination of PaaA. Crystals of selenomethionine PaaA were grown at room temperature using the hanging-drop diffusion method by mixing the protein (15 mg/mL) with reservoir solution containing 12% PEG4000, 0.15 M $MgSO_4$, and 10% glycerol. Crystals were briefly soaked in a solution containing 12% PEG4000, 0.15 M $MgSO_4$, and 20% glycerol before plunging into liquid nitrogen in preparation for x-ray data collection. Diffraction data were collected on the 23-IDD beam line at GM/CA-XSD (Advanced Photon Source, Argonne National Laboratory). The selenomethionine crystals exhibited at $P2_12_12$ space group, and the asymmetric unit contained two monomers. The PHENIX software suite (AutoSol) was utilized to locate heavy atom sites, and the model was built by hand in Coot using the SAD data. Monomers A and B both exhibited clearly interpretable electron density within the C-terminal domains; by contrast, the N-terminal domains were less ordered, which is expected in this apo structure without PaaP bound. Data collection and refinement statistics are presented in **Table S1**.

Site-directed Mutagenesis. PaaA mutants were created using PCR mutagenesis in which K134, R174, or K187 was replaced with alanine. Primers were synthesized by Integrated DNA Technologies. Primer sequences are as follows:

R174A. Forward: 5'-CGAGTTGTCAAACATCAAC**GCA**CAGGTGTTGTATCGCACA -3'
Reverse: 5'-TGTGCGATACAACACCTG**TGC**GTTGATGTTTGACAACTCG-3'

K187A. Forward: 5'-CAGATGATGTAGGCAAGAAC**GCA**GTTGATGCGGCCAAAGATA-3'
Reverse: 5'-TATCTTTGGCCGCATCAAC**TGC**GTTCTTGCCTACATCATCTG-3'

R134A. Forward: 5'-GTATCAGGAACGGCTAAAGCAAAGT**GCG**GTAGCCATCTTC-3'

Reverse: 5'-GAAGATGGCTAC**CGC**ACTTTGCTTTAGCCGTTCCTGATAC-3'

Plasmids from isolated colonies were sequenced by Eton Biosciences Inc. to confirm the correct mutations were introduced into the gene. Expression and purification was performed as for the wild-type enzyme (see above).

Solid-phase Peptide Synthesis of PaaP and Truncates. Peptide synthesis was carried out using solid-phase peptide synthesis on a Biotage Initiator+ Alstra microwave peptide synthesizer using either the Rinkamide resin from Chempep Inc. or Rinkamide Chemmatrix resin from Biotage. The peptides were synthesized from the C to the N-terminus with a C-terminus amide functional group using standard fluorenylmethoxycarbonyl (Fmoc) chemistry. This process employed N-α-Fmoc-L-amino acids, with the coupling agents HATU and HOAt as 0.4 M solutions in DMF, and 5 min microwave irradiation was used for the addition of each amino acid at 75 °C. Fmoc deprotection was achieved using 20% (v/v) piperidine/DMF for 3 min and 10 min, with DMF wash at the end of each deprotection step. After the final deprotection to obtain a primary amine at the N-terminus of the synthesized peptide, the resin was washed with dichloromethane and dried. The final peptide product was cleaved using a cleavage cocktail containing 94 or 95% (v/v) TFA, 2.5% (v/v) water, 2.5% (v/v) triisopropylsilane (TIPS), and 1 or 0% (v/v) ethanedithiol (EDT). EDT was added when the peptide sequence contained at least one methionine. Cleavage reactions were carried out at room temperature for about 3 hours in syringes fitted with frits. The cleaved peptide was precipitated by the dropwise addition of the cleavage reaction mixture into cold diethyl ether. The diethyl ether was evaporated using a speedvac; the dried solid was dissolved in DMSO, and diluted with a mixture of 50% v/v acetonitrile-water. This crude product was then purified using preparatory HPLC, and the fractions were analyzed using the UV trace at 220 nm and LC-MS. In general, the synthesized peptides eluted in a range between 33%-41% buffer B. Fractions containing the relatively pure product were dried using Rotavap rotary evaporator, redissolved in methanol, and transferred to clean eppendorf tubes. The product was then dried in the speedvac, weighed and stored at -20 °C.

Enzymatic Assays of PaaA using LC-MS. In general, 25 μM PaaP was incubated with 2.5 μM PaaA for ~16 hours in the presence of 50 mM HEPES pH 7.5, 4 mM $MgCl_2$, 1 mM ATP, and 0.5 mM TCEP, alongside a control containing no PaaA. These solutions were analyzed using LC-MS. Assays for comparing the activity of PaaA variants were carried out by incubating 25 μM PaaP with 1 μM PaaA for ~6 hours in the presence of 50 mM HEPES pH 7.5, 4 mM $MgCl_2$, and 0.5 mM ATP.

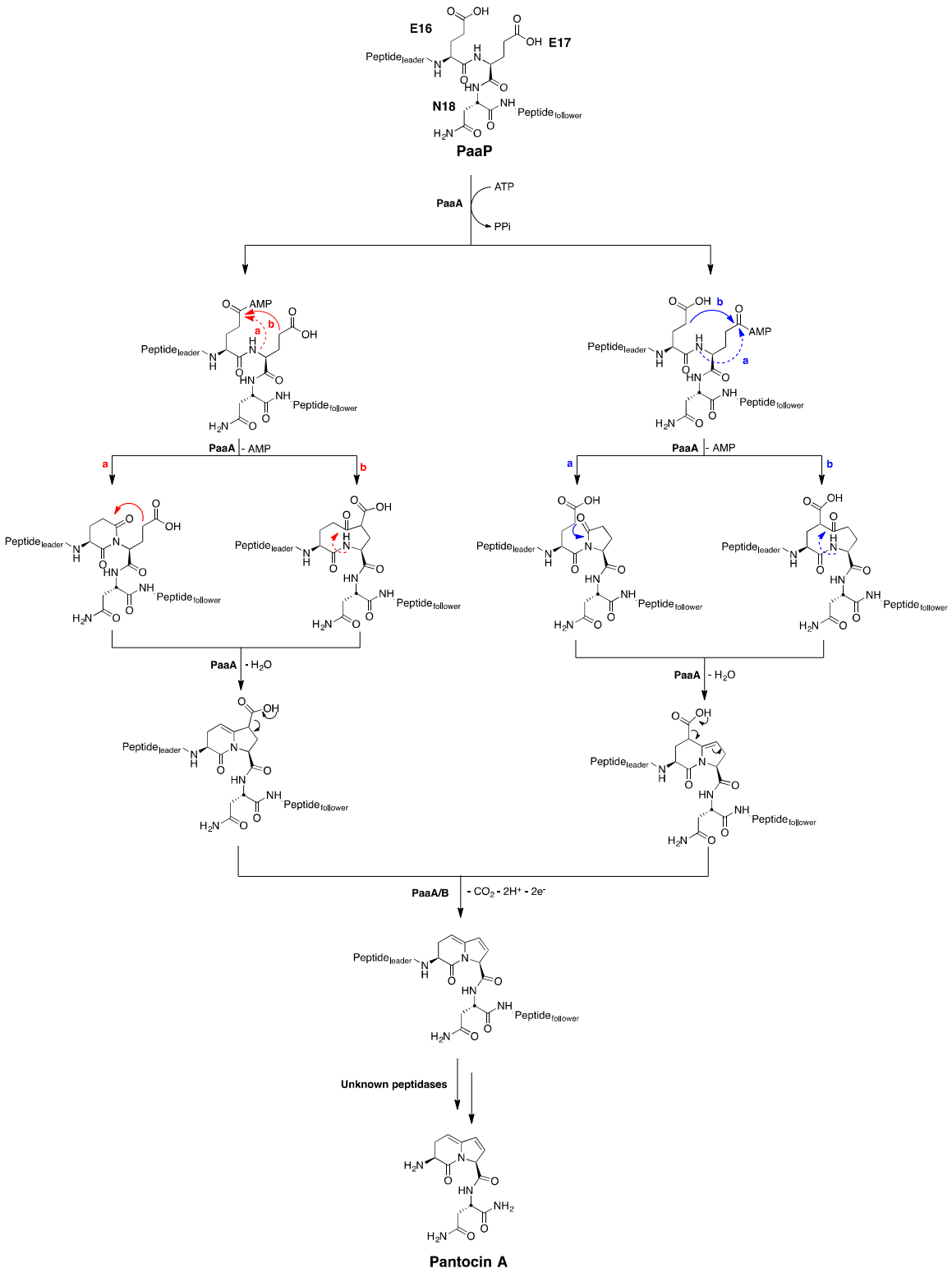
Limited Proteolysis of PaaA. Protease digestion reactions (40 μL) were carried

out at room temperature using a final PaaA concentration of 0.6 mg/mL in Buffer C (20 mM Tris, pH 7.5, 150 mM NaCl). For reactions containing the peptide ligand, M1-S30 or M1-Q25 was added in 5:1 molar ratio to PaaA. Stocks of trypsin and chymotrypsin (1 mg/mL) were prepared in 0.1 mM HCl and added to a final concentration of 1.4 μg/mL to initiate the reactions. The reactions were quenched by the addition of phenylmethylsulfonyl fluoride (PMSF, 100 mM) and Laemmli loading dye at 0, 1, 5, 15, 30, 60, and 120 minutes. Samples from each time point were assessed by SDS-PAGE (**Figure S8**).

