nature chemical biology

ARTICLES

https://doi.org/10.1038/s41589-018-0043-2

Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein

Ben E. Clifton[1], Joe A. Kaczmarski[1], Paul D. Carr[1], Monica L. Gerth[2,4], Nobuhiko Tokuriki[3] and Colin J. Jackson[1*]

The emergence of enzymes through the neofunctionalization of noncatalytic proteins is ultimately responsible for the extraordinary range of biological catalysts observed in nature. Although the evolution of some enzymes from binding proteins can be inferred by homology, we have a limited understanding of the nature of the biochemical and biophysical adaptations along these evolutionary trajectories and the sequence in which they occurred. Here we reconstructed and characterized evolutionary intermediate states linking an ancestral solute-binding protein to the extant enzyme cyclohexadienyl dehydratase. We show how the intrinsic reactivity of a desolvated general acid was harnessed by a series of mutations radiating from the active site, which optimized enzyme–substrate complementarity and transition-state stabilization and minimized sampling of noncatalytic conformations. Our work reveals the molecular evolutionary processes that underlie the emergence of enzymes de novo, which are notably mirrored by recent examples of computational enzyme design and directed evolution.

Much of the functional diversity observed in modern enzyme superfamilies originates from molecular tinkering with existing enzymes[1]. New enzymes frequently evolve from enzymes with latent, promiscuous activities[2] and often inherit key features of the ancestral enzyme, retaining conserved catalytic groups and stabilizing analogous intermediates or transition states[3]. Experimental evolutionary biochemistry has yielded considerable insight into the evolution of new enzymes from existing enzymes[4]; however, the emergence of catalytic activity de novo remains poorly understood. Although certain enzymes are thought to have evolved from noncatalytic proteins[5–7], the mechanisms underlying these complete evolutionary transitions have not been described.

To explore the evolutionary processes underlying the emergence of enzymes by neofunctionalization of noncatalytic proteins, we investigated the evolution of cyclohexadienyl dehydratase (CDT; EC 4.2.1.51, 4.2.1.91). This enzyme catalyzes the cofactor-independent Grob-type fragmentation of prephenate and L-arogenate to yield phenylpyruvate and L-phenylalanine, respectively[8] (Fig. 1a,b). CDT is one of several enzymes that appear to have evolved from solute-binding proteins (SBPs), which comprise an abundant and adaptable superfamily of extracytoplasmic receptors that are mainly involved in solute transport and chemotaxis in association with bacterial ATP-binding cassette (ABC) importers and chemotactic receptors[9] (Supplementary Table 1). The relationship between CDTs and SBPs was initially recognized on the basis of sequence similarity between CDTs and polar amino acid–binding proteins (AABPs)[5]. More recently, crystal structures of CDT from *Pseudomonas aeruginosa* (PaCDT; PDB ID 3KBR) and a putative AABP from *Wolinella succinogenes* (Ws0279, 26% sequence identity; PDB ID 3K4U) from structural genomics projects have further supported the close evolutionary relationship between CDTs and AABPs. The periplasmic binding protein-like II fold shared by PaCDT and Ws0279 consists of two α/β domains connected by two flexible hinge strands, and the ligand binding site is located at the interface of the two domains (Fig. 1c).

In this work, we used ancestral protein reconstruction[10] to investigate the biophysical and biochemical mechanisms underlying the evolutionary transition between SBPs and CDTs. By analyzing the evolutionary trajectory between reconstructed ancestors and extant proteins, we show that the emergence and optimization of catalytic activity involves several distinct processes. The emergence of CDT activity was potentiated by the incorporation of a desolvated general acid into the ancestral binding site, which provided an intrinsically reactive catalytic motif, and reshaping of the ancestral binding site, which facilitated enzyme–substrate complementarity. Catalytic activity was subsequently gained via the introduction of hydrogen bonding networks that positioned the catalytic residue precisely and contributed to transition-state stabilization. Finally, catalytic activity was enhanced by remote substitutions that refined the active site structure and reduced sampling of noncatalytic states.

Results

Evolutionary history of CDT. Ws0279, the SBP with the highest sequence identity to CDT that has been functionally or structurally characterized, has been annotated as a lysine-binding protein on the basis of sequence homology. We evaluated the binding specificity of Ws0279 using differential scanning fluorimetry (DSF), thereby confirming that the protein is an AABP that is specific for the cationic amino acids L-lysine and, to a lesser extent, L-arginine (Supplementary Fig. 1a).

To reconstruct the evolutionary history of CDT, we inferred the maximum-likelihood phylogeny of 113 homologs of Ws0279 and PaCDT and used ancestral protein reconstruction to infer the most likely amino acid sequence for each ancestral node in the phylogeny (Fig. 1d and Supplementary Fig. 2). We selected five ancestral nodes, designated AncCDT-1 to AncCDT-5, for experimental characterization based on patterns of sequence conservation in the extant sequences (Fig. 1d). AncCDT-1 represents the last common ancestor of Ws0279 and PaCDT, whereas the other ancestral nodes represent intermediates in the evolution of PaCDT from AncCDT-1.

[1]Research School of Chemistry, Australian National University, Canberra, ACT, Australia. [2]Department of Biochemistry, University of Otago, Dunedin, New Zealand. [3]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada. [4]Present address: School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand. *e-mail: colin.jackson@anu.edu.au

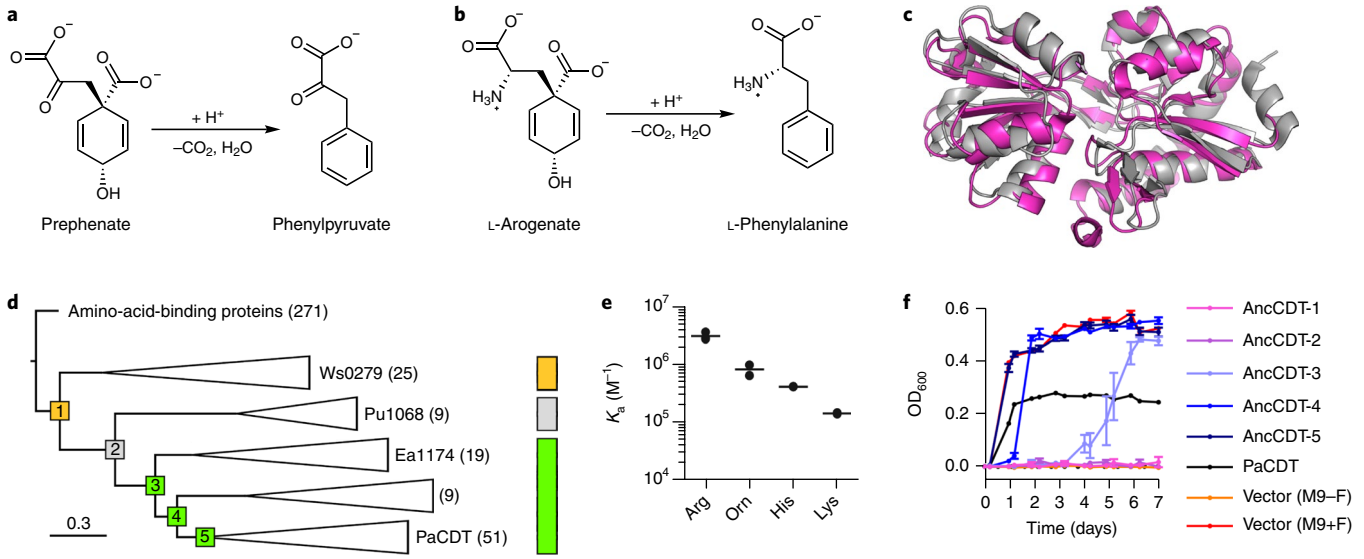

Fig. 1 | Functional evolution of CDT. a,**b**, Fragmentation reactions of prephenate (**a**) and L-arogenate (**b**) catalyzed by CDT. **c**, Structural similarity between PaCDT (gray; PDB ID 3KBR) and Ws0279 (pink; PDB ID 3K4U) (r.m.s. deviation of 2.25 Å for backbone atoms). **d**, Condensed maximum-likelihood phylogeny of CDT homologs. The scale bar represents the mean number of substitutions per site. The five compressed clades are labeled with the corresponding number of sequences and the representative extant protein characterized in this work. The five ancestral nodes that were characterized experimentally (AncCDT-1 to AncCDT-5) are labeled and colored according to function (gold, amino acid binding; gray, binding of unknown solute; green, CDT). **e**, Affinity of AncCDT-1 for cationic amino acids, determined by ITC (Orn, L-ornithine). The mean and individual values for two (Orn, Lys) or three (Arg, His) titrations are shown. **f**, Growth of auxotrophic *E. coli* Δ*pheA* cells complemented with ancestral proteins or PaCDT in selective M9 – F (M9 without L-phenylalanine) media. Growth curves of empty vector transformants in selective M9 – F media and unselective M9 + F (M9 with L-phenylalanine) media are shown as negative and positive controls, respectively. Results are mean ± s.e.m. for four replicate cultures (AncCDT-5) or five replicate cultures (otherwise). OD_{600}, optical density at 600 nm.

