ACS Catalysis

Research Article

Cite This: ACS Catal. 2018, 8, 2402−2412

pubs.acs.org/acscatalysis

Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization

Pavel Vanacek,[†,‡,∥] Eva Sebestova,[†,∥] Petra Babkova,[†,‡] Sarka Bidmanova,[†,‡] Lukas Daniel,[†,‡] Pavel Dvorak,[†] Veronika Stepankova,[†,‡,§] Radka Chaloupkova,[†,‡,§] Jan Brezovsky,[†,‡] Zbynek Prokop,*[,†,‡,§] and Jiri Damborsky*[,†,‡]

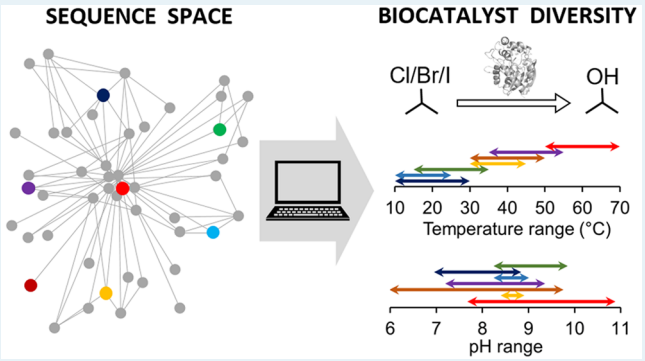
[†]Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

[‡]International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic

[§]Enantis Ltd., Biotechnology Incubator INBIT, Kamenice 34, 625 00 Brno, Czech Republic

Supporting Information

ABSTRACT: Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Here, we describe an integrated system for automated in silico screening and systematic characterization of diverse family members. The workflow consists of (i) identification and computational characterization of relevant genes by sequence/structural bioinformatics, (ii) expression analysis and activity screening of selected proteins, and (iii) complete biochemical/biophysical characterization and was validated against the haloalkane dehalogenase family. The sequence-based search identified 658 potential dehalogenases. The subsequent structural bioinformatics prioritized and selected 20 candidates for exploration of protein functional diversity. Out of these 20, the expression analysis and the robotic screening of enzymatic activity provided 8 soluble proteins with dehalogenase activity. The enzymes discovered originated from genetically unrelated Bacteria, Eukaryota, and also Archaea. Overall, the integrated system provided biocatalysts with broad catalytic diversity showing unique substrate specificity profiles, covering a wide range of optimal operational temperature from 20 to 70 °C and an unusually broad pH range from 5.7 to 10. We obtained the most catalytically proficient native haloalkane dehalogenase enzyme to date ($k_{cat}/K_{0.5}$ = 96.8 mM^{-1} s^{-1}), the most thermostable enzyme with melting temperature 71 °C, three different cold-adapted enzymes showing dehalogenase activity at near-to-zero temperatures, and a biocatalyst degrading the warfare chemical sulfur mustard. The established strategy can be adapted to other enzyme families for exploration of their biocatalytic diversity in a large sequence space continuously growing due to the use of next-generation sequencing technologies.



KEYWORDS: *diversity, sequence space, bioinformatics, biocatalyst, biochemical characterization, activity, substrate specificity, haloalkane dehalogenases*

INTRODUCTION

The postgenomic era is characterized by an exponential increase in the number of protein sequences,[1] which represent an immense treasure of novel enzyme catalysts with unexplored structural-functional diversity. Despite their enormous promise for biological and biotechnological discovery, experimental characterization has been performed on only a small fraction of the available sequences.[2] This "big data" problem is further extended by continuous genome and metagenome sequencing projects which employ powerful next-generation sequencing technologies.[3,4] Traditional biochemical techniques are time-demanding, cost-ineffective, and low-throughput, providing insufficient capacity for the exploitation of genetic diversity.[5] In response to these limitations, high-throughput experimental techniques employing miniaturization and automation have been developed in order to keep track of the ever-growing sequence information.[6−8] Although fluorescent biochemical assays implemented in microformats provide easily measurable signals, enzyme activities from determined end point measurements often differ from those obtained using native substrates ("You get what you screen for.")[6] These microformat techniques are powerful tools for the prefiltering of large libraries; however, they must be followed by additional assays with the target substrates. Robotic platforms using the microtiter plate format are therefore employed to provide quantitative kinetic data.[9] Despite these innovations, the

Received: October 16, 2017
Revised: January 28, 2018

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

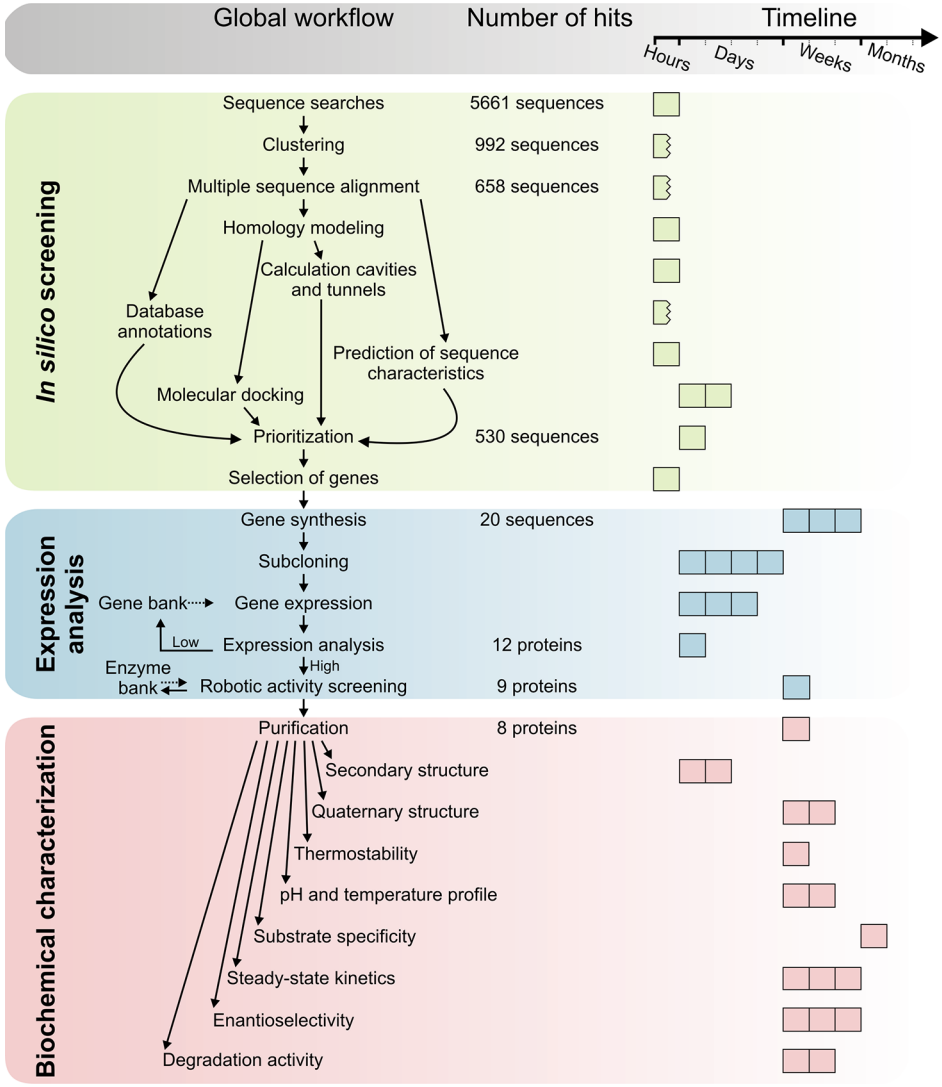


Figure 1. Workflow of an integrated system for the exploitation of the protein structural and functional diversity. Different colors highlight three distinct phases of the workflow: (i) automated sequence and structural bioinformatics (green), (ii) protein production and robotic activity screening (blue) and (iii) biochemical characterization (red). The timeline shows the periods required for data collection and analysis, divided into intervals of 4 h, 4 days, 3 weeks and 3 months column for clarity. A cut square indicates a time requirement of less than 1 h.

existing experimental methods do not provide sufficient capacity for the full biochemical/biophysical characterization of proteins spanning the ever-increasing sequence space, and new technical solutions are therefore required.

