BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Covalent Docking Predicts Substrates for Haloalkanoate Dehalogenase Superfamily Phosphatases

Nir London,[†] Jeremiah D. Farelli,[‡] Shoshana D. Brown,[†,§] Chunliang Liu,[⊥] Hua Huang,[⊥] Magdalena Korczynska,[†,∥] Nawar F. Al-Obaidi,[#] Patricia C. Babbitt,[†,§] Steven C. Almo,[#] Karen N. Allen,*[,‡] and Brian K. Shoichet*[,†,∥]

[†]Department of Pharmaceutical Chemistry, and [§]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, United States

[‡]Department of Chemistry, Boston University, 590 Commonwealth Avenue, Room 290, Boston, Massachussetts 02215, United States

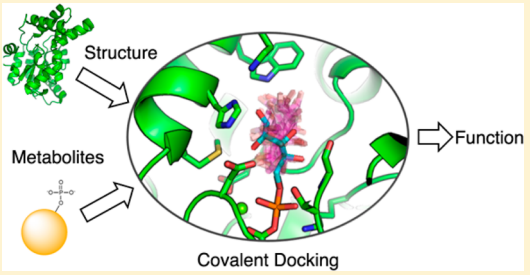
[⊥]Department of Chemistry and Biology, University of New Mexico, Albuquerque, New Mexico 87131, United States

[∥]Faculty of Pharmacy & Ontario Institute for Cancer Research, University of Toronto, Toronto, Canada

[#]Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York, New York 10461, United States

Supporting Information

ABSTRACT: Enzyme function prediction remains an important open problem. Though structure-based modeling, such as metabolite docking, can identify substrates of some enzymes, it is ill-suited to reactions that progress through a covalent intermediate. Here we investigated the ability of covalent docking to identify substrates that pass through such a covalent intermediate, focusing particularly on the haloalkanoate dehalogenase superfamily. In retrospective assessments, covalent docking recapitulated substrate binding modes of known cocrystal structures and identified experimental substrates from a set of putative phosphorylated metabolites. In comparison, noncovalent docking of high-energy intermediates yielded nonproductive poses. In prospective predictions against seven enzymes, a substrate was identified for five. For one of those cases, a covalent docking prediction, confirmed by empirical screening, and combined with genomic context analysis, suggested the identity of the enzyme that catalyzes the orphan phosphatase reaction in the riboflavin biosynthetic pathway of Bacteroides.



With the explosion of protein sequences, protein functional assignment has emerged as a key problem of the postgenomic era.[1] Despite much progress,[2] sequence-based bioinformatics approaches are mostly limited to annotation transfer of known functions.[3] Meanwhile, function prediction using structure alone is also challenging,[4−6] in part due to the multiple chemical reactions catalyzed by enzymes sharing the same folds.[7−9] Structure-based methods have had most success when they have been combined with ligand chemistry, often via molecular docking.[10−17] In these calculations, libraries of candidate substrates are fit into active sites. Noncovalent complementarity between the protein and the ligand is calculated, using either high-energy intermediate[10,11] or ground-state[18,19] forms of the candidate substrates. Whereas this method suffers from the well-known weaknesses of docking,[20,21] it has nevertheless succeeded in predicting the activities of several families of enzymes, and a much larger number of individual enzymes by annotation transfer.

A key gap in this docking approach has been the reliance on modeling noncovalent fit between a substrate and an enzyme, using modifications of methods first developed for inhibitor discovery.[22−26] Whereas this has proven effective for metalloenzymes such as those in the amidohydrolase and enolase superfamilies, many enzymes proceed through a covalent intermediate that does not lend itself readily to noncovalent modeling. For instance, serine proteases[27] and esterases[28] proceed through an acyl-enzyme intermediate, as do β-lactamases,[29] while decarboxylases and transaminases often form covalent adducts with PLP cofactors.[30] Indeed, some have speculated that many enzymes undergo covalent reactions in the key recognition step along the reaction coordinate.[31] For these enzymes, noncovalent docking of candidate substrates is problematic, as the bond-length approach of the covalent intermediate, and the constraints of the new covalent bond, are poorly modeled by the noncovalent terms of standard docking.

We were thus inspired to investigate the application of a new covalent docking screening method, DOCKovalent,[32] to substrate prediction for enzymes that proceed through covalent intermediates. The method combines covalent bond-length and angle constraints with noncovalent complementarity, drawn from standard docking, and enables large-scale library screens. As with classical, noncovalent docking, the method makes

Received: September 10, 2014
Revised: December 11, 2014
Published: December 16, 2014

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

important approximations and adds new ones. Most importantly, it does not calculate the energy of the covalent terms (bond length and angle terms are ignored, as are new torsional energies) but relies exclusively on restraints to model the covalent adduct and complementarity energies from the noncovalent terms. Whereas this has advantages—preventing, for instance, the dominance of covalent terms—the approximation is substantial; as is true with any docking method, it must be tested experimentally before it can be shown to be useful. While covalent docking was used in the past retrospectively to predict substrates of glutathione transferases[33] and predict the chain length of polyprenyl transferases substrates,[34] to our knowledge it was never used in large scale against an enzyme family with a diverse substrate range.

Here we describe the testing of this covalent screening approach against enzymes of the haloalkanoate dehalogenase (HAD) superfamily (HADSF), a superfamily with almost 80 000 sequences in the Structure−Function Linkage Database.[35] Largely dominated by phosphatases, HAD enzymes have wide substrate diversity,[36] with substrates ranging from phosphoserine and histidinol-phosphate through sugar and nucleotide phosphates. We undertake parallel docking and screening campaigns of exactly the same library of phosphate-bearing candidate substrates, using a recently developed empirical screening method to experimentally characterize the substrate specificity of multiple HADSF enzymes. Evaluation of the covalent docking approach by both retrospective enrichment of known substrates and prospective docking for new substrates is considered. Guided by the docking, empirical screens, and by genomic context, we suggest a probable enzyme to fulfill an orphan reaction in the riboflavin biosynthetic pathway, the catalysis of phosphate hydrolysis from 5-amino-6-(5-phospho-D-ribitylamino) uracil. This reaction was hypothesized decades ago, and only recently it was suggested that HADSF members might fill this role[37] in flavinogenic microorganisms.

METHODS

Library Generation. The 167 phosphate substrates (Supplementary Dataset 1, Supporting Information) were represented as isomeric smiles strings. The oxygen atom connecting the substrates to the target phosphate was converted to a "dummy" atom to represent the location of the covalent bond. Ligand conformations were generated using Omega[38] as described in ref 32. Corina[39] (Molecular Networks, Erlangen, Germany) was used to generate initial 3D structures and stereoisomers. EPIK[40] (Schrodinger software, Catsville, NY) was used for protonation and tautomer assignment. AMSOL[41] was used to assign partial charges and solvation energies.

Docking. DOCKovalent[32] is a covalent adaptation of DOCK3.6.[42,43] Given a pregenerated set of ligand conformation and a covalent attachment point, it exhaustively samples ligand conformations around the covalent bond and selects the lowest energy pose using a physics-based energy function. For the docking reported in this work, a pentavalent triagonal bipyrimidal phosphate was first modeled covalently attached to the catalytic aspartate residue, either by docking or by manual placement. The substrate library was then covalently docked to the remaining axial phosphate oxygen with a bond length of 1.4 ± 0.3 Å sampled in 0.1 Å increments, bond angle (P−O−ligand) of 120 ± 10° in 2.5° increments, and bond angle (O-ligand covalent attachment point-rest of ligand) of 109.5 ± 10°, also in 2.5° increments. Scoring was as described[32] using a physics-based energy function which uses precalculated van der Waals, electrostatics (calculated with DELPHI[44]), and solvent-excluded desolvation[42] grids, with dampening of the electrostatic potential of the phosphate oxygens to avoid excessive interactions of the ligand with its own phosphate. The receptor is kept fixed throughout the docking simulation.