We experimentally characterized the five ancestral proteins, using isothermal titration calorimetry (ITC) to test for amino acid binding and genetic complementation to test for enzymatic activity; in the genetic complementation assay, expression of CDT rescues the growth of *Escherichia coli* L-phenylalanine auxotrophs that lack prephenate dehydratase encoded by the gene *pheA*[8]. AncCDT-1 is an AABP displaying high affinity and broad specificity for cationic amino acids, including L-arginine ($K_d = 0.32\,\mu M$), L-ornithine (1.2 μM), L-histidine (2.3 μM) and L-lysine (6.7 μM) (Fig. 1e and Supplementary Fig. 1b). Neither AncCDT-2 nor any subsequent ancestral protein exhibited binding of proteinogenic amino acids. AncCDT-3, AncCDT-4, and AncCDT-5 have sufficient CDT activity to rescue growth of *E. coli* Δ*pheA* cells in minimal media (Fig. 1f).

We next tested the robustness of the predicted ancestral sequences to variations in the phylogenetic analysis. An alternative phylogeny was reconstructed using an alternative substitution model, and the resulting phylogeny was used together with the alternative substitution model to reconstruct different versions of the ancestral proteins, designated AncCDT-1W to AncCDT-5W (Supplementary Figs. 2b and 3). The alternative ancestral proteins differed from the maximum-likelihood ancestral proteins at 4%–9% of sites, and these sequence differences reflect uncertainty in both the phylogenetic tree and the parameters of the substitution model (Supplementary Table 2). The alternative ancestral proteins gave results that were qualitatively similar to those of the maximum-likelihood ancestral proteins in genetic complementation assays, supporting the qualitative conclusion that CDT activity evolved between AncCDT-2 and AncCDT-3 (Supplementary Fig. 4a). However, some substantial quantitative differences in growth were observed; most importantly, AncCDT-3W transformants exhibited faster growth than AncCDT-3 transformants. Recombination of the two genes using staggered extension PCR followed by genetic selection showed that a single substitution (P188L) in AncCDT-3 was sufficient to recapitulate the higher growth rate associated with AncCDT-3W (Supplementary Fig. 4b). Spectrophotometric kinetic assays in vitro confirmed that AncCDT-3 and AncCDT-3(P188L), but not AncCDT-2, have prephenate dehydratase activity (Supplementary Fig. 4d–h). Altogether, characterization of both sets of ancestral proteins supported the conclusions that ancestral amino acid–binding activity was lost between AncCDT-1 and AncCDT-2, that CDT activity was gained between AncCDT-2 and AncCDT-3, and that AncCDT-2 apparently had neither CDT activity nor binding affinity toward amino acids.

An intermediate function between AABPs and CDTs. To test whether AncCDT-2 was rendered nonfunctional by an error in its reconstructed sequence or if it had a function distinct from AncCDT-1 and AncCDT-3, we examined representatives of the previously uncharacterized evolutionary clades consisting of extant descendants of AncCDT-2 and AncCDT-3: Pu1068 from "*Candidatus Pelagibacter* ubique" and Ea1174 from *Exiguobacterium antarcticum* (Fig. 1d). Genetic complementation experiments showed that Ea1174, but not Pu1068, has CDT activity (Supplementary Fig. 4c), and DSF experiments showed that Pu1068 is not an AABP (Supplementary Fig. 1c). Analysis of the genomic context of *Pu1068* and several of its orthologs revealed that these genes, like the SBP gene *Ws0279*, are adjacent to genes encoding transmembrane components of ABC importers, suggesting that *Pu1068* encodes an SBP rather than an enzyme (Supplementary Table 3).

We used two strategies that have previously been used for functional annotation of SBPs[11] to identify the physiological ligands of Pu1068 and AncCDT-2: crystallization of Pu1068 with co-purified ligands bound during heterologous expression of the protein

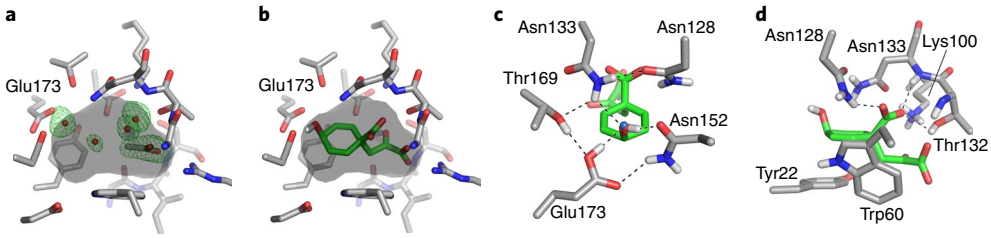

Fig. 2 | Crystal structure of PaCDT. a, Active site of the PaCDT–acetate complex. The surface of the active site is shown in gray. Electron density for water and acetate molecules is shown by an omit mF_o – DF_c map contoured at +3σ. **b**, Structure of the PaCDT–prephenate complex predicted by computational docking. Docking with L-arogenate yielded a similar pose. **c**, Glu173 is poised for proton donation to the departing hydroxyl group of prephenate by hydrogen bonding interactions with neighboring residues. The position predicted to be occupied by the hydroxyl group of prephenate is occupied by a water molecule in the unliganded PaCDT structure (blue sphere). **d**, π-stacking interactions with Tyr22 and Trp60 and polar interactions with Lys100, Asn128, Thr132, and Asn133 could contribute to transition-state stabilization.

in *E. coli* and DSF screening of Pu1068 and AncCDT-2 against metabolite libraries. Although crystallization and structure determination of Pu1068 were successful, the protein crystallized in the unliganded state, suggesting the absence of high-affinity ligands of the protein in the *E. coli* metabolome. DSF experiments to identify the ligands of Pu1068 and AncCDT-2 were performed using several hundred potential metabolites from libraries and rationally selected metabolites with plausible physiological importance for oceanic bacteria such as *C. Pelagibacter* ubique (Supplementary Fig. 5 and Supplementary Dataset 1). Although the exact physiological ligands of AncCDT-2 and Pu1068 could not be identified, we found that these proteins have some affinity for a variety of carboxylates (Supplementary Fig. 5) and for some sulfonates, such as the sulfobetaine NDSB-221, which binds Pu1068 with a K_d of 0.53 mM (Supplementary Fig. 6). Regardless of the specific physiological ligands of AncCDT-2 and Pu1068, the functional properties of the various extant clades (Ws0279: cationic AABP; Pu1068: SBP of unknown function; Ea1174 and PaCDT: CDTs) agreed with those expected based on functional characterization of the ancestral proteins, supporting a likely evolutionary trajectory from a cationic AABP, to a carboxylic-acid-binding protein, and finally to CDT, an enzyme with carboxylic acid substrates (Fig. 1d).

Emergence of catalytic activity. To establish the molecular basis for the functional transition from binding protein to enzyme, we first attempted to rationalize the catalytic activity of the extant enzyme PaCDT in structural terms. PaCDT has previously been crystallized (PDB ID 3KBR) in complex with the nonphysiological ligand HEPES, which shares some fortuitous similarities with the cyclohexadienol substrates of the enzyme. We therefore attempted to solve the crystal structure of the native, unliganded enzyme, and obtained a structure in which the active site cavity was occupied by only an acetate molecule from the crystallization buffer and four ordered water molecules (Fig. 2a). Unlike that in the structure of the PaCDT–HEPES complex, the active site of the PaCDT–acetate complex was fully occluded from solvent and highly complementary to its cyclohexadienol substrates (Fig. 2a and Supplementary Fig. 7). Docking of prephenate and L-arogenate into the PaCDT–acetate structure implied a binding mode in which Glu173 is positioned adjacent to the departing hydroxyl group of the substrate, suggesting that the enzyme mechanism involves general acid catalysis by Glu173 (Fig. 2b and Supplementary Fig. 8a). Consistent with its proposed role as a general acid, Glu173 is partially desolvated and predicted by PROPKA to be protonated at neutral pH (pK_a = 7.75), and the substitution E173Q abolishes prephenate dehydratase activity with minimal impact on secondary structure and thermostability (Supplementary Fig. 8b–d). The active site of PaCDT is pre-organized for protonation and elimination of the departing hydroxyl group of the substrate by an intricate hydrogen bonding network extending from Glu173 (Fig. 2c). Other active site residues most likely contribute to stabilization of the departing carboxylate group and delocalized electrons in the developing π system in the transition state (Fig. 2d).

We next solved the crystal structures of AncCDT-1 and AncCDT-3(P188L). Comparison of these structures with the structures of the extant proteins PaCDT and Pu1068 revealed the contribution of historical amino acid substitutions to remodeling, functionalization, and refinement of the ancestral amino acid binding site (Fig. 3a–e). First, mutations that occurred between AncCDT-1 and AncCDT-2 caused two important structural changes that potentiated the emergence of catalytic activity: the substitution V173E introduced a general acid that is positioned appropriately for general acid catalysis (Fig. 3b), whereas substitutions D19T and A20G allowed a conformational change for Trp60, reshaping the ancestral binding site and facilitating steric complementarity between CDT and its substrates (Fig. 3c). These substitutions can be considered potentiating because the structural features associated with them are also observed in Pu1068, and they initially enabled binding of a different ligand rather than CDT activity (Fig. 3d). Indeed, each residue associated with these structural changes was reconstructed with high statistical confidence in the noncatalytic protein AncCDT-2 (Supplementary Fig. 2c). Thus, the evolution of CDT from AABPs required the acquisition of a new binding function, resulting in the introduction of amino acids that potentiated the structure for subsequent evolution of catalytic function.