Computational approaches offer an adequate capacity for in silico screening of a large pool of sequence entries to facilitate the identification and rational selection of attractive targets for experimental testing. The recently published genomic mining strategy employing molecular modeling and structural bioinformatics has demonstrated the identification of enzymes catalyzing the targeted reaction in a synthetic pathway. This exciting approach led to the discovery of decarboxylases from more than 239 selected hits without expensive and laborious enzyme-engineering efforts.[10] The main benefit of this study lies in the effective sampling of particular enzymatic activity from sequence databases. Developing automated computational workflows and integrating them with experimental platforms is thus essential if the effective discovery of novel proteins is desired. In addition to their applicability in finding new members of protein families, they can identify a wealth of functional novelty when a homologue is found in an unexpected biological setting (such as a new species or environment) or co-occurring with other proteins.[11] Likely hotspots of functional novelty in sequence space may be uncovered either in under-sampled phyla from the tree of life or by finding functional shifts in sequence motifs or domain architecture.[12] Moreover, in silico analysis of structural and functional properties can reveal evolutionary changes in the enzymatic machinery.

Here, we describe an integrated system comprising automated in silico screening protocol and experimental procedures for characterization and the exploitation of the structural and functional diversity of an entire enzyme family (Figure 1). As proof of concept, we used this system to explore the diversity of microbial enzymes haloalkane dehalogenases (EC 3.8.1.5, HLDs). HLDs have been identified in a broad spectrum of microorganisms inhabiting soil, water, animal tissues, and symbiotic plants.[13−21] These α/β-hydrolase fold enzymes, which belong to one of the largest protein superfamilies (>100000 members), catalyze the hydrolytic

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

dehalogenation of a wide range of organohalogens. HLDs can be employed for the biocatalysis of optically pure building blocks,[22−24] the bioremediation of environmental pollutants,[25−27] the decontamination of chemical warfare agents,[28,29] the biosensing of pollutants,[30−32] and molecular imaging.[33−35] This diversity of reported practical applications is especially astonishing if one bears in mind that only two dozen HLDs have been biochemically characterized during the last 30 years, although the sequences of hundreds of putative HLDs are available in genomic databases.[20,36] The proposed integrated system effectively explored the sequence diversity and delivered eight novel biocatalyst possessing unique properties. Particularly, the most catalytically proficient HLD enzyme to date, the most thermostable biocatalyst and the extremophile-derived enzymes are promising for various biotechnological applications. The strategy was critically evaluated by validation of in silico predictions against experimentally verified results.

EXPERIMENTAL SECTION

In Silico Screening. The sequences of three experimentally characterized HLDs were used as queries for two iterations of PSI-BLAST v2.2.28+[37] searches against the NCBI nr database (version 25-9-2013)[38] with *E*-value thresholds of 10^{-20}. Information about the source organisms of all putative HLDs was collected from the NCBI Taxonomy and Bioproject databases.[38] A multiple sequence alignment of all putative full-length HLD sequences was constructed by Clustal Omega v1.2.0.[39] The homology modeling was performed using Modeler v9.11.[40] Pockets in each homology model were calculated and measured using the CASTp program[41,42] with a probe radius of 1.4 Å. The CAVER v. 3.01 program[43] was then used to calculate tunnels in the ensemble of all homology models. The three-dimensional structures of 34 halogenated compounds, which are environmental pollutants, artificial sweeteners, chemical warfare agents, or their surrogates and disinfectants, were constructed in Avogadro[44] and docked to the catalytic pockets using AutoDock 4.2.3. Each local search was based on the pseudo Solis and Wets algorithm with a maximum of 300 iterations per search.[45] The chance for soluble expression in *E. coli* of each protein was predicted based on the revised Wilkinson-Harrison solubility model.[46,47] Detailed bioinformatics protocols are described in the Supporting Information.

Expression Analysis and Activity Screening. Codon-optimized genes encoding 20 putative HLDs were designed and commercially synthesized. The synthetic genes were subcloned individually into the expression vector pET21b between the *Nde*I/*Xho*I restriction sites. *E. coli* BL21(DE3), *E. coli* ArcticExpress(DE3), and *E. coli* Rosetta-gami B(DE3) pLysS competent cells were transformed with DNA constructs using the heat-shock method and expressed in lysogeny broth (LB) or EnPresso B medium. Biomass was harvested at the end of the cultivation, washed, and disrupted using a homogenizer. The activity of cell-free extract toward 1-iodobutane, 1,2-dibromoethane, and 4-bromobutyronitrile substrates was robotically screened at 10, 37, and 55 °C. Detailed experimental protocols are described in the Supporting Information.

Biochemical and Biophysical Characterization. Enzymes were purified using single-step nickel affinity chromatography. Secondary structure was evaluated using circular dichroism spectroscopy at room temperature. Size-exclusion chromatography with static light scattering, refractive index, ultraviolet and differential viscometer detectors was used to analyze protein quaternary structure, molecular weights, hydrodynamic radius, and intrinsic viscosities. Thermal stability was analyzed by circular dichroism spectroscopy and robotic differential scanning calorimetry. The thermal unfolding was monitored by change in the ellipticity or heat capacity over the temperature range from 20 to 90 °C. The temperature profile was determined as an effect of temperature on enzymatic activity toward 1,3-diiodopropane at pH 8.6 over the temperature range from 5 to 80 °C. The pH profile was determined as an effect of pH on enzymatic activity toward 1,3-dibromopropane at the pH ranging from 4 to 12 at 10, 37, or 55 °C. Substrate specificity toward a set of 30 halogenated compounds was analyzed at 10, 37, or 55 °C. The specific activity data toward 30 substrates were analyzed by Principal Component Analysis (PCA). The steady-state kinetics of the novel HLDs toward 1,2-dibromoethane were measured using an isothermal titration calorimeter at either 10, 37, or 55 °C. Enantioselectivity was evaluated from kinetic resolution of 2-bromopentane or ethyl 2-bromopropionate at 20 °C. Enzymatic activity toward chemical warfare agent sulfur mustard was measured using fluorescent assay. The degradation of selected environmental pollutants, 1,3-dichloropropene, γ-hexachlorocyclohexane, and hexabromocyclododecane was analyzed using robotic GC/MS. Detailed experimental protocols are described in the Supporting Information section.

RESULTS

In Silico Screening. The developed automated workflow for the in silico identification and characterization of HLDs provides a useful tool for the selection of interesting proteins for experimental characterization. Sequence database searches led to the identification of 5661 sequences representing putative HLDs and their close relatives. In order to automatically distinguish between putative HLD sequences and sequences of proteins from other families, average-link hierarchical clustering based on pairwise sequence distances was applied. After removing 39 artificial sequences, 953 putative HLDs were retained (333 from HLD-I, 295 from HLD-II, 314 from HLD-III, and 11 from HLD-IIIb). On the basis of multiple sequence alignment, 117 incomplete and 178 degenerate sequences were excluded from the data set. The substitution of halide-stabilizing residues was the most common reason for exclusion (Table S2). The remaining 658 putative HLDs were subjected to homology modeling and in silico characterization. The following data were gathered for each putative HLD: (i) sequence annotations, (ii) the taxonomy of the source organism, (iii) extremophilic properties of the source organism, (iv) a list of highly similar proteins and the most closely related known HLDs, (v) the HLD subfamily, (vi) composition of the catalytic pentad, (vii) the domain composition, (viii) the predicted solubility, and if applicable, (ix) suitable template for homology modeling and (x) constructed homology model, (xi) the volume and area of the catalytic pocket, (xii) characteristics of access tunnels, and (xiii) structures of enzyme−substrate complexes.