Retrospective Assessment. For the pose recapitulation benchmark, the phosphate oxygen was used as an anchor point. RMSD was calculated for all substrate heavy atoms excluding the phosphate. The screening hit-rate was defined as the number of substrates divided by the size of the tested library. The docking hit-rate was defined as the number of substrates ranking in the top 20 of the docking hit list divided by 20. Docking enrichment was calculated by dividing the docking hit rate by the screening hit rate; random selection gives an enrichment of 1.

Noncovalent Docking of High-Energy Intermediates. Here the phosphate reactive center is also represented in a trigonal bipyramidal geometry. Three-dimensional models of the ligands are prepared with Corina,[39] and the molecules are then transformed to their intermediate state using OEChem TK, version 1.7.4, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, www.eyesopen.com, 2010. In these reactions, first the phosphate is attacked with a hydroxide and second the phosphate protonation states are enumerated, creating multiple high-energy intermediates for each molecule. Next the molecules are prepared as described previously.[10,45]

Receptor structures were prepared for noncovalent docking as described.[10,45,46] Chain A from each structure was chosen, except for 3DDH where chain B was used. 1RKU, 3PGV, 3N07, 2OBB, 3GYG, 3MMZ, and 3N1U were prepared as dimers. Only the highest occupancy rotamer of each residue was kept, and incomplete side chains were replaced by using rotamer libraries.[47] The charge of the Mg^{2+} ion was reduced to +1.4 and two water molecules were placed to coordinate the metal ion in the same idealized geometry for all structures. Parallel docking experiments were run with the wild type enzyme, and with a mutation of the catalytic aspartate to a serine residue, in which the hydroxyl was oriented to coordinate the Mg^{2+}. The high-energy library was then docked to each active site with DOCK3.6.[42] The docking was run using receptor and ligand bin sizes of 0.4 Å, an overlap of 0.1 Å, a distance tolerance of 1.5 Å, color matching turned off, and 250 cycles of rigid-body minimization.

Genome Neighborhood Networks. Beginning with gil 29346380, a set of similar protein sequences (seed sequences) were collected by performing a BLAST[48] search against the NCBI GenPept database[49] at an e-value cutoff of 10^{-20}. For each protein BLAST hit, the analogous gene was determined, and the proteins corresponding to the 10 genes upstream and downstream to each were collected. All of these proteins were combined, and an all-by-all BLAST was performed at an E-value cutoff of 10^{-14} against a custom database containing only the proteins of interest. A cytoscape[50] network was created from these BLAST results. Each node in the network represents a single protein sequence, and each edge represents the pairwise connection with the most significant BLAST E-value (better than the cutoff) connecting the two sequences. Connections between nodes are only shown if the E-value of the best Blast hit between two sequences is at least as good as the specified E-value cutoff. Lengths of edges are not meaningful except that sequences in tightly clustered groups are more similar to each other than sequences with few connections. The nodes were arranged using the yFiles organic layout provided with Cytoscape version 2.8. Annotation information retrieved from Swiss-Prot[51] (functional

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

annotation) and NCBI[49] (taxonomy information; position of the corresponding gene relative to the seed gene) was associated with each node as applicable.

Empirical Screen. The target protein was diluted to 10 μM in a volume of 2.5 μL in assay buffer (20 mM HEPES, pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 2 mM DTT) and added to each of 167 putative substrates diluted in assay buffer to 2 mM, in a volume of 2.5 μL (see Supplementary Dataset 1, Supporting Information for the substrate list) in duplicate wells using 384-well plates (Corning 384 Well low Volume Black Clear Bottom - Cat. No. 3540) and incubated for 30 min followed by addition of BioMol Green (13.5 μL; Prod. No. BML-AK111), a dye that is sensitive to the presence of free phosphate. The mixture was incubated for an additional 45 min and absorbance at 650 nm was read using an EnVision multilabel reader (product number: 2104-0010).

Preparation of Recombinant *Escherichia coli* RibA. The DNA encoding the *ribA* gene from *E. coli* was amplified by polymerase chain reaction (PCR) using *E. coli* genomic DNA (ATCC 29148D), Pfu Turbo DNA polymerase, and oligonucleotide primers (5′-AATCATATGCAGCTTAAACGTGTG and 5′-GGTTATTTTGGATCCGGCAAGC) containing restriction endonuclease cleavage sites *Nde*I and *Bam*HI. The pET-15b TEV vector, cut by restriction enzymes *Nde*I and *Bam*HI, was ligated to the PCR product that had been purified and digested with the same restriction enzymes. The ligation product was used to transform to NEB Express Iq Competent *E. coli* cells that were then grown on a ampicillin-containing agar plate. A selected colony was checked for RibA expression, and the isolated plasmid was sequenced to verify the correct gene sequence. For RibA preparation, the transformed cells were grown at 37 °C with agitation at 225 rpm in 1 L Terrific Broth containing 100 μg/mL ampicillin to an OD_{600} of 0.6 and then induced for 12 h at 16 °C with 0.4 mM isopropyl beta-D-thiogalactopyranoside (IPTG). The cells were harvested by centrifugation (6500 rpm for 15 min at 4 °C) to yield 16 g/L of culture medium. The cell pellet was suspended (1 g of wet cells/10 mL) in ice-cold buffer A (20 mM HEPES (pH 7.5), 5 mM $MgCl_2$, and 10 mM imidazole). The cell suspension was passed through a French press at 1200 psi before centrifugation at 20 000 rpm at 4 °C for 20 min. The supernatant was loaded onto 5 mL Ni-NTA agarose column at 4 °C. After the column had been washed with 100 mL of buffer B (20 mM HEPES (pH 7.5), 300 mM NaCl, and 20 mM imidazole), the enzyme was eluted with 200 mL of elution buffer C [20 mM HEPES (pH 7.5), 50 mM NaCl, and 500 mM imidazole]. The column fractions were analyzed by SDS−PAGE, and the desired fractions were combined before dialysis at 4 °C against 6 L OF triethanolamine hydrochloride (TEA) buffer (20 mM TEA (7.5) and 100 mM NaCl). The final yield was 20 mg of RibA/g of wet cells.

Preparation of Recombinant *Thermotoga maritima* RibD. The plasmid (PSI biology) containing the *T. maritima ribD* gene was transformed to BL21(DE3)pLysS competent cells and then grown on an ampicillin-containing agar plate. A selected colony was checked for RibD expression. For RibD preparation, the transformed cells were grown at 37 °C with agitation at 225 rpm in 2 L of Terrific Broth containing 100 μg/mL ampicillin to an OD_{600} of 1.0 and then induced for 12 h at 25 °C with 0.6 mM IPTG. The cells were harvested by centrifugation (6500 rpm for 15 min at 4 °C) to yield 12 g/L of culture medium. The cell pellet was suspended (1 g of wet cells/10 mL) in ice-cold buffer A (50 mM HEPES (pH 8.0), 50 mM NaCl, and 10 mM imidazole). The cell suspension was passed through a French press at 1200 psi before centrifugation at 20 000 rpm at 4 °C for 20 min. The supernatant was loaded onto 5 mL Ni-NTA agarose column at 4 °C. After the column had been washed with 100 mL of buffer B (50 mM HEPES (pH 8.0), 300 mM NaCl, and 40 mM imidazole, 10% glycerol), the enzyme was eluted with 200 mL of elution buffer C (50 mM HEPES (pH 8.0), 300 mM imidazole, 10% glycerol). The column fractions were analyzed by SDS−PAGE, and the desired fractions were combined and then dialyzed at 4 °C against 6 L 50 mM HEPES (pH 8.0), 50 mM NaCl. The final yield was 4.9 mg of RibD/g of wet cells.