Structural analysis indicated that functionalization of the ancestral binding site for catalytic activity occurred by subsequent mutations that fixed either between AncCDT-1 and AncCDT-2 or between AncCDT-2 and AncCDT-3. The substitutions Q100K, Q128N, and S133N introduced the hydrogen bonding network that positions the catalytic group precisely and contributes to transition-state stabilization through interactions with the departing hydroxyl and carboxylate groups of the substrate (Fig. 3b). Additionally, the substitutions Q100K and L198K likely contributed to dual specificity for α-amino and α-keto acid substrates (i.e., L-arogenate and prephenate) via electrostatic shielding of Asp170 (Fig. 3e). However, AncCDT-2 contains each of these four active site substitutions (except L198K, which is not itself sufficient to introduce catalytic activity) (Fig. 3a), implying that additional substitutions between AncCDT-2 and AncCDT-3 were required for the emergence of CDT activity. To identify these substitutions, we performed site-directed mutagenesis and three rounds of directed evolution, which resulted in the isolation of an AncCDT-2 variant with only six substitutions (CDT-M5) that allowed slow growth of *E. coli* L-phenylalanine auxotrophs and exhibited prephenate dehydratase activity in vitro (Fig. 3f and Supplementary Figs. 4h and 9). Although three of these substitutions (T131G, A155I, and L198K) are present in AncCDT-3, the other three substitutions (F25L, G99S, and P102L) represent an

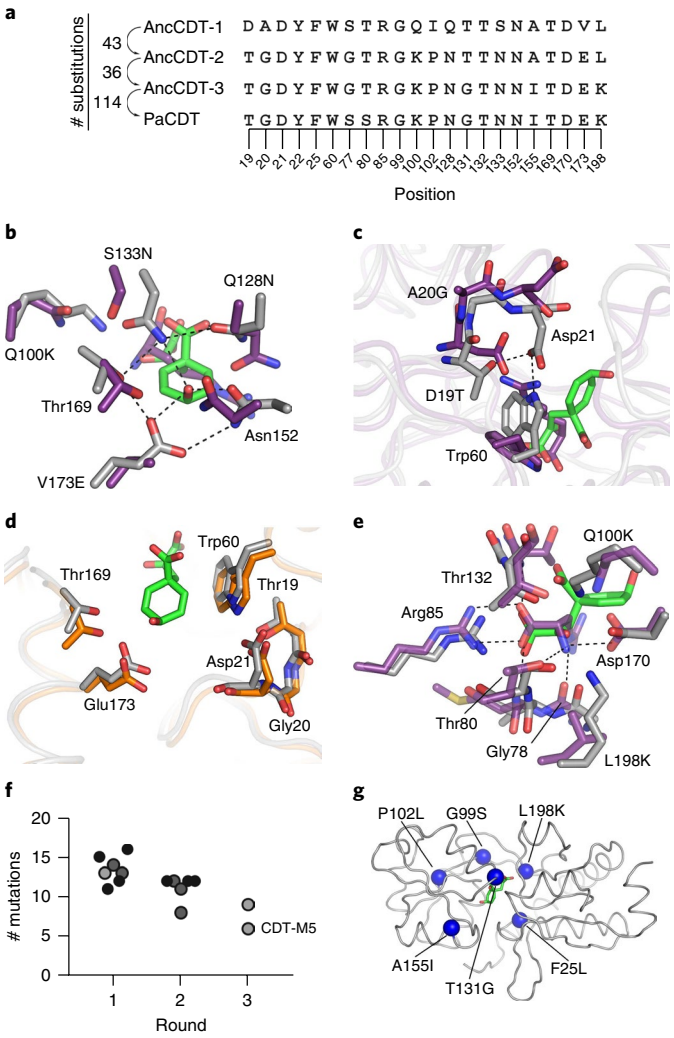

Fig. 3 | Structural and mutational basis for evolution of CDT activity. **a**, Multiple sequence alignment of ancestral proteins and PaCDT at positions important for CDT activity. The number of substitutions between each sequence in this evolutionary trajectory is shown. **b–e**, Comparison of the AncCDT-1–L-arginine (purple), Pu1068 unliganded (orange), and PaCDT–acetate (gray) structures. Positions are labeled with the corresponding residue in AncCDT-1 and AncCDT-3, if conserved in both proteins, or with the corresponding substitution between the two proteins. The green structures represent the position of L-arogenate that has been docked into the active site of PaCDT. **b**, Functionalization of the ancestral binding site introduced the general acid Glu173 and residues required for substrate binding and transition-state stabilization. **c**, The ancestral binding site was remodeled by a conformational change of Trp60. D19T introduces a hydrogen bond with Asp21, and A20G enables a conformation disfavored for nonglycine residues but necessary for the interaction between Thr19 and Asp21. **d**, Structural similarities between Pu1068 and PaCDT. The two domains of Pu1068 were superimposed separately on the structure of PaCDT. **e**, CDT inherited the α-amino acid–binding motif from AncCDT-1, with two substitutions (Q100K and L198K) that also enable binding of the α-keto acid prephenate. **f**, Introduction of CDT activity into AncCDT-2 by directed evolution. Each point represents a unique clone, and the color gives a qualitative indication of activity (black, high activity; dark gray, moderate activity; light gray, low activity). See also, Supplementary Fig. 9. **g**, Positions of six substitutions sufficient to introduce CDT activity into AncCDT-2. F25L, G99S, P102L and A155I are located in the second or third shells of the active site.

alternative evolutionary trajectory toward higher catalytic activity. While the T131G substitution removes a steric clash between the enzyme and the departing carboxylate group of the substrate and the L198K substitution assists the binding of the ketone group, the other four substitutions are located in the second or third shells of the active site and must have indirect effects on catalysis (Fig. 3g). The introduction of additional mutations in various combinations supported faster growth of L-phenylalanine auxotrophs (Fig. 3f and Supplementary Fig. 9d). These results show that there are multiple mutational pathways to higher CDT activity via remote substitutions following the introduction of key active site residues.

Evolution of an efficient enzyme. Although AncCDT-3(P188L) has CDT activity, its second order rate constant (k_{cat}/K_M) is ~6,000-fold lower than that of PaCDT despite their active sites being virtually identical (Supplementary Figs. 4h and 10). We therefore investigated the role of structural dynamics in the evolutionary process. Upon ligand binding, SBPs undergo domain-scale open–closed conformational changes that are essential for function[12], and these are exemplified by the unliganded and arginine-bound crystal structures of AncCDT-1 (Fig. 4a). The open–closed conformational equilibrium of an SBP controls binding affinity[13] and the rate of solute transport[12], suggesting that the position of this equilibrium must be adjusted for optimization of solute transport in a cellular context. On the other hand, efficient enzyme catalysis depends on pre-organization of the active site; unproductive conformational sampling has been shown to constrain the catalytic efficiency of recently evolved enzymes[14,15]. The closed conformation of CDT is the catalytically competent conformation; the open–closed conformational change would be necessary only to the extent needed to enable substrate binding and product release from the occluded active site. These considerations suggest a possible point of adaptive conflict between solute binding and catalytic activity that may have necessitated changes in structural dynamics during the evolution of CDT.

The unliganded SBPs AncCDT-1 and Pu1068 and the inefficient ancestral enzyme AncCDT-3(P188L), whose structures were solved in this work, crystallized in an open conformation (Fig. 4a). This is consistent with previous studies showing that unliganded AABPs sample closed or semi-closed conformations only transiently[12,16], as well as with previously reported crystal structures of unliganded AABPs, of which only 1/14 crystallized in a closed conformation (Supplementary Table 4). By contrast, PaCDT crystallized in a closed conformation in the absence of substrate or substrate analogs in multiple, differently packed crystals, suggesting that the closed conformation of the enzyme is unusually stable for this protein fold (Fig. 4a and Supplementary Fig. 7a). It should be noted that the adventitiously bound acetate molecule observed in the active site of the PaCDT–acetate structure makes several polar interactions with the large domain of the enzyme (via Arg85 and Ser80), but no polar interactions with the small domain, with Thr132 being the only residue within potential hydrogen bonding distance (Supplementary Fig. 11). The acetate molecule is therefore unlikely to make a substantial contribution to stabilization of the closed conformation of the PaCDT–acetate complex.

To further investigate the structural dynamics of PaCDT, we initialized molecular dynamics (MD) simulations of the PaCDT trimer from the PaCDT–acetate structure with the acetate molecule removed. These simulations indicated that the open conformation is accessible in PaCDT, although most of the subunits remained closed throughout each 170 ns trajectory (Fig. 4b,c and Supplementary Fig. 12g). Additional simulations using a different initial structure (PaCDT–HEPES structure, with the HEPES molecule removed) or a different force field gave similar results (Supplementary Fig. 12a,b). In contrast, MD simulations of AncCDT-1 initialized from unliganded closed and unliganded open structures showed that the ancestral protein had dynamical properties typical of SBPs; in each simulation, the unliganded closed structure transitioned to an open conformation within 150 ns (Supplementary Fig. 12d,f,i), whereas the unliganded open structure remained in an open conformation for the duration of the simulations (Supplementary Fig. 12c,e,h).