The majority of sequences in HLD-I, HLD-II, and HLD-IIIb were correctly annotated (75%, 80%, and 71%, respectively), while this was the case for only 26% of the HLD-III sequences (Table S3). More than half of the sequences in HLD-III were annotated generally as α/β-hydrolase fold enzymes (45%) or hydrolases (10%). Due to the presence of multidomain proteins, 3% of the HLD-I sequences and 11% of the HLD-III sequences were annotated as CMP deaminases and acyl-

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

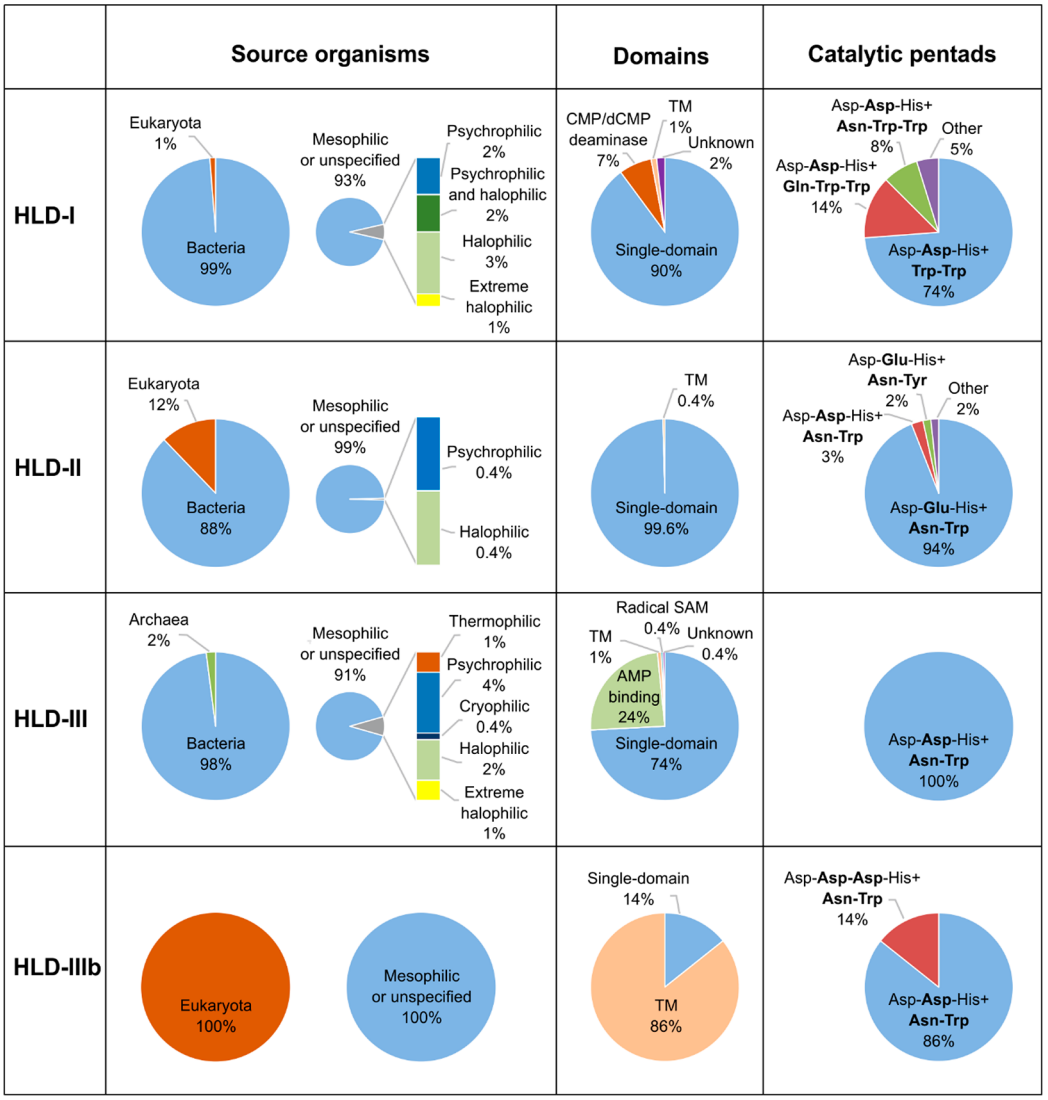


Figure 2. Overview of putative HLDs identified. Most putative HLD sequences are composed of one α/β-hydrolase domain (single-domain). Additionally, some sequences contain the N-terminal cytidine and deoxycytidylate deaminase domain (CMP/dCMP deaminase), the N-terminal radical SAM domain (radical SAM), the C-terminal AMP binding domain (AMP binding), or transmembrane helices (TM). The catalytic pentad in HLDs is composed of nucleophile-catalytic acid−base+halide-stabilizing residues. The nucleophile and catalytic base are conserved in all family members, whereas the catalytic acid and halide-stabilizing residues are variable (highlighted in bold).

CoA synthetases, respectively. Miss-annotations of single-domain proteins were rare (2 proteins). An overview of source organisms, catalytic pentads, and domain compositions of the putative HLDs identified is provided in Figure 2. The sequences of putative HLDs were identified in the genomes of organisms from all three domains of life. The source organisms included 3 thermophilic, 1 cryophilic, 13 psychrophilic, 3 psychrophilic-moderate halophilic, 12 moderate halophilic, and 4 extreme halophilic strains. The majority of the putative HLDs of extremophilic origin were found in subfamilies HLD-I and HLD-III. The prevalent compositions of the catalytic pentads agreed with those described previously.[36] Potential alternative catalytic pentad compositions were predicted for 26% of HLD-I, 6% of HLD-II, and 14% of HLD-IIIb sequences.

Even though all HLDs that have been experimentally characterized to date are single-domain proteins, we identified a number of multidomain proteins in the HLD-I and HLD-III subfamilies. The N-terminal cytidine and deoxycytidylate deaminase domain were detected in 12 HLD-I sequences, while 60 HLD-III sequences had a C-terminal AMP binding domain and one HLD-III sequence had the N-terminal radical SAM domain. N-Terminal transmembrane helices were predicted for 6 out of 7 HLD-IIIb sequences, while they were present in only a small proportion of sequences from the other three subfamilies. Since the majority of structurally characterized HLDs can be found in the HLD-II subfamily, the most reliable homology models could be constructed for members of this subfamily (Figure S1). No protein structures are currently available for members of the HLD-III/b subfamilies, limiting the possibility of homology modeling in these subfamilies (10% for HLD-III and none for HLD-IIIb; Figure S2). Overall, homology models were built and evaluated for 275 sequences.

The predicted volumes of catalytic pockets ranged from 126 Å^3 to 1981 Å^3 (Figure S3). Generally, the HLD-I subfamily members were predicted to have smaller pockets (median volume 554 Å^3) than HLD-II (716 Å^3) and HLD-III (777 Å^3). Tunnels were calculated for the ensemble of all superimposed