Preparation of Recombinant *E. coli* Putative 5-Amino-6-(5-phospho-D-ribitylamino) Uracil Phosphatase. The cell stock [BL21 (Ros2), EFI:501083, UniProt: Q8A947] was grown at 37 °C with agitation at 225 rpm in 10 mL of Terrific Broth containing 100 μg/mL ampicillin overnight, and then the 10 mL culture was transferred to 2 L of Terrific Broth containing 100 μg/mL ampicillin and grown at 37 °C with agitation at 225 rpm to an OD_{600} of 0.8, and then induced for 12 h at 16 °C with 0.5 mM IPTG. The cells were harvested by centrifugation (6500 rpm for 15 min at 4 °C) to yield 9 g/L of culture medium. The cell pellet was suspended (1 g of wet cells/10 mL) in ice-cold buffer A (25 mM HEPES (pH 7.5), 50 mM NaCl, and 5 mM $MgCl_2$, 5% glycerol, 10 mM imidazole). The cell suspension was passed through a French press at 1200 psi before centrifugation at 20 000 rpm at 4 °C for 20 min. The supernatant was loaded onto 5 mL Ni-NTA agarose column at 4 °C. After the column had been washed with 100 mL of buffer B (25 mM HEPES (pH 7.5), 300 mM NaCl, 5 mM $MgCl_2$ and 40 mM imidazole), the enzyme was eluted with 200 mL of buffer C (25 mM HEPES (pH 7.5), 5 mM $MgCl_2$, 500 mM imidazole). The column fractions were analyzed by SDS−PAGE, and the desired fractions were combined and then dialyzed at 4 °C against 6 L 25 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM $MgCl_2$, 5% glycerol. The final yield was 8.6 mg of protein/g of wet cells.

Enzymatic Synthesis of 5-Amino-6-(5-phospho-D-ribitylamino) Uracil. A reaction mixture containing 50 mM HEPES (pH 7.9), 5 mM $MgCl_2$, 4.05 mM GTP, 4.54 mM NADPH, 90 μM RibA, and 175 μM RibD in a final volume of 222 μL was incubated at 37 °C for 1.5 h. The concentration of 5-amino-6-(5-phospho-D-ribitylamino) uracil was determined by the decrease in absorbance at 340 nm, indicative of NADPH oxidation (ε_{340} = 6220 M^{-1} cm^{-1}).

Determination of Steady-State Kinetic Constants. Initial velocities for putative 5-amino-6-(5-phospho-D-ribitylamino) uracil phosphatase-catalyzed hydrolysis of sugar phosphate were measured at 25 °C using assay solutions that contained 1 mM $MgCl_2$, 0.1 mM sodium azide 1.0 unit/mL purine nucleoside phosphorylase, and 0.2 mM MESG in 50 mM Tris-HCl (pH 7.5). Absorbance changes were monitored at 360 nm ($\Delta\varepsilon$ = 9.8 mM^{-1} cm^{-1}). The steady-state kinetic parameters (K_m and k_{cat}) were determined by fitting the initial velocity data measured at varying substrate concentrations (ranging from 0.5 K_m to 5 K_m) to the equation: $V_0 = (V_{max}[S])/([S] + K_m)$ where V_0 is the initial velocity, V_{max} the maximum velocity, [S] the substrate concentration, and K_m the Michaelis constant for the substrate, using the SigmaPlot Enzyme Kinetics Module. The k_{cat} values were calculated from V_{max} and [E] according to the equation $k_{cat} = V_{max}/[E]$, where [E] is the enzyme concentration.

RESULTS

Overview of the Method. Catalysis by HAD phosphohydrolases proceeds via an aspartylphosphate intermediate[52] (Supplementary Figure 1, Supporting Information). Since the phosphate-binding mode is conserved in the HAD catalytic core

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

domain[36] we begin by covalently attaching a pentavalent, trigonal-bipyrimidal phosphate[53] to the catalytic aspartate (O_δ–P bond length = 1.9 ± 0.3Å; C_γ–O_δ–P angle = 120 ± 10°; Figure 1). This can be done manually or by applying the covalent docking protocol described below to a model phosphorylated molecule. Once the pentavalent phosphate is placed, we covalently dock a library of candidate substrates to the axial phosphate oxygen (Figure 1; See Methods for geometric parameters). The substrate library of 167 phosphorylated molecules mirrors that of the library used in the empirical screens (Supplementary Data set 1, Supporting Information). For each ligand, we sample ligand conformations with respect to the covalent bond to the fixed phosphate, constrained by ideal bond lengths and angles, in 20° increments around the two newly formed torsion angles, with conformations for the remaining portions of the molecule precalculated. Each sampled conformation is scored using a physics-based scoring function in DOCK3.6,[42] which evaluates the ligand van der Waals and electrostatic interactions, and corrects for ligand desolvation. The best scoring pose for each ligand is saved. The library is then ranked by this docking score.

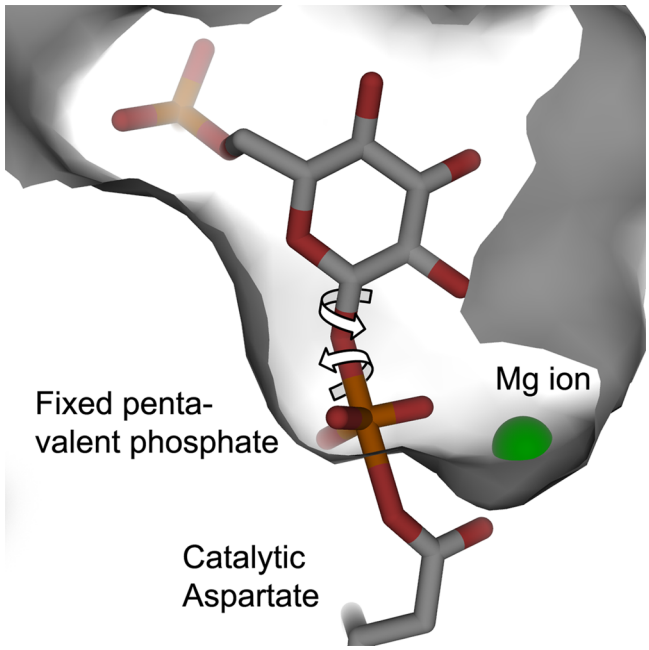


Figure 1. Overview of covalent docking to the HADSF. An illustration of the sampling in the HADSF covalent docking. A pentavalent trigonal-bipyramidal phosphate is modeled as covalently attached adduct to the catalytic aspartate, either manually or via covalent docking of the phosphate. Substrates are then covalently docked to the phosphate axial oxygen, exhaustively sampling the two indicated dihedral angles as well as pregenerated ligand conformations. The best pose is kept for each substrate and substrates are than ranked based on their scores.

Inadequacy of Noncovalent Docking. Previously, we used noncovalent docking of high-energy intermediates to predict substrate recognition by amidohydrolase family enzymes,[10−13] which do not proceed via a covalent intermediate. To investigate whether this approach could model enzymes like those of the HAD superfamily, which do proceed by such an intermediate, we docked high-energy intermediates of the phosphatase reaction against both the unmodified enzyme structure and against an artificial form of that structure truncated at the nucleophilic aspartate (substituting the aspartate with a serine) so as to allow close approach of the substrate and enzyme.

While in several of the cases substrates indeed ranked in the top of the docking hit list (Supplementary Table 1, Supporting Information), it was often the case that the predicted docking pose was not competent for the reaction (Supplementary Figure 2, Supporting Information). This supported efforts at modeling the reaction using the covalent docking approach, which by design produces only competent poses.