Bioinformatic Analysis of PaaA and Related Enzymes. The sequence similarity network diagram of enzyme sequences closely related to PaaA was generated using the EFI-EST (Enzyme function initiative- enzyme similarity tool) online tool. PaaA sequence was used as a query for BLAST search by the EFI-EST server, which generated a network diagram containing ~4500 redundant enzyme sequences similar to PaaA ranging between 230-420 amino acids in length. A smaller, and more manageable network diagram containing 635 sequences was chosen for study in which each node represented sequences bearing ≥60% sequence identity to each other. A BLAST E-value of 1×10^{-45} was arbitrarily chosen for representation. Representative sequences from each cluster of nodes were used as a query in the HHPred online database to verify whether they possessed the N-terminal RRE or leader-peptide recognition domain. A list of the enzyme sequence identifiers that were determined to possess an RRE domain fused to an E-1 domain are provided in **Table S4**, in conjunction with **Figure S10**.

NMR analysis of the product of PaaA-catalyzed modification of PaaP. A 32 mL reaction containing 25 μM PaaP, 2 μM PaaA, 4 mM $MgCl_2$, and 0.5 mM ATP in 50 mM HEPES buffer, pH 7.5, was incubated at room temperature for ~16 hours. The solution was subjected to LC-MS analysis to confirm formation of the product. The reaction mixture was split and transferred into three 20-mL scintillation vials and was lyophilized. The solid obtained was treated with ~0.8 mL DMSO to redissolve the peptide product. The DMSO solution was centrifuged to separate the undissolved solids and the supernatant was further diluted to ~3 mL using 1:4 mixture of acetonitrile-water. The peptide product was purified using preparatory-scale HPLC, using a procedure similar to the purification of peptides after solid phase synthesis. Eluent was monitored based on the absorbance at 220 nm, and the eluted fraction (~38% buffer B) displaying significant absorbance was selected. Presence of product was confirmed by LC-MS. The fraction was partially concentrated in a rotary evaporator, followed by flash freezing and lyophilization to obtain the purified solid product of PaaA reaction (~1.5 mg solid). This solid was redissolved in 180 μL DMSO-d_6 for NMR analysis. NMR data was acquired

using Bruker Avance III-HD NMR spectrometers with a magnetic field of 850 MHz (for one-dimensional ^{1}H-NMR spectroscopy) or 700 MHz (for ^{13}C-HSQC spectroscopy) respectively.


E16
E17
N18
PaaP
PaaA
ATP
PPi
AMP
PaaA - AMP
PaaA - H2O
PaaA/B
- CO2 - 2H+ - 2e-
Unknown peptidases
Pantocin A

Scheme S1.

Table S1. Crystallographic statistics for the selenomethionine PaaA structure

Resolution (highest shell), Å	29.4-2.93 (3.11-2.93)
Space group	$P2_12_12$
Unit cell, Å	68.9, 170.4, 64.3
Total reflections	427,046
Unique reflections	16,878
Multiplicity	25.3 (25.1)
Completeness, %	99.7 (99.4)
Mean I/sigma (I)	24.9 (6.9)
Wilson B-factor, $Å^2$	78.7
R_{sym}	0.11 (0.60)
R	0.210
R_{free}	0.289
No. of atoms	4,760
rms bonds, Å	0.009
rms angles, °	1.30
Ramachandran favored, %	93.4
Ramachandran outliers, %	0.99
Clash score	19.2
Average B-factor, $Å^2$	66.3

Table S2. Disordered regions (residue numbers) not displayed in the Se-Met PaaA structure.

Chain A	**Chain B**
1-10	1-5
53-57	17-55
66-68	67-72
233-241	233-241
287-310	289-298
375-381	373-381

Table S3. Average B-factors of the N-terminal and C-terminal domains in the Se-Met PaaA structure.

Domain*	Average B-factor**
Chain A, residues 1-98	64.7
Chain A, residues 99-381	53.9
Chain B, residues 1-98	68.3
Chain B, residues 99-381	56.5

* Residues 1-98 constitute the RRE domain while 99-381 constitute the adenylation domain.
**Residues that were disordered within each domain did not contribute to the average B-factor.

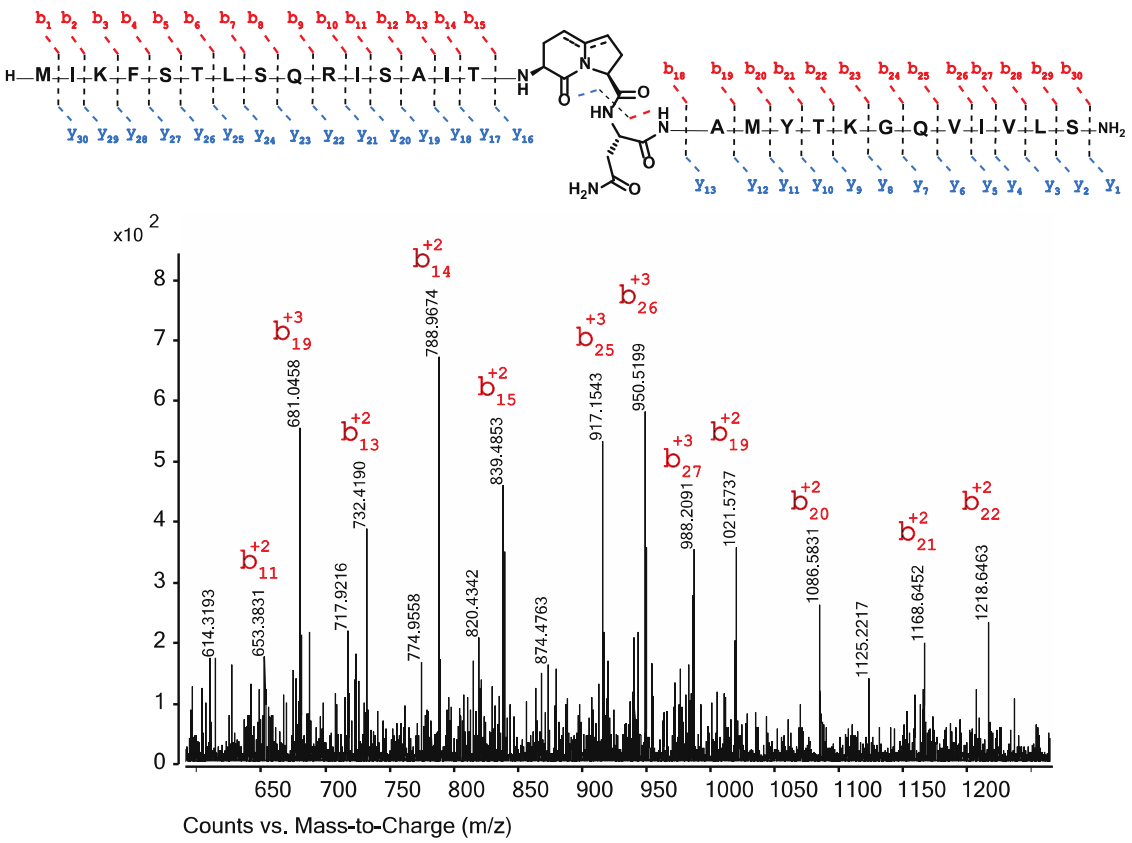


Figure S1. LC-MS/MS analysis of the product of PaaA-catalyzed transformation of PaaP. $[M + 4H^+]^{+4}$ = 820.2010 ion was selected for fragmentation using a collision energy of 30 eV. The b-ions are shown in red while the y-ions are indicated in blue.