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

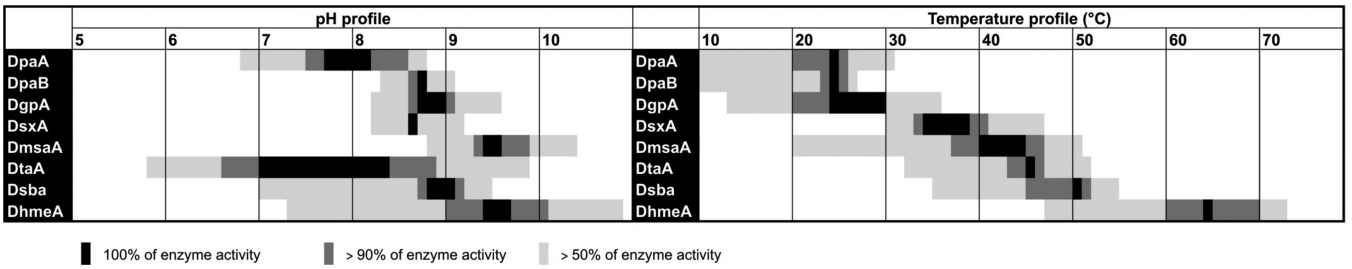
Table 1. Annotations and Predicted Properties of 20 Putative HLDs Selected for Experimental Characterization[a]

	source organism								
name	organism	taxonomy	extremophile	subfamily	predicted catalytic pentad	presence of additional protein domain	volume of active site cavity [Å^3]	radii of transport tunnels p1/p2/p3 [Å]	number of predicted active compounds
DacA	*Aspergillus clavatus* NRRL 1	eukaryota[b]		IIIb[b]	2 catalytic acids[b]	TM helices[b]	N/A	N/A	N/A
DssA	*Shewanella sediminis* HAW-EB3	bacteria	psychrophile[b]	I	3 halide-stabilizing residues[b]		N/A	N/A	N/A
DcaA	*Chloroflexus aurantiacus* J-10-fl	bacteria	thermophile[b]	III	standard		N/A	N/A	N/A
DdaA	*Desulfatibacillum alkenivorans* AK-01	bacteria		III	standard	radical SAM[b]	N/A	N/A	N/A
DamA	*Amycolatopsis mediterranei* S699	bacteria		II	standard		1507[b]	2.1/1.6/1.0	5
DmtA	*Microbacterium testaceum* StLB037	bacteria		III	standard	AMP-binding[b]	N/A	N/A	N/A
DtaA	*Trichoderma atroviride* IMI 206040	eukaryota[b]		II	standard		348	1.3/1.4/–	8[b]
DsbA	*Streptomyces bingchenggensis* BCW-1	bacteria		II	standard		796	2.2/1.3/–	9[b]
DhmeA	*Haloferax mediterranei* ATCC 33500	archaea[b]	moderate halophile[b]	III	standard		N/A	N/A	N/A
DadA	*Alcanivorax dieselolei* B5	bacteria		II	standard		1218	1.5/1.6/–	5[b]
DmgA	marine gamma proteobacterium HTCC2148	bacteria		I	3 halide-stabilizing residues[b]		N/A	N/A	N/A
DpaB	*Paraglaciecola agarilytica* NO_2	bacteria	moderate halophile + psychrophile[b]	I	standard		396	1.9/–/–	4
DpaA	*Paraglaciecola agarilytica* NO_2	bacteria	moderate halophile + psychrophile[b]	I	standard		943	1.9/–/1.4	1
DmsaA	*Marinobacter santoriniensis* NKSG1	bacteria	extreme halophile[b]	I	standard		928	1.9/1.2/1.2	3
DgpA	gamma proteobacterium NOR5-3	bacteria	psychrophile[b]	I	3 halide-stabilizing residues[b]		N/A	N/A	N/A
DcsA	*Caenispirillum salinarum* AK4	bacteria		II	standard		N/A	N/A	N/A
DsxA	*Sandarakinorhabdus* sp. AAP62	bacteria		I	standard		126[b]	1.8/–/–	7
DlxA	*Limnohabitans* sp. Rim47	bacteria		I	3 halide-stabilizing residues[b]	dCMP_cyt_deam[b]	456	2.1/1.1/1.1	7
DncA	*Nonomuraea coxensis*	bacteria		II	standard		1981[b]	2.5/1.6/1.6[b]	1
DmmaA	*Mycobacterium marinum* str. Europe	bacteria		II	standard		545	2.1/1.1/1.1	8[b]

[a]N/A, prediction not available due to lack of homology model. [b]Parameters used as selection criteria.

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

Table 2. Temperature and pH Profiles of 8 Biochemically Characterized HLDs



homology models and then clustered, enabling automated and direct comparison of the tunnels identified in different proteins. The three top-ranked tunnel clusters, corresponding to the p1, p2, and p3 tunnels, were further analyzed. Given that the probe radius used was 1 Å, the p1 and p2 tunnels were found in the majority of the models analyzed (99% and 84%, respectively), while the p3 tunnel was identified in only 50% of models (Figure S4). Finally, 34 potential HLD substrates were docked to homology models and evaluated (Table S4). Generally, the largest number of substrates in the reactive orientation was detected for HLD-II (Figure S5). Almost 15% of HLD-II members were found to have more than 5 substrates in the reactive orientation, compared to 5% and 4% of the HLD-I and HLD-III subfamily members, respectively. One or no substrates were identified for 67% of the HLD-III members, compared to 43% of the HLD-I and 43% of HLD-II members.

The subsequent semirational selection was employed to prioritize the computationally characterized set of 658 putative HLDs based on their sequence and structural characteristics as well as the annotations available. To further enrich the sequence diversity for this fold family, we excluded putative HLDs that had sequence identity lower than 90% to any that had been experimentally characterized. Gathered data for the remaining 530 sequentially distinct putative HLDs were compiled into the data set, on which the selection criteria were applied. The initial criterion was aimed to maximize the diversity of the target properties within the HLD family. Further we prioritized the putative HLDs predicted as soluble with high probability, those with lowest identity to known HLDs, or if the homology model was available, those predicted as likely active HLDs with high confidence. Simultaneously, we strived to filter out or minimize the candidates from HLD-III subfamily, since they are often difficult-to-express. This procedure was followed until the required number of hits with the given properties was received as illustrated in Table S5. It enabled sampling of putative HLDs with the most diverse or completely novel properties. The set of 20 selected genes encoding putative HLDs consisted of (i) bacterial, eukaryotic, and archaeal enzymes, (ii) single- and multidomain enzymes, (iii) enzymes originating from extremophilic organisms, (iv) enzymes belonging to four subfamilies, (v) enzymes with alternative composition of the catalytic pentads, (vi) enzymes with small and large active-site cavity, and (vii) enzymes possessing HLD activity with high confidence (Table 1, Table S6). Putative HLDs are denoted according to the previously established convention (first letter is "D" for dehalogenase; second and third letters are the initials of the source organism; and a last letter "A" or "B" refers to the first or second HLD from a single source organism).

Expression Analysis and Activity Screening. The expression analysis of putative HLD genes was performed in three different *E. coli* strains, with two types of cultivation media and under five expression conditions (Figure S6). In total, 15 (75%) out of 20 target genes were overexpressed and 12 (60%) genes provided proteins in a soluble form that enabled screening of enzymatic activity. We observed some agreement between the theoretically predicted and the experimentally determined solubility of the 12 HLDs that were successfully produced in *E. coli* (Table S7). A robotic platform employing a 96-well microtiter plate format was setup for fast screening of the HLD activity. Altogether, the activities of cell-free extracts of 12 soluble putative HLDs were screened against 3 diverse halogenated substrates (1-iodobutane, 1,2-dibromoethane, and 4-bromobutyronitrile) in an all-enzyme versus an all-substrate screen format at 3 temperatures regime (10, 37, and 55 °C) to maximize the chance of detecting HLD activity. Out of these 12 soluble HLDs, 9 exhibited hydrolytic activity toward at least one of the substrates tested (Table S8). Eight out of nine novel HLDs were successfully purified and subjected to detailed biochemical and biophysical characterization.

Biochemical Characterization. Biochemical and biophysical characterization included the (i) determination of folding and secondary structure, (ii) quaternary structure, (iii) thermostability, (iv) temperature and pH optima, (v) substrate specificity, (vi) steady-state kinetics, (vii) enantioselectivity, and (viii) activity toward selected environmental pollutants and warfare agents.