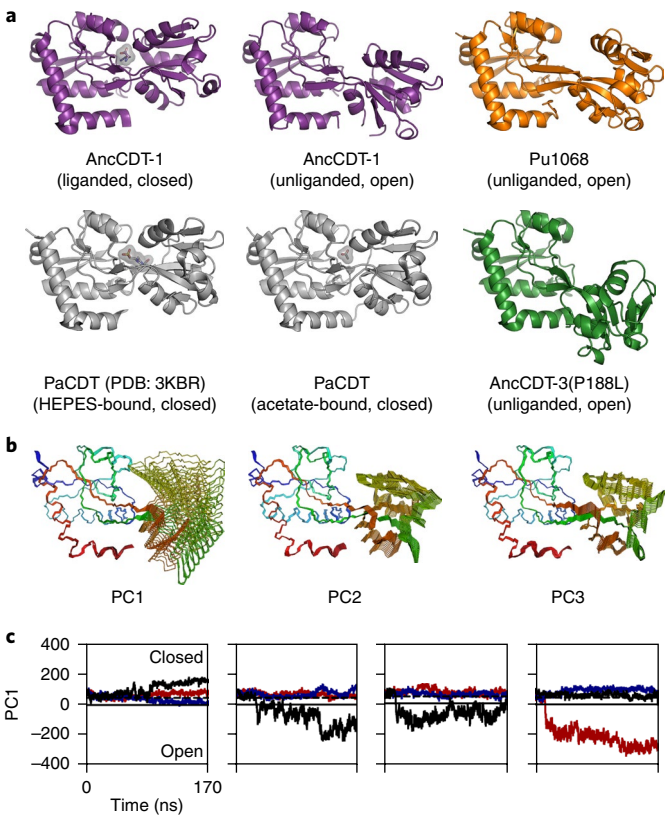

Fig. 4 | Structural dynamics of CDT. a, Open and closed structures of AncCDT-1 (purple), Pu1068 (orange), PaCDT (gray), and AncCDT-3(P188L) (green). Unusually, PaCDT adopts a closed structure in the absence of a substrate or substrate analog. **b**, Principal component analysis of MD simulations of PaCDT ($n = 8$ independent trajectories of the PaCDT homotrimer). The structures illustrating the physical interpretation of the first three principal components (PCs) were generated by interpolating between structures at the extremities of each principal component axis. These motions represent hinge-bending and hinge-twisting motions typical of AABPs[19,20]. **c**, Open–closed conformational dynamics in 4 × 170 ns simulations of PaCDT, initialized from the PaCDT–acetate structure using the GROMOS 53a6 force field. Projections of the trajectories of individual PaCDT subunits onto the PC1 axis are shown. Each color represents a subunit of the PaCDT homotrimer. The dotted line represents the crystallographic conformation observed in the PaCDT–acetate structure.

The domain-scale conformational fluctuations that did occur in the PaCDT MD simulations were characteristic of SBPs; principal component analysis showed that hinge-bending and hinge-twisting motions typical of AABPs[17,18] accounted for >85% of conformational variance (Fig. 4b). Indeed, these domain-scale motions were similar to those observed in the simulations initialized from the unliganded, closed AncCDT-1 structure. Furthermore, the open structure of AncCDT-3(P188L), which provided experimental evidence for sampling of the open conformation in CDTs, resembled the simulated open conformation of PaCDT (Supplementary Fig. 12g). Thus, the characteristic domain-scale dynamics of the SBP fold are retained in CDTs and are indeed necessary for substrate or product diffusion from the occluded active site. However, the unusual stability of the closed conformation of PaCDT suggests that the conformational landscape of the enzyme has evolved between AncCDT-3(P188L) and PaCDT to minimize unproductive sampling of the noncatalytic open conformation, contributing to improvements in catalytic efficiency toward the end of the evolutionary trajectory.

Discussion

In this work, we have outlined the steps required for a noncatalytic protein to not only evolve some initial catalytic activity, but also to become a proficient, bona fide enzyme (for example, PaCDT, k_{cat}/K_M ~$10^6\,M^{-1}\,s^{-1}$). One of the most notable aspects of the evolutionary trajectory of CDT is the extent to which the emergence of catalysis depended on the repurposing of structural features of the ancestral SBPs AncCDT-1 and AncCDT-2. As an obvious example, the ancestral AABP AncCDT-1 provided an amino acid binding motif that was exploited for substrate binding in CDT. Additionally, some residues that are important for CDT activity, but not the ancestral amino acid binding activity, are observed in AncCDT-1 (Asp21, Asn152, and Thr169). The presence of these residues in AncCDT-1 shows that they were compatible with the ancestral function and could have evolved neutrally. Moreover, our phylogenetic and functional analyses showed that the evolution of CDT, which has enzymatic activity on the carboxylic acid prephenate, from an AABP required the prior acquisition of a new binding function, represented by AncCDT-2 and the extant protein Pu1068, which have affinity for carboxylic acids. The evolution of this new binding function resulted in fixation of two structural changes that were required for the evolution of CDT: the conformational change of Trp60 and the incorporation of Glu173, which later became the catalytic acid in CDT, into the binding site. Altogether, these results suggest that the major change in function between AABPs and CDT was contingent on various preexisting structural features, both essential and nonessential with respect to the ancestral functions.

Directed evolution of ancestral CDT variants highlighted the existence of multiple mutational pathways to higher catalytic activity following the introduction of key catalytic residues, raising the possibility that the evolution of CDT activity could have occurred gradually and nondeterministically. Previous studies of enzyme evolution have provided examples of evolutionary trajectories that are deterministic and constrained, in which the occurrence of a particular mutation determines the nature and occurrence of subsequent mutations[19,20]. In contrast, our directed evolution experiments demonstrated that there were alternative mutational pathways to higher catalytic activity in ancestral CDT variants that were distinct from the historical mutational pathway between AncCDT-2 and AncCDT-3, indicating either that the evolutionary trajectory of CDT was nondeterministic or that the historical mutations were fixed instead of the alternative mutations because of pleiotropy and/or adaptive conflict (i.e., because the alternative mutations had negative effects on fitness independent of their effect on enzymatic activity). One potential reason for the existence of numerous mutations that increase the enzymatic activity of ancestral CDT variants is the relatively simple chemistry associated with the enzyme mechanism; CDT has a simple one-step (or possibly two-step) mechanism that depends on a single catalytic residue, and the substrate of CDT is a high-energy metabolic intermediate that is predisposed to the reaction catalyzed by the enzyme. Therefore, higher catalytic activity might be accessed by a larger number of mutational pathways in CDT than in enzymes that catalyze more complex reactions with higher activation energy barriers, such as hydrolysis of unactivated substrates, for which more precise positioning of active site residues is likely necessary.

Our structural analysis of evolutionary intermediates and extant proteins demonstrates in molecular detail how the evolution of highly specialized and efficient CDTs from noncatalytic ancestors occurred in several distinct stages. Incorporation of the desolvated general acid Glu173 into the binding pocket of an ancestral SBP most likely provided sufficient chemical reactivity for initial, promiscuous CDT activity. Indeed, the intrinsic reactivity of desolvated acidic and basic residues has been exploited similarly in enzymes that have evolved recently in response to anthropogenic substrates[21] and in enzymes engineered via single substitutions in noncatalytic proteins[22]. Following the introduction of a reactive general acid, optimization of enzyme–substrate complementarity and introduction of hydrogen

bonding networks to position the catalytic residue precisely and stabilize the departing carboxylate group of the substrate appear to have occurred. Further improvements in catalytic efficiency could have been gained by second- and third-shell substitutions that refine the structure of the active site and optimize conformational sampling to favor catalytically relevant conformations. Similar mutational patterns have been documented in directed evolution experiments[14,23]. Additionally, adaptation of protein dynamics has been shown to occur analogously in the evolution of a binding protein from an enzyme, in which the catalytically relevant conformation was disfavored by the function-switching mutation[24].

Although some computationally designed protein structures have been made with atomic-level accuracy[25], and various strategies have been developed to introduce catalytic activity into arbitrary protein scaffolds[22,26,27], replicating the catalytic proficiency of natural enzymes using computational design remains a major challenge[28,29]. The evolutionary trajectory of CDT has striking similarities with the optimization of rationally designed enzymes by directed evolution; catalytic activity can be initialized by computationally guided grafting of a reactive catalytic motif (for example, a desolvated carboxylate) into a protein scaffold that can accommodate the transition state for a given reaction, and directed evolution can be used to introduce additional stabilizing interactions, optimize positioning of catalytic groups, improve enzyme-transition state complementarity, and optimize conformational sampling, frequently via remote substitutions[30,31]. Thus, the strategies that have been used to improve catalytic activity in computational design and directed evolution experiments appear to mirror those that drove the emergence of an enzyme from a non-catalytic protein by natural evolution.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41589-018-0043-2.