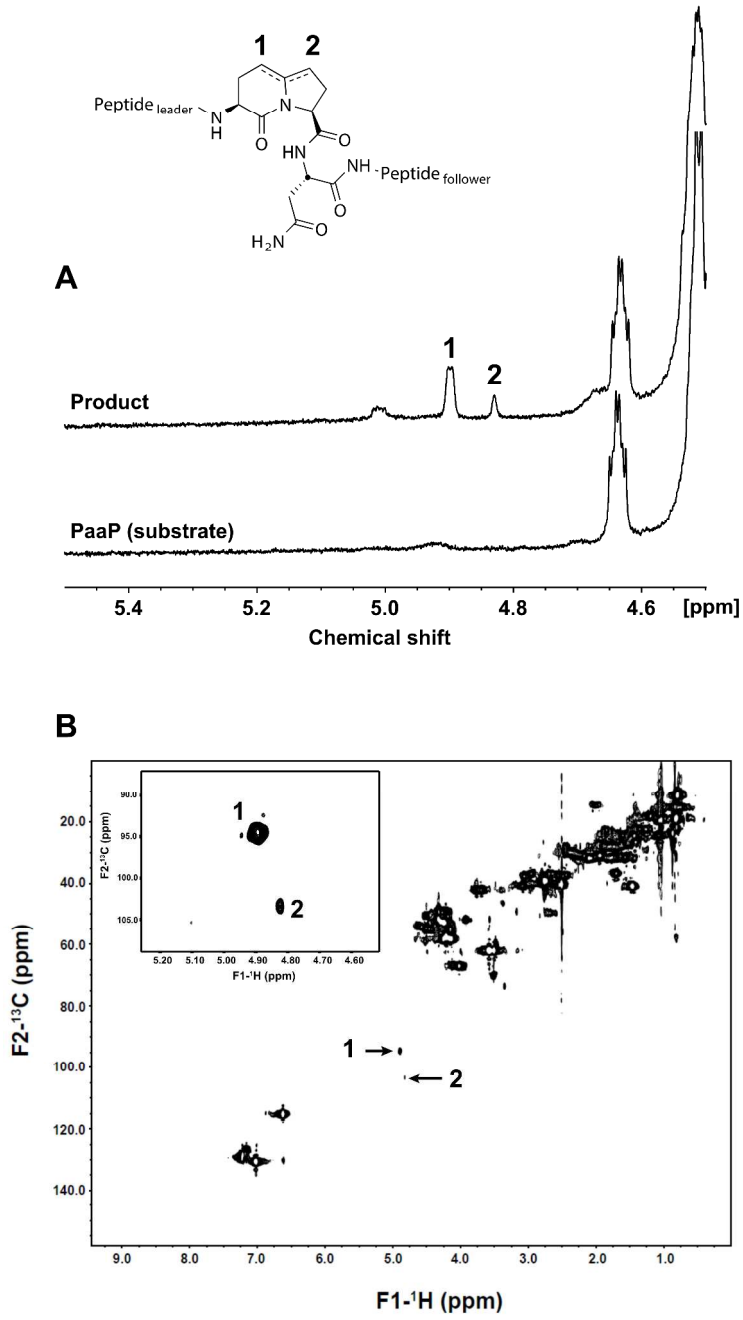


Figure S2. NMR analysis of the purified product of PaaA-catalyzed transformation of PaaP. A) One-dimensional ^{1}H-NMR spectrum of the product and substrate of PaaA. **B)** ^{13}C-HSQC spectrum of the product of PaaA-catalyzed reaction. The new resonance signals are shown in the inset.

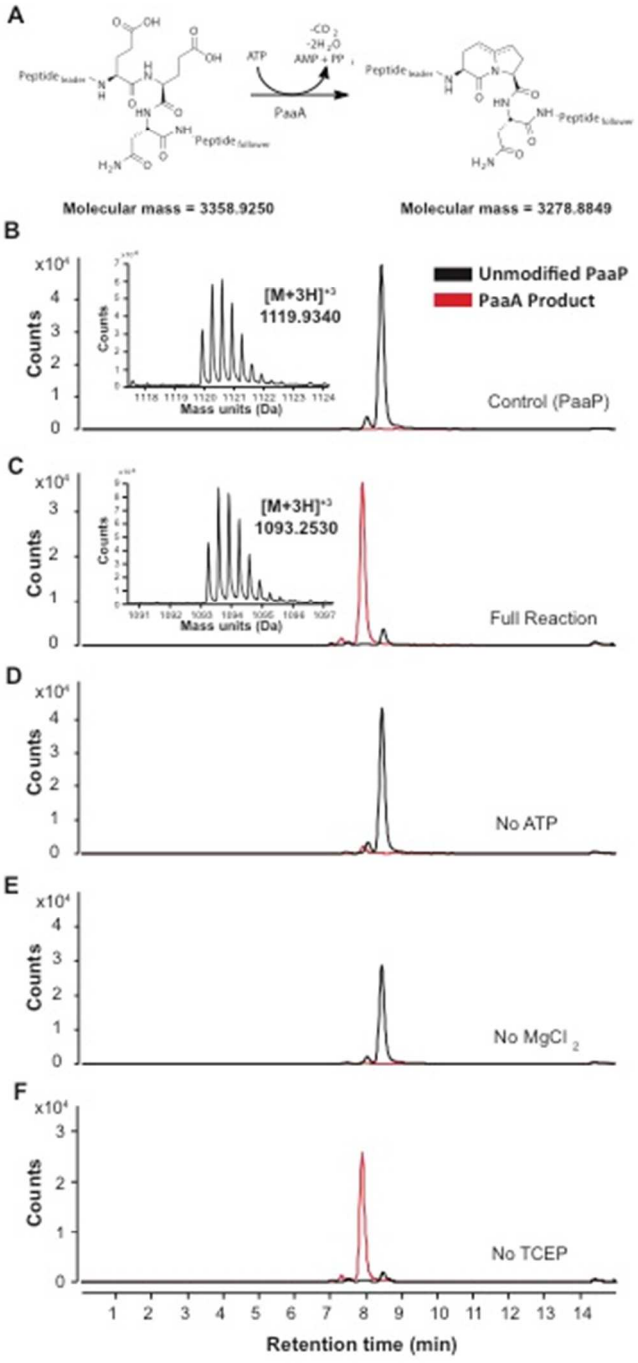


Figure S3. Extracted ion chromatograms for PaaA enzymatic modification of synthetic PaaP. A. Reaction scheme. **B.** Control PaaP (25 µM) without enzyme. **C.** PaaP (25 µM) after incubation with PaaA (2.5 µM), 4 mM $MgCl_2$, 1 mM ATP and 0.5 mM TCEP for 15 hours. **D**, **E**, and **F** are reactions without ATP, $MgCl_2$, or TCEP respectively.