Circular dichroism spectroscopy was used to assess the correctness of folding and secondary structure composition. All enzymes exhibited CD spectra characteristic of α-helical content and proper folding[48] (Figure S7). The quaternary structure was examined by size-exclusion chromatography. DpaB, DsxA, DgpA, DtaA, and DsbA are monomeric proteins with calculated molecular weights (M_w) ranging between 31.2 and 36.9 kDa. DpaA is as a dimer with a determined M_w = 70.6 kDa and DhmeA is an oligomer with M_w > 2 MDa (Table S9). The oligomeric state of DmsaA could not be determined due to protein precipitation under the conditions tested. Thermally induced denaturation of novel HLDs was tested by monitoring ellipticity or heat capacity (Table S10). Three enzymes originating from psychrophilic organisms (DpaA, DpaB, and DgpA) exhibited significantly lower melting temperatures T_m = 35.2−38.2 °C in comparison to the enzymes of mesophilic origin with T_m = 47.2−54.6 °C. The highest melting temperature T_m = 70.6 °C was determined for the archaeal enzyme DhmeA, making this enzyme the most thermostable wild-type HLD ever reported.

The temperature and pH profiles were determined using a high-throughput robotic assay. Altogether, the novel HLDs exhibited high diversity of operational temperature optima ranging from 20 to 70 °C, while the majority of previously

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

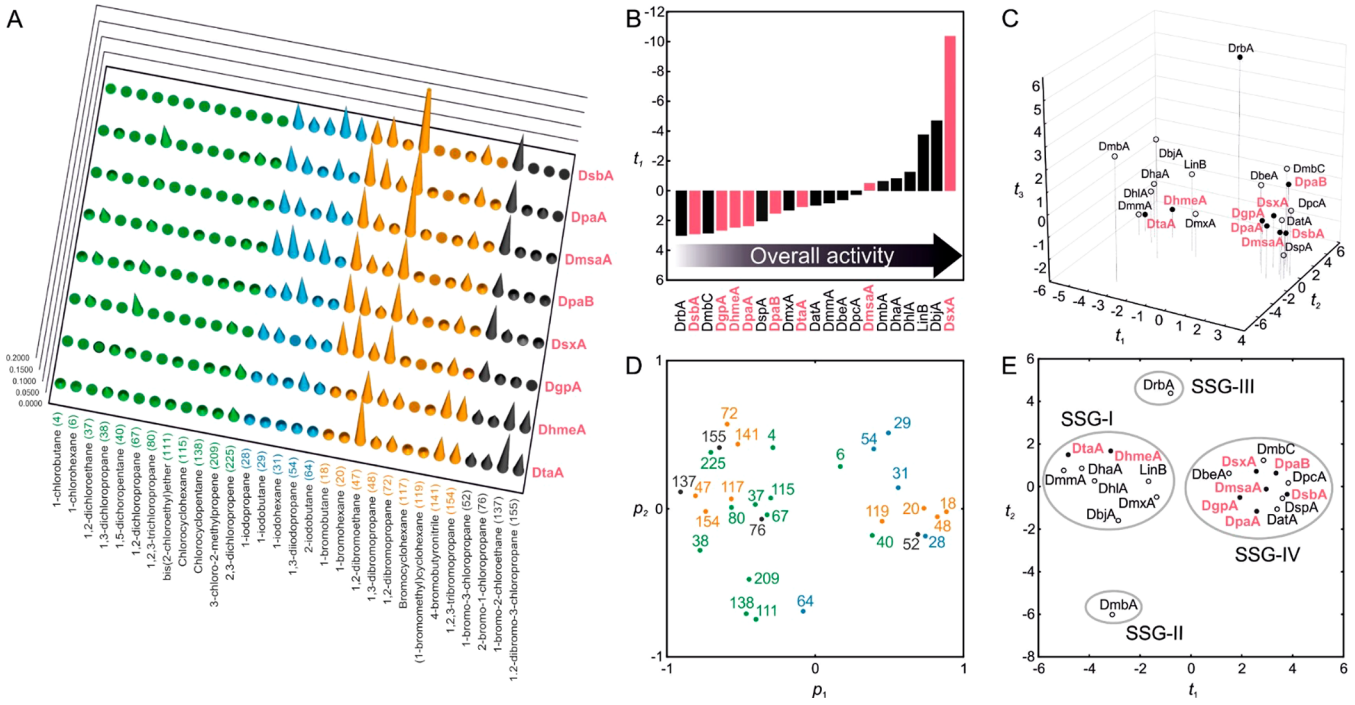


Figure 3. Comparison of the substrate specificities and overall catalytic activities of novel HLDs with previously characterized enzymes using multivariate statistics. (A) Substrate specificity profiles of newly discovered HLDs toward 30 halogenated substrates. (B) The score-contribution plot t_1 shows differences in overall activities of individual HLDs and explains 40% of the variance in the untransformed data set. (C) The score plot $t_1/t_2/t_3$ from PCA of the transformed data set, which suppressed differences in absolute activities and allowed substrate specificity profiles to be compared. The score plot, which describes 55.5% of the variance in the data set, shows enzymes clustered in individual substrate-specificity groups (SSGs). Objects (HLDs) with similar properties (specificity profiles) are colocalized. (D+E) The loading plot p_1/p_2 and the corresponding score plot t_1/t_2 from PCA of the transformed data set. The score plot describes 45.0% of the variance in the data set. The loading plot shows the main substrates for each SSG. Brominated substrates are shown in amber, chlorinated in green, iodinated in blue, and brominated and chlorinated in black.

characterized HLDs[35] possess a narrow range of temperature optima from 30 to 50 °C (Table 2). Moreover, DhmeA showed nearly 90% of its maximum activity even at 70 °C (Figure S8). In correspondence to previously characterized HLDs, the optimal pH conditions ranges between neutral to moderate alkaline condition for most of the novel HLDs variants (Figure S9). Remarkably wide pH tolerance was observed in the case of DhmeA which maintain significant activity in alkaline conditions to nearly pH 11 and DtaA covering unusually broad pH range between 5.7 and 10.0.

Substrate specificity was assessed with a panel of 30 halogenated hydrocarbons (Figure S10). All enzymes exhibited a preference for halogenated hydrocarbons in the following order: brominated > iodinated ≫ chlorinated (Table S11). The optimal length of the alkyl-chain for the substrate was between three and four carbon atoms (Figure 3A). The majority of enzymes possessed rather narrow substrate specificities, converting 12−22 out of the 30 halogenated hydrocarbons. The broadest substrate specificity profile was detected for DtaA, which showed activity toward 27 of the substrates, including the recalcitrant environmental pollutant 1,2,3-trichloropropane. PCA of the transformed data clustered the novel enzymes within substrate specificity groups (SSG) based on their overall substrate specificity profiles (Figure 3, panels C and E). The novel HLDs are spread across different SSGs, indicating significant diversity. DhmeA and DtaA were clustered in SSG-I, whereas the rest of enzymes in SSG-IV. PCA of untransformed specific activities compared the overall activities of HLDs (Figure 3B). Many of the new enzymes exhibited moderate or low activities, indicating that the natural substrates for the new family members are not adequately covered by the used set of 30 representative compounds. The striking exception is the enzyme DsxA, which showed outstanding activity toward most of the tested substrates. This enzyme exhibited the highest specific activity 493.7, 661.2, and 444.6 nmol s^{-1} mg^{-1} toward 1-bromobutane, 1,3-dibromopropane, and 1-bromo-3-chloropropane, respectively. This is 2−3 times higher than those recorded for any previously identified HLDs.