Retrospective Assessment of Covalent Docking. We first investigated the ability of the covalent docking method to recapitulate the crystallographic poses of HAD−substrate complexes. We assembled a small benchmark of seven liganded structures covering a diverse range of substrates and HAD subtypes (Supplementary Table 2). In five of the seven cases covalent docking recapitulated the substrate binding pose to less than 2Å RMSD (Figure 2), while in the other two structures the overall binding pose was recapitulated but not with atomic accuracy (RMSD = 2.14 Å and 3.1 Å). A more stringent test is docking to unliganded structures. When using ligand free structures (or in one case, a complex with an alternative ligand bound) performance decreased, although still in three cases poses were recovered to less than 2 Å RMSD (Supplementary Table 2).

To investigate the ability to find HAD substrates using unliganded structures, we calculated docking structures, in covalent high-energy intermediate form, of the library of 167 candidate phosphate substrates that was used in empirical screens for HAD function. We compared docking performance to that of the empirical screen on a set of 20 HADSF members with a solved unliganded crystal structure (Table 1). For each enzyme we considered the top 20 predictions, out of 167 possible, as putative substrates. We considered two alternatives as true substrates from the empirical screen. The poor definition

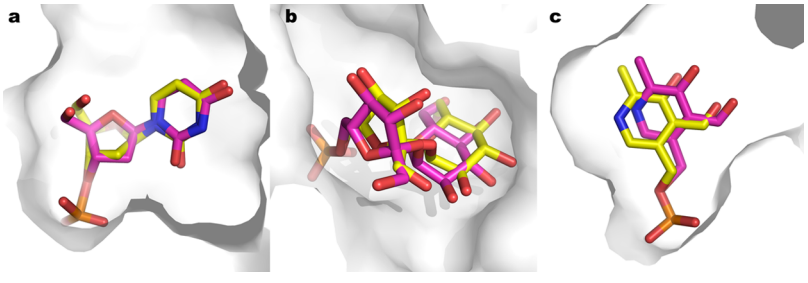


Figure 2. Substrate pose recovery by covalent docking. Examples of covalent docking pose predictions (magenta) for known HAD (white)/substrate (yellow) complexes. (a) Deoxyribonucleotidase in complex with deoxyuridine (PDB: 2I7D); (b) sucrose-phosphatase in complex with sucrose-6P; (c) human pyridoxal phosphate phosphatase in complex with pyridoxal-5P.

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

Table 1. Retrospective Assessment of HAD Substrate Prediction

	enrichment[a]		examples of correctly predicted substrates (docking rank)[b]	
PDB	good	poor	good	poor
2b82[c]	3.30	1.98	IMP (17)	GMP (10)
				dUMP (16)
4dcc	3.30	1.89	riboflavin-5-phosphate-(FMN) (9)	L-sorbose-1-phosphate (5)
				D-iditol-6-phosphate (11)
				arabinose-5-phosphate (19)
4dfd	3.30	1.89	riboflavin-5-phosphate-(FMN) (4)	L-sorbose-1-phosphate (9)
				D-allitol-3-phosphate(15)
				D-galactitol-6-phosphate(17)
3d6j	3.30	2.50	D-tagatose-6-phosphate (4)	arabinose-5-phosphate (1)
			2-deoxyribose-5-phosphate (5)	D-ribose-5-phosphate(10)
			D-2-deoxy-ribitol-5-phosphate (6)	erythrose-4-phosphate (13)
1nrw	2.93	2.16	isoerythritol-4-phosphate (7)	*O*-phosphorylethanolamine (3)
			2-deoxyribose-5-phosphate (18)	D-ribulose-5-phosphate (4)
			glycerol-3-phosphate (19)	D-threitol-4-phosphate (5)
			ribitol-5-phosphate (20)	glycerol-phosphate-(GP) (6)
1rku	2.40	0.60	*O*-phospho-L-serine (7)	arabinose-5-phosphate (16)
1te2	2.10	2.15	mannose-6-phosphate (11)	glucosamine-6-phosphate (2)
			ribitol-5-phosphate (14)	D-2-keto-glucose-6-phosphate (5)
			2-deoxy-D-glucose-6-phosphate (18)	L-ribitol-5-P (6)
3pgv(AC)	1.55	1.52	glucosamine-6-phosphate (2)	allose-6-phosphate (8)
			D-sedoheptulose-7-phosphate (6)	IMP (9)
			D-psicose-6-phosphate (13)	dGMP (12)
			CMP (20)	D-allitol-6-phosphate (14)
4eek	1.20	1.13	UMP (6)	D-altronate-6-phosphate (14)
			D-allonate-6-phosphate (12)	CMP (16)
				D-allitol-6-phosphate (17)
3r4c	1.20	0.83	dAMP (20)	CMP (7)
				UMP (10)
				arabinose-5-phosphate (17)
3n07(B)	1.20	0.27	D-fructose-1,6-bisphosphate (17)	
3niw	0.80	0.44	arabinose-5-phosphate (7)	mannose-6-phosphate (11)
2obb	0.00	2.20		DL-glyceraldehyde-3-phosphate (16)
2hx1	0.00	1.47		D-3-deoxy-glucose-6-phosphate (14)
				mannitol-6-phosphate (17)
3gyg[d]	0.00	1.47		DL-glyceraldehyde-3-phosphate (3)
				D-ribulose-5-phosphate (7)
				L-sorbose-1-phosphate (9)
3s6j	0.00	1.44		erythrose-4-phosphate (2)
				DL-glyceraldehyde-3-phosphate (3)
				arabinose-5-phosphate (6)
3mmz[d]	0.00	0.69		D-fructose-1,6-bisphosphate (2)
				6-phosphogluconic-acid (5)
				mannose-6-phosphate (14)
3n1u[d]	0.00	0.65		6-phosphogluconic-acid (4)
				α-D-glucose-1,6-bisphosphate (13)
				arabinose-5-phosphate (14)
1z5g	0.00	0.00		
3ddh(B)	0.00	0.00		

[a]Docking enrichment calculated separately for good and poor substrates as defined in the text. An enrichment value of 1 corresponds to random prediction. Values larger than 1 indicate successful docking predictions. [b]Examples of predicted substrates (docking rank indicated in parentheses) that were empirically shown to serve as good or poor substrates for the indicated enzyme. [c]Unless indicated otherwise chain A was used for prediction. [d]Crystallographic symmetry was applied to model the "biological" binding site.

refers to any molecule that showed a signal larger than 0.1 absorbance units in the empirical screen. The good definition refers to substrates that showed a signal larger than one standard deviation from the mean of all poor substrates signal. While poor substrates are turned over by the enzyme, the cellular function of the enzyme is more likely related to good substrates.

In 12 out of the 20 cases, we find one or more good substrate among the top 20 docking predictions, in 11 of these we also observed enrichment of these good substrates 20−230% better than expected by random (Table 1). In six other cases, a poor substrate is detected by docking (four of them with substantial enrichment), and in two cases no substrate is detected by

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

docking. Although these enrichments are substantially better than a random predictor, the enrichment reached statistical significance (p-value < 0.05, as evaluated by Fisher's exact test; Supplementary Table 3) in only four of these cases, possibly due to the large number of substrates for many of the HAD phosphatases.