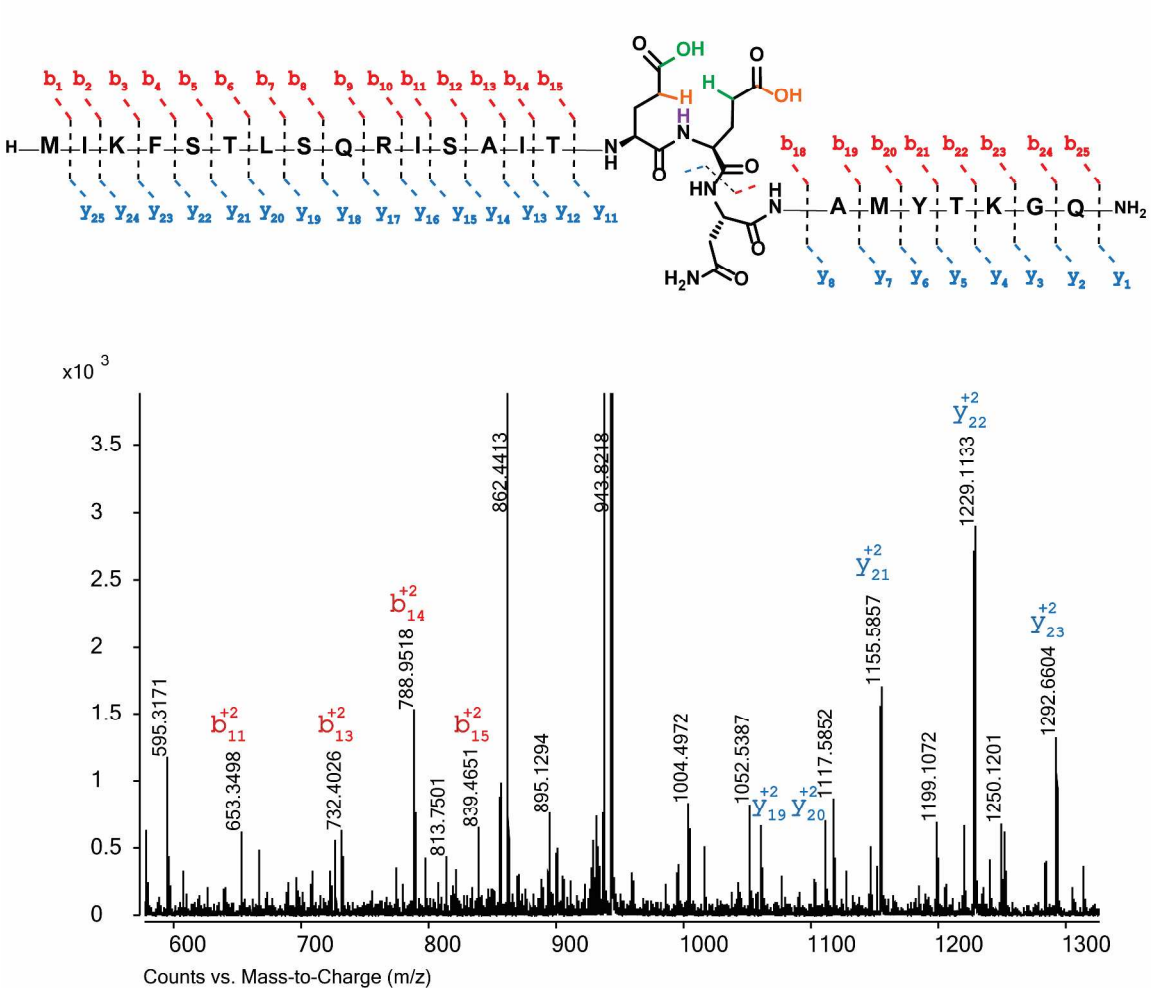


Figure S4. LC-MS/MS analysis of the product of PaaA-catalyzed transformation of M1-Q25 PaaP. $[M + 3H^+]^{+3}$ = 943.5000 ion was selected for fragmentation using a collision energy of 35 eV. The b-ions are shown in red while the y-ions are indicated in blue. Due to a lack of knowledge of the mechanistic details of the PaaA-catalyzed enzymatic reaction, it is presently unclear as to which water molecule is eliminated. The two possibilities of a Claisen-type condensation between the two glutamate side-chains to give a 9-membered ring structure is shown using orange and green colors, while the possibility that the backbone nitrogen atom acting as a nucleophile resulting in either a 5- or 6-membered ring structure depending on which of the two glutamate side-chains is activated using ATP, is indicated by the purple H-atom.

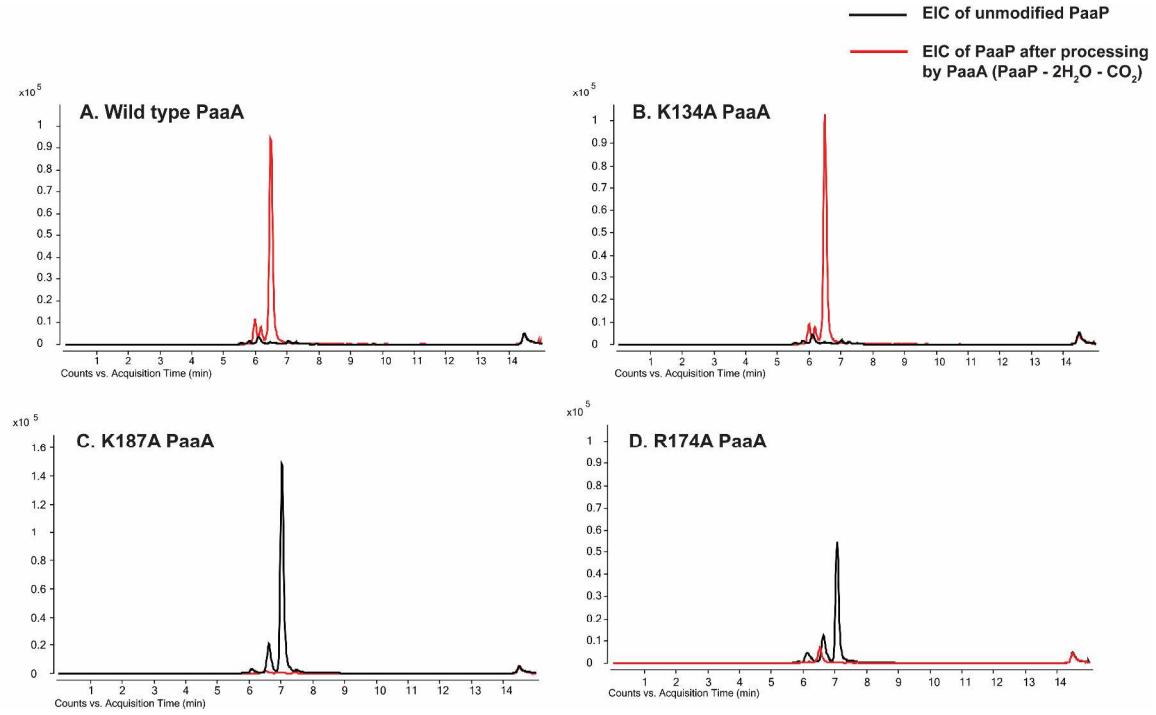


Figure S5. Relative enzymatic activity of PaaA active site variants with point mutations of select residues predicted to bind ATP. 25 µM PaaP was incubated with 1 µM enzyme in each case, in the presence of 50 mM HEPES pH 7.5, 0.5 mM ATP, and 4 mM $MgCl_2$. The reaction was allowed to proceed for ~6 hours, and then analyzed by LC-MS. K134 is a residue away from the active site, and was chosen as a control point mutation.