Received: 31 July 2017; Accepted: 1 March 2018;

References

1. Baier, F., Copp, J. N. & Tokuriki, N. Evolution of enzyme superfamilies: comprehensive exploration of sequence-function relationships. *Biochemistry* **55**, 6375–6388 (2016).
2. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
3. Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M. & Orengo, C. A. Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *J. Mol. Biol.* **428** 2 Pt A, 253–267 (2016).
4. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).
5. Tam, R. & Saier, M. H. Jr. A bacterial periplasmic receptor homologue with catalytic activity: cyclohexadienyl dehydratase of *Pseudomonas aeruginosa* is homologous to receptors specific for polar amino acids. *Res. Microbiol.* **144**, 165–169 (1993).
6. Ngaki, M. N. et al. Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. *Nature* **485**, 530–533 (2012).
7. Ortmayer, M. et al. An oxidative N-demethylase reveals PAS transition from ubiquitous sensor to enzyme. *Nature* **539**, 593–597 (2016).
8. Zhao, G. S., Xia, T. H., Fischer, R. S. & Jensen, R. A. Cyclohexadienyl dehydratase from *Pseudomonas aeruginosa*. Molecular cloning of the gene and characterization of the gene product. *J. Biol. Chem.* **267**, 2487–2493 (1992).
9. Berntsson, R. P.-A., Smits, S. H. J., Schmitt, L., Slotboom, D.-J. & Poolman, B. A structural classification of substrate-binding proteins. *FEBS Lett.* **584**, 2606–2617 (2010).
10. Hochberg, G. K. A. & Thornton, J. W. Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).
11. Vetting, M. W. et al. Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. *Biochemistry* **54**, 909–931 (2015).
12. Gouridis, G. et al. Conformational dynamics in substrate-binding domains influences transport in the ABC importer GlnPQ. *Nat. Struct. Mol. Biol.* **22**, 57–64 (2015).
13. Marvin, J. S. & Hellinga, H. W. Manipulation of ligand binding affinity by exploitation of conformational coupling. *Nat. Struct. Mol. Biol.* **8**, 795–798 (2001).
14. Campbell, E. et al. The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944–950 (2016).
15. Bar-Even, A., Milo, R., Noor, E. & Tawfik, D. S. The moderately efficient enzyme: futile encounters and enzyme floppiness. *Biochemistry* **54**, 4969–4977 (2015).
16. Bermejo, G. A., Strub, M.-P., Ho, C. & Tjandra, N. Ligand-free open-closed transitions of periplasmic binding proteins: the case of glutamine-binding protein. *Biochemistry* **49**, 1893–1902 (2010).
17. Silva, D.-A., Domínguez-Ramírez, L., Rojo-Domínguez, A. & Sosa-Peinado, A. Conformational dynamics of l-lysine, l-arginine, l-ornithine binding protein reveals ligand-dependent plasticity. *Proteins* **79**, 2097–2108 (2011).
18. Chu, B. C. H., Chan, D. I., DeWolf, T., Periole, X. & Vogel, H. J. Molecular dynamics simulations reveal that apo-HisJ can sample a closed conformation. *Proteins* **82**, 386–398 (2014).
19. Salverda, M. L. M. et al. Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7**, e1001321 (2011).
20. Kaltenbach, M., Jackson, C. J., Campbell, E. C., Hollfelder, F. & Tokuriki, N. Reverse evolution leads to genotypic incompatibility despite functional and active site convergence. *eLife* **4**, e06492 (2015).
21. Sugrue, E., Carr, P. D., Scott, C. & Jackson, C. J. Active site desolvation and thermostability tradeoffs in the evolution of catalytically diverse triazine hydrolases. *Biochemistry* **55**, 6304–6313 (2016).
22. Moroz, Y. S. et al. New tricks for old proteins: single mutations in a non-enzymatic protein give rise to various enzymatic activities. *J. Am. Chem. Soc.* **137**, 14905–14911 (2015).
23. Tokuriki, N. et al. Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nat. Commun.* **3**, 1257 (2012).
24. Anderson, D. P. et al. Evolution of an ancient protein function involved in organized multicellularity in animals. *eLife* **5**, e10147 (2016).
25. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
26. Burton, A. J., Thomson, A. R., Dawson, W. M., Brady, R. L. & Woolfson, D. N. Installing hydrolytic activity into a completely de novo protein framework. *Nat. Chem.* **8**, 837–844 (2016).
27. Röthlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
28. Mak, W. S. & Siegel, J. B. Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* **27**, 87–94 (2014).
29. Korendovych, I. V. & DeGrado, W. F. Catalytic efficiency of designed catalytic proteins. *Curr. Opin. Struct. Biol.* **27**, 113–121 (2014).
30. Blomberg, R. et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
31. Khersonsky, O. et al. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci. USA* **109**, 10358–10363 (2012).

Acknowledgements

B.E.C. and J.A.K. were supported by Australian Postgraduate Awards. B.E.C. was also supported by a Rod Rickards PhD scholarship and an Alan Sargeson scholarship. This research was undertaken with the assistance of resources, services, and staff from the Australian National Computational Infrastructure (NCI), the Australian Synchrotron, and the CSIRO Collaborative Crystallisation Centre, and funding from the Australian Research Council Discovery Project scheme (C.J.J.). We thank A. Saeed, P. Yates, L. Tan and S. Warring for additional technical contributions. We thank H. Janovjak (IST Austria) for gifting us the pDOTS7 plasmid.

Author contributions

B.E.C. and C.J.J. conceived the study; B.E.C. and J.A.K. performed computational analysis; J.A.K., B.E.C., and M.L.G. performed experimental characterization of proteins; B.E.C., J.A.K., P.D.C., and C.J.J. solved the crystal structures; N. T. and C.J.J. supervised students; B.E.C., J.A.K., and C.J.J. wrote the paper. All authors contributed to experimental design, editing of the paper, and interpretation of results.

Competing interests

The authors declare no competing interests

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-018-0043-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.J.J.

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

Methods

Materials. pDOTS7 is a derivative of pQE-82L (Qiagen) modified to enable Golden Gate cloning[32] and was created by removal of the SapI site from pQE-82L and introduction of two reciprocal SapI sites following the His_6 tag, with the SapI sites separated by a 28 bp stuffer fragment. This vector was obtained from H. Janovjak (IST Austria). The Δ*pheA* strain of *E. coli* K-12 from the Keio collection (strain JW2580-1) was obtained from the Coli Genetic Stock Center (Yale University, CT).

Phylogenetic analysis and ancestral sequence reconstruction. The protein sequences of 113 homologs of Ws0279 and PaCDT were collected from the NCBI reference sequence database using the BLAST server. The sequences were aligned in MUSCLE[33]. The alignment was edited to remove N-terminal signal peptides and large insertions and combined with a subset of a previous alignment of representative AABP sequences[34] by profile–profile alignment in MUSCLE, which yielded an outgroup of 271 AABP sequences. Phylogenetic trees were inferred using the maximum-likelihood (ML) method implemented in PhyML[35]. Evaluation of BIONJ trees reconstructed using different amino acid substitution models, using the Akaike information criterion as implemented in ProtTest[36], supported the use of the WAG substitution matrix with gamma-distributed rate heterogeneity, a fixed proportion of invariant sites, and equilibrium amino acid frequencies estimated from the data (WAG + I + Γ + F model). Phylogenies were reconstructed in PhyML by optimization of an initial BIONJ tree using the nearest-neighbor interchange as well as subtree pruning and regrafting algorithms. Robustness of the resulting tree topology to the substitution model was assessed by repeating the analysis using the LG and JTT substitution matrices (LG/JTT + I + Γ + F models), and convergence to the ML tree was checked by repeating the analyses with ten randomized initial trees. Although the resulting trees had essentially identical topologies, the tree inferred using the LG + I + Γ + F model had the highest likelihood and was therefore taken as the ML tree. Ancestral protein sequences were reconstructed using the empirical Bayes method implemented in PAML[37]. The ancestral sequences AncCDT-1 to AncCDT-5 were reconstructed using the LG substitution matrix together with the ML tree inferred using the LG + I + Γ + F model, and the ancestral sequences AncCDT-1W to AncCDT-5W were reconstructed using the WAG substitution matrix together with the tree inferred using the WAG + I + Γ + F model (Supplementary Fig. 2).

Cloning and mutagenesis. Codon-optimized synthetic genes encoding the ancestral proteins, Ws0279 (UniProt: Q7MAG0; residues 24–258), Pu1068 (UniProt: Q4FLR5; residues 19–255), Ea1174 (UniProt: K0ABP5; residues 31–268), and PaCDT (UniProt: Q01269; residues 26–268) were cloned into the pDOTS7 vector using the Golden Gate method[32]. Site-directed mutagenesis was achieved using Gibson assembly[38]: gene fragments with ~30 bp overlap were synthesized by PCR using complementary primers encoding the desired mutation and assembled together with the linearized pDOTS7 vector using Gibson assembly. Successful cloning and mutagenesis was confirmed by Sanger sequencing of the vector insert.