The catalytic properties of the novel enzymes were assessed by measuring steady-state kinetics toward the model substrate 1,2-dibromoethane (Table S12). With the exception of DmsaA, which provided classical hyperbolic kinetics, all novel enzymes exhibited positive cooperativity with Hill coefficients from 1.3 to 2.4. All newly discovered HLDs have shown high $K_{0.5}$ values with 1,2-dibromoethane (≥2.1 mM). The kinetics of DsxA was determined also with 1,3-dibromopropane. Significantly lower $K_{0.5}$ = 0.17 mM and high k_{cat} = 16.5 s^{-1} results in exceptionally high catalytic efficiency ($k_{cat}/K_{0.5}$ = 96.82 mM^{-1} s^{-1}), which is approximately 5.5-fold higher in comparison to the most efficient native HLD of today DadB.[17]

The enantioselectivity of novel HLDs was assayed by determining the kinetic resolution of model *β*-bromoalkane (2-bromopentane) and model *α*-bromoester (ethyl 2-bromopropionate). High enantioselectivity (*E*-value > 200) toward ethyl 2-bromopropionate was observed with DpaA, DpaB, DsxA, and DtaA (Figures S11 and S12), moderately enantioselective DmsaA and DhmeA provided *E*-value 54 and 79, and negligible enantioselectivity was shown by DgpA and DsbA (Table S13). In most cases, only weak enantioselectivity (*E*-value < 15) was recorded with 2-bromopentane. The only

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

exception was DpaA, which exhibited moderate enantioselectivity with *E*-value 85. All enzymes showed (*R*)-enantiopreference which is in correspondence with previously characterized HLDs.[22]

Degradation of halogenated environmental pollutants and warfare agents is one of the main applications of HLDs (Table S14). The activity of novel HLDs has been tested with 1,3-dichloropropene, hexabromocyclododecane, and γ-hexachlorocyclohexane. The majority of enzymes showed significant catalytic activities toward 1,3-dichloropropene (0.5−149.3 nmol s^{-1} mg^{-1}), a synthetic compound introduced into the environment through its use as a fumigant. Activity to other tested environmental pollutants has not been detected. Novel enzymes were also screened with the chemical weapon sulfur mustard. Sulfur mustard is a prominent warfare chemical which has been shown to be transformed to the nontoxic product through the action of HLDs.[28] The screening identified significant activity of DtaA toward sulfur mustard (1.46 nmol s^{-1} mg^{-1}), which makes this enzyme a potential candidate for use in decontamination mixtures or biodetection devices.

DISCUSSION

In this study, we bring our contribution to the big data challenge in the postgenomics era by the development of an automated in silico screening protocol for the exploitation of the protein functional diversity within an enzyme family. Since the rapidly growing genomic databases may contain vague sequence annotations and miss-annotations, our sequence-based search was employed to identify the new members of a protein family based on their evolutionary relationships to other known family members. The putative HLD sequences were automatically identified using global pairwise sequence identities and average-link hierarchical clustering. Furthermore, we cut the hierarchy of sequences at the level of individual HLD-subfamilies, this minimizing the risk of selecting non-HLD sequences.

The sequence analysis identified a number of putative HLDs whose catalytic pentads had alternative compositions: (i) three halide-stabilizing residues, (ii) two catalytic acids, (iii) the HLD-I/II members which have halide-stabilizing Gln/Tyr, (iv) and the HLD-II members with Asp serving as the catalytic acid. It is unclear whether proteins containing these abnormalities represent functional HLDs. Only 58% of the putative HLDs were functionally annotated. While miss-annotations were rare, many proteins were annotated as α/β-hydrolases or hypothetical proteins. The largest number of annotation problems occurred in the HLD-III subfamily, which contains only three experimentally characterized enzymes.[16,20] All 658 putative HLDs were characterized computationally to provide criteria for candidate selection exploring the diversity within the identified sequence space. The selection procedure was based on the mapping of the sequence and structural characteristics as well as annotations. The procedure yielded candidate proteins originating from new species, environments, and under-sampled phyla; proteins with novel domain combinations; proteins with alternative composition of the catalytic pentad; and proteins belonging to newly identified subfamily.

To critically evaluate the effectiveness and the main bottlenecks of the platform, we performed the validation of predictions against experimental data. The expression analysis was performed in three different *E. coli* strains, with two types of cultivation media under five different conditions. Although available solubility prediction tools have been employed during the selection of the candidates and a number of diverse expression conditions have been tested, we attained only 60% success rate for the production of soluble proteins. This is in an agreement with previously published large-scale expression trials demonstrating that 50−80% of bacterial proteins and 15−20% of nonbacterial proteins can be produced in *E. coli* in a soluble form.[49−51] The production of soluble proteins for experimental characterization remains a challenging, "hit-or-miss" affair, and currently represents the biggest bottleneck in studies of this type. With regard to time requirements and cost effectiveness, a more reasonable strategy is to apply expression screening to a larger number of candidates from protein databases rather than wasting time and resources on optimizing the production of "difficult-to-express" proteins. Robust expression systems must constitute an indispensable component of studies of this type.[50] Prediction of protein solubility using software tools represents an attractive future perspective with a challenge toward development of methods achieving higher reliability of estimates.

The initial robotic activity screening performed with cell-free extracts revealed 9 active enzymes out of 12 tested, implying a 75% success rate. Lack of activity in the case of 3 enzymes may have been due to low levels of HLD in the cell-free extract, requirements for specific conditions or preferences for unknown substrate. The latter may indicate that the previously identified "universal" substrates may not be preferred for all existing HLDs. In order to minimize the risk of missing interesting biocatalysts, the substrates used in the activity screening step should be carefully considered. Thanks to the miscellaneous origins and selection approach oriented to maximize diversity of selected candidates, the identified enzymes exhibited a wide range of characteristics with several unique properties. They originated from various phylogenetically unrelated organisms belonging to the domains of Bacteria, Eukaryota, and newly also Archaea. This first archaeal HLD with melting temperature 71 °C represents the most thermostable wild-type HLD known to date. On the contrary, DpaA, DpaB, and DgpA, originating from psychrophilic organisms with melting temperatures below 40 °C open the possibility for the operation at near-to-zero temperatures, which is attractive mainly for environmental applications (e.g., biodegradation or biosensing).[35] The eminent diversity of the novel variants with a wide range of optimal temperature from 20 to 70 °C and broad pH range from 6 to 11 offers expanding operational window to biotechnological applications. Observed extremophilic characteristics were reliably predicted by in silico protocol.

The majority of the newly isolated HLDs exhibited moderate or low enzymatic activities toward 30 halogenated compounds, suggesting that the currently used representative set of substrates for this enzyme family may lack some relevant native substrates. One notable exception is DsxA from *Sandarakinorhabdus* sp. AAP62, which showed exceptionally high activity toward many brominated substrates, particularly those with alkyl-chains containing 2−4 carbon atoms. This observation corresponds with the restricted volume of the active-site cavity predicted by homology modeling. Importantly, DsxA represents the most catalytically efficient member of this enzyme family ever isolated. Crystallographic and cryo-electron microscopy analysis of several newly isolated HLDs is ongoing in our laboratory. DhmeA from the HLD-III subfamily and DgpA with unique 30 amino acid insertion in the N-terminal part of the cap domain are particularly interesting targets.

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

Solving the very first structure from subfamily-III will provide a template for predicting the 3D structures of other subfamily members. In subsequent work, we will implement and release the computational part of the workflow as a user-friendly web tool. The experimental testing process will be extended by integration of miniaturized lab-on-chip assays, requiring only tiny fractions of a protein material and providing increase in the throughput.

CONCLUSIONS

In summary, we have demonstrated that the enormous wealth of genomic sequences available in public databases can be efficiently explored by in silico mapping of proteins structural and functional diversity within protein families. Integration of sequence/structural bioinformatics with experimental procedures enabled us to narrow down the number of enzyme candidates under consideration and allowed their catalytic properties to be explored with reasonable expenditures of time and effort. Although examples of sequence-mining platforms have previously been reported, here we describe integrated platform for the computational analysis of the structural and functional diversity of an entire enzyme family, coupled to full biochemical and biophysical characterization of the identified hits.[10,52] Using our platform, number of novel HLDs with potential practical uses have been identified, characterized, and made available to the community in industry and academia. Further application of our platform to other enzyme families will expand our knowledge in the field of enzymology and will lead to the discovery of novel biocatalysts for the biotechnological and pharmaceutical industries.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscatal.7b03523.