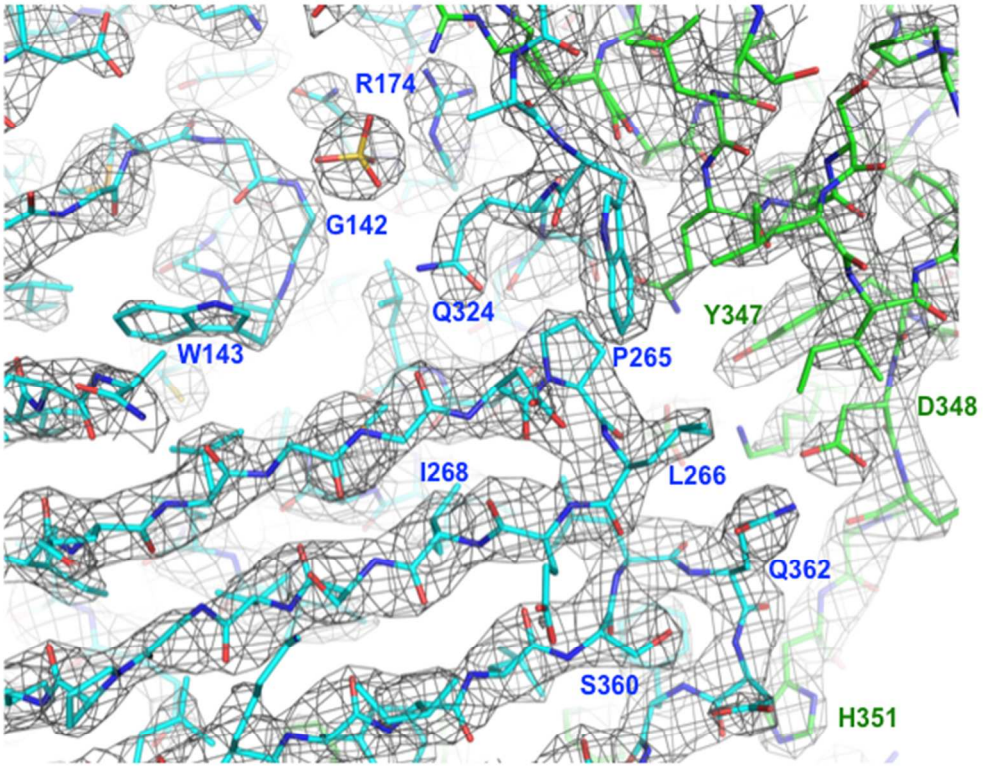


Figure S6. Representative 2Fo-Fc electron density of Se-Met PaaA contoured at 1.5σ level. The two protomers are shown in stick representation in green (chain A) and cyan (chain B).

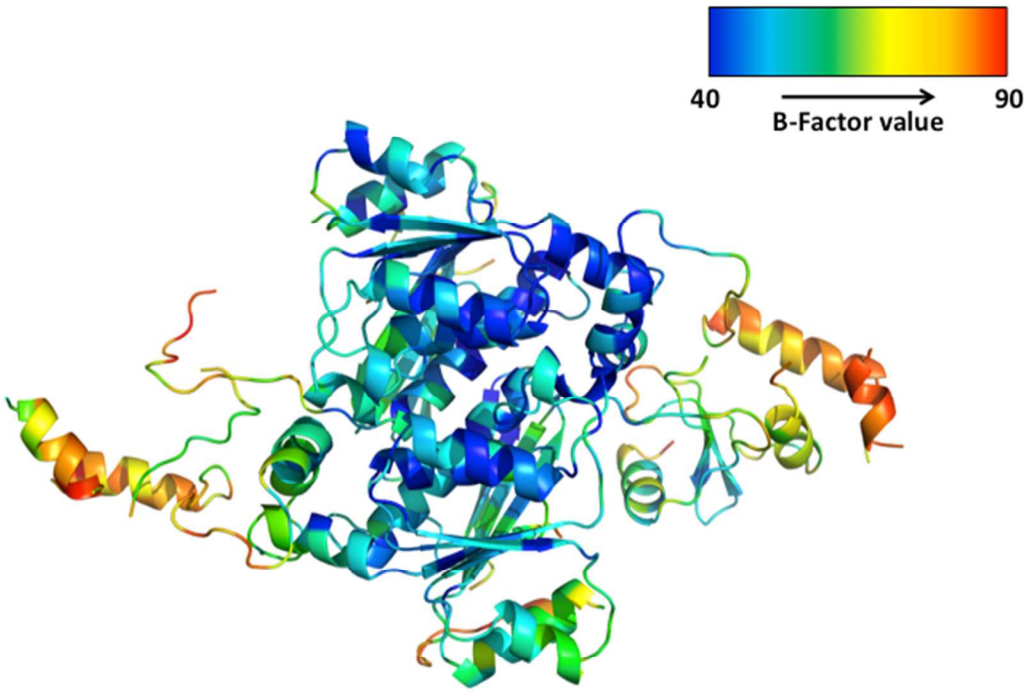


Figure S7. B-factor presentation of the Se-Met PaaA structure. The warmer the color, the higher the B-factor of the alpha carbon.

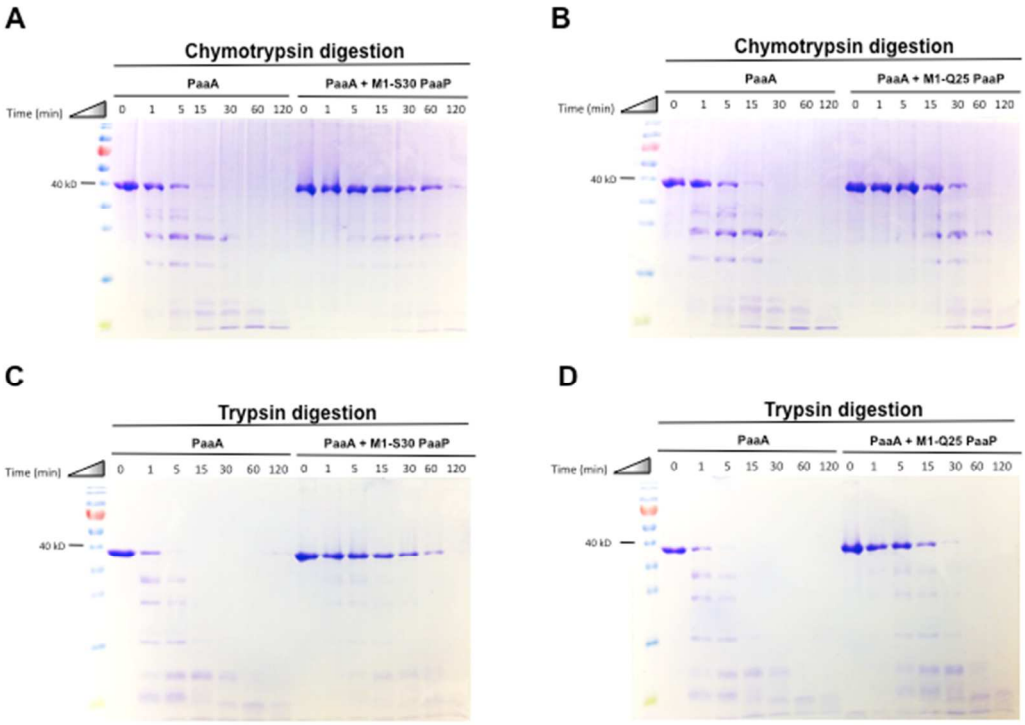


Figure S8. Limited proteolysis of PaaA in the presence and absence of substrates M1-S30 (full length) PaaP, and M1-Q25 PaaP. Results indicate that the peptide substrates protect PaaA from proteolysis by the two proteases. Experimental details are provided in the materials and methods section.

