Co-occurrence of enzyme domains guides the discovery of an oxazolone synthetase

Tristan de Rond, Julia E. Asay and Bradley S. Moore

Multidomain enzymes orchestrate two or more catalytic activities to carry out metabolic transformations with increased control and speed. Here, we report the design and development of a genome-mining approach for targeted discovery of biochemical transformations through the analysis of co-occurring enzyme domains (CO-ED) in a single protein. CO-ED was designed to identify unannotated multifunctional enzymes for functional characterization and discovery based on the premise that linked enzyme domains have evolved to function collaboratively. Guided by CO-ED, we targeted an unannotated predicted ThF-nitroreductase di-domain enzyme found in more than 50 proteobacteria. Through heterologous expression and biochemical reconstitution, we discovered a series of natural products containing the rare oxazolone heterocycle and characterized their biosynthesis. Notably, we identified the di-domain enzyme as an oxazolone synthetase, validating CO-ED-guided genome mining as a methodology with potential broad utility for both the discovery of unusual enzymatic transformations and the functional annotation of multidomain enzymes.

Our knowledge of nature’s diversity of enzymatic transformations is crucial to advancing research in a multitude of disciplines. For instance, our ability to predict metabolic capacity from genome sequences enables new insights in human health and ecology, while the development of bioprocesses to produce chemicals in an environmentally benign fashion relies on the availability of a well-stocked biocatalytic toolbox. Billions of years of evolution have resulted in immense natural genetic diversity, which we are rapidly starting to uncover using modern sequencing technologies. However, functional assignment of the enzymes encoded by this sequence diversity is lagging. Many enzymes catalyzing chemical transformations previously not known to occur in nature are still being discovered, and there is no end in sight of unannotated and misannotated enzymes in genomic databases. The search for new enzymes in this genomic wilderness can fulfill the dual purpose of functional gene annotation and biocatalyst discovery.

Computational searches for biosynthetic enzymes that underlie ‘genome-mining’ campaigns are most commonly based on gene homology to known core biosynthetic genes of well-established classes of medicinal natural products such as polyketides, nonribosomal peptides and terpenoids. This approach, while likely to identify gene clusters that produce bioactive products, is prone to the rediscovery of known enzymology. Therefore, we set out to develop a genome-mining strategy that could guide enzyme discovery in a complementary manner.

One notable feature among specialized metabolism is an over-abundance of enzymes harboring multiple catalytic domains (Supplementary Fig. 1). Multidomain enzymes are thought to arise through gene fusion of two or more single-domain enzymes, affording catalytic advantages such as coupled temporal and spatial regulation, a fixed active site stoichiometry and channeling of reactive intermediates. Over time, a multidomain enzyme may even evolve to orchestrate the reactivity of its constituent active sites to such an extent that it acquires a new function. Classic examples of multidomain enzymes are fatty acid synthase, polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) assembly-line enzymes, some of which consist of dozens of domains. More recently, multidomain enzymes have been shown to catalyze epimerization and N-nitrosation reactions.

Here, we present a genome-mining paradigm in which we leverage the evolved co-occurrence of enzyme domains in proteins to inform enzyme discovery. Through CO-ED analysis, we discovered a series of new oxazolone-containing natural products, as well as an oxazolone synthetase, a new bifunctional cyclodehydratase–oxidoreductase.

Results

CO-ED-guided identification of a di-domain enzyme candidate.

The CO-ED workflow takes a query set of proteins and generates a network representation of co-occurring enzymatic domains (Fig. 1). A node is drawn for each of a curated set of enzymatic Pfam domains found in the query proteins, and an edge is drawn between two nodes for each two-domain combination of domains found together in at least one query protein. In parallel, a set of characterized enzymes, derived from the BRENDA, MIBiG and UniProt databases, is subjected to the same analysis. If an edge in the network represents a domain pair found in a previously characterized multidomain enzyme, the edge is color coded according to the originating database. Remaining uncolored edges, therefore, represent enzyme domain combinations that are likely uncharacterized. A public web server allowing for CO-ED analysis on user-supplied datasets is available at http://enzyme-analysis.org.