Prospective Substrate Prediction. Encouraged by the covalent docking performance in recovering the geometries of bound substrates, and enriching for known vs decoy substrates in retrospective screens, we next covalently docked *prospectively* against structures of HADSF members for which no function was known. We docked the library of 167 phosphorylated metabolites against the structures of seven such enzymes (Table 2) and in parallel screened the same library against

Table 2. Prospective Prediction of HAD Substrates via Covalent Docking

	enrichment[a]		examples of correctly predicted substrates (docking rank)[b]	
PDB	good	poor	good	poor
1nf2[c]	2.09	1.19	2-deoxy-D-glucose-6-phosphate (14)	acetyl-phosphate (5)
				2-deoxyribose-5-phosphate (13)
4gxt	1.79	1.28	D-tagatose-6-phosphate (11)	6-phosphogluconic-acid (2)
			ribitol-5-phosphate (12)	L-sorbose-1-phosphate (6)
			mannitol-6-phosphate (20)	D-glycero-beta-D-manno-heptose-1,7-bisphosphate (7)
4jb3	1.67	1.34	ribitol-5-phosphate (10)	L-ribose-5-P (6)
			meso-erythritol-4-phosphate (14)	L-ribitol-5-P (11)
				L-lyxitol-5-P (13)
3dv9	0.70	1.92	D-threitol-4-phosphate (18)	D-xylose-5-P (1)
				L-xylose-5-P (2)
				D-2-keto-glucose-6-phosphate (3)
2b0c	0.38	0.97	α-D-glucose-1-phosphate (16)	allose-6-phosphate (3)
				D-tagatose-6-phosphate (5)
				D-glucose-6-phosphate (6)
2fi1	0.00	0.60		dTTP (13)
1ydf	0.00	0.00		

[a]Docking enrichment calculated separately for good and poor substrates as defined in the text. An enrichment value of 1 corresponds to random prediction. Values larger than 1 indicate successful docking predictions. [b]Examples of predicted substrates (docking rank indicated in parentheses) that were empirically shown to serve as good or poor substrates for the indicated enzyme. See Supplementary Data set 1 for additional substrates discovered for these targets. [c]Chain A was used for prediction for all targets.

them empirically. Covalent docking was able to predict good substrates in five cases and poor substrates in six cases, with a better than random enrichments in 4/7 cases (Table 2; full empirical screening results are in Supplementary Dataset 1). For instance, D-tagatose-6P ranked 11 by docking for the orphan HAD enzyme EFI-508415 (PDB: 4gxt), forming extensive polar contacts with Asp52, Arg339, and Asp 317 (Supplementary Figure 3a). It was shown by the empirical screen to be a good substrate for this enzyme. L-Xylose-5-P ranked second by docking for the orphan HAD enzyme EFI-502344 (PDB: 3dv9) based on good binding site complementarity and hydrogen bonding network with His38, Trp42, Asp28 and the backbone of Gly129 (Supplementary Figure 3b). Empirically it was observed to be a robust substrate. D-Ribitol-5P ranked 10 by docking to the orphan EFI-501083 (PDB: 4jb3, UniProt: Q8A947), displaying good geometric fit supplemented with two hydrogen bonds to Glu62 (Supplementary Figure 3c). In the empirical screen, it was the best substrate for this enzyme. We decided to further investigate this enzyme, for which the docking prediction combined with genomic context analysis hinted at an interesting biological function.

EFI-501083 Catalyzes an Orphan Reaction in the Riboflavin Biosynthesis Pathway. EFI-501083 (PDB: 4jb3, UniProt: Q8A947) is a HAD enzyme from *Bacteroides thetaiotaomicron* with unknown function, whose structure was determined as part of a larger effort to determine enzyme function.[54] Empirical screens against this enzyme revealed it was a broad range 5−6 carbon alcohol and aldol sugar phosphatase (Supplementary Data set 1). Consistent with this observation, 5 of the top 20 prospective docking predictions were confirmed as substrates by the empirical screen, with the best substrate, as mentioned, D-ribitol-5-phosphate, ranked 10 by docking (Supplementary Figure 3c).

Bacterial enzymatic pathways are often organized in operons such that enzymes of the same pathway are found in sequence proximity in the genome. For this reason the genomic context, here defined as 10 genes upstream and downstream of the target gene, can provide important clues for an enzyme's function.[15] Genome neighborhood networks[55] expand this notion and incorporate the genomic context of closely related sequences (see Methods).

The enzyme EFI-501083 (UniProt: Q8A947) was used to seed a genome neighborhood network calculation. One of the closest clusters in the average genomic neighborhood contained enzymes annotated as participating in the riboflavin biosynthesis pathway (Supplementary Figure 4). Specifically, one cluster contained RibF, a bifunctional riboflavin kinase, and FMN adenylyltransferase, which is comprised of two enzymes (fused in most bacteria) of the common FMN/FAD pathway.[56,57] Another cluster contained RibD, a reductase that catalyzes the conversion of 5-amino-6-(5-phospho-D-ribosylamino) uracil (compound **1**) to 5-amino-6-(5-phospho-D-ribitylamino) uracil (compound **2**; Figure 3a). The next step in riboflavin biosynthesis is the removal the phosphoryl group from compound **2** by a putative, unidentified, phosphatase.[58,59] This genomic context hint, combined with the docking prediction and screening evidence that 5-phospho-D-ribitol, a substructure of **2** (Figure 3a; marked in red), is a substrate for EFI-501083, led us to suspect that it is in fact the missing phosphatase of this pathway in *B. thetaiotaomicron*.

While docking **2** to the structure (PDB: 4jb3) did not provide a fit in the enzyme binding pocket, repacking of the side-chains of Phe83 and Ile16 and subsequent minimization resulted in an enzyme structure that could accommodate **2** in the docking pose of D-ribitol-5-phosphate (Figure 3b). This prompted us to assess **2** as a substrate for EFI-501083 in vitro.

In a phosphatase assay against EFI-501083, D-ribitol-5-phosphate and the related D-ribitol-1-phosphate were found to be enzyme substrates, with k_{cat}/K_M of 1.4×10^3 M^{-1} s^{-1} and 1.7×10^3 M^{-1} s^{-1}, respectively (Table 3). Next we measured the kinetics of phosphoryl hydrolysis of **2** generated by enzymatic synthesis using RibD.[60] The resulting k_{cat}/K_M of 1.6×10^2 M^{-1} s^{-1} (Table 3; Figure 3c) is within 10-fold of the rates