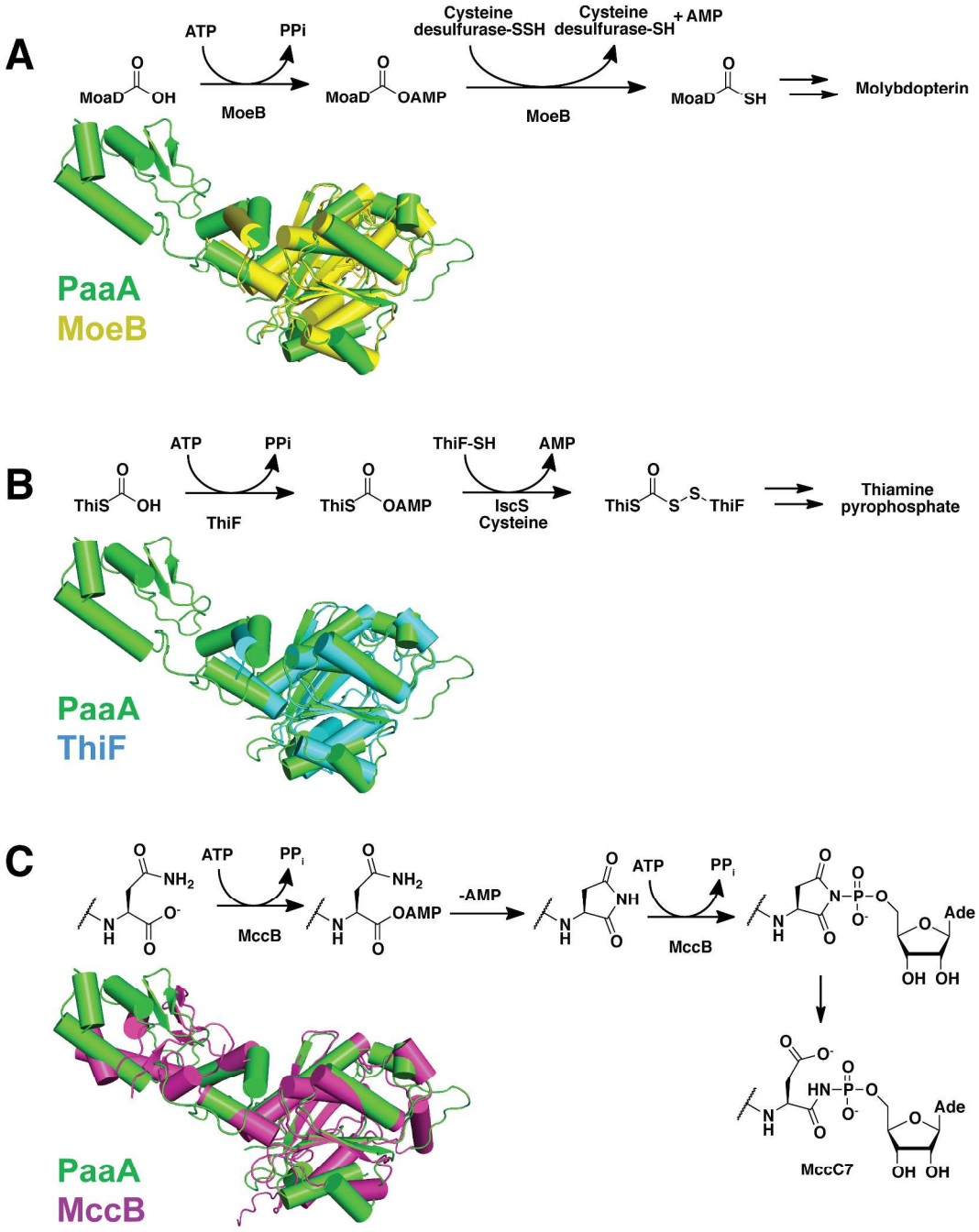


Figure S9. Structural alignment of Se-Met PaaA (chain A, green) with its closest homologs; (A) MoeB (PDB: 1JW9, yellow), **(B)** ThiF (PDB: 1ZKM, cyan), and **(C)** MccB (PDB: 3H5A, magenta). The respective reactions are shown.

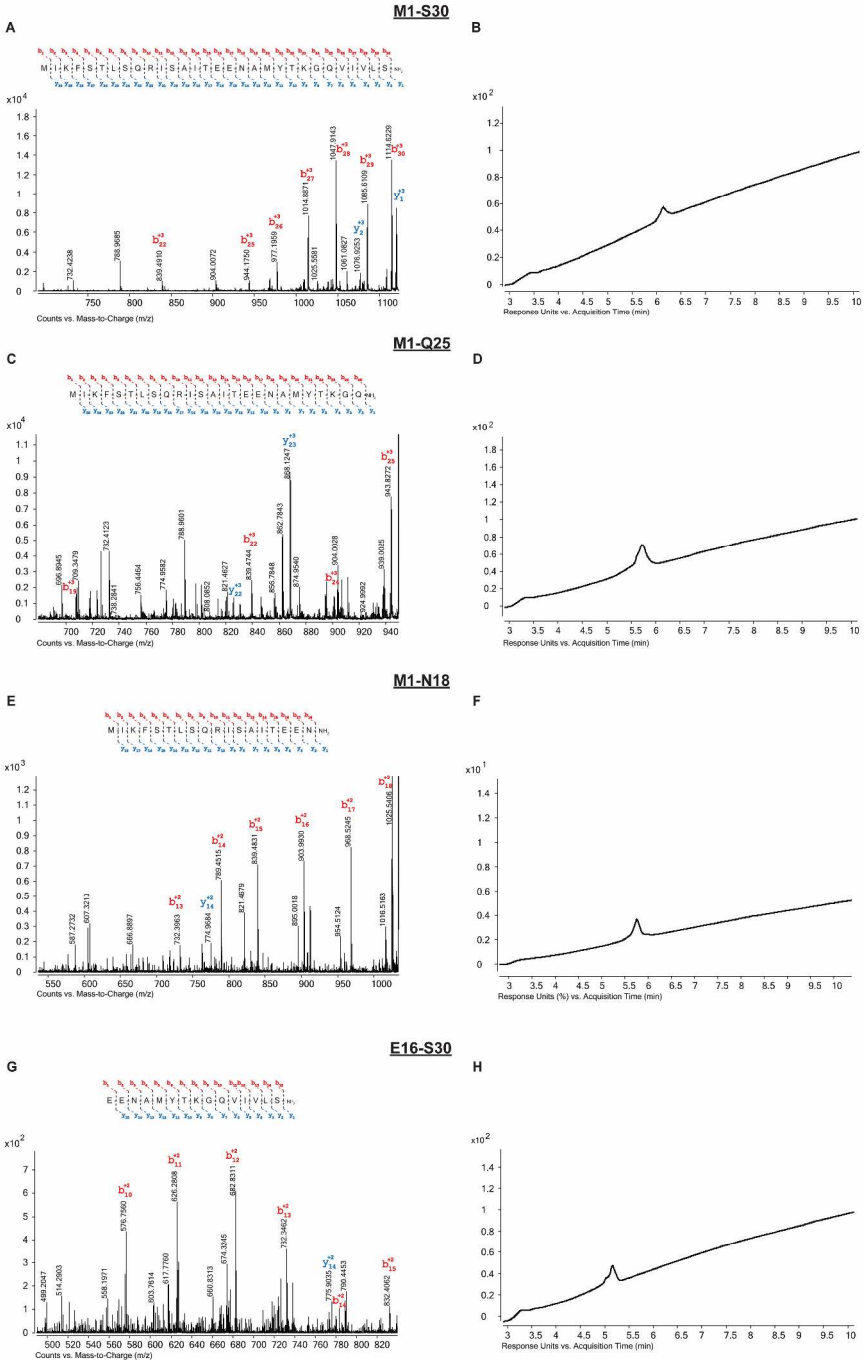


Fig S10. LC-MS/MS analysis of the synthesized peptides (A, C, E, G), along with the respective UV traces (λ = 220 nm) (B, D, F, G). Details of MS/MS: M1-S30, $[M + 3H^+]^{+3}$ = 1119.9340, CID = 35.0 eV; M1-Q25, $[M + 3H^+]^{+3}$ = 949.4901, CID = 40.0 eV; M1-N18, $[M + 2H^+]^{+2}$ = 1034.0529, CID = 35 eV; E16-S30, $[M + 2H^+]^{+2}$ = 840.9355, CID = 30.0 eV.

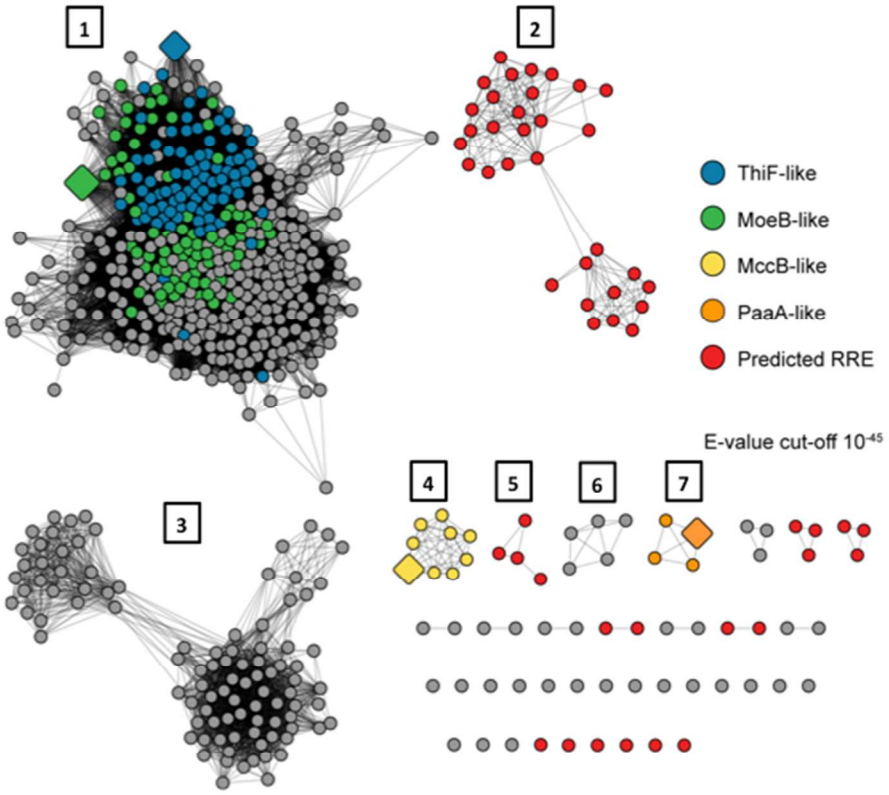


Figure S11. Sequence similarity network of enzyme sequences closely related to PaaA (practically identical to Figure 4C in the main manuscript). The sequence clusters have been given arbitrary numbers for identification. A list of enzyme sequences (Uniprot IDs) that were identified to possess a N-terminal RRE-domain fused to a C-terminal E1-domain is provided in **Table S4**. A histogram of number of enzyme sequences versus sequence length showing the bimodal distribution of sequence length among the above enzymes is provided in **Figure S12**.