Protein expression and purification. Proteins were generally expressed in *E. coli* (BL21)DE3 cells, except for enzyme assays, in which case they were expressed in Δ*pheA* cells to exclude the possibility of contamination with endogenous prephenate dehydratase. Cells were typically grown in Luria-Bertani (LB) or Terrific Broth (TB) media at 37 °C to OD_{600} 0.8, induced with 0.5 mM β-D-1-isopropylthiogalactopyranoside and incubated for a further 20 h at 37 °C. Cells were pelleted and stored at −80 °C before protein purification. For most applications, proteins were purified under native conditions by nickel-nitrilotriacetic acid (Ni-NTA) affinity chromatography and size-exclusion chromatography (SEC). Cells were thawed, resuspended in equilibration buffer (50 mM NaH_2PO_4, 500 mM NaCl, 20 mM imidazole, pH 7.4), lysed by sonication, and fractionated by ultracentrifugation (24,200 × *g* for 1 h at 4 °C). The supernatant was filtered through a 0.45 µm filter and loaded onto a 5 mL HisTrap HP column (GE Healthcare) equilibrated with equilibration buffer. The column was washed with 50 mL equilibration buffer and 25 mL wash buffer (50 mM NaH_2PO_4, 500 mM NaCl, 44 mM imidazole, pH 7.4), and the target protein was eluted in 25 mL elution buffer (50 mM Na_2HPO_4, 500 mM NaCl, 500 mM imidazole, pH 7.4). For ITC experiments, proteins were subjected to on-column refolding during the affinity chromatography step to remove endogenously bound ligands, as described previously[34]. Proteins were concentrated using a centrifuge filter (Amicon Ultra-15 filter unit with 10 kDa cut-off) and purified by SEC on a HiLoad 26/600 Superdex 200 column (GE Healthcare), typically eluting in SEC buffer (20 mM Na_2HPO_4, 150 mM NaCl, pH 7.4). Protein purity was confirmed by SDS–PAGE, and protein concentrations were measured spectrophotometrically using molar absorption coefficients calculated in ProtParam (http://expasy.org/tools/protparam.html).

Analytical size-exclusion chromatography. The size-exclusion column (HiLoad 26/600 Superdex 200, GE Healthcare) was calibrated using a set of standard proteins (Gel Filtration HMW Calibration Kit, GE Healthcare) in SEC buffer. The partition coefficient (K_{av}) of each protein was calculated using the equation $K_{av} = (v_e - v_o)/(v_c - v_o)$, where v_e is the elution volume, v_o is the column void volume, and v_c is the geometric column volume; this was used to construct a calibration curve of K_{av} versus log(molecular mass).

Differential scanning fluorimetry. Differential scanning fluorimetry (DSF) experiments to test Ws0279, AncCDT-1, Pu1068, and AncCDT-2 for binding of amino acids and other metabolites were performed using a ViiA 7 (Thermo Scientific) or 7900HT Fast (Applied Biosystems) real-time PCR instrument. Reaction mixtures contained 5 µM protein in DSF buffer (50 mM Na_2HPO_4, 150 mM NaCl, pH 7.6), 5× SYPRO orange dye (Sigma-Aldrich) and ligand (1 mM or 10 mM for amino acids, ≥10 mM for other metabolites) in a total volume of 20 µL, and were dispensed onto a 384-well PCR plate in at least triplicate. At least eight replicates of ligand-free control were also included on each plate. Fluorescence intensities were monitored continuously as the samples were heated from 20 °C to 99 °C at a rate of 0.05 °C/s, with excitation at 580 nm and emission measured at 623 nm. Melting temperatures (T_M) were determined by fitting the data to a Boltzmann function, $F = AT + B + (CT + D)/(1 + \exp((T_M - T)/E))$, where F is fluorescence and T is temperature. The parameters A and C, accounting for the slopes of the pre- and post-transition baselines, were fixed at zero if possible.

Pu1068, AncCDT-1, and AncCDT-2 were also screened against a subset of Biolog Phenotype Microarray (PM) plates (Biolog, Hayward, CA, USA), as described previously[39]. Libraries of biologically relevant potential ligands were generated by dissolving each compound in 50 µL water, resulting in concentrations of approximately 10–20 mM in the assay (the exact concentrations vary from well to well, and are not released by the manufacturer). Plates PM1–PM5 contain single concentrations of each compound, whereas plate PM9 contains a series of concentrations of each compound. Fluorescence intensities were measured on a Lightcycler 480 real-time PCR instrument (Roche Diagnostics). Initial hits were further tested using known concentrations (0–600 mM) of each potential ligand to confirm binding. An additional in-house screen consisted of a subset of the Solubility and Stability Screen (Hampton Research), which was tested by the CSIRO Collaborative Crystallization Centre (http://www.csiro.au/C3), Melbourne, Australia. For this screen, the reaction mixtures contained 0.3 µg Pu1068, 3.75× SYPRO orange and 5 µL ligand in a total volume of 20 µL, in a 96-well plate format; each ligand was tested at three concentrations and three replicates of a ligand-free control were also included. Fluorescence intensities were measured on a Bio-Rad CFX384 real-time PCR instrument with excitation at 490 nm and emission at 570 nm. The temperature was ramped from 20 °C to 100 °C at a rate of 0.05 °C/s, and the fluorescence intensity was measured at 0.5 °C intervals. Melting temperatures were taken as the temperature at the minimum of the first derivative of the melt curve, which was determined by fitting the data to a quadratic function in the vicinity of the melting temperature using GraphPad Prism 7 software.

Isothermal titration calorimetry. ITC experiments were performed using a Nano-ITC low-volume calorimeter (TA Instruments); details of instrument calibration have been described previously[34]. ITC experiments were performed at 25 °C with stirring at 200 r.p.m. Protein and ligand solutions were prepared in matched SEC buffer and degassed before use. Amino acid solutions were prepared volumetrically from commercial samples (Sigma-Aldrich, Alfa Aesar) with stated purity ≥98%. Ancestral proteins were tested for binding of proteinogenic amino acids via screening experiments in which 45 µL of 0.844 mM ligand was injected continuously into 164 µL of 50 µM protein over 300 s. In some cases, ligands were tested in mixtures of structurally related amino acids. For quantitative titrations, 100 µM protein was generally titrated with 1 × 1 µL and then 28 × 1.6 µL injections of 0.69 mM ligand at 300 s intervals. The background heat was estimated as the average heat associated with each injection in a control titration of ligand into buffer and subtracted from each protein–ligand titration. Association constants (K_a) were determined by fitting the integrated heat data to the independent binding sites model in NanoAnalyze software (TA Instruments).

Genetic complementation. *E. coli* strain JW2580-1 (Δ*pheA*) cells were transformed with the appropriate plasmid by electroporation, plated on LB agar supplemented with 100 mg/L ampicillin (LBA agar), and incubated at 37 °C overnight. Single colonies were used to inoculate 20 mL M9 minimal media supplemented with L-tyrosine, ampicillin and IPTG (M9 − F; per L: 6 g Na_2HPO_4, 3 g KH_2PO_4, 0.5 g NaCl, 1 g NH_4Cl, 20 mL 20% (w/v) glucose, 2 mL 1 M $MgCl_2$, 0.1 mL 1 M $CaCl_2$, 2 mL 2.5 mg/mL L-tyrosine, 1 mL 100 mg/mL ampicillin, 0.2 mL 1 M IPTG). The cultures were incubated at 37 °C with shaking at 180 r.p.m., and OD_{600} was measured periodically. We confirmed that the observed differences in growth rates could not be explained by differences in protein expression by culturing each clone in M9 − F media supplemented with 20 µg/mL L-phenylalanine (M9 + F media) and assessing protein expression by SDS–PAGE of the soluble fraction of the crude cell lysate from each culture.

Preparation of sodium prephenate. Sodium prephenate was prepared from barium chorismate (Sigma, 60–80% purity). Barium chorismate (40 mM in H_2O) was mixed with an equimolar amount of 1 M Na_2SO_4. An equal volume of 100 mM Na_2HPO_4 (pH 8.0) was added to the mixture, and the $BaSO_4$ precipitate was removed by centrifugation. Sodium prephenate was obtained by heating the resulting sodium chorismate solution at 70 °C for 1 h[40]. Aliquots were stored

at −80 °C. The concentration of prephenate was measured by quantitative conversion of prephenate to phenylpyruvate under acidic conditions (0.5 M HCl, 15 min, 25 °C) and spectrophotometric determination of phenylpyruvate concentration, as described previously[41].

Prephenate dehydratase assay. Prephenate dehydratase activity was determined by spectrophotometric measurement of phenylpyruvate formation, as described previously[41]. Protein solutions were prepared in 20 mM Na_2HPO_4, 150 mM NaCl (pH 7.4), and prephenate solutions were prepared in 50 mM Na_2HPO_4 (pH 8.0). After equilibration at room temperature (20–25 °C) for 5 min, the reaction was initiated by mixing equal volumes of protein and substrate solutions. Aliquots (50 µL or 100 µL) were regularly removed from the reaction mixture and quenched by addition of an equal volume of 2 M NaOH. Absorbance at 320 nm was measured using an Epoch Microplate Spectrophotometer (BioTek), and phenylpyruvate concentrations were determined assuming a molar extinction coefficient of 17,500 M^{-1} cm^{-1}. Reaction times and enzyme concentrations were adjusted to ensure <20% conversion of prephenate to phenylpyruvate. The rate of nonenzymatic turnover was subtracted from the observed rate of enzyme-catalyzed turnover.

Circular dichroism spectroscopy. Circular dichroism (CD) experiments were performed using a Chirascan spectropolarimeter (Applied Photophysics) with a 1-mm path length quartz cuvette. Proteins were diluted to 0.3 mg/mL in water (for recording CD spectra) or SEC buffer (for thermal denaturation experiments) and degassed before measurements. CD spectra were recorded at 20 °C between 190 nm and 260 nm, with a bandwidth of 0.5 nm and a scan rate of 3 s per point, with adaptive sampling. For thermal denaturation experiments, CD was monitored at 222 nm over a temperature range of 20 °C to 90 °C, heating at 1 °C min^{-1}. T_M values were determined by fitting the data to a two-state model:

$$y_{obs} = (y_n + m_n T + (y_u + m_u T) \times \exp((\Delta H_{vH}/R) \times ((1/T) - (1/T_M)))) / (1 + \exp((\Delta H_{vH}/R) \times ((1/T) - (1/T_M))))$$

where y_{obs} is ellipticity at 222 nm, y_n, m_n, y_u, and m_u describe the pre-transition and post-transition baselines, T is temperature, R is the gas constant, and ΔH_{vH} is the apparent van't Hoff enthalpy of unfolding.