To validate the CO-ED workflow, we first applied it to the genome of the model organism Escherichia coli K12. The E. coli network (Fig. 2a,b, Supplementary Figs. 2 and 3 and Supplementary Data 1) revealed 19 co-occurring pairs of domains, including some well-studied multifunctional enzymes including penicillin-binding proteins, the aldehyde dehydrogenase–alcohol dehydrogenase AdhE, the UDP-1-Ara4PN biosynthesis enzyme ArnA, several bifunctional amino acid biosynthesis enzymes and the enterobactin NRPS EntF. Every edge is colored, suggesting that the function of every multidomain enzyme in E. coli detected by CO-ED has likely already been established.
We next applied CO-ED analysis to the genome of the non-model marine γ-proteobacterium *Pseudoalteromonas rubra* DSM 6842, chosen because of our experience with heterologously expressing enzymes from this and closely-related species [22][23]. The *P. rubra* CO-ED network (Fig. 2c, Supplementary Fig. 4 and Supplementary Data 1) revealed numerous NRPSs, as can be inferred from the edge indicating that the AMP-binding (also known as adenylation) and 'condensation' Pfam domains co-occur in 38 proteins, including two hybrid PKS–NRPS enzymes, as known as adenylation) and 'condensation' Pfam domains co-occur in 38 proteins, including two hybrid PKS–NRPS enzymes.

A *P. rubra* protein harboring a 'ThiF'–'nitroreductase' domain pair (outlined in the bottom right of Fig. 2c) caught our attention, because each of these domains is known to catalyze a diverse set of chemical transformations (Extended Data Fig. 1), yet their combination is unprecedented. Enzymes of the ThiF protein family catalyze various vital reactions that proceed through carboxylate adenylation, including the activation of ubiquitin and ubiquitin-like proteins by E1 enzymes [24], the cyclization of N*-threonylcarbamoyladenosine in tRNA maturation [25] and the post-translational modification of ribosomally synthesized peptide antibiotics [26]. ThiF-family enzymes also play a role in sulfur-incorporation machinery for thiamin, the molybdenum cofactor and some natural products [27][28]. To date, all characterized substrates of ThiF-family enzymes have been polypeptides or tRNA. The nitroreductase family consists of flavoenzymes that catalyze a remarkable variety of redox reactions, such as reductive deiodination of thyroid hormones [29], dehydrogenation of oxazolines and thiazolines [30], fragmentation of flavin mononucleotide to form the vitamin B12 lower ligand 5,6-dimethylbenzimidazole [31], enol methylene reduction in cofactor F420 biosynthesis [32], diketopiperazine desaturation in albonoursin biosynthesis [33] and reductive detoxification of various nitro functional groups [34].

Querying the UniProt database with the di-domain ThiF–nitroreductase enzyme from *P. rubra*, which we named OxxB, revealed 56 proteins originating from α-, β- and γ-proteobacteria (Extended Data Fig. 1b and Supplementary Fig. 5) isolated from a variety of (predominantly aquatic) environments (Extended Data Fig. 1c). In cases in which the genomic context of the oxyB homolog was known, it was typically accompanied by an N-acetyltransferase-encoding homolog (oxxA). In *P. rubra* and many other species, oxxA is immediately upstream of oxxB on the same strand, forming an apparent two-gene oxxAB operon (Fig. 3a and Extended Data Fig. 1d).

**Heterologous expression of oxxAB produces new metabolites.** To investigate the function of the oxxAB genes, we amplified oxxAB by PCR from *P. rubra*, *Rheinheimera pacifica*, *Colwellia chukchiensis* (all γ-proteobacteria), *Skermanella aerolata* (α-proteobacterium) and *Undibacterium pigrum* (β-proteobacterium) and heterologously expressed the gene pairs in *E. coli*. Colonies on solid medium and cell pellets after growth in liquid medium both turned visibly yellow (Supplementary Fig. 6). The color would briefly intensify when incubated above pH 9, suggesting the possibility of a phenolic, base-labile product (Supplementary Figs. 6 and 7). HPLC–UV–MS analysis of extracts of the pellets revealed several chromatographic peaks with absorbance in the 300–400-nm range that were not produced by *E. coli* native (Fig. 3b). Product profiles of heterologously expressed oxxAB genes originating from different species shared many constituents, and the inclusion of genes found up- or downstream of oxxAB did not substantially change the product profiles (Supplementary Fig. 8).

Through a combination of mass spectrometry and spectroscopic techniques, with notable insight provided by 1,1-ADEQUATE NMR, [1H–15N] HMBC NMR and the identification of degradation products formed in basic methanol, we were able to assign the products of OxxAB as a series of oxazolones (Fig. 3c, Supplementary Figs. 9 and 10 and Supplementary Note). We named these seemingly tyrosine- and phenylalanine-derived molecules tyrazolones and phanazolones, respectively. The major products were heptyl- nonyl-, undecyl- and ω-6-undecenyltyrazolone (2, 4, 5, 6 and 7, respectively), which were found to occur in an equilibrium of (E) and (Z) isomers, with the latter being predominant, and (Z)-heptylphenazolone (9). These products have not been previously reported in the literature; however, an oxazolone that appears to be derived from tryptophan, alomazol (Fig. 3d), was isolated from a red alga [35], and its structure was confirmed through synthesis. Alomazolone similarly exists in an (E)–(Z) equilibrium, forms degradation products analogous to those formed by tyrazolones and phanazolones and has similar spectroscopic properties, lending credence to our structural assignments.