Supporting Methods: in silico screening, expression analysis, biochemical and biophysical characterization; Supporting Tables: list of known HLDs used for definition of HLD clusters, catalytic residues in putative degenerated HLD sequences, annotations of putative HLD sequences, list of the chemical formulas of 34 halogenated compounds, overview of the selection process with respective criteria, accession numbers of putative HLDs, overview of expression analysis of 20 putative HLDs, robotic screening of enzymatic activity, the parameters related to size, shape, and oligomeric state of HLDs, melting temperatures of HLDs by CD spectroscopy and robotic DSC, specific activities of HLDs toward 30 halogenated substrates, steady-state kinetic parameters of HLDs, enantioselectivities of HLDs toward 2-bromopentane and ethyl 2-bromopropionate, activities of HLDs toward warfare agent and environmental pollutants; Supporting Figures: representative 3D structure of HLD, availability of homology modeling templates in individual HLD subfamilies, distributions of predicted volumes of catalytic pockets in HLD subfamilies, distributions of predicted bottleneck radii of p1, p2, and p3 tunnels in HLD subfamilies, distributions of mechanism-based geometric criteria for reactivity in HLD subfamilies, expression analysis of the set of 20 putative HLDs, far-UV circular dichroism spectra of HLDs, temperature profiles, pH profiles, the set of 30 halogenated substrates, kinetic resolution of 2-bromopentane, and kinetic resolution of ethyl-2-bromopropionate (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: zbynek@chemi.muni.cz. Tel: +420-5-4949-6667.
*E-mail: jiri@chemi.muni.cz. Tel: +420-5-4949-3467.

ORCID

Jiri Damborsky: 0000-0002-7848-8216

Author Contributions

‖P.V. and E.S. contributed equally.

Notes

The authors declare the following competing financial interest(s): Dr. Veronika Stepankova, Dr. Radka Chaloupkova and Dr. Zbynek Prokop work part-time at the University biotechnology spin-off company Enantis Ltd. Enzymes isolated throughout this study will be produced in sufficient quantities and distributed upon request to the community.

ACKNOWLEDGMENTS

The work was supported by the Grant Agency of the Czech Republic (GA16-06096S, GA16-07965S, and GA16-24223S), the Technological Agency of the Czech Republic (TH02010219), the Ministry of Education, Youth, and Sports of the Czech Republic (LQ1605 and LO1214) and the European Union (720776 and 722610). Computational resources were provided by CESNET (LM2015042) and the CERIT Scientific Cloud (LM2015085), as part of the "Projects of Large Research, Development, and Innovations Infrastructures" program (LM2015051, CZ.02.1.01/0.0/0.0/16_013/0001761, LM2015047, LM2015055).

REFERENCES

(1) The UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204−D212.10.1093/nar/gku989

(2) Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput. Biol.* **2009**, *5*, e1000605.

(3) Check Hayden, E. The $1,000 Genome 2006. *Nature* **2014**, *507*, 294−295.

(4) Khosla, C. Quo Vadis, Enzymology? *Nat. Chem. Biol.* **2015**, *11*, 438−441.

(5) Furnham, N.; Garavelli, J. S.; Apweiler, R.; Thornton, J. M. Missing in Action: Enzyme Functional Annotations in Biological Databases. *Nat. Chem. Biol.* **2009**, *5*, 521−525.

(6) Bornscheuer, U. T. Protein Engineering: Beating the Odds. *Nat. Chem. Biol.* **2016**, *12*, 54−55.

(7) Colin, P.-Y.; Kintses, B.; Gielen, F.; Miton, C. M.; Fischer, G.; Mohamed, M. F.; Hyvönen, M.; Morgavi, D. P.; Janssen, D. B.; Hollfelder, F. Ultrahigh-Throughput Discovery of Promiscuous Enzymes by Picodroplet Functional Metagenomics. *Nat. Commun.* **2015**, *6*, 10008.

(8) Chen, B.; Lim, S.; Kannan, A.; Alford, S. C.; Sunden, F.; Herschlag, D.; Dimov, I. K.; Baer, T. M.; Cochran, J. R. High-Throughput Analysis and Protein Engineering Using Microcapillary Arrays. *Nat. Chem. Biol.* **2016**, *12*, 76−81.

(9) Dörr, M.; Fibinger, M. P. C.; Last, D.; Schmidt, S.; Santos-Aberturas, J.; Böttcher, D.; Hummel, A.; Vickers, C.; Voss, M.; Bornscheuer, U. T. Fully Automatized High-Throughput Enzyme Library Screening Using a Robotic Platform. *Biotechnol. Bioeng.* **2016**, *113*, 1421−1432.