Crystallization and structure determination. Crystal structures of AncCDT-1 (complexed with L-arginine), Pu1068 (unliganded), AncCDT-3(P188L), and PaCDT (complexed with acetate) were solved and refined at resolutions between 1.6 Å and 2.6 Å. An additional low-resolution structure of the PaCDT–acetate complex (3.2 Å) shows an alternate crystal packing arrangement; except where explicitly noted, the high-resolution PaCDT–acetate structure was used for structural analysis. A low-resolution (3.4 Å) structure of unliganded AncCDT-1, which illustrates the domain-scale conformational change resulting from ligand binding, was also solved. Finally, Pu1068 was also co-crystallized with NDSB-221 (3-(1-methylpiperidinium-1-yl)propane-1-sulfonate); this low-affinity ligand was identified by DSF and confirmed by fluorescence spectroscopy to bind with a K_d of 0.53 mM (Supplementary Fig. 6).

AncCDT-1, AncCDT-3(P188L), Pu1068, and PaCDT were crystallized using the vapor diffusion method at 18 °C. Crystals were cryoprotected and flash frozen in a nitrogen stream at 100 K. Diffraction data were collected at 100 K on the MX1 or MX2 beamline of the Australian Synchrotron[42]. The data were indexed and integrated in iMOSFLM[43] or XDS[44], and scaled in Aimless[45]. Structures were solved by molecular replacement in Phaser[46] and refined by real space refinement in Coot[47] and reciprocal space refinement in REFMAC5[48] and/or PHENIX[49]. Full details of crystallization and structure determination for each protein are given in Supplementary Tables 5–8. Data collection and refinement statistics are given in Supplementary Tables 9–12. Representative electron density for the active/binding site of each structure is given in Supplementary Fig. 13.

Computational docking. The PaCDT–acetate structure was prepared for computational docking in Maestro (Schrödinger). Missing side chains were rebuilt. Glu173 was protonated, and other residues were assigned the appropriate protonation state at pH 7.0. Asn, Gln, and His side chains were flipped, and Ser, Thr, Tyr, and water hydroxyl groups were reoriented to optimize hydrogen bonding networks. The structure was energy-minimized under the OPLS3 force field, with heavy atoms restrained within 0.3 Å of their initial position. Water and acetate molecules were removed from the structure after energy minimization. The structures of the PaCDT–prephenate and PaCDT–L-arogenate complexes were modeled by computational docking in Glide (Schrödinger) using the standard precision mode with default parameters for docking and scoring. The resulting complexes were energy minimized using the OPLS3 force field. In their respective highest scoring poses, L-arogenate and prephenate adopted the expected orientation, with the α-amino acid and α-keto acid moieties binding at the conserved structural motif that recognizes the same functional groups in AABPs.

Staggered extension process. AncCDT-3 and AncCDT-3W were recombined using the staggered extension process (StEP) following a literature protocol[50]. The StEP reaction mixture contained 5 µL 10× Taq buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 75 fmol of each template plasmid, 30 pmol of each primer, and 2.5 U Taq polymerase (New England BioLabs) in a total volume of 50 µL. The primers used in the reaction were the 5′ flanking primer P7XF and the 3′ flanking primer P7XR (Supplementary Table 13), which amplify ~100 bp on either side of the SapI site of the pDOTS7 vector. The thermocycling program consisted of 80 cycles of (i) a denaturation step for 30 s at 95 °C; and (ii) an annealing/extension step for 5 s at 52 °C. 2 µL of the resulting PCR product was incubated with 10 U DpnI (Thermo Scientific) in a reaction volume of 10 µL at 37 °C for 1 h to digest the parental plasmid DNA. 5 µL of the DpnI-digested StEP product was then amplified in a nested PCR reaction using Taq polymerase, in a total volume of 100 µL. The primers used for the nested PCR reaction, P7NF and P7NR (Supplementary Table 13), target the EcoRI site on the 5′ strand and the HindIII site on the 3′ strand of the pDOTS7 vector, respectively. The nested PCR product was run on a 1% agarose gel and purified by gel extraction.

Incorporation of synthetic oligonucleotides via gene reassembly. Incorporation of synthetic oligonucleotides via gene reassembly (ISOR) was achieved following literature protocols[51,52]. The template gene was amplified by PCR using Phusion Hot Start II Polymerase (Thermo Scientific) using the primers P7XF and P7XR (Supplementary Table 13). The purified PCR product was digested with DNAse I (New England BioLabs) in a reaction mixture containing 100 mM TRIS pH 7.5, 10 mM $MnCl_2$, 4 µg PCR product and 0.3 U DNAse I in a total volume of 40 µL. The reaction mixture was incubated at 37 °C for 1–2 min and quenched by the addition of 20 µL 0.1 M EDTA pH 8.0 pre-incubated at 80 °C, which was followed by heat inactivation at 80 °C for 15 min. The digested PCR product was run on a 2% agarose gel, and fragments 50–250 bp in size were excised from the gel and purified using the Wizard SV Gel and PCR Clean-Up System (Promega). The fragments were reassembled using Taq polymerase: each reaction contained 40 ng gene fragments, 2 µL 10× buffer, 0.2 mM dNTPs, 1.25 U Taq polymerase and varied concentrations of equimolar mutagenic oligonucleotides (5–800 nM total concentration) in a volume of 20 µL (see Supplementary Table 13 for a list of oligonucleotides included in each round). The thermocycling protocol consisted of (i) an initial denaturation step at 95 °C for 2 min; (ii) 40 cycles of a denaturation step at 95 °C for 30 s, then 13 hybridization steps from 65 °C to 41 °C in 2 °C steps, each for 90 s (total 13.5 min), then an extension step at 72 °C for 1 min; and (iii) a final extension step at 72 °C for 7 min. 0.5 µL of the unpurified assembly reaction mixture was amplified in a 50 µL nested PCR reaction using Taq polymerase and the primers P7NF and P7NR (Supplementary Table 13). The nested PCR product was run on a 1% agarose gel and purified by gel extraction.

Library creation and selection. Purified PCR products (0.5 µg) from StEP or ISOR reactions were digested with 2.5 µL each of HindIII FD and EcoRI FD (Thermo Scientific) in a 50 µL reaction at 37 °C for 30 min. The reaction mixture was purified immediately using a PCR purification kit. The pDOTS7 vector containing the AncCDT-2 insert (2.5 µg) was digested using 2.5 µL each of HindIII FD, EcoRI FD, and PstI FD (which cuts within the AncCDT-2 insert) in a 50 µL reaction at 37 °C for 30 min. The digested vector was purified immediately using a PCR purification kit, and then run on a 1% agarose gel and purified by gel extraction. Ligation reaction mixtures contained 100 ng pDOTS7 vector, a three-fold molar excess of insert, 2 µL 10× T4 DNA ligase buffer, and 5 U T4 DNA ligase (Thermo Scientific) in a volume of 20 µL, and were incubated at room temperature for 1 h. Following purification of the ligation reaction mixture using a PCR purification kit, electrocompetent *E. coli* strain JW2580-1 (Δ*pheA*) cells were transformed with 1 µL ligation product by electroporation and plated on LBA agar. Following overnight incubation of the plates at 37 °C, colonies were scraped into LB media, then resuspended in 20 mL fresh LBA media. 100 µL of the resulting cell suspension was used to inoculate 20 mL fresh LBA media, which was then incubated at 37 °C until the OD_{600} reached ~0.5. A 1 mL aliquot of the culture was washed twice with 1 mL M9 salts (6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl), and resuspended in 1 mL M9 salts. Serial dilutions of the cell suspension were made in M9 salts, plated on M9 – F agar, and incubated at 37 °C. The resulting colonies were streaked onto LBA agar, and their plasmid DNA was amplified by PCR using the sequencing primers P7SF and P7SR (Supplementary Table 13). The resulting PCR products were sequenced by GENEWIZ (South Plainfield, NJ, USA) or the Biomolecular Resource Facility at ANU. Single colonies from the streaked LBA plates were used to confirm growth of the clone in liquid M9 – F media, as described above, and to inoculate LBA cultures, from which plasmid DNA was extracted.