---

**Fig. 1** | **Outline of the CO-ED workflow.** A CO-ED network consists of nodes representing a curated set of enzymatic domains present in a query set of proteins and edges connecting two domains if they co-occur in the same protein. Edges are colored based on whether any known enzymes contain the pair of domains that the edge represents. Known enzymes were compiled from BRENDA, MIBiG and a curated set of entries from UniProt [36].
In vitro reconstitution of oxazolone biosynthesis by OxzAB. Considering the enzymatic domains contained within OxzAB, we hypothesized oxazolone biosynthesis to proceed as follows: OxzA forms an N-acyl amino acid using acyl-CoA derived from the natural fatty acyl-CoA β-oxidation pool. To form oxazolone, OxzB then catalyzes cyclization and oxidation of N-acyl amino acid, in either order (Supplementary Fig. 11). The ThiF-family enzyme TdA is known to catalyze cyclodehydration of an amino acid moiety, and the nitroreductase-family enzyme AlbA was shown to oxidize amino acid-derived substrates (Extended Data Fig. 1a). It is thus conceivable that OxzB, which harbors both of these domains, could act to form the oxazolone heterocycle.

To test our biosynthetic hypothesis, we heterologously expressed and purified *P. rubra* OxzA and OxzB in *E. coli* as N-terminal His6 and mannose-binding protein (MBP) fusions proteins, respectively (Supplementary Fig. 12), and assayed their activity by HPLC (Fig. 4). When incubated with L-tyrosine (12) and decanoyl-CoA (13), OxzA catalyzed the formation of N-decanoyl-L-tyrosine (14). OxzB was able to catalyze the formation of nonyltyrazolone both from synthetic N-decanoyltyrosine as well as when combined with OxzA, tyrosine and decanoyl-CoA. OxzB activity was only observed in the presence of ATP, as is typical for ThiF-family enzymes. (Z)-nonyltyrazolone, the favored isomer of the (E)–(Z) equilibrium, was the major product upon extended incubation of the enzymatic reactions; however, shorter incubation times revealed a proportionally greater amount of (E)-tyrazolone (Supplementary Fig. 13), suggesting that the immediate product of OxzB is (E)-nonyltyrazolone.

Analysis of oxazolone production in native hosts. To determine whether oxazolones are true natural products or merely artifacts of heterologous expression, we attempted to detect these molecules in the five bacteria for which we heterologously tested oxzAB pairs above. Under standard culturing conditions, oxazolones could be reliably detected only in *C. chukchiensis* and *P. rubra* DSM 6842. Some overlapping edge labels are omitted for clarity. Many edges are unannotated (gray), suggesting that there is much potential for exploring multidomain enzymes for new biosynthetic capacity in this organism. The ThiF-nitroreductase domain pair investigated further in this study is outlined. Only nodes with edges are shown (b,c). Pfam domain names are shown.

**Discussion**

Using a CO-ED-guided genome-mining strategy, we discovered a series of new oxazolone natural products as well as a unique
CO-ED unites these concepts to specifically guide the discovery of enzymes catalyzing new transformations. The availability of a web tool at http://enzyme-analysis.org will allow any user to easily apply CO-ED to identify unstudied multidomain enzymes present in their own datasets.

While CO-ED analysis has revealed that all multidomain enzymes in E. coli likely have known functions, we wished to explore whether this is the case for other model bacteria, particularly those that are promising sources of new biocatalysts. To this end, we applied CO-ED to the genomes of Streptomyces coelicolor A3(2), Salinispora tropica CNB-440 and Pseudomonas fluorescens Pf-5, model organisms with remarkable biosynthetic capacities that are well studied. We found over a dozen unannotated domain pairs in each of these species (Supplementary Figs. 15–17). Investigating the functions of these multidomain enzymes may reveal yet more biochemical newness even in these well-studied organisms.