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

(10) Mak, W. S.; Tran, S.; Marcheschi, R.; Bertolani, S.; Thompson, J.; Baker, D.; Liao, J. C.; Siegel, J. B. Integrative Genomic Mining for Enzyme Function to Enable Engineering of a Non-Natural Biosynthetic Pathway. *Nat. Commun.* **2015**, *6*, 1−9.
(11) Lobb, B.; Doxey, A. C. Novel Function Discovery through Sequence and Structural Data Mining. *Curr. Opin. Struct. Biol.* **2016**, *38*, 53−61.
(12) Studer, R. A.; Dessailly, B. H.; Orengo, C. A. Residue Mutations and Their Impact on Protein Structure and Function: Detecting Beneficial and Pathogenic Changes. *Biochem. J.* **2013**, *449*, 581−594.
(13) Jesenska, A.; Pavlova, M.; Strouhal, M.; Chaloupkova, R.; Tesinska, I.; Monincova, M.; Prokop, Z.; Bartos, M.; Pavlik, I.; Rychlik, I.; Möbius, P.; Nagata, Y.; Damborsky, J. Cloning, Biochemical Properties, and Distribution of Mycobacterial Haloalkane Dehalogenases. *Appl. Environ. Microbiol.* **2005**, *71*, 6736−6745.
(14) Sato, Y.; Monincova, M.; Chaloupkova, R.; Prokop, Z.; Ohtsubo, Y.; Minamisawa, K.; Tsuda, M.; Damborsky, J.; Nagata, Y. Two Rhizobial Strains, *Mesorhizobium loti* MAFF303099 and *Bradyrhizobium japonicum* USDA110, Encode Haloalkane Dehalogenases with Novel Structures and Substrate Specificities. *Appl. Environ. Microbiol.* **2005**, *71*, 4372−4379.
(15) Jesenska, A.; Bartos, M.; Czernekova, V.; Rychlik, I.; Pavlik, I.; Damborsky, J. Cloning and Expression of the Haloalkane Dehalogenase Gene *dhmA* from *Mycobacterium avium* N85 and Preliminary Characterization of DhmA. *Appl. Environ. Microbiol.* **2002**, *68*, 3724−3730.
(16) Jesenska, A.; Monincova, M.; Koudelakova, T.; Hasan, K.; Chaloupkova, R.; Prokop, Z.; Geerlof, A.; Damborsky, J. Biochemical Characterization of Haloalkane Dehalogenases DrbA and DmbC, Representatives of a Novel Subfamily. *Appl. Environ. Microbiol.* **2009**, *75*, 5157−5160.
(17) Li, A.; Shao, Z. Biochemical Characterization of a Haloalkane Dehalogenase DadB from *Alcanivorax dieselolei* B-5. *PLoS One* **2014**, *9*, e89144−89153.
(18) Hesseler, M.; Bogdanović, X.; Hidalgo, A.; Berenguer, J.; Palm, G. J.; Hinrichs, W.; Bornscheuer, U. T. Cloning. Functional Expression, Biochemical Characterization, and Structural Analysis of a Haloalkane Dehalogenase from *Plesiocystis pacifica* SIR-1. *Appl. Microbiol. Biotechnol.* **2011**, *91*, 1049−1060.
(19) Hasan, K.; Fortova, A.; Koudelakova, T.; Chaloupkova, R.; Ishitsuka, M.; Nagata, Y.; Damborsky, J.; Prokop, Z. Biochemical Characteristics of the Novel Haloalkane Dehalogenase DatA, Isolated from the Plant Pathogen *Agrobacterium tumefaciens* C58. *Appl. Environ. Microbiol.* **2011**, *77*, 1881−1884.
(20) Fung, H. K. H.; Gadd, M. S.; Drury, T. a.; Cheung, S.; Guss, J. M.; Coleman, N. V.; Matthews, J. M. Biochemical and Biophysical Characterisation of Haloalkane Dehalogenases DmrA and DmrB in *Mycobacterium* Strain JS60 and Their Role in Growth on Haloalkanes. *Mol. Microbiol.* **2015**, *97*, 439−453.
(21) Fortova, A.; Sebestova, E.; Stepankova, V.; Koudelakova, T.; Palkova, L.; Damborsky, J.; Chaloupkova, R. DspA from *Strongylocentrotus purpuratus*: The First Biochemically Characterized Haloalkane Dehalogenase of Non-Microbial Origin. *Biochimie* **2013**, *95*, 2091−2096.
(22) Prokop, Z.; Sato, Y.; Brezovsky, J.; Mozga, T.; Chaloupkova, R.; Koudelakova, T.; Jerabek, P.; Stepankova, V.; Natsume, R.; Van Leeuwen, J. G. E.; Janssen, D. B.; Florian, J.; Nagata, Y.; Senda, T.; Damborsky, J. Enantioselectivity of Haloalkane Dehalogenases and Its Modulation by Surface Loop Engineering. *Angew. Chem.* **2010**, *122*, 6247−6251.
(23) Schober, M.; Faber, K. Inverting Hydrolases and Their Use in Enantioconvergent Biotransformations. *Trends Biotechnol.* **2013**, *31*, 468−478.
(24) Swanson, P. E. Dehalogenases Applied to Industrial-Scale Biocatalysis. *Curr. Opin. Biotechnol.* **1999**, *10*, 365−369.
(25) Dvorak, P.; Bidmanova, S.; Damborsky, J.; Prokop, Z. Immobilized Synthetic Pathway for Biodegradation of Toxic Recalcitrant Pollutant 1,2,3-Trichloropropane. *Environ. Sci. Technol.* **2014**, *48*, 6859−6866.
(26) Dravis, B. C.; Lejeune, K. E.; Hetro, A. D.; Russell, A. J. Enzymatic Dehalogenation of Gas Phase Substrates with Haloalkane Dehalogenase. *Biotechnol. Bioeng.* **2000**, *69*, 235−241.
(27) Lal, R.; Pandey, G.; Sharma, P.; Kumari, K.; Malhotra, S.; Pandey, R.; Raina, V.; Kohler, H. E.; Holliger, C.; Jackson, C.; Oakeshott, J. G. Biochemistry of Microbial Degradation of Hexachlorocyclohexane and Prospects for Bioremediation. *Microbiol. Mol. Biol. Rev.* **2010**, *74*, 58−80.
(28) Prokop, Z.; Oplustil, F.; DeFrank, J.; Damborsky, J. Enzymes Fight Chemical Weapons. *Biotechnol. J.* **2006**, *1*, 1370−1380.
(29) Bidmanova, S.; Steiner, M.-S.; Stepan, M.; Vymazalova, K.; Gruber, M. A.; Duerkop, A.; Damborsky, J.; Prokop, Z.; Wolfbeis, O. S. Enzyme-Based Test Strips for Visual or Photographic Detection and Quantitation of Gaseous Sulfur Mustard. *Anal. Chem.* **2016**, *88*, 6044−6049.
(30) Bidmanova, S.; Chaloupkova, R.; Damborsky, J.; Prokop, Z. Development of an Enzymatic Fiber-Optic Biosensor for Detection of Halogenated Hydrocarbons. *Anal. Bioanal. Chem.* **2010**, *398*, 1891−1898.
(31) Bidmanova, S.; Kotlanova, M.; Rataj, T.; Damborsky, J.; Trtilek, M.; Prokop, Z. Fluorescence-Based Biosensor for Monitoring of Environmental Pollutants: From Concept to Field Application. *Biosens. Bioelectron.* **2016**, *84*, 97−105.
(32) Campbell, D. W.; Müller, C.; Reardon, K. F. Development of a Fiber Optic Enzymatic Biosensor for 1,2-Dichloroethane. *Biotechnol. Lett.* **2006**, *28*, 883−887.
(33) Los, G. V.; Encell, L. P.; Mcdougall, M. G.; Hartzell, D. D.; Karassina, N.; Zimprich, C.; Wood, M. G.; Learish, R.; Ohana, R. F.; Urh, M.; Simpson, D.; Mendez, J.; Zimmerman, K.; Otto, P.; Vidugiris, G.; Zhu, J.; Darzins, A.; Klaubert, D. H.; Bulleit, R. F.; Wood, K. V. HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis. *ACS Chem. Biol.* **2008**, *3*, 373−382.
(34) Ohana, R. F.; Encell, L. P.; Zhao, K.; Simpson, D.; Slater, M. R.; Urh, M.; Wood, K. V. HaloTag7: A Genetically Engineered Tag That Enhances Bacterial Expression of Soluble Proteins and Improves Protein Purification. *Protein Expression Purif.* **2009**, *68*, 110−120.
(35) Koudelakova, T.; Bidmanova, S.; Dvorak, P.; Pavelka, A.; Chaloupkova, R.; Prokop, Z.; Damborsky, J. Haloalkane Dehalogenases: Biotechnological Applications. *Biotechnol. J.* **2013**, *8*, 32−45.
(36) Chovancova, E.; Kosinski, J.; Bujnicki, J. M.; Damborsky, J. Phylogenetic Analysis of Haloalkane Dehalogenases. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 305−316.
(37) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(38) N.R. Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2013**, *41*, D8−D20.
(39) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
(40) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(41) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins with Structural and Topographical Mapping of Functionally Annotated Residues. *Nucleic Acids Res.* **2006**, *34*, W116−W118.
(42) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884−1897.
(43) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **2012**, *8*, e1002708.
(44) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412

Chemical Editor, Visualization, and Analysis Platform. *J. Cheminf.* **2012**, *4*, 17.
(45) Solis, F. J.; Wets, R. J.-B. Mnimization by Random Search Techniques. *Math. Oper. Res.* **1981**, *6*, 19−30.
(46) Harrison, R. G. Expression of Soluble Heterologous Proteins via Fusion with NusA Protein. *InNovations* **2000**, *11*, 4−7.
(47) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in *Escherichia coli*. *Nat. Biotechnol.* **1991**, *9*, 443−448.
(48) Woody, R. W. *Circular Dichroism and the Conformational Analysis of Biomolecules*, 1st ed.; Fasman, G. D., Ed.; Plenum Press: New York, 1996.
(49) Gräslund, S.; Nordlund, P.; Weigelt, J.; Hallberg, B. M.; Bray, J.; Gileadi, O.; Knapp, S.; Oppermann, U.; Arrowsmith, C.; Hui, R.; Ming, J.; et al. Protein Production and Purification. *Nat. Methods* **2008**, *5*, 135−146.
(50) Braun, P.; LaBaer, J. High Throughput Protein Production for Functional Proteomics. *Trends Biotechnol.* **2003**, *21*, 383−388.
(51) Pacheco, B.; Crombet, L.; Loppnau, P.; Cossar, D. A Screening Strategy for Heterologous Protein Expression in *Escherichia coli* with the Highest Return of Investment. *Protein Expression Purif.* **2012**, *81*, 33−41.
(52) Bastard, K.; Smith, A. A. T.; Vergne-Vaxelaire, C.; Perret, A.; Zaparucha, A.; De Melo-Minardi, R.; Mariage, A.; Boutard, M.; Debard, A.; Lechaplais, C.; Pelle, C.; Pellouin, V.; Perchat, N.; Petit, J.-L.; Kreimeyer, A.; Medigue, C.; Weissenbach, J.; Artiguenave, F.; De Berardinis, V.; Vallenet, D.; Salanoubat, M. Revealing the Hidden Functional Diversity