Article

Structure-Guided Discovery of New Deaminase Enzymes

Daniel S. Hitchcock, Hao Fan, Jungwook Kim, Matthew W. Vetting, Brandan Hillerich, Ronald D. Seidel, Steven C. Almo, Brian K. Shoichet, Andrej Sali, and Frank Michael Raushel

J. Am. Chem. Soc., **Just Accepted Manuscript** • DOI: 10.1021/ja4066078 • Publication Date (Web): 22 Aug 2013

Structure-Guided Discovery of New Deaminase Enzymes

Daniel S. Hitchcock[ϕ,θ], Hao Fan[ψ,ζ,Σ,θ], Jungwook Kim[ω], Matthew Vetting[ω], Brandan Hillerich[ω], Ronald D. Seidel[ω], Steven C. Almo[ω], Brian K. Shoichet[ζ,Ω] Andrej Sali[ψ,ζ,Ω], and Frank M. Raushel[ϕ,π,2]

[ϕ]Department of Biochemistry & Biophysics, [π]Department of Chemistry, Texas A&M University, College Station, Texas 77843; [ψ]Department of Bioengineering and Therapeutic Sciences, [ζ]Department of Pharmaceutical Chemistry, and [Σ]California Institute for Quantitative Biosciences, University of California, San Francisco; [ω]Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461

ABSTRACT

A substantial challenge for genomic enzymology is the reliable annotation for proteins of unknown function. Described here is an interrogation of uncharacterized enzymes from the amidohydrolase superfamily using a structure-guided approach that integrates bioinformatics, computational biology and molecular enzymology. Previously, Tm0936 from *Thermotoga maritima* was shown to catalyze the deamination of S-adenosylhomocysteine (SAH) to S-inosylhomocysteine (SIH). Homologues of Tm0936 homologues were identified, and substrate profiles were proposed by docking metabolites to modeled enzyme structures. These enzymes were predicted to deaminate analogues of adenosine including SAH, 5'-methylthioadenosine (MTA), adenosine (Ado), and 5'-deoxyadenosine (5'-dAdo). Fifteen of these proteins were purified to homogeneity and the three-dimensional structures of three proteins were determined by X-ray diffraction methods. Enzyme assays supported the structure-based predictions and identified subgroups of enzymes with the capacity to deaminate various combinations of the adenosine analogues, including the first enzyme (Dvu1825) capable of deaminating 5'-dAdo. One subgroup of proteins, exemplified by Moth1224 from *Moorella thermoacetica*, deaminates guanine to xanthine and another subgroup, exemplified by Avi5431 from *Agrobacterium vitis* S4, deaminates two oxidatively damaged forms of adenine: 2-oxoadenine and 8-oxoadenine. The sequence and structural basis of the observed substrate specificities was proposed and the substrate profiles for 834 protein sequences were provisionally annotated. The results highlight the power of a multidisciplinary approach for annotating enzymes of unknown function.

INTRODUCTION

The rate at which new genes are being sequenced greatly exceeds our ability to correctly annotate the functional properties of the corresponding proteins.[1] Annotations based primarily on sequence identity to experimentally characterized proteins are often misleading because closely related sequences can have different functions, while highly divergent sequences can have identical functions.[2,3] Unfortunately, our understanding of the principles that dictate the catalytic properties of enzymes, based on protein sequence alone, is often insufficient to correctly annotate proteins of unknown function. New methods must therefore be developed to define the sequence boundaries for a given catalytic activity and new approaches must be formulated to identify those proteins that are functionally distinct from their close sequence homologues. To address these problems, we have developed a comprehensive strategy for the functional annotation of newly sequenced genes using a combination of structural biology, bioinformatics, computational biology, and molecular enzymology.[4-6] The power of this multidisciplinary approach for discovering new reactions catalyzed by uncharacterized enzymes is being tested using the amidohydrolase superfamily (AHS) as a model system.[5,7]

The AHS is an ensemble of evolutionarily related enzymes capable of hydrolyzing amide, amine, or ester functional groups at carbon and phosphorus centers.[7,8] More than 24,000 unique protein sequences have been identified in this superfamily and they have been segregated into 24 clusters of orthologous groups (COG).[9] One of these clusters, cog0402, catalyzes the deamination of nucleic acid bases.[5,7] Previously, we successfully predicted that Tm0936, an enzyme from *Thermotoga maritima,* would catalyze the deamination of *S*-

adenosylhomocysteine (SAH) to *S*-inosylhomocysteine (SIH).[5] Here we significantly expand the scope of these efforts by addressing the functional and specificity boundaries for more than 1000 proteins homologous to Tm0936, resulting in the prediction and discovery of novel substrate profiles for neighboring enzyme subgroups. To do so, we have integrated a physical library screen with the computational docking of high-energy reaction intermediates to homology models of fifteen previously uncharacterized proteins. To identify enzymes most closely related to Tm0936, we retrieved all of the protein sequences that correlated with a BLAST E-value cutoff better than 10^{-36}.[10] This procedure identified 1358 proteins that were further sorted into smaller subgroups through the construction of a sequence similarity network at a BLAST E-value cutoff of 10^{-100} (**Figure 1**).[11] The minimal sequence identity between any two proteins in this network is 23% and twelve representative subgroups (sg-1a through sg-11) were arbitrarily defined, colored-coded, and numbered.

The three-dimensional structure of Tm0936 (from sg-8) was previously determined in the presence of the product SIH. The most salient structural features for substrate recognition include Glu-84, Arg-136, Arg-148 and His-173 (**Figure 2**). These residues form electrostatic interactions with the 2'- and 3'-hydroxyls from the ribose sugar, the α-carboxylate of the homocysteine moiety, and N3 of the purine ring.[5] The catalytic machinery is composed of a zinc ion that is coordinated by three histidine residues (His-55, His-57, His-173) and an aspartate (Asp-279), while proton transfer reactions are facilitated by Glu-203 and His-228. The six residues required for metal binding and proton transfers are fully conserved in all 1358 proteins (**Figure 1**). However, only those proteins within sg-8 fully conserve the four residues that are utilized in the recognition of SIH; sg-1 through sg-7 lack one or both of the carboxylate-

binding arginine residues, while sg-9, sg-10, and sg-11 lack the two adenine/ribose recognition residues, histidine and glutamate. All of these proteins are therefore anticipated to contain unique substrate profiles and to catalyze the deamination of unanticipated substrates.