CO-ED analysis of all proteins in the UniProt17 database (Supplementary Fig. 18 and Supplementary Data 1) shows that, while the most abundantly distributed co-occurring domain pairs are represented by enzymes annotated in BREnda, MiBiG or UniProt, there is a still a plethora of less-widespread multidomain enzymes, the functions of which are unassigned, with some notable examples shown in Supplementary Fig. 18b. Of the 252 enzymatic

---

**Fig. 3** Heterologous expression of oxzAB results in the production of a series of oxazolones. a, Organization of oxzAB in the genomes of the five proteobacteria for which these genes were tested heterologously. b, Reversed-phase HPLC profiles of extracts of E. coli expressing oxzAB from five proteobacteria. An absorbance range of 300–400 nm was chosen to visualize all oxazolones reported in this study (under our chromatography conditions (water–acetonitrile with 0.1% formic acid), the λmax of (E)-tyrazolones is 360 nm, that of (Z)-tyrazolones is 358 nm and that of (Z)-phenazolones is 330 nm). c, Structures of tyrazolones and phenazolones, which are found as a series with varying alkyl tails. Tyrazolones 4–7 and phenazolone 9 were characterized by NMR, while all others were inferred by exact mass and MS/MS fragmentation patterns. Tyrazolones exist as (E)-(Z) isomers in equilibrium. d, Almazolone, an oxazolone natural product previously isolated from an alga, also occurs in an (E)-(Z) equilibrium.

oxazolone synthetase harboring ThiF and nitroreductase domains. This proof-of-principle experiment showcases the promise of CO-ED in selecting orphan multidomain enzymes for targeted functional discovery. Many computational workflows for the prioritization of unannotated genes for genome mining rely on identifying gene clusters based on sequence homology to known biosynthetic genes, and downstream analysis tools in turn often depend on these gene-cluster annotations. This approach risks overlooking genes that do not either themselves have homology to known biosynthetic genes, or are clustered with such genes. Alternatively, there exist workflows that allow the user to explore enzyme families related to a query sequence or domain, such as EnzymeMiner and the tools developed by the Enzyme Function Initiative, but these might miss protein families that the user did not consider searching for. CO-ED is complementary to both of the aforementioned approaches in that it considers unannotated enzyme domain co-occurrences in a relatively unbiased fashion.

The analysis of co-occurring protein domains has previously been applied in contexts other than genome mining, such as in evolutionary studies, to help identify antifungal drug targets and in functional gene annotation (in which it is sometimes called the ‘Rosetta stone’ method). Parallel to this, efforts have been made to connect protein domains to enzymatic reactions.
Articles NaTurE CHEMICal BIOlOGy

Potential to exhibit better functional group tolerance than synthetic production have been reported. Biocatalytic approaches have the oxazolone-containing natural products (Supplementary Fig. 19). may be part of biosynthetic gene clusters coding for more complex genes are clustered with genes encoding pre-

Fig. 4 | In vitro characterization of oxazolone biosynthesis by P. rubra
OxzA and OxzB. Black HPLC chromatograms show absorbance at 280 nm (N-decanoyl-L-tyrosine); red chromatograms show absorbance at 360 nm (tyrazolones) and are scaled down 4.5-fold in relation to the 280-nm signal. Spont., spontaneous.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-021-00808-4.

References
1. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. Genome Biol. 20, 121 (2019).
2. Bornscheuer, U. T. The fourth wave of biocatalysis is approaching. Philos. Trans. A Math. Phys. Eng. Sci. 376, 20170063 (2018).
3. Gerlt, J. A. et al. The enzyme function initiative. Biochemistry 50, 9950–9962 (2011).
4. Hanson, A. D., Pribat, A., Walter, J. C. & de Crécy-Lagard, V. Unknown proteins and orphan enzymes: the missing half of the engineering parts list—and how to find it. Biochem. J. 425, 1–11 (2009).
5. Scott, T. A. & Piel, J. The hidden enzymology of bacterial natural product biosynthesis. Nat. Rev. Chem. 3, 404–425 (2019).
6. Ellens, K. W. et al. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. 45, 11495–11514 (2017).
7. Medema, M. H., de Rond, T. & Moore, R. S. Mining genomes to illuminate the specialized chemistry of life. Nat. Rev. Genet. https://doi.org/10.1038/s41576-021-00363-7 (2021).
8. Michael, A. J. Evolution of biosynthetic diversity. Biochem. J. 474, 2277–2299 (2017).
9. Hagel, J. M. & Facchin, P. J. TYING the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond. J. Exp. Bot. 68, 4029–4043 (2017).
10. Bashtoon, M. & Chothia, C. The generation of new protein functions by the combination of domains. Structure 15, 85–99 (2007).
11. Weissman, K. J. The structural biology of biosynthetic megaenzymes. Nat. Chem. Biol. 11, 660–670 (2015).
12. Winzer, T. et al. Plant science. Morphinan biosynthesis in opium poppy requires a P450–oxidoreductase fusion protein. Science 349, 309–312 (2015).
13. Ng, T. L., Rohac, R., Mitchell, A. J., Boal, A. K. & Balskus, E. P. An N-nitrosating metalloenzyme constructs the pharmacophore of streptozotocin. Nature 566, 94–99 (2019).
14. El-Gebali, S. et al. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D427–D432 (2019).
Methods

Computational methods. A CO-ED analysis web tool is available at http://enzyme-analysis.org. For offline CO-ED analysis, code written in Python version 3.6.10, Jupiter notebook version 6.0.3 and pandas version 1.0.3 is publicly available at https://github.com/tderond/CO-ED. Aside from Python code, the Jupyter notebooks contain explanations of the CO-ED workflow's steps and example command-line instructions.