Table S4. List of identifiers of enzyme sequences that are predicted, from this study, to possess an N-terminal RRE domain fused to an E1-domain, similar to those found in MccB and PaaA.

Cluster #2

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
W7J032	W7J032_9PSEU	*Actinokineospora spheciospongiae*.
A0A093AL49	A0A093AL49_9PSE	*Amycolatopsis lurida* NRRL 2430.
A0A093ARQ3	A0A093ARQ3_9PSE	*Amycolatopsis lurida* NRRL 2430.
A0A066U4V6	A0A066U4V6_9PSE	*Amycolatopsis rifamycinica*.
C3DUH3	J8FGE9_BACCE C3DUH3_BACTS J3X0F0_BACTU A0A0B5R0H0_BACT A0A0D0H915_BACT B7IJ69_BACC2 C3IWZ0_BACTU J7W1T1_BACCE Q3EYC7_BACTI R8C5U0_BACCE R8IEP8_BACCE R8SIK1_BACCE R8YEI9_BACCE	*Bacillus cereus* MSX-A1. *Bacillus thuringiensis* serovar sotto str. T04001. *Bacillus thuringiensis* HD-771. *Bacillus thuringiensis* HD1002. *Bacillus thuringiensis* subsp. morrisoni. *Bacillus cereus* (strain G9842) *Bacillus thuringiensis* IBL 4222. *Bacillus cereus* VD022. *Bacillus thuringiensis* serovar israelensis ATCC 35646. *Bacillus cereus* str. Schrouff. *Bacillus cereus* K-5975c. *Bacillus cereus* HuB4-4. *Bacillus cereus* TIAC219.
A7GP75	A7GP75_BACCN	*Bacillus cereus* subsp. cytotoxis (strain NVH 391-98)
A0A0E1MCL2	A0A0E1MCL2_BACC Q4MHV5_BACCE	*Bacillus cereus*. *Bacillus cereus* G9241.

A0A084GSD0	A0A084GSD0_9BAC A0A084H489_9BAC	*Bacillus indicus* LMG 22858. *Bacillus cibi*.
I8UH01	I8UH01_9BACI	*Bacillus macauensis* ZFHKF-1.
A6CTC6	A6CTC6_9BACI	*Bacillus sp.* SG-1.
A0A084HAW3	A0A084HAW3_9BAC	*Bacillus sp.* SJS.
H0UFH8	H0UFH8_BRELA A0A075QWJ5_BREL A0A0F7EFJ7_BREL	*Brevibacillus laterosporus* GI-9. *Brevibacillus laterosporus* LMG 15441. *Brevibacillus laterosporus*
J2H7P5	J2H7P5_9BACL C0Z4C4_BREBN	*Brevibacillus sp.* BC25. *Brevibacillus brevis* (strain 47 / JCM 6285 / NBRC 100599)
F3JLL6	A0A060PD12_9BUR F3JLL6_PSESX W6VQI7_9PSED	*Burkholderia sp.* RPE67. *Pseudomonas syringae* pv. aceris str. M302273. *Pseudomonas sp.* GM30.
A0A010ZV91	A0A010ZV91_9ACT	*Cryptosporangium arvum* DSM 44712.
D3HLN5	D3HLN5_LEGLN D5T7R6_LEGP2	*Legionella longbeachae serogroup 1* (strain NSW150) *Legionella pneumophila serogroup 1* (strain 2300/99 Alcoy)
C2KHC2	C2KHC2_LEUMC	*Leuconostoc mesenteroides* subsp. cremoris ATCC 19254.
W2ESC3	W2ESC3_9ACTN	*Microbispora sp.* ATCC PTA-5024.
W7W6X4	W7W6X4_9ACTN	*Micromonospora sp.* M42.
A0A0A2VAT0	A0A0A2VAT0_9BAC	*Pontibacillus chungwhensis* BH030062.
D2NQP8	D2NQP8_ROTMD G5EPD7_9MICC	*Rothia mucilaginosa* (strain DY-18) *Rothia mucilaginosa* M508.

K0JU89	K0JU89_SACES R1G7G7_9PSEU	*Saccharothrix espanaensis* (strain ATCC 51144 / DSM 44229 / JCM 9112 /NBRC 15066 / NRRL 15764) *Amycolatopsis vancoresmycina* DSM 44592.
K8YM84	K8YM84_STRIT	*Streptococcus intermedius* BA1.
D3H7P5	D3H7P5_STRM6 G0IBY4_STRES	*Streptococcus mitis* (strain B6) *Streptococcus pseudopneumoniae* (strain IS7493)
I0SMH4	I0SMH4_STRMT	*Streptococcus mitis* SK616.
V8IJD7	V8IJD7_9STRE A0A064C2F9_STRE A0A0E2PP63_9STR G0IC56_STRES	*Streptococcus pseudopneumoniae* 5247. *Streptococcus pneumoniae*. *Streptococcus pseudopneumoniae* 22725. *Streptococcus pseudopneumoniae* (strain IS7493)
V6KJT2	H1Q667_9ACTN V6KJT2_STRRC A0A0F5AFI8_9ACT K4R3M9_9ACTN	*Streptomyces coelicoflavus* ZG0656. *Streptomyces roseochromogenus* subsp. oscitans DS 12.976. *Streptomyces sp.* MUSC164. *Streptomyces davawensis* JCM 4913.
S5UN22	S5UN22_STRCU S5V1N6_STRCU B5HV74_9ACTN H1QHZ1_9ACTN	*Streptomyces collinus* Tu 365. *Streptomyces sviceus* ATCC 29083. *Streptomyces coelicoflavus* ZG0656.
A0A0F7N6W8	A0A0F7N6W8_9ACT	*Streptomyces sp.* CNQ-509.
A0A0F4J813	A0A0F4J813_9ACT	*Streptomyces sp*. NRRL S-495.
A0A077NK48	A0A077N775_XENB A0A077NK48_XENB	*Xenorhabdus bovienii* str. feltiae Florida. *Xenorhabdus bovienii* str. puntauvense.
K1ZWN5	K1ZWN5_9BACT	uncultured bacterium.

Cluster #4: MccB-like

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
A0A0B2T253	A0A0B2T253_9ENT	*Dickeya solani*.
I2WYG2	Q47506_ECOLX I2WYG2_ECOLX Q2KKH8_ECOLX H9AXU7_ECOLX Q83Y58_ECOLX	*Escherichia coli*. *Escherichia coli* 4.0967.
A0A097H3S3	A0A097H3S3_MORC	*Moraxella catarrhalis*.
A0A094SZ85	A0A094SZ85_PECC	*Pectobacterium carotovorum* subsp. odoriferum.
Q7N6N1	Q7N6N1_PHOLL	*Photorhabdus luminescens* subsp. laumondii (strain TT01)
A0A081RU21	A0A081RU21_PHOT W3VAQ7_PHOTE	*Photorhabdus temperata* subsp. temperata Meg1. *Photorhabdus temperata* subsp. khanii NC19.
A0A068Z1N7	A0A068Z1N7_9ENT	*Serratia symbiotica*.
D7HFZ3	D7HFZ3_VIBCL	*Vibrio cholerae* RC385.
A0A068QXZ7	A0A068QXZ7_9ENT	*Xenorhabdus doucetiae*.