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

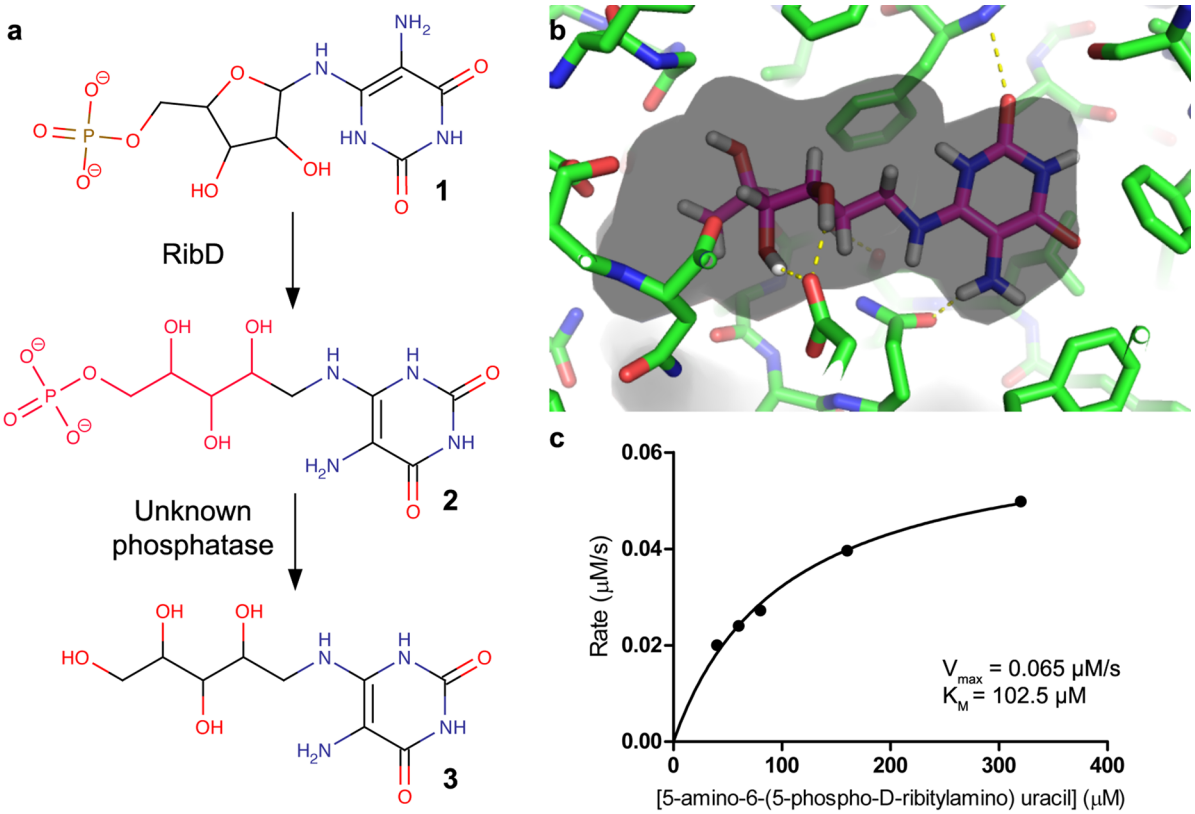


Figure 3. EFI-501083 (UniProt: Q8A947) may be the missing enzyme for the orphan reaction in the riboflavin biosynthetic pathway. (a) Two consecutive intermediate steps in the riboflavin biosynthetic pathway.[58] The first is a reduction catalyzed by RibD, which is found in the genome neighborhood network of EFI-501083. The second is a phosphohydrolase reaction that may be catalyzed by EFI-501083. Ribitol-5P — a substructure of **2** (marked in red) was predicted by docking as a substrate for this enzyme and was a good substrate discovered by empirical screening. (b) A model of **2** in the binding pocket of EFI-501083 suggests it can indeed bind the full substrate. (c) Phosphatase catalytic activity of EFI-501083 on substrate **2**.

Table 3. Steady-State Kinetic Constants for Putative 5-Amino-6-(5-phospho-D-ribitylamino) Uracil Phosphatase[a]

substrate	k_{cat} (s^{-1})	K_m (μM)	k_{cat}/K_m (M^{-1} s^{-1})
D-ribitol-1-P[b]	1.8 ± 0.1	1030 ± 180	1.7×10^3
D-ribitol-5-P[b]	1.5 ± 0.1	1080 ± 140	1.4×10^3
5-amino-6-(5-phospho-D-ribitylamino)uracil[c]	0.016 ± 0.001	100 ± 10	1.6×10^2

[a]Catalyzed hydrolysis of sugar phosphate at 25 °C and pH 7.5 (see Methods). [b]Substrate concentration were determined by full conversion of free phosphate catalyzed by EFI-501083. [c]Substrate concentration was determined by decrease in absorbance at 340 nm, indicative of NADPH oxidation (ε_{340} = 6220 M^{-1} cm^{-1}) during the enzymatic synthesis.

measured for the preceding step in the riboflavin pathway[60] consistent with the hypothesis that EFI-501083 may serve as the missing phosphatase of this biosynthesis pathway.

DISCUSSION

Two key observations emerge from this study. First, covalent docking captures the substrate recognition for multiple subtypes of HAD superfamily enzymes, which have a key covalent intermediate along their reaction coordinate; this approach should be applicable to enzymes from other superfamilies that go through such an intermediate. Such substrate recognition, both retrospective and prospective, was better captured by the new covalent docking approach than by classical, noncovalent docking. Second, when combined with genome neighborhood networks, the method illuminated the identity of a long hypothesized phosphatase that completes the riboflavin biosynthesis pathway.

Noncovalent docking was proven successful in various modeling and substrate prediction applications[11−13] including those performed on dehalogenases.[61,62] However, it requires no great leap to imagine that classical, noncovalent docking will struggle to model the covalent intermediates that feature in so many enzyme reaction coordinates. For these reactions, the enzyme active site is preorganized to stabilize such a covalent intermediate;[63,64] hence any noncovalent positioning of the intermediate will be suboptimal. Naturally, just because noncovalent docking might be expected to struggle does not mean that a covalent docking method will succeed. The ability to accurately recover substrate-bound crystallographic structures, even using unliganded receptor structures in some cases, and to identify true substrates both retrospectively and prospectively, supports the use of this method, and likely many covalent docking approaches, for prospective substrate prediction. For instance, the accuracy of pose recapitulation for known cocrystal complexes (Figure 2; Supplementary Table 2) leads us to believe that highly scoring covalent docking poses are in productive, on-path conformations, in contrast to the incompatible poses predicted by noncovalent docking (Supplementary Figure 2).

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

The correct prediction of good substrates for 17 out of 27 overall different proteins, five of them prospectively, and poorer substrates in seven additional cases, speaks to the pragmatic applicability of this approach.

The discovery that EFI-501083 can fulfill the role of the long-hypothesized phosphatase in the riboflavin biosynthesis pathway illuminates the potential of this method for new biological discovery. Since the 1950s, when the observation was made that the yield of vitamin B2 could be increased during fermentation of *Eremothecium ashbyii* by the addition of purines to the culture medium,[65] numerous studies have demonstrated that the atoms of the purine ring system are incorporated into riboflavin and have elucidated the related biochemical transformations (for review see ref 59). However, although the intermediacy of the pyrimidine **2** is well established, as is the use of the dephosphorylation product **3** by lumazine synthase, the identity of the phosphatase itself has remained elusive. Recently, a clue to the involvement of HADSF members in this phosphatase activity came from plants, which use the same pathway as eubacteria, in which a HADSF member purified from chloroplasts was shown to have FMN hydrolase activity.[66] Newly published work in *E. coli* presented a chemi-enzymatic synthesis of **2** and evidence that both the HADSF members YigB and YbjI (17% sequence identity) can support turnover of **2**, with K_m values in the physiological range, 20 and 70 μM, respectively, and specific activity similar to other enzymes in the riboflavin pathway.[37] The single deletion strains of the *yigB* or *ybjI* genes grow normally in the absence of riboflavin, a finding that is ascribed to the fact that either can carry out the phosphatase activity and on the observed activity on FMN of other *E. coli* HADSF enzymes.[67] Here, the candidate substrates prioritized by covalent docking, together with empirical screens and genome context, identified enzyme EFI-501083 from *B. thetaiotaomicron* as the long sought phosphatase. This enzymatic function could not have been inferred by using each of these prediction methods separately, nor was this function clear from sequence identity alone, as an isofunctional orthologue of either *yigB* or *ybjI*, given the sequence identities of only 17% and 5%, respectively. We note that in *B. thetaiotaomicron*, a second member EFI-501088 with 33% identity to EFI-501083, has activity against FMN[68] with similar low identity to *yigB* or *ybjI* (sequence identity 17% and 16%, respectively).

While an advantage of covalent over noncovalent docking is its dramatic reduction in the number of sampled degrees of freedom, it shares the limitations of most docking approaches and must overcome several challenges. We do not pretend here to have met all of these challenges here —key problems such as accurate accounting for the covalent-bond energetics and binding-site flexibility remain unresolved. HAD enzymes can be found in an "open" or "closed" conformation. While for this study we focused on structures solved in the closed conformation, being able to model this conformational change would increase the coverage of putative targets for our approach. Finally, the approach, like docking in general, suffers from false positive predictions. Indeed, compared to its application to inhibitor discovery,[32] which was able to discover potent, reversible, and highly ligand efficient inhibitors of several enzymes with high hit rates, the covalent docking method often only had modest hit rates with many false negatives. We suspect that this reflects the challenges of recognizing phosphorylated ligands more than intrinsic challenges with substrate versus inhibitor prediction, but this remains an area of ongoing research.