Briefly, the CO-ED workflow has the following stages. (1) Annotate domains (Domain annotation) in a set of 'query' protein sequences and a set of 'known enzymes' (Selection of known enzymes). (2) For every domain in the curated set of non-redundant enzymatic or catalytic domains ('Curation of enzyme Domains for CO-ED'), find its occurrences in each of the proteins in the query set. (3) Draw a node for each domain found in at least one protein in the query set and label the number of proteins that the domain was found in. (4) For every combination of two domains in the curated set of non-redundant enzymatic or catalytic domains, find the occurrences of each combination in the query set and in the 'known enzymes' sets. (5) Draw an edge for each combination of domains found in at least one protein in the query set, label it with the number of occurrences and color the edge based on which set(s) of 'known enzymes' contain at least one entry with this combination of domains. Networks were rendered in Cytoscape 3.5.3 or 3.8 (ref. 1).

Query sequences. Protein sequences derived from the following genomes were downloaded from NCBI GenBank: E. coli K12 (assembly ASM584 version 21), P. rubra DSM 6842 (assembly Prub 2.0), S. coelicolor A3(3) (assembly SM2D383 version 1), S. tropica CNB-440 (ASM1642 version 1) and P. fluorescens PF-5 (ASM1226 version 1).

Domain annotation. Pfam-A domain annotations were either retrieved from UniProt, as was implemented for the 'known enzymes' and for the all-of-UniProt analysis or annotated using PfamScan 1.6 (which in turn made use of HHMER 3.3 (ref. 3)), as was carried out for the single-genome queries, which gives the CO-ED Domain annotation.

Selection of known enzymes. For UniProt, all proteins in UniProtKB that are either manually annotated as having catalytic activity (that is, match the query annotation type=catalytic activity evidence=manual) or listed in UniProt's pathway:TXT file (https://www.uniprot.org/docs/pdbtaxa) were selected. This set has overlaps with but is not identical to 'Swiss-Prot' (the subset of UniProt marked 'reviewed'); it contains some 'unreviewed' entries and does not include Swiss-Prot entries for which catalytic activity was automatically assigned by homology. For BRENDA, all entries in the BRENDA database that refer to UniProt accessions were selected. For each entry in MIBiG, an NCBI nucleotide region is defined. NCBI protein identifiers for all proteins in this region were obtained and mapped to UniProt accessions. A small number of proteins that could not be mapped to UniProt were not included in the final set.

Curation of enzyme domains for CO-ED. The set of domains used to conduct CO-ED is intended to be enzymatic and non-redundant. By 'enzymatic', we mean that only catalytic enzyme domains are included, and, for example, purely structural, regulatory or docking domains are omitted. By 'non-redundant', we mean that we wish to avoid 'trivial' domain pairs in which two crystallographic Pfam domains would be used to catalyze a single reaction (such as 'Terpene_synth' and 'Terpene_synth_C' or 'G6PD_N' and 'G6PD_C'). This section describes how this set of curated domains was established.

A set of Pfam domains was compiled by taking all domains annotated for entries in UniProt (https://www.uniprot.org/) that are also annotated in MIBiG (https://mibig.secondarymetabolites.org/, all proteins), or in UniProt's pathway:TXT file. Non-catalytic domains were removed. For pseudocatalytic domains that together catalyze one reaction (often detected by performing CO-ED analysis on all proteins in UniProt and finding domains that co-occur with one at a high percentage of time), the more abundant domain was included. Overlapping domains with similar catalytic functions are often members of the same Pfam clan, causing only the best-matching domain to be annotated by PfamScan, but, in some cases in which both are annotated in a high proportion of proteins in UniProt (for example, because clan assignment has yet not been completed for those domains), only one of the proteins was included in our set. Lastly, many enzymatic domains acting on macromolecules and domains with unknown functions were annotated as such, and the analysis can be run with or without their consideration. Annotation categories are as follows: 'm', nucleases, topoisomerases, transposases, helicases, polymerases, proteases, protein kinases and phosphatases, ATP-dependent transporters; 's', glycolytransferases, glycolyl hydrolases (cellulases, amylases, etc.); 'e', enzymes in electron transport chains (oxidative phosphorylation, photosynthesis, etc.); 'd', domains with unknown function; 'y', all other enzymes but only those transporters that couple transport to a reaction besides ATP hydrolysis; 'n', determined to either be catalytic or to comprise a catalytic domain together with a domain annotated in one of the above categories. For analyses shown in this paper, annotation categories 'y', 'm', 's' and 'e' were considered, totaling 1,745 domains. We realize that the domain-curation process is somewhat subjective, and hence the CO-ED web tool allows the set of domains considered in the submission interface to be edited, and the Jupiter notebook allows this by editing the pfamD_ to_name_desc_longdesc.csv file.