Cluster #5

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
A0JTI2	A0JTI2_ARTS2	*Arthrobacter sp.* (strain FB24)
Q3AGP4	Q3AGP4_SYNSC	*Synechococcus sp.* (strain CC9605)
Q05XM9	Q05XM9_9SYNE	*Synechococcus sp.* RS9916.
I4B9Q3	I4B9Q3_TURPD	*Turneriella parva* (strain ATCC BAA-1111 / DSM 21527 / NCTC 11395 / H)

Cluster #7: PaaA-like

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
A0A066TVR9	A0A066TVR9_9PSE	Amycolatopsis rifamycinica.
A0A089YR18	A0A059IDB8_ENTA Q9ZAR3_ENTAG A0A0A3Z577_9ENT E1SC02_PANVC C6CMV8_DICZE A0A089YR18_9PSE	*Pantoea agglomerans* Eh318. *Enterobacter agglomerans* *Enterobacter cancerogenus*. *Pantoea vagans* (strain C9-1) *Dickeya zeae* (strain Ech1591) *Pseudomonas rhizosphaerae*.
B6VML4	B6VML4_PHOAA K8W971_9ENTR A0A0A2R5Y1_MORM W3YH34_9ENTR	*Photorhabdus asymbiotica* subsp. asymbiotica (strain ATCC 43949 /3105-77) *Providencia burhodogranariea* DSM 19968. *Morganella morganii* *Providencia alcalifaciens* PAL-3.
A0A023Y5D6	A0A023Y5D6_9GAM	*Stenotrophomonas rhizophila*.

	A0A072CQG0_9RHI	*Sinorhizobium americanum* CCGM7.

Triads

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
A0A0C7GDE2	A0A0C7GDE2_CLOS	*Clostridium sordellii*.
X8H5B6	R7N0E0_9FIRM X8H5B6_9FUSO	*Megasphaera elsdenii* CAG:570. *Fusobacterium sp.* CM22.
A0A0F9JH24	A0A0F9JH24_9ZZZ	marine sediment metagenome.
A8MGD8	A0A0C7GFF0_CLOS A8MGD8_ALKOO	*Clostridium sordellii*. *Alkaliphilus oremlandii* (strain OhILAs)
X8H5E2	X8H5E2_9FUSO	*Fusobacterium sp.* CM22.
R7MYS1	R7MYS1_9FIRM	*Megasphaera elsdenii* CAG:570.

Doublets

Uniprot ID of the displayed representative node	Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)	Organism
A0A087VVN7	A0A087VVN7_9BIF A0A0A8NJI7_BIFL B7GUS2_BIFLS	*Bifidobacterium indicum* LMG 11587 = DSM 20214. *Bifidobacterium longum* subsp. infantis. *Bifidobacterium longum* subsp. infantis (strain ATCC 15697 / DSM 20088/ JCM 1222 / NCTC 11817 / S12)

L2HAB5	L2HAB5_ENTFC R3KWK3_ENTFC R4EEX5_ENTFC	*Enterococcus faecium* EnGen0012. *Enterococcus faecium* EnGen0371. *Enterococcus faecium* EnGen0174.

H5SWR4	H5SWR4_LACLL	*Lactococcus lactis* subsp. lactis IO-1.
F6DIW7	F6DIW7_THETG	*Thermus thermophilus* (strain SG0.5JP17-16)

Singlets

Uniprot ID of the displayed representative node	**Uniprot IDs of all redundant sequences (sequence identity ≥ 60%)**	**Organism**
A7BJB3	A7BJB3_BACNA	*Bacillus subtilis* subsp. natto.
J1A7E5	J1A7E5_BARTA J1IU97_9RHIZ J0Q0K9_9RHIZ	*Bartonella taylorii* 8TBB. *Bartonella alsatica* IBS 382. *Bartonella sp.* DB5-6.
G7M965	G7M965_9CLOT	*Clostridium sp.* DL-VIII.
A0A0B6EMG0	A0A0B6EMG0_9COR	*Corynebacterium singulare*.
A0A081S0W5	A0A081S0W5_PHOT A0A0F7LLG9_PHOT T0PAZ7_PHOTE U7R6K5_PHOTE Q7N4X2_PHOLL	*Photorhabdus temperata* subsp. temperata Meg1. *Photorhabdus temperata* subsp. thracensis. *Photorhabdus temperata* subsp. temperata M1021. *Photorhabdus temperata* J3. *Photorhabdus luminescens* subsp. laumondii (strain TT01)
A0A0B3AQI5	A0A0B3AQI5	archaeon GW2011_AR19

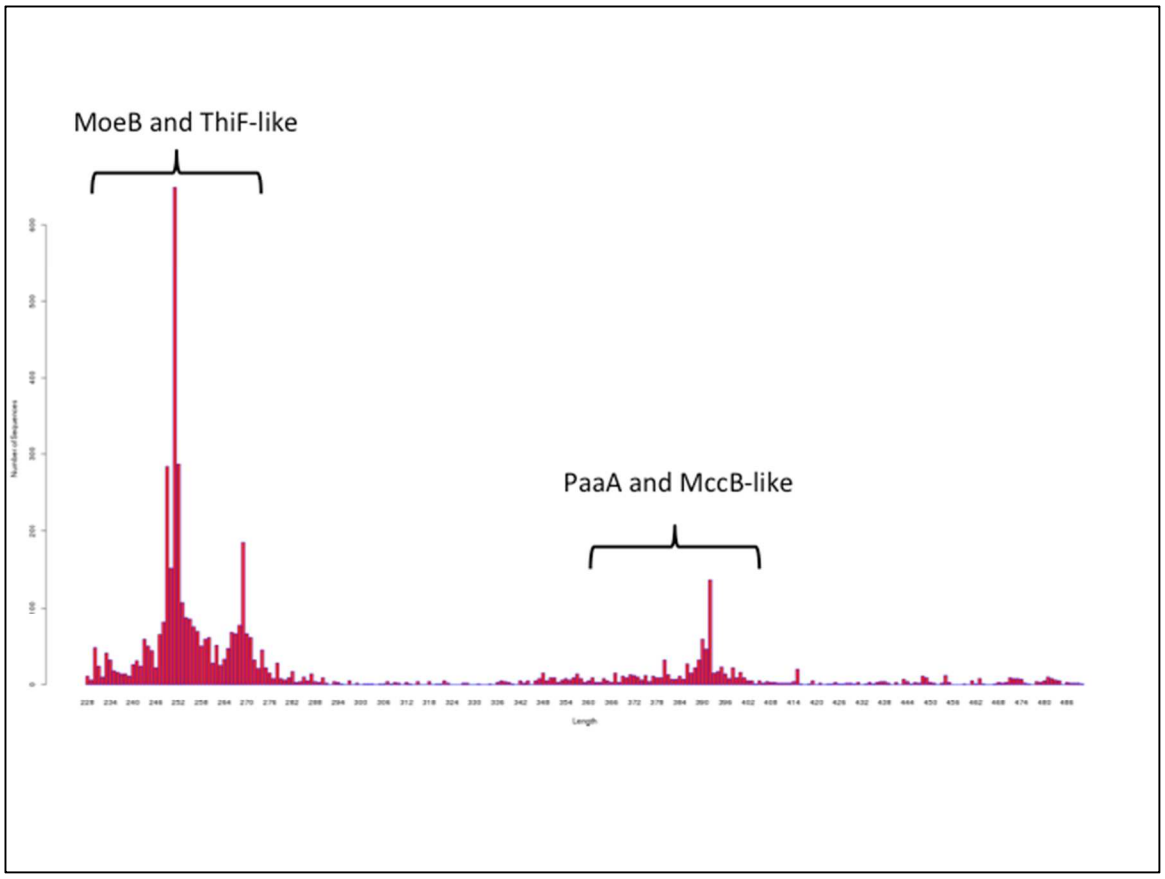


Fig S12. Bimodal distribution of the lengths of proteins related to PaaA and MccB. This figure was generated as a part of the output from the EFI-EST online tool, and has been reproduced here with suitable labels.