METHODS

Cloning and Purification of Mm2279, Spn3210, Moth1224, Avi5431, Tvn0515 Caur1903, and Cthe1199. The genes for Mm2279, Spn3210, Moth1224, Avi5431, and Tvn0515 were cloned into a pET-30a expression vector with C-terminal 6x His-tags. Caur1903 was cloned into pET-42a with an N-terminal GST tag. Cthe1199 was cloned as a tagless construct into pET-30a. *E. coli* BL-21 (DE3) cells were transformed with the recombinant plasmids and plated onto LB-agarose containing 50 µg/mL kanamycin. 1 L cultures (LB broth, kanamycin 50 µg/mL) were inoculated from the resulting colonies and grown at 37 °C until the optical density at 600 nm reached 0.6. Protein expression was induced with 1 mM IPTG, and the cultures were incubated for 20 hours at 25 °C while shaking. The cultures were centrifuged at 7000 RPM for 15 minutes and the cell pellets were disrupted by sonication in 35 mL of running buffer and PMSF. DNA was precipitated by protamine sulfate. Proteins with a 6x His-tag were loaded onto a 5 mL HisTrap column (GE HEALTHCARE, USA) with running buffer (20 mM HEPES, 250 mM NaCl, 250 mM NH_4SO_4, 20 mM imidazole, pH 7.5) and eluted with a linear gradient of elution buffer (20 mM HEPES, 250 mM NaCl, 250 mM NH_4SO_4, 500 mM imidazole, pH 7.5).

Caur1903 was loaded onto a 5 mL GSTrap column (GE HEALTHCARE, USA) with running buffer (20 mM HEPES, 10% [w/v] glycerol, 200 mM NaCl, 200 mM NH_4SO_4, pH 8.0) and eluted

with a gradient of elution buffer (20 mM HEPES, 10% [w/v] glycerol, 200 mM NaCl, 200 mM NH_4SO_4, 10 mM reduced glutathione, pH 8.0). Cthe1199 was precipitated with 70% saturation NH_4SO_4 and the pellet resuspended in 4 mL of 50 mM HEPES, pH 7.5. The solution was loaded onto a HiLoad 26/60 Superdex 200 gel filtration column. Active fractions were loaded onto an anion exchange column with binding buffer (20 mM HEPES, pH 7.5), and eluted with a gradient of 20 mM HEPES, 1 M NaCl, pH 7.5. All proteins were concentrated with a 10 kDa molecular weight cutoff filter. Moth1224 and Mm2279 were almost entirely insoluble. Both proteins were verified by N-terminal sequencing.

Cloning and Purification of Cv1032, Pfl4080, Ef1223, Cpf1475, Dvu1825, and Bc1793. The genes for these proteins were PCR amplified and ligated into pSGX3, a derivative of pET26b,[12] yielding a protein with Met-Ser-Leu followed by the PCR product and a C-terminal hexa-histidine tag. E. coli BL21 (DE3)-Condon Plus 1 RIL cells (Invitrogen) were transformed with the recombinant plasmids and selenomethionine-labeled protein was expressed in 3 L of HY media with 50 μg/mL of kanamycin and 35 μg/mL of chloramphenicol. Expression was induced with 0.4 mM IPTG at an OD 600 of 1.0 and grown for 21 hours whereupon the cells were collected by centrifugation, resuspended in 3x the volume with 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 25 mM imidizole, 0.1% (*v/v*) Tween20, lysed by sonication, and clarified by centrifugation at 4° C. The supernatant was applied to a 5-mL HisTrapHP column (GE HEALTHCARE, USA) pre-equilbrated with 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 25 mM imidizole, 10% (w/v) glycerol. The column was washed with 10 column volumes (CV) of 20 mM Tris–HCl pH 8.0, 500 mM NaCl, 10% (w/v) glycerol, and 40 mM imidazole, and subsequently step-eluted with 2 CV of same buffer with an imidazole concentration of 250 mM.

Concentrated fractions were pooled and subsequently passed over a 120-mL Superdex 200 column (GE HEALTHCARE, USA) equilibrated with 10 mM HEPES pH 7.5, 150 mM NaCl, 10% (w/v) glycerol, and 5 mM DTT. Proteins exhibiting purity on SDS-GELS greater than 95% were pooled and concentrated to > 10 mg/mL using AMICON centrifugal filters. Samples were snap-cooled in liquid N_2, and stored at -80 °C.

Cloning and Purification of Pa3170 and Xcc2270. The genes for Pa3170 and Xcc2270 were amplified from their respective genomic DNA using primers that included a ligation independent cloning sequence and the N- and C-terminal sequences of each respective gene. In general, PCR was performed using KOD Hot Start DNA Polymerase (NOVAGEN). The amplified fragments for Pa3170, and Xcc2270 were cloned into the C-terminal TEV cleavable StrepII-6x-His-tag containing vector CHS30, a derivative of pET30 by ligation-independent cloning.[13] *E. coli* BL21 (DE3)-Condon Plus 1 RIL cells (Invitrogen) were transformed by the recombinant plasmids and protein was expressed by autoinduction in 2 L of ZYP-5052 media with 50 µg/mL of kanamycin and 35 µg/mL of chloramphenicol. Cells were grown at 37 °C to an OD of 2-4 and then reduced to 22 °C for a period of 12-16 hours whereupon the cells were pelleted by centrifugation. Cells were resuspended in 3x the volume with 50 mM Hepes, pH 7.5, 150 mM NaCl, 10% (*w/v*) glycerol, 20 mM imidazole, lysed by sonication, and centrifuged at 4 °C. The supernatant was applied to a 1-mL HisTrapHP column (GE HEALTHCARE) pre-equilbrated with 50 mM Hepes, pH 7.5, 150 mM NaCl, 25 mM imidizole, 10% (*w/v*) glycerol. The column was washed with 10 column volumes of 50 mM Hepes, pH 7.5, 150 mM NaCl, 10% (w/v) glycerol, 25 mM imidazole and subsequently step-eluted with the same buffer with an imidazole concentration of 250 mM directly onto a 120-mL Superdex 200 column (GE