Bacterial culture. Bacterial stocks were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen: P. rubra DSM 6842, C. chukchienensis DSM 22576 and R. pacifica DSM 17816, cultured on Marine Agar or Broth 2216 at 30 °C; and U. piscium DSM 15972 and S. aerolata DSM 18479, cultured on Reasoner's 2A Agar or Broth at 30 °C.

Molecular cloning. Genomic DNA was isolated from overnight liquid cultures using the Zymo Research Quick-DNA Miniprep kit. oxz genes were amplified by PCR from this genomic DNA using NEB Q5 High-Fidelity polymerase. Cycling conditions were as follows: an initial denaturation temperature of 98 °C for 30 s, followed by 35 cycles of a denaturing temperature of 98 °C for 10 s, an annealing temperature of 68 °C for 30 s and an extension at 72 °C for 5 min. After cycling, a final extension at 72 °C for 2 min was performed. All amplicons were purified by gel extraction. oxzAB gene pairs were cloned into the Thermofisher Zero Blunt TOPo Kit, and Pr_oxzA and Pr_oxzB were cloned into pET28a without and with the sequence for MBP, respectively, using the NEB HiFi Assembly master mix. All plasmids were verified by Sanger sequencing.

Metabolite analysis of heterologously expressed oxazolones. E. coli BLR(DE3) containing pTOPO-Pr_oxzAB or pTOPO-Sa_oxzAB was grown in 3x 1.51 TB with 50 µM-1 kanamycin and 10 µM-1 glycinamycin at 30 °C. Expression was induced with 0.1 mM IPTG at an OD600 of 0.8, and the culture was left to shake overnight at 30 °C. Cells were pelleted at 10,000 × g for 20 min, and pellets were frozen and lyophilized for 2 d. Dried biomass was crushed, extracted overnight in 100 ml ethyl acetate, filtered and dried under reduced pressure. The residue was subjected to flash chromatography using hexane:ethyl acetate on a Combiflash EZ Prep system using a 24-g RediSep gold silica gel column. UV active peaks were then further subjected to preparative reverse-phase HPLC (Phenomenex Luna C18 column (100 mm × 2.1 mm, 5-µm particle size) using isocratic conditions between 80:20 and 90:10 acetonitrile:water with 0.1% formic acid, and samples were evaporated by rotary evaporation followed by lyophilization, yielding between 1 mg and 9 mg purified oxazolone, for which compound characterization data can be found in the Supplementary Note.

Methanol adduct formation. Two mg or 9 mg was dissolved in 1 ml methanol, and 10 mg K2CO3 was added, upon which the reaction with 4 turned bright yellow (the reaction with 9 stayed colorless). Reactions were left to react for 5 min at room temperature, filtered and purified by preparative reverse-phase HPLC as described above, under isocratic conditions with 78:22 acetonitrile:water and 0.1% formic acid, compound characterization data for which can be found in the Supplementary Note.

Protein purification and enzyme assays. Overnight cultures of E. coli BLR (DE3) harboring pET28a-His6-PrOxzA or pET28a-MBP-PrOxzB, respectively, were diluted into 1.51 TB with 50 µg ml−1 kanamycin and 10 µl ml−1 glycinamycin at 30 °C. When an OD600 of 0.8 was reached, flasks were cooled to 18 °C, expression was induced with 50 µM IPTG, cultures were left to shake at 18 °C overnight, Cells were pelleted at 4 °C for 15–20 min at 15,000 × g and resuspended in 40 ml cold 50 mM Tris, 200 mM NaCl and 10% glycerol at pH 7 (Tisys buffer) and sonicated at 50% amplitude, 15 s on, 15 s off for 5 min. Proteins were purified using loose resin in 50 ml centrifuge tubes. For His6-PrOxzA, 2.5 ml nickel-IDA resin was used, and the resin was washed twice with 40 ml lysis buffer with 50 mM imidazole, eluted with 5 ml lysis buffer with 500 mM imidazole and concentrated using a 15-kDa MWCO filter. For MBP-PrOxzB, 2.5 ml NEB amylase resin was used, and the resin was washed twice with 40 ml lysis buffer, eluted in 5 ml lysis buffer with 10 mM maltose and concentrated using a 50-kDa MWCO filter. MBP-PrOxzB is visibly yellow, which suggests that it binds a flavin cofactor, as expected for nitroreductase–family enzymes.