Molecular dynamics simulations. MD simulations of PaCDT were initialized from the PaCDT–HEPES and PaCDT–acetate structures. The structure of the PaCDT trimer was generated from the monomer structure by application of the crystallographic three-fold rotation operation. Small molecules, including the active site HEPES and acetate molecules, were removed from the structures, and missing side chains and a missing residue (Gln190) in the PaCDT–HEPES structure were modeled in MODELLER[53]. N-terminal acetyl caps and C-terminal

amide caps were added using MODELLER and Coot[47]. MD simulations of AncCDT-1 were initialized from the L-arginine-bound and unliganded structures. Small molecules were removed from the structures, and missing side chains and residues were modeled in Desmond (Schrödinger) and Coot[47]. N-terminal acetyl caps and C-terminal amide caps were added using Desmond (Schrödinger). GROMOS MD simulations were performed using GROMACS version 4.5.5 (ref. [54]) for the PaCDT–HEPES structure and GROMACS version 4.6.5 for the PaCDT–acetate structure and AncCDT-1 structures, using the GROMOS 53a6 force field[55] in all cases. Each protein was solvated in a rhombic dodecahedron with SPC water molecules, such that the minimal distance of the protein to the periodic boundary was 15 Å, and an appropriate number of ions were added to neutralize the system (15 Na^+ ions for PaCDT simulations, 1 Cl^- ion for AncCDT-1 simulations). Energy minimization was achieved using the steepest descent algorithm. A 100 ps isothermal (NVT) MD simulation with position restraints on the protein was used to equilibrate the system at 300 K. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using Berendsen's thermostat ($\tau_T = 0.1$ ps), and the pressure was maintained at 1 bar using Berendsen's barostat ($\tau_p = 0.5$ ps, compressibility $= 4.5 \times 10^{-5}$ bar^{-1}). All protein bonds were constrained with the LINCS algorithm; water molecules were constrained using the SETTLE algorithm; the time step for numerical integration was 2 fs; the cut-offs for short-range electrostatics and van der Waals forces were 9 Å and 14 Å, respectively; the Particle-Mesh Ewald method was used to evaluate long-range electrostatics; neighbor lists were updated every 10 steps. Following a 1 ns equilibration phase, which was not considered in the analysis, the four simulations of the PaCDT–HEPES structure were continued for 100 ns, and the four simulations of the PaCDT–acetate structure were continued for 170 ns. The three AncCDT-1 simulations were continued for 150 ns.

Additional 150 ns simulations were performed in Desmond version 4.8 (Schrödinger 2016-4)[56] using the OPLS3 force field[57]. These simulations were initiated from the PaCDT–acetate trimer, unliganded AncCDT-1, and L-arginine-bound AncCDT-1 structures, with all small molecules removed. Desmond was used to add the N-terminal acetyl caps and C-terminal amide caps in each case, and for energy minimization of the protein structure. The protein was solvated in an orthorhombic box (15 Å periodic boundary) with SPC water molecules, and the system was neutralized as described above. Energy minimization was achieved using a hybrid method of the steepest descent algorithm and the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (maximum of 2,000 iterations and a convergence threshold of 1 kcal/mol/Å). The system was relaxed using the default relaxation procedure in Desmond. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using a Nosé–Hoover thermostat ($\tau_T = 1.0$), and the pressure was maintained at 1.01 bar ($\tau_p = 2.0$) using a Martyna–Tobias–Klein barostat. Otherwise, default Desmond options were used. Following relaxation of the system, each simulation was run for 150 ns.

Structural analysis. Residues in extant CDT homologs (Ws0279, Pu1068, Ea1174, PaCDT) are numbered according to the equivalent position in the ancestral proteins. Bio3D[58] was used for r.m.s. deviation, radius of gyration, and interdomain angle calculations, and principal component analysis. These analyses were performed on the PaCDT–HEPES and PaCDT–acetate-GROMOS simulations using protein backbone atoms (N, C, and Cα) of individual protein subunits at 0.1 ns intervals. The PaCDT–acetate-OPLS simulations were analyzed separately and projected onto the principal components derived from the PaCDT–HEPES and PaCDT–acetate GROMOS simulations. The AncCDT-1 simulations were also analyzed separately and projected onto the principal components derived from simulations initialized from closed AncCDT-1. The interdomain angle was calculated as the angle between the centers of mass of three groups of backbone atoms: the large domain (residues 2–97 and 196–234), the hinge region (residues 96–98 and 196–198) and the small domain (residues 98–195). Hinge axes for rigid-body domain displacements were determined using DynDom[59] (Supplementary Fig. 7d). PROPKA version 3.0[60] was used for pK_a prediction.

Intrinsic tryptophan fluorescence spectroscopy. Intrinsic tryptophan fluorescence spectra were recorded using a Cary Eclipse fluorimeter. Pu1068 was prepared at a concentration of 5 µM in DSF buffer. The excitation wavelength was 280 nm, and emission was measured between 300 nm and 400 nm. Following addition of each aliquot of NDSB-221, the sample was incubated at ambient temperature for 1 min before the fluorescence spectrum was recorded. The K_d for the Pu1068/NDSB-221 interaction was calculated by fitting the fluorescence data to a hyperbolic binding curve: $F = F_0 + (F_{max} - F_0) \times [L]/(K_d + [L])$, where F is fluorescence, F_0 and F_{max} are initial and final fluorescence, and [L] is ligand concentration.

Statistics. For DSF experiments, one-way analysis of variance (ANOVA) with Dunnett's test for multiple comparisons was used to assess the statistical significance of differences in T_M between untreated and ligand-treated samples. ΔT_M values were taken to be indicative of binding if $\Delta T_M > 2$ °C and $P < 0.05$.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary

Data availability. Structure factors and coordinates for the crystal structures solved in this work (Figs. 2–4) have been deposited in the Protein Data Bank under accession codes 5HPQ (PaCDT, acetate complex, space group *H*3), 6BQE (PaCDT, acetate complex, space group $P4_322$), 5WJP (Pu1068, unliganded), 5KKW (Pu1068, NDSB-221 complex), 5TUJ (AncCDT-1, unliganded), 5T0W (AncCDT-1, L-arginine complex), and 5JOS (AncCDT-3(P188L)). Source data for Supplementary Fig. 5 is included in Supplementary Dataset 1. The remaining data produced and analyzed in this work are available from the corresponding author upon reasonable request.

References

32. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647 (2008).
33. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
34. Clifton, B. E. & Jackson, C. J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol.* **23**, 236–245 (2016).
35. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
36. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
37. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
38. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
39. McKellar, J. L., Minnell, J. J. & Gerth, M. L. A high-throughput screen for ligand binding reveals the specificities of three amino acid chemoreceptors from *Pseudomonas syringae* pv. *actinidiae*. *Mol. Microbiol.* **96**, 694–707 (2015).
40. Gibson, F. Chorismic acid: purification and some chemical and physical studies. *Biochem. J.* **90**, 256–261 (1964).
41. Gibson, M. I. & Gibson, F. Preliminary studies on the isolation and metabolism of an intermediate in aromatic biosynthesis: chorismic acid. *Biochem. J.* **90**, 248–256 (1964).
42. McPhillips, T. M. et al. Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. *J. Synchrotron Radiat.* **9**, 401–406 (2002).
43. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 271–281 (2011).
44. Kabsch, W. XDS. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 125–132 (2010).
45. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
46. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
48. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).
49. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
50. Zhao, H. & Zha, W. In vitro 'sexual' evolution through the PCR-based staggered extension process (StEP). *Nat. Protoc.* **1**, 1865–1871 (2006).
51. Herman, A. & Tawfik, D. S. Incorporating Synthetic Oligonucleotides via Gene Reassembly (ISOR): a versatile tool for generating targeted libraries. *Protein Eng. Des. Sel.* **20**, 219–226 (2007).
52. Rockah-Shmuel, L., Tawfik, D. S. & Goldsmith, M. in *Directed Evolution Library Creation: Methods and Protocols* (eds. Gillam, E. M. J., Copp, J. N. & Ackerley, D. F.) Vol. 1179, 129–137 (Springer-Verlag, 2014).
53. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
54. Pronk, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
55. Oostenbrink, C., Villa, A., Mark, A. E. & van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25**, 1656–1676 (2004).
56. Bowers, K. et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE SC Conf. Supercomput. (SC06)* (ACM, Tampa, Florida, 2006).

57. Harder, E. et al. OPLS3: A force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).
58. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695–2696 (2006).
59. Hayward, S. & Berendsen, H. J. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* **30**, 144–154 (1998).
60. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical p*K*a calculations. *J. Chem. Theory Comput.* **7**, 525–537 (2011).

nature research

Corresponding author(s): Colin Jackson

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Sample sizes were not predetermined. We generally used sample sizes of 3-4 technical replicates, which are standard for in vitro experiments using purified proteins.

2. Data exclusions

   Describe any data exclusions.

   Outliers representing individual technical replicates were excluded from DSF, ITC, and kinetic analysis as appropriate. In the case of DSF analysis, outliers generally arose from a poor fit of the data to the Boltzmann equation, while in the case of ITC analysis, outliers generally arose from baseline instability (resulting from an unclean instrument sample cell or the formation of bubbles during the experiment). In the case of kinetic analysis, outliers generally arose from the formation of bubbles in the plate wells, resulting in noisy readings. Exclusion criteria were not predetermined.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   All attempts at replication were successful. The number of replicates for each experiment is stated in the text.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Randomization was not applicable because we did not measure the effects of different treatments on distinguishable individuals.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not applicable because the study was an in vitro study that did not involve group allocation.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ 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 any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	MUSCLE v. 3.6, PhyML v. 3.0, ProtTest v. 3.2, PAML v. 4.7, GraphPad v. 7.02, iMOSFLM v. 1.0.7, XDS May 1 2016 build, Aimless v. 0.1.29, Phaser v. 2.5.2, REFMAC v. 5.7.0032, PHENIX v. 1.8.1, Glide v. 6.9, GROMACS v. 4.5.5 and 4.6.5, Desmond v. 4.8, Bio3D v. 2.2, GraphPad Prism 7. See citations in text for further details.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	Unique materials are available from the authors on reasonable request.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No research animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The research did not involve human research participants.