HEALTHCARE) equilibrated with 10 mM HEPES pH 7.5, 150 mM NaCl, 10% (v/v) glycerol, and 5 mM DTT (protein storage buffer). Proteins exhibiting purity on SDS-GELS greater than 95% were pooled, concentrated to >10 mg/mL using AMICON centrifugal filters. Samples were snap-cooled in liquid N_2, and stored at -80 °C.

Substrate Characterization. Substrate screening was performed by monitoring the UV-vis spectral change from 235 to 320 nm. Potential substrates at a concentration of 100 μM were incubated overnight in a 96-well quartz plate with 1 μM enzyme in 50 mM HEPES, pH 7.5. Those showing a spectral change were subjected to more detailed kinetic assays. The kinetic constants were determined in 20 mM HEPES buffer at pH 7.5 by a direct UV assay at 30 °C. Deamination of adenosine, 5'-dAdo, MTA, SAH and *S*-adenosylthiopropylamine (TPA) were monitored at 263 nm ($\Delta\varepsilon$ = 7900 $M^{-1}\,cm^{-1}$) or 274 nm ($\Delta\varepsilon$ = 2600 $M^{-1}\,cm^{-1}$), depending on the concentration of substrate. Non-adenosine compounds were as follows: guanine (245 nm, $\Delta\varepsilon$ = 5200 $M^{-1}\,cm^{-1}$), 2-hydroxyadenine ($\Delta\varepsilon$ = 294 nm, 6600 $M^{-1}\,cm^{-1}$), 8-oxoadenine ($\Delta\varepsilon$ = 272 nm, 3700 $M^{-1}\,cm^{-1}$). Values of k_{cat} and k_{cat}/K_m were determined by fitting the data to equation 1 using SigmaPlot 11, where v is the initial velocity, E_t is enzyme concentration, and A is the substrate concentration.

$$v/E_t = k_{cat}\,A/(A + K_a) \quad (1)$$

Homology Modeling. Two structures (PDB code: 2PLM and 1P1M) have been solved for Tm0936 from *Thermotoga maritima* in the ligand-bound (holo) and ligand-free (apo) states, respectively. The holo structure of Tm0936 was used as a template to build homology models for 15 proteins from 11 subgroups. Because Moth1224 from sg-9 has a sequence identity similar to that of guanine deaminase (PDB code: 2UZ9) as it does to Tm0936, the structure of

guanine deaminase was also used as template for Moth1224. For each target-template pair, the same procedure of homology modeling was applied. First, the sequence alignment was computed by MUSCLE (Multiple Sequence Comparison by Log-Expectation). Second, a total of 500 homology models were generated with the standard "automodel" class in MODELLER,[14] and the model with the best DOPE[15] score was selected to begin the model refinement. Third, side chains of active site residues that are within a distance of 5 Å from the bound ligand in the template structure, were optimized using the "side chain prediction" protocol in PLOP,[22] resulting in one representative model. The bound ligand from the template structure was included in the second step for construction of the initial homology model, but was removed in the model refinement.

Virtual Screening. The high-energy intermediate (HEI) library[17,18] that contains 57,672 different intermediate forms of 6440 KEGG (Kyoto Encyclopedia of Genes and Genomes)[19,20] molecules was screened against the crystal structure of Tm0936 and each of the refined homology models using the docking program DOCK 3.6.[21] The computed poses were subjected to a distance cutoff to make sure that the O^- of the HEI portion of the molecule is found within 4 Å of the metal ions in the active site. Of the molecules that satisfied this constraint, the top 500 compounds ranked by DOCK score (consisting of van der Waals, Poisson–Boltzmann electrostatic, and ligand desolvation penalty terms) were inspected visually to ensure the compatibility of the pose with the amidohydrolase reaction mechanism. The details of the HEI docking library preparation, the molecular docking procedure, and the docking results analysis have been previously described.[17,18,22,23]

RESULTS

Target Selection and Structure Determination. Representative examples from each of the twelve major subgroups contained within the sequence similarity network of proteins (**Figure 1**) related to the initial protein target, Tm0936, were selected for interrogation (**Table S1**). Proteins from sg-1a (Cthe1199), sg-1b (Mm2279), sg-2 (Cv1032, Pfl4080, Xcc2270, and Pa3170), sg-3 (Ef1223 and Cpf1475), sg-4 (Bc1793), sg-5 (Spn13210), sg-6 (Dvu1825), sg-7 (Caur1903), sg-8 (Tm0936), sg-9 (Moth1224), sg-10 (Avi5431), and sg-11 (Tvn0515) were purified to homogeneity and the three-dimensional structures of three proteins were determined (**Table S2**). The structure of Cv1032 was determined in the presence of inosine (PDB id: 4F0S) and also with 5'-methylthioadenosine (MTA) in an unproductive complex (PDB id: 4F0R). The structures of Xcc2270 (PDB id: 4DZH) and Pa3170 (PDB id: 4DYK) were determined with zinc bound in the active site.

Predictions of Enzyme Function. Homology models were constructed for each of the purified proteins using the X-ray structure of Tm0936 (PDB id: 2PLM) as the initial structural template.[24] A virtual library of 57,672 high-energy reaction intermediates (HEI) was screened against the homology models, as well as the X-ray structure of Tm0936.[6] For the thirteen targets from sg-1 through sg-8, analogues of adenosine (5'-deoxyadenosine, adenosine, 5'-methylthioadenosine, and SAH) were predicted as the most probable substrates (**Table 1**).