In vitro assays with His6–OxzA and MBP–OxzB were conducted in 200 µl 50 mM potassium phosphate, pH 7, 200 mM NaCl, 10% DMSO, with reagents at the following final concentrations: 150 µM ATP, 150 µM decanoyl-CoA, 100 µM L-tyrosine and 100 µM N-decanoyl-L-tyrosine (synthesis is described in Nature Chemical Biology | www.nature.com/naturechemicalbiology
the Supplementary Information). After 1 h, reactions were quenched with 100 µl acetonitrile, filtered through 0.2-µM filters and analyzed by LC–MS as described in the Supplementary Information.

**Analysis of oxazolone production in *P. rubra* and *C. chukchiensis***. Cells were grown as a lawn on Marine Agar 2216 with or without a drop of antibiotic stock (see Supplementary Table 2 for amounts) in the middle of the plate. After 2 d of growth, biomass adjacent to the zone of inhibition was collected by scraping the plate. An effort was made to collect a roughly equivalent amount of biomass from each plate. Biomass was lyophilized, extracted with ethyl acetate for 2 h, filtered through glass pipette cotton filters, evaporated under a stream of nitrogen gas, redissolved in 200 µl 35:65 acetonitrile:water, filtered through a 0.2-µm filter and analyzed by HPLC–UV–MS as described in the Supplementary Information.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

All data generated during this study are included in this article and its Supplementary Information. CO-ED networks used to generate Fig. 2 and several Supplementary Figs. showing CO-ED networks are available in Supplementary Data 1. The curated set of enzyme domains for CO-ED can be found as the node table of the ‘all_of_uniprot’ network in Supplementary Data 1, as well as at [https://github.com/tderond/CO-ED/blob/master/pfamID_to_name_desc_longdesc.tsv](https://github.com/tderond/CO-ED/blob/master/pfamID_to_name_desc_longdesc.tsv). This same set of domains is also the default setting for the web tool. NMR spectra of newly reported structures are available in the Supplementary Note. Tandem MS spectra of newly reported structures are available in the Supplementary Note and were also deposited to the GNPS spectral library at the URLs shown in the Supplementary Note. The following bioinformatic databases were employed in this study: BRENDA ([https://www.brenda-enzymes.org/](https://www.brenda-enzymes.org/)), MIBiG ([https://mibig.secondarymetabolites.org/](https://mibig.secondarymetabolites.org/)), UniProt ([https://www.uniprot.org/](https://www.uniprot.org/)) and Pfam-A ([http://pfam.xfam.org/](http://pfam.xfam.org/)). Source data are provided with this paper.

**Code availability**

Jupyter notebooks containing Python code for CO-ED analysis and for generating the statistics shown in Supplementary Figs. 1 and 18 are publicly available at [https://github.com/tderond/CO-ED](https://github.com/tderond/CO-ED).

**References**

51. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* 20, 1160–1166 (2019).

52. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504 (2003).

53. Blattner, F. R. et al. The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1462 (1997).

54. Xie, B.-B. et al. Genome sequence of the cycloprodigiosin-producing bacterial strain *Pseudoalteromonas rubra* ATCC 29570(T). *J. Bacteriol.* 194, 1637–1638 (2012).

55. Bentley, S. D. et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417, 141–147 (2002).

56. Udvari, D. W. et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc. Natl Acad. Sci. USA* 104, 10376–10381 (2007).

57. Paulsen, I. T. et al. Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat. Biotechnol.* 23, 873–878 (2005).

58. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* 7, e1002195 (2011).

**Acknowledgements**

We thank our UCSD colleagues J. Li, V. Shende, T. Fallon and B. Duggan for helpful discussions. This work was supported by National Institutes of Health awards F32GM129960 to T.d.R. and R01GM085770 to B.S.M., as well as an American Society for Pharmacognosy Undergraduate Research Award and a UC San Diego ‘Eureka’ Undergraduate Research Scholarship to J.E.A.