These limitations should not obscure the main observations of this study: against 17 different HAD enzymes, covalent docking captured the essential features of substrate recognition, both retrospectively and, more compellingly, prospectively. The prediction that ribitol-5P is a substrate of EFI-501083, combined with genomic context and kinetic analysis implicates it as the enzyme that catalyzes the orphan phosphatase reaction in the riboflavin biosynthetic pathway sought for over 50 years, illustrating the potential applications of the method to pathway discovery and deeper biology.

■ ASSOCIATED CONTENT

Supporting Information

Results of the empirical screens for seven prospective targets, the performance of noncovalent docking for the retrospective benchmark, including examples of nonproductive poses. The numeric results for the pose-recapitulation benchmark, pose predictions for the prospectively discovered substrates, and the genome neighborhood network for EFI-501083. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*(K.A.) E-mail: drkallen@bu.edu.
*(B.S.) E-mail: bshoichet@gmail.com.

Funding

This work was supported by NIH U54 GM093342 to P.C.B., S.C.A, K.N.A., and B.S. N.L. was also supported by an EMBO long-term fellowship (ALTF 1121-2011).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to Debra Dunaway-Mariano for advice. We thank OpenEye scientific for the free use of the Omega and OEChem through an academic license.

■ REFERENCES

(1) Friedberg, I. (2006) Automated protein function prediction–the genomic challenge. *Briefings Bioinf. 7*, 225−242.

(2) Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Toronen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Honigschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjorne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J., Skunca, N., Supek, F., Bosnjak, M., Panov, P., Dzeroski, S., Smuc, T., Kourmpetis, Y. A., van Dijk, A. D., ter Braak, C. J., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D., and Friedberg, I. (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods 10*, 221−227.

(3) Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol. 9*, e1003063.

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

(4) Watson, J. D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R. A., and Thornton, J. M. (2007) Towards fully automated structure-based function prediction in structural genomics: a case study. *J. Mol. Biol. 367*, 1511−1522.
(5) Wang, Z., Yin, P., Lee, J. S., Parasuram, R., Somarowthu, S., and Ondrechen, M. J. (2013) Protein function annotation with Structurally Aligned Local Sites of Activity (SALSAs). *BMC Bioinf. 14* (Suppl 3), S13.
(6) Han, G. W., Ko, J., Farr, C. L., Deller, M. C., Xu, Q., Chiu, H. J., Miller, M. D., Sefcikova, J., Somarowthu, S., Beuning, P. J., Elsliger, M. A., Deacon, A. M., Godzik, A., Lesley, S. A., Wilson, I. A., and Ondrechen, M. J. (2011) Crystal structure of a metal-dependent phosphoesterase (YP_910028.1) from Bifidobacterium adolescentis: Computational prediction and experimental validation of phosphoesterase activity. *Proteins 79*, 2146−2160.
(7) Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol. 321*, 741−765.
(8) Cuff, A., Redfern, O., Dessailly, B., and Orengo, C. (2011) Exploiting Protein Structures to Predict Protein Functions, in *Protein Function Prediction for Omics Era*, pp 107−123, Springer, Berlin.
(9) Miles, Z. D., Roberts, S. A., McCarty, R. M., and Bandarian, V. (2014) Biochemical and Structural Studies of 6-Carboxy-5,6,7,8-tetrahydropterin Synthase Reveal the Molecular Basis of Catalytic Promiscuity within the Tunnel-fold Superfamily. *J. Biol. Chem. 289*, 23641−23652.
(10) Hermann, J. C., Ghanem, E., Li, Y., Raushel, F. M., Irwin, J. J., and Shoichet, B. K. (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc. 128*, 15882−15891.
(11) Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature 448*, 775−779.
(12) Fan, H., Hitchcock, D. S., Seidel, R. D., 2nd, Hillerich, B., Lin, H., Almo, S. C., Sali, A., Shoichet, B. K., and Raushel, F. M. (2013) Assignment of pterin deaminase activity to an enzyme of unknown function guided by homology modeling and docking. *J. Am. Chem. Soc. 135*, 795−803.
(13) Hitchcock, D. S., Fan, H., Kim, J., Vetting, M., Hillerich, B., Seidel, R. D., Almo, S. C., Shoichet, B. K., Sali, A., and Raushel, F. M. (2013) Structure-guided discovery of new deaminase enzymes. *J. Am. Chem. Soc. 135*, 13927−13933.
(14) Kalyanaraman, C., Bernacki, K., and Jacobson, M. P. (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry 44*, 2059−2071.
(15) Zhao, S., Kumar, R., Sakai, A., Vetting, M. W., Wood, B. M., Brown, S., Bonanno, J. B., Hillerich, B. S., Seidel, R. D., Babbitt, P. C., Almo, S. C., Sweedler, J. V., Gerlt, J. A., Cronan, J. E., and Jacobson, M. P. (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature 502*, 698−702.
(16) Favia, A. D., Nobeli, I., Glaser, F., and Thornton, J. M. (2008) Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. *J. Mol. Biol. 375*, 855−874.
(17) Macchiarulo, A., Nobeli, I., and Thornton, J. M. (2004) Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol. 22*, 1039−1045.
(18) Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S. D., Imker, H. J., Song, L., Fedorov, A. A., Fedorov, E. V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y., Vetting, M. W., Nair, S. K., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Nat. Acad. Sci. U. S. A. 109*, 4122−4127.
(19) Rakus, J. F., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Mills-Groninger, F. P., Toro, R., Bonanno, J., Bain, K., Sauder, J. M., Burley, S. K., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2009) Computation-facilitated assignment of the function in the enolase superfamily: a regiochemically distinct galactarate dehydratase from Oceanobacillus iheyensis. *Biochemistry 48*, 11546−11558.
(20) Schneider, G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery 9*, 273−276.
(21) Klebe, G. (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today 11*, 580−594.
(22) DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. (1988) Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem. 31*, 722−729.
(23) Powers, R. A., Morandi, F., and Shoichet, B. K. (2002) Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure 10*, 1013−1023.
(24) Totrov, M., and Abagyan, R. (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* No. Suppl 1, 215−220.
(25) Goodsell, D. S., and Olson, A. J. (1990) Automated docking of substrates to proteins by simulated annealing. *Proteins 8*, 195−202.
(26) Leach, A. R. (1994) Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol. 235*, 345−356.
(27) Kraut, J. (1977) Serine proteases: structure and mechanism of catalysis. *Annu. Rev. Biochem. 46*, 331−358.
(28) Holmquist, M. (2000) Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr. Protein Peptide Sci. 1*, 209−235.
(29) Knott-Hunziker, V., Petursson, S., Waley, S. G., Jaurin, B., and Grundstrom, T. (1982) The acyl-enzyme mechanism of beta-lactamase action. The evidence for class C Beta-lactamases. *Biochem. J. 207*, 315−322.
(30) Schneider, G., Kack, H., and Lindqvist, Y. (2000) The manifold of vitamin B6 dependent enzymes. *Structure 8*, R1−6.
(31) Zhang, X., and Houk, K. N. (2005) Why enzymes are proficient catalysts: beyond the Pauling paradigm. *Acc. Chem. Res. 38*, 379−385.
(32) London, N., Miller, R. M., Krishnan, S., Uchida, K., Irwin, J. J., Eidam, O., Gibold, L., Cimermancic, P., Bonnet, R., Shoichet, B. K., and Taunton, J. (2014) Covalent docking of large libraries for the discovery of chemical probes. *Nat. Chem. Biol. 10*, 1066−1072.
(33) Dong, G. Q., Calhoun, S., Fan, H., Kalyanaraman, C., Branch, M. C., Mashiyama, S. T., London, N., Jacobson, M. P., Babbitt, P. C., Shoichet, B. K., Armstrong, R. N., and Sali, A. (2014) Prediction of substrates for glutathione transferases by covalent docking. *J. Chem. Inf. Model. 54*, 1687−1699.
(34) Wallrapp, F. H., Pan, J. J., Ramamoorthy, G., Almonacid, D. E., Hillerich, B. S., Seidel, R., Patskovsky, Y., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Poulter, C. D. (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Nat. Acad. Sci. U. S. A. 110*, E1196−1202.
(35) Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L., and Babbitt, P. C. (2014) The Structure-Function Linkage Database. *Nucleic Acids Res. 42*, D521−530.
(36) Allen, K. N., and Dunaway-Mariano, D. (2009) Markers of fitness in a successful enzyme superfamily. *Curr. Opin. Struct. Biol. 19*, 658−665.
(37) Haase, I., Sarge, S., Illarionov, B., Laudert, D., Hohmann, H. P., Bacher, A., and Fischer, M. (2013) Enzymes from the haloacid dehalogenase (HAD) superfamily catalyse the elusive dephosphorylation step of riboflavin biosynthesis. *ChemBioChem 14*, 2272−2275.
(38) Hawkins, P. C., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model. 50*, 572−584.
(39) Gasteiger, J., Rudolph, C., and Sadowski, J. (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol. 3*, 537−547.
(40) Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., and Uchimaya, M. (2007) Epik: a software program for pK(a)