The histidine residue that interacts with N3 from the purine ring of SAH in the structure of Tm0936 is conserved in all of the proteins except for the enzymes from sg-9 (Moth1224), sg-10 (Avi5431), and sg-11 (Tvn0515). In Moth1224 (sg-9), this histidine residue is replaced by an

arginine (Arg-192), which has previously been observed in the active site of guanine deaminase.[4] The sequence identity (33%) between Moth1224 and Tm0936 is similar to that between Moth1224 and human guanine deaminase (31%) (PDB id: 2UZ9). Therefore, a homology model was constructed for Moth1224 based on each one of these two templates separately. For the model based on Tm0936, the docking hits were predominantly adenosine analogues (**Table 1**). However, for the model based on guanine deaminase, the primary docking hits were analogues of guanine. The highest ranking of these compounds included guanine (rank 10), 8-oxoguanine (rank 29), and thioguanine (rank 35).

In Avi5431 (sg-10) the sequence alignment predicts that a glutamate (Glu-189) replaces the histidine that interacts with N3 of the adenine moiety in the Tm0936 template. However, in the modeled active site of Avi5431, His-190 occupies the same physical position as His-173 in Tm0936, while the side chain of Lys-308 is located on the other side of the pocket, which has no available space for the binding of the ribose moiety of adenosine-like compounds. The primary docking hits for Avi5431 included small aromatic amines that are not attached to a ribose group. Among the best ranking compounds are mercazole (rank 3), 6-hydroxymethyl 7,8-dihydropterin (rank 6), 8-oxoadenine (rank 21), melamine (rank 22), and guanine (rank 28).

Tvn0515 from sg-11 is unusual as proteins within this subgroup have a conserved glutamine in place of the histidine that binds to N3 of the adenine moiety in Tm0936. Originally, it was hypothesized that this substitution would enable the binding of 2-oxoadenosine in a manner similar to the way the homologous carbonyl group binds in cytosine, pterin, and guanine deaminases.[25] However, virtual screening against the homology model of

Tvn0515 suggested adenosine analogues as the most probable substrates; SAH was ranked higher than adenosine, MTA, and 5'-dAdo (**Table 1**).

Physical Library Screen and Kinetic Assays. A physical library of more than 100 potential substrates was assembled and tested for catalytic activity using the 16 enzymes isolated for this investigation by monitoring the change in absorbance from 240-300 nm. The full list of the compounds tested is provided in Supporting Information and the structures of the best substrates are provided in **Figure 3**. Compounds that exhibited catalytic activity with any of the 16 proteins were subjected to detailed kinetic assays using a direct spectrophotometric assay**.** Analogues of adenosine, including 5'-dAdo, MTA, and SAH, were deaminated by proteins from sg-1a, sg-1b, sg-2, sg-4 through sg-8, and sg-11. For most targets, the substrates with the highest activity had k_{cat}/K_m values that ranged from 10^5 - 10^7 M^{-1} s^{-1}, except for Cpf1475 and Ef1223 from sg-3 for which no substrates could be identified. The proteins from sg-1a, sg-1b, sg-8, and sg-11 preferentially deaminated SAH. MTA was deaminated by proteins from sg-1 and sg-8, but not from sg-11. The four proteins purified from sg-2 deaminated MTA, 5'-dAdo, and adenosine, but not SAH. The proteins purified from sg-4 through sg-7 preferentially deaminated adenosine and 5'-dAdo; MTA is either not a substrate or it is deaminated 4-orders of magnitude less efficiently. For the proteins purified from sg-4, sg-5, and sg-7, adenosine and 5'-dAdo are equally good substrates, whereas for Dvu1825 from sg-6, 5'-dAdo is a substantially better substrate than is adenosine. The values of k_{cat}/K_m for adenosine, 5-dAdo, MTA, and SAH are presented in **Table 1** and the values of K_m and k_{cat} are provided in **Table S3**.

Moth1224 (sg-9) and Avi5431 (sg-10) were subjected to more extensive library screening because these proteins lacked many of the conserved residues found in Tm0936.

Guanine was the only substrate identified for Moth1124 with values of k_{cat}/K_m and k_{cat} of 1.2 x 10^5 M^{-1} s^{-1} and 0.48 s^{-1}, respectively. Avi5431 deaminated two adenine related compounds, 8-oxoadenine and 2-oxoadenine, with values of k_{cat}/K_m of 1.8 x 10^5 and 2.3 x 10^4 M^{-1} s^{-1}, respectively. The values of k_{cat} for these two compounds are 3.4 and 13 s^{-1}. This enzyme did not deaminate adenine.

DISCUSSION

Substrate Profiles. In this investigation we have successfully characterized fifteen proteins related in sequence to the first enzyme (Tm0936) found capable of deaminating SAH to SIH. Using a combination of homology modeling, virtual screening and physical library screening, six unique substrate profiles were established for twelve subgroups of proteins homologous to Tm0936. Tvn0515 (sg-11) deaminates only SAH, while enzymes from three subgroups (sg-1a, sg-1b, and sg-8) efficiently deaminate both SAH and MTA. Enzymes from sg-2 deaminate MTA, adenosine, and 5'-dAdo equally well, whereas proteins from sg-4, sg-5, and sg-7 cannot deaminate MTA but will deaminate adenosine and 5'-dAdo equally well. Enzymes from sg-6 deaminate 5'-dAdo about 100 times faster than adenosine. All of these substrates are linked to the recycling of reaction products from the utilization of *S*-adenosylmethionine in the cell. Not only was the correct family of substrates identified for these enzymes from full library screens, but much of the specificity among the adenosine analogues was also predicted, at least by relative rank. The substrate specificities for the 1000 proteins contained within sg-1 through sg-11 are provided in **Table S4**.