**Author contributions**

T.d.R. and B.S.M. designed research; T.d.R. and J.E.A. performed research; T.d.R. analyzed data; T.d.R. and B.S.M. wrote the paper.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Extended data is available for this paper at [https://doi.org/10.1038/s41589-021-00808-4](https://doi.org/10.1038/s41589-021-00808-4).

The online version contains supplementary material available at [https://doi.org/10.1038/s41589-021-00808-4](https://doi.org/10.1038/s41589-021-00808-4).

Peer review information *Nature Chemical Biology* thanks A. James Link, Maude Pupin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).
Extended Data Fig. 1 | Survey of transformations known to be catalyzed by the ThiF and nitroreductase domains, and properties of organisms harboring a ThiF-nitroreductase di-domain enzyme identified by Co-ED. a, OxzB exhibits homology to both the ThiF and nitroreductase enzyme families. Shown is a selection of characterized members of these enzyme families along with the transformations they catalyze, along with midpoint-rooted gene trees showing the relationships between each other and OxzB. ThiF family enzymes are known to catalyze ATP-dependent carboxylate activating reactions, while nitroreductase family enzymes catalyze a variety of redox reactions. Inferred phylogenies were generated from protein sequences using neighbor joining on the MAFFT web server and midpoint-rooted. Scale bars designate 1 substitution per site. b,c, Phylogenetic distribution (b) and habitat (c) of the host organisms harboring genes encoding OxzB proteins represented in the Uniprot database. Organisms with unknown habitats are not included. d, Genomic context of oxzB genes as determined using the Enzyme Function Initiative Genome Neighborhood Tool. Most oxzB homologs are accompanied by oxzA, which codes for an N-acyltransferase. The arrow labeled ‘?’ represents genes unrelated to oxzA or oxzB. Organisms with unknown oxzB genomic context (for example, at the edge of a contig) are not included.
Extended Data Fig. 2 | Induction of oxazolone production in *P. rubra* and *C. chukchiensis* by various antibiotics. Bacteria were grown as a lawn on Marine Agar 2216 with a drop of antibiotic. A consistent amount of biomass adjacent to the zone of inhibition was harvested, extracted and analyzed by HPLC. Dots indicate summed peak areas between wavelengths of 300 nm and 400 nm. Three biological replicates were analyzed for each condition.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

☐ n/a Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

HPLC-UV-MS data was collected using Agilent Masshunter 6, NMR spectra were collected using Jeol Delta 5 or Bruker Topspin 3.6

Data analysis

Jupyter notebooks for CO-ED analysis are publicly available at https://github.com/tderond/CO-ED written in Python v3.6.10, Jupyter notebook v6.0.3, and pandas v1.0.3. The workflow also made use of PfamScan 1.6, which in turn made use of HMMER 3.3. A public web server is publicly available at http://enzyme-analysis.org. CO-ED networks were drawn with Cytoscape 3.5 or 3.8. NMR spectra were analyzed using Mnova 12.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated during this study are included in this article and its supplementary information files.
- The CO-ED networks used to generate Figure 2 and the several supplemental figures showing CO-ED networks, are available as Supplementary Data 1
- The data used to generate Extended Data Figs. 1 and 2 are available as Source Data
- The curated set of enzyme domains for CO-ED can be found as the node table of the all_of_uniprot network in Supplementary Data 1, as well as at https://github.com/tderond/CO-ED/blob/master/ofamID_to_name_desc_longdesc.tsv. This same set of domains is also the default setting for the web tool.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-list.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

The only experiment where sample size is relevant is the one shown in Extended Data Fig. 2, wherein three independent biological replicates were analyzed for each condition. This data is intended to highlight qualitative differences and hence triplicate was deemed sufficient. No statistical methods used to predetermine sample size were used in this study.

### Data exclusions

No excluded data

### Replication

All attempts at replication as part of ongoing follow-up studies have been successful. In vivo oxazolone production has been independently replicated with similar results four times, purification of His6-OxzA and MBP-OxzB has been independently carried out with similar results twice, and in vitro enzyme assays have been replicated with similar results at least five times.

### Randomization

Randomization was not applicable to this study as for all samples in an experiment, the same bacterial culture was used, and the samples were analyzed in identical fashion.

### Blinding

Blinding was not applicable to this study because no manual counting, scoring or sorting was employed

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a  | n/a |
| ☑ Antibodies  | ☑ ChIP-seq |
| ☑ Eukaryotic cell lines  | ☑ Flow cytometry |
| ☑ Palaeontology and archaeology  | ☑ MRI-based neuroimaging |
| ☑ Animals and other organisms  | |
| ☑ Human research participants  | |
| ☑ Clinical data  | |
| ☑ Dual use research of concern  | |