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537

prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des. 21*, 681−691.
(41) Li, J., Zhu, T., Hawkins, G. D., Winget, P., Liotard, D. A., Cramer, C. J., and Truhlar, D. G. (1999) Extension of the platform of applicability of the SM5. 42R universal solvation model. *Theor. Chem. Acc. 103*, 9−63.
(42) Mysinger, M. M., and Shoichet, B. K. (2010) Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model. 50*, 1561−1573.
(43) Lorber, D. M., and Shoichet, B. K. (2005) Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem. 5*, 739−749.
(44) Gilson, M. K., Sharp, K. A., and Honig, B. H. (1988) Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem. 9*, 327−335.
(45) Xiang, D. F., Kolb, P., Fedorov, A. A., Meier, M. M., Fedorov, L. V., Nguyen, T. T., Sterner, R., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2009) Functional annotation and three-dimensional structure of Dr0930 from Deinococcus radiodurans, a close relative of phosphotriesterase in the amidohydrolase superfamily. *Biochemistry 48*, 2237−2247.
(46) Korczynska, M., Xiang, D. F., Zhang, Z., Xu, C., Narindoshvili, T., Kamat, S. S., Williams, H. J., Chang, S. S., Kolb, P., Hillerich, B., Sauder, J. M., Burley, S. K., Almo, S. C., Swaminathan, S., Shoichet, B. K., and Raushel, F. M. (2014) Functional Annotation and Structural Characterization of a Novel Lactonase Hydrolyzing d-Xylono-1,4-lactone-5-phosphate and l-Arabino-1,4-lactone-5-phosphate. *Biochemistry 53*, 4727−4738.
(47) Dunbrack, R. L., Jr. (2002) Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol. 12*, 431−440.
(48) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389−3402.
(49) Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res. 40*, D13−25.
(50) Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics 27*, 431−432.
(51) UniProt, C. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res. 40*, D71−75.
(52) Collet, J. F., Stroobant, V., Pirard, M., Delpierre, G., and Van Schaftingen, E. (1998) A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. *J. Biol. Chem. 273*, 14107−14112.
(53) Lu, Z., Dunaway-Mariano, D., and Allen, K. N. (2008) The catalytic scaffold of the haloalkanoic acid dehalogenase enzyme superfamily acts as a mold for the trigonal bipyramidal transition state. *Proc. Nat. Acad. Sci. U. S. A. 105*, 5687−5692.
(54) Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P., Minor, W., Poulter, C. D., Raushel, F. M., Sali, A., Shoichet, B. K., and Sweedler, J. V. (2011) The Enzyme Function Initiative. *Biochemistry 50*, 9950−9962.
(55) Zhao, S., Sakai, A., Zhang, X., Vetting, M. W., Kumar, R., Hillerich, B., San Francisco, B., Solbiati, J., Steves, A., Brown, S., Akiva, E., Barber, A., Seidel, R. D., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2014) Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, e03275.
(56) Mack, M., van Loon, A. P., and Hohmann, H. P. (1998) Regulation of riboflavin biosynthesis in Bacillus subtilis is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. *J. Bacteriol. 180*, 950−955.
(57) Gerdes, S. Y., Scholle, M. D., D'Souza, M., Bernal, A., Baev, M. V., Farrell, M., Kurnasov, O. V., Daugherty, M. D., Mseeh, F., Polanuyer, B. M., Campbell, J. W., Anantha, S., Shatalin, K. Y., Chowdhury, S. A., Fonstein, M. Y., and Osterman, A. L. (2002) From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol. 184*, 4555−4572.
(58) Bacher, A., Eberhardt, S., Fischer, M., Kis, K., and Richter, G. (2000) Biosynthesis of vitamin b2 (riboflavin). *Annu. Rev. Nutr. 20*, 153−167.
(59) Fischer, M., and Bacher, A. (2010) 7.02 - Riboflavin Biosynthesis, in *Comprehensive Natural Products II* (Liu, H.-W., and Mander, L., Eds.) pp 3−36, Elsevier, Oxford.
(60) Magalhaes, M. L., Argyrou, A., Cahill, S. M., and Blanchard, J. S. (2008) Kinetic and mechanistic analysis of the Escherichia coli ribD-encoded bifunctional deaminase-reductase involved in riboflavin biosynthesis. *Biochemistry 47*, 6499−6507.
(61) Pandey, R., Lucent, D., Kumari, K., Sharma, P., Lal, R., Oakeshott, J. G., and Pandey, G. (2014) Kinetic and sequence-structure-function analysis of LinB enzyme variants with beta- and delta-hexachlorocyclohexane. *PloS one 9*, e103632.
(62) Wijma, H. J., Marrink, S. J., and Janssen, D. B. (2014) Computationally efficient and accurate enantioselectivity modeling by clusters of molecular dynamics simulations. *J. Chem. Inf. Model. 54*, 2079−2092.
(63) Kamerlin, S. C., Chu, Z. T., and Warshel, A. (2010) On catalytic preorganization in oxyanion holes: highlighting the problems with the gas-phase modeling of oxyanion holes and illustrating the need for complete enzyme models. *J. Org. Chem. 75*, 6391−6401.
(64) Warshel, A. (1998) Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem. 273*, 27035−27038.
(65) Mac, L. J. (1952) The effects of certain purines and pyrimidines upon the production of riboflavin by *Eremothecium ashbyii*. *J. Bacteriol. 63*, 233−241.
(66) Rawat, R., Sandoval, F. J., Wei, Z., Winkler, R., and Roje, S. (2011) An FMN hydrolase of the haloacid dehalogenase superfamily is active in plant chloroplasts. *J. Biol. Chem. 286*, 42091−42098.
(67) Kuznetsova, E., Proudfoot, M., Gonzalez, C. F., Brown, G., Omelchenko, M. V., Borozan, I., Carmel, L., Wolf, Y. I., Mori, H., Savchenko, A. V., Arrowsmith, C. H., Koonin, E. V., Edwards, A. M., and Yakunin, A. F. (2006) Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family. *J. Biol. Chem. 281*, 36149−36161.
(68) Barelier, S., Cummings, J. A., Rauwerdink, A. M., Hitchcock, D. S., Farelli, J. D., Almo, S. C., Raushel, F. M., Allen, K. N., and Shoichet, B. K. (2014) Substrate deconstruction and the nonadditivity of enzyme recognition. *J. Am. Chem. Soc. 136*, 7374−7382.

DOI: 10.1021/bi501140k
Biochemistry 2015, 54, 528−537