Two of the subgroups that are closely related to Tm0936 deaminate quite different substrates. Moth1224 (sg-9) deaminates guanine while Avi5431 (sg-10) deaminates 8-oxoadenine. Avi5431 is the first enzyme ever reported to deaminate oxidatively damaged adenine. In addition, Dvu1825 (sg-6) is the first protein found capable of deaminating 5'-deoxyadenosine.

Deamination of SAH. In the crystal structure of Tm0936, the carboxylate of the homocysteine moiety of SAH interacts with two arginine residues, Arg-136 and Arg-148 (**Figure 2**). For Cthe1199 from sg-1a, the first arginine residue is conserved (Arg-147), while the second arginine is replaced with an aspartate (Asp-159). In the modeled active site of Cthe1199, these two residues interact with the carboxylate and amino groups of the homocysteine moiety of SAH, respectively (**Figure S1**). For Mm2279 (sg-1b), the two arginine residues are substituted with histidine (His-169) and aspartate (Asp-183), respectively. In the modeled active site of Mm2279, the amino group of the homocysteine moiety of SAH interacts with His-169 and Asp-183 (**Figure S2**). Sequence alignments for the remaining proteins in sg-1a and sg-1b suggest that catalytic activity with SAH is a consequence of an arginine (or lysine) residue aligned to Arg-136 in Tm0936, and a charged residue (Arg, Lys, Asp, Glu) aligned to Arg-148 in Tm0936. For proteins from sg-1b there is usually a tyrosine or histidine at residue position 136 and a glutamate or aspartate at residue position 148 from Tm0936.

This sequence-structure-activity relationship is not adequate to explain the SAH specificity for all of the proteins examined in this investigation. Bc1793 from sg-4 preserves the arginine (Arg-150) in the position aligned to Arg-136 in Tm0936, but SAH was not prioritized by docking against the homology model and not active in subsequent biochemical assays. The lack

of SAH activity can be rationalized by the modeled active site of Bc1793, where an inserted loop from residue 126-129 prevents Arg-150 from interacting with potential substrates (**Figure S3**).

In Tvn0515 from sg-11, the two critical arginine residues from Tm0936 are replaced by the nonpolar residues Trp-140 and Pro-154. Surprisingly, SAH ranked high in docking to the homology model of Tvn0515 and this compound was efficiently deaminated. The carboxylate group from the homocysteine moiety of SAH apparently forms a hydrogen bond with the side chain from Lys-145 in the model of Tvn0515, while the amino group from the homocysteine moiety is exposed to the protein surface and forms a cation-π interaction with the side chain from Phe-85 (**Figure S4**). Uniquely, catalytic activity for Tvn0515 could not be detected with any compound other than SAH.

Deamination of MTA. The four enzymes from sg-2 can deaminate MTA but not SAH. In the modeled active site of Pa3170, the methylthio group of the docked MTA resides in a hydrophobic pocket surrounded by Met-132, Tyr-133, Phe-134, Pro-155, Leu-157 and His-194. This model is consistent with the crystal structure of Pa3170 determined in the presence of 5'-methylthiocoformycin, an analogue of MTA.[26] A superposition of the homology model for Cv1032 from sg-2 and the two x-ray structures of this protein are presented in **Figure 4A**. The modeled active site occupied by the HEI form of adenosine has an all-atom RMSD of 1.8 Å with respect to the crystal structure complexed with inosine, while the docking pose of adenosine is nearly identical to the crystal structure of inosine with a RMSD of 0.8 Å (**Figure 4B**).

Deamination of Ado and 5'-dAdo. The enzymes from sg-4 through sg-7 catalyze the deamination of adenosine and 5'-deoxyadenosine, but not MTA and SAH. The exclusion of MTA from the active sites of enzymes from these subgroups is likely the result of a hydrophobic

amino acid residue found directly after β-strand 3 (**Figure S5**). In Tm0936, Val-139 is positioned near the C5'-carbon of the ribose moiety. This residue apparently permits Tm0936 and all proteins within sg-1a and sg-1b (Ile, Val) and sg-2 (Leu, Ile) to accommodate the thiomethyl substituent. In sg-4, sg-5, and sg-6 the residue equivalent to Val-139 is replaced with a bulky phenylalanine and in sg-7 this residue is replaced with methionine. These changes likely restrict the binding of only adenosine or 5'-dAdo within the active site. The lack of adenosine binding in the active site of Dvu1825 (sg-6) is less obvious, but it may be related to the residue corresponding to Gly-137 in Tm0936. In sg-4, sg-5, and sg-7 a conserved threonine is at this position. The crystal structure of Cv1032 (sg-2) suggests that a hydrogen bond is formed between this residue (Ser-153) and the 5'-hydroxyl of the bound inosine (PDB id: 4F0S). Dvu1825 contains an alanine (Ala-152) at this residue position and thus a hydrogen bond is not possible.

Convergent Evolution of Guanine Deaminase. Moth1224 from sg-9 was active as a deaminase only for guanine. The observed substrate specificity can be rationalized by the identity of the conserved residues located within the active site of this enzyme. Arg-192 and Gln-68 from Moth1224 align with residues Arg-213 and Gln-68 in human guanine deaminase (GuaD), which form critical hydrogen bonds to the substrate in the active site. The docking pose of guanine in the active site of Moth1224 (based on the GuaD structure) is presented in **Figure 5**. Although the protein is more similar to the entire group of enzymes related to Tm0936, the active site of Moth1224 bears the hallmark of the GuaD family. This example of convergent evolution underscores the importance of observing conservation of active site residues when predicting enzyme functions based on homology.

8-Oxoadenine Deaminase. Avi5431 from sg-10 was able to deaminate only two substrates, 2-oxoadenine and 8-oxoadenine. To the best of our knowledge this is the first time that an enzyme has been identified that can deaminate these compounds. Both of these substrates are oxidatively damaged adenine within DNA.[27,28] The deaminated products, 8-oxohypoxanthine and xanthine, can be further metabolized to uric acid.[29] We had earlier identified the first enzyme capable of deaminating 8-oxoguanine (4). In the docking pose of 8-oxoadenine to the modeled Avi5431 active site, the N3 nitrogen and the C8 carbonyl oxygen from the purine ring interact with the side chains of His-190 and Lys-308, respectively (**Figure S6**).

Functional Annotation. Using a combination of bioinformatics, homology modeling, computational docking, structural biology, and molecular enzymology, we have interrogated the catalytic properties of a large cluster of proteins homologous to Tm0936, a protein from *Thermotoga maritima*, which was previously shown to deaminate S-adenosyl homocyteine to S-inosyl homocysteine. Subsets of enzymes were found to deaminate the catabolic metabolites of S-adenosyl methionine, including S-adenosyl homocyteine, methylthioadenosine, adenosine, and 5'-deoxy-adenosine. Not only were all of the actual substrates ranked within the top 100 of the metabolites in the virtual library, but the overall substrate specificities were also captured. We did not anticipate this level of agreement between prediction and experiment given the well-known complexities of docking. Remarkably, these predictions were based on homology models of the targets, often templated at the limit of reliable primary sequence alignments. These results support the feasibility of prioritizing substrates and substrate-specificity from large library docking screens against homology models and presage large-scale

functional annotation efforts that will significantly expand the range of metabolic diversity present in nature.

ASSOCIATED CONTENT

Supporting Information

Physical substrate library, crystallization and X-ray diffraction data collection methods; Tables S1 through S6; Figures S1 through S6. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author:

Frank M. Raushel, E-mail: raushel@tamu.edu; telephone: 979-845-3373;

Department of Chemistry, Texas A&M University, College Station, TX

Author Contributions

$^{\theta}$These authors contribute equally to this work.

Notes

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

This work was supported in part by the Robert A. Welch Foundation (A-840) and the National Institutes of Health (GM 071790, GM 093342, and GM 094662). X-ray diffraction data for this study were measured at beamlines X29 of the National Synchrotron Light Source (NSLS),

Brookhaven National Laboratory, and Lilly Research Laboratory Collaborative Access Team (LRL-CAT) at the Advanced Photon Source, Argonne National Laboratory.

REFERENCES

(1) Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C. *PLoS Comput. Biol.* **2009**, *5*, e1000605.
(2) Seffernick, J. L.; de Souza, M. L.; Sadowsky, M. J.; Wackett, L. P. *J. Bacteriol.* **2001**, *183*, 2405-2410.
(3) Glasner, M. E.; Fayazmanesh, N.; Chiang, R. A.; Sakai, A.; Jacobson, M. P.; Gerlt, J. A.; Babbitt, P. C. *J. Mol. Biol.* **2006**, *360*, 228-250.
(4) Hall, R. S.; Fedorov, A. A.; Marti-Arbona, R.; Fedorov, E. V.; Kolb, P.; Sauder, J. M.; Burley, S. K.; Shoichet, B. K.; Almo, S. C.; Raushel, F. M. *J. Am. Chem. Soc.* **2010**, *132*, 1762-1763.
(5) Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. *Nature* **2007**, *448*, 775-779.
(6) Hermann, J. C.; Ghanem, E.; Li, Y.; Raushel, F. M.; Irwin, J. J.; Shoichet, B. K. *J. Am. Chem. Soc.* **2006**, *128*, 15882-15891.
(7) Seibert, C. M.; Raushel, F. M. *Biochemistry* **2005**, *44*, 6383-6391.
(8) Holm, L.; Sander, C. *Proteins Struct. Funct. Bioinformat.* **1997**, *28*, 72-82.
(9) Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. *Science* **1997**, *278*, 631-637.
(10) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403-410.
(11) Cline, M. S.; Smoot, M.; Cerami, E.; Kuchinsky, A.; Landys, N.; Workman, C.; Christmas, R.; Avila-Campilo, I.; Creech, M.; Gross, B.; Hanspers, K.; Isserlin, R.; Kelley, R.; Killcoyne, S.; Lotia, S.; Maere, S.; Morris, J.; Ono, K.; Pavlovic, V.; Pico, A. R.; Vailaya, A.; Wang, P.-L.; Adler, A.; Conklin, B. R.; Hood, L.; Kuiper, M.; Sander, C.; Schmulevich, I.; Schwikowski, B.; Warner, G. J.; Ideker, T.; Bader, G. D. *Nat. Protoc.* **2007**, *2*, 2366-2382.
(12) Sauder, M. J.; Rutter, M. E.; Bain, K.; Rooney, I.; Gheyi, T.; Atwell, S.; Thompson, D. A.; Emtage, S.; Burley, S. K. *Methods Mol. Biol.* **2008**, *426*, 561-575.
(13) Aslanidis, C.; de Jong, P. J. *Nucleic Acids Res.* **1990**, *18*, 6069-6074.
(14) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779-815.
(15) Shen, M. Y.; Sali, A. *Protein Sci.* **2006**, *15*, 2507-2524.
(16) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534-553.
(17) Hermann, J. C.; Ghanem, E.; Li, Y. C.; Raushel, F. M.; Irwin, J. J.; Shoichet, B. K. *J Am Chem Soc* **2006**, *128*, 15882-15891.
(18) Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. *Nature* **2007**, *448*, 775-U772.
(19) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27-30.
(20) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354-D357.
(21) Mysinger, M. M.; Shoichet, B. K. *J. Chem. Inf. Model.* **2010**, *50*, 1561-1573.
(22) Kamat, S. S.; Fan, H.; Sauder, J. M.; Burley, S. K.; Shoichet, B. K.; Sali, A.; Raushel, F. M. *J Am Chem Soc* **2011**, *133*, 2080-2083.
(23) Goble, A. M.; Fan, H.; Sali, A.; Raushel, F. M. *ACS Chem. Biol.* **2011**, *6*, 1036-1040.
(24) Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. *J. Chem. Inf. Model.* **2009**, *49*, 2512-2527.

(25) Fan, H.; Hitchcock, D. S.; Seidel, R. D.; Hillerich, B.; Lin, H.; Almo, S. C.; Sali, A.; Shoichet, B. K.; Raushel, F. M. *J. Am. Chem. Soc.* **2012**, *135*, 795-803.
(26) Guan, R.; Ho, M.-C.; Fröhlich, R. F. G.; Tyler, P. C.; Almo, S. C.; Schramm, V. L. *Biochemistry* **2012**, *51*, 9094-9103.
(27) Malins, D. C.; Polissar, N. L.; Ostrander, G. K.; Vinson, M. A. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 12442-12445.
(28) Kamiya, H.; Kasai, H. *FEBS Letters* **1996**, *39*, 113-116.
(29) Krenitsky, T. A.; Neil, S. M.; Elion, G. B.; Hitchings, G. H. *Arch. Biochem. Biophys.* **1972**, *150*, 585-599.

Table 1. Enzyme kinetic constants at pH 7.5, 30 °C, and docking ranks.

Protein			Substrate k_{cat}/K_m (M^{-1} s^{-1})			
Label	**Subgroup**	**Locus Tag**	**5'-dAdo**	**Ado**	**MTA**	**SAH**
(a)	1a	Cthe1199	5.0×10^4 (6)	1.6×10^4 (5)	2.9×10^6 (9)	1.4×10^7 (1)
(b)	1b	Mm2279	1.2×10^7 (16)	2.2×10^5 (12)	1.7×10^7 (10)	1.3×10^7 (11)
(c)	2	Cv1032	3.4×10^6 (26)	6.6×10^6 (13)	1.2×10^6 (20)	<10 (80)
(d)	2	Pfl014080	8.9×10^6 (13)	4.6×10^5 (5)	1.3×10^7 (2)	<10 (316)
(e)	2	Xcc2270	4.7×10^6 (16)	3.4×10^6 (10)	6.8×10^6 (4)	<10 (60)
(f)	2	Pa3170	4.5×10^7 (23)	9.6×10^6 (9)	8.2×10^7 (19)	<10 (361)
(g)	3	Cpf1475	<10 (66)	<10 (50)	<10 (135)	<10 (212)
(h)	3	Ef1223	<10 (82)	<10 (44)	<10 (37)	<10 (373)
(i)	4	Bc1793	3.6×10^5 (12)	3.7×10^5 (6)	3.4×10^1 (32)	<10 (na)[a]
(j)	5	Spn13210	1.5×10^7 (34)	7.2×10^6 (23)	<10 (52)	<10 (334)
(k)	6	Dvu1825	1.2×10^7 (10)	2.6×10^5 (5)	6.2×10^2 (12)	<10 (na)
(l)	7	Caur1903	2.6×10^5 (6)	2.6×10^5 (3)	<10 (12)	<10 (14)
(m)	8	Tm0936	3.7×10^4 (6)	9.2×10^3 (3)	1.5×10^6 (2)	1.1×10^7 (1)
(n)	9	Moth1224	<10 (na)	<10 (na)	<10 (na)	<10 (na)
(o)	10	Avi5431	<10 (na)	<10 (na)	<10 (na)	<10 (na)
(p)	11	Tvn0515	<10 (175)	<10 (131)	<10 (106)	2.6×10^5 (28)

[a]na, the docking pose was not in a productive orientation for catalysis. The docking ranks, relative to the 57,672 HEI metabolite library, are given in parentheses.

Figure Legends

Figure 1. Sequence similarity network for proteins related to Tm0936 from *Thermotoga maritima*. A node (dot) represents an enzyme from a bacterial species and an edge (a connecting line) indicates that the two proteins are related by a BLAST E-value of 10^{-100} or better. Proteins sharing sequence similarity with Tm0936 cluster into apparent subgroups and 12 of these have been arbitrarily numbered and color-coded based on the network diagram. Subgroups are predicted to be functionally similar, and representatives from each subgroup, denoted by the letters **a-p**, were selected for purification and functional characterization.

Figure 2. The active site of Tm0936. The crystal structure of Tm0936 (PDB id: 2PLM) in the presence of *S*-inosylhomocysteine (green) highlights the four residues (dark gray) that are important for substrate binding. Arg-148 and Arg-136 bind to the carboxylate moiety of SAH; His-173 and Glu-84 interact with N3 of the purine ring and the 2', 3' hydroxyls of the ribose moiety, respectively. Faded residues denote the six residues that bind the zinc or facilitate proton transfer reactions during the catalytic transformations. These residues are conserved in all of the proteins depicted in **Figure 1**.

Figure 3: The structures of substrates for the enzymes related to Tm0936.

Figure 4: Comparison of the crystallographic and modeled structures of Cv1032. **(A)** The crystal structure of Cv1032 (PDB: 4F0S) in a catalytically productive state (white ribbon), the crystal structure of Cv1032 (PDB: 4F0R) in a catalytically unproductive state (cyan ribbon), and the homology model of Cv1032 based on the X-ray structure of Tm0936 (PDB: 2PLM) (yellow ribbon). The inosine bound in the active site of Cv1032 (PDB: 4F0S) is highlighted (white stick).

(**B)** The crystal structure of inosine (white stick) in the active site of Cv1032 (transparent white stick), and the docking pose of adenosine (yellow stick) in the modeled active site (transparent green stick) composed of the same set of residues as the active site.

Figure 5: The binding pose of guanine in the homology model of Moth1224. The docking pose of guanine in a high-energy intermediate state (yellow stick) in the modeled active site of Moth1224 (transparent white stick) is presented.

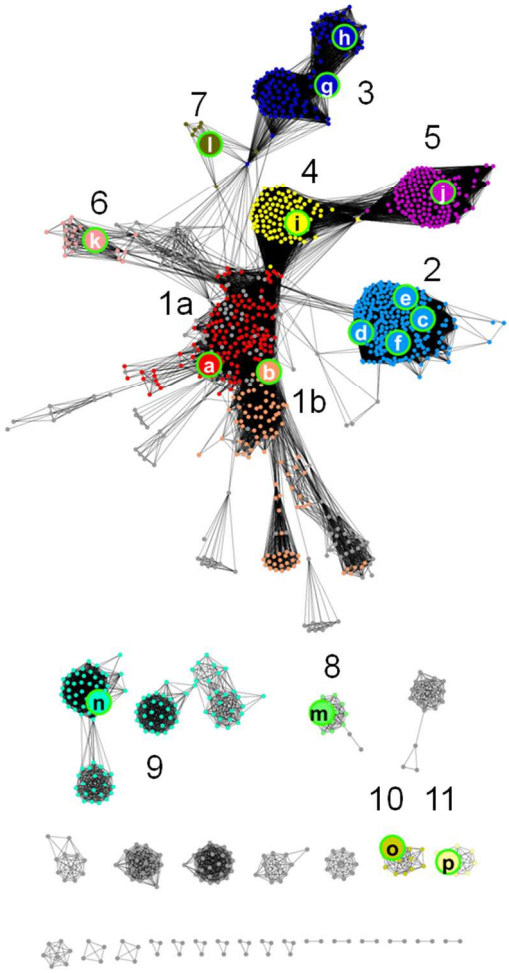

Figure 1

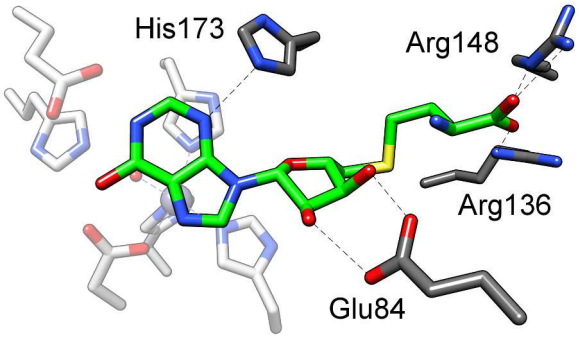

Figure 2

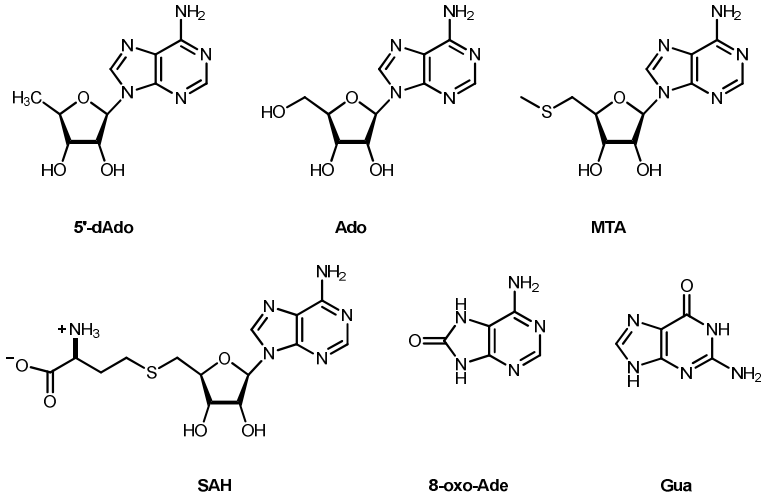

Figure 3

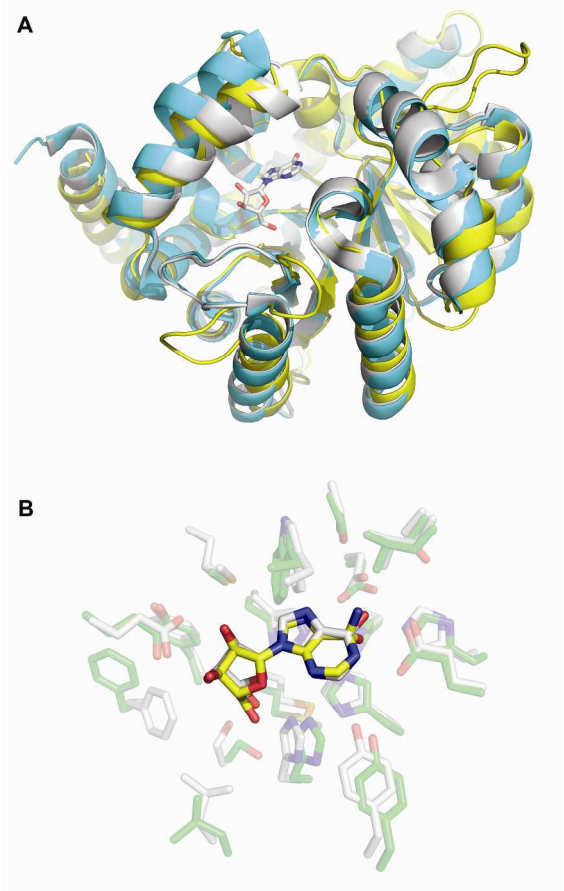

Figure 4

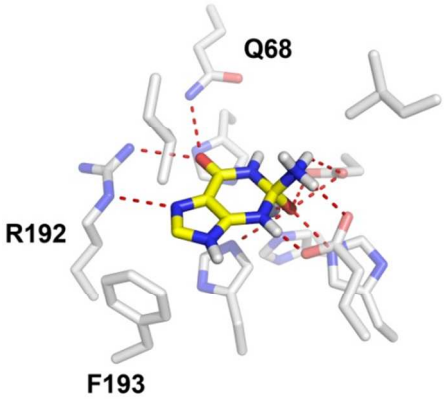

Figure 5

Table of Contents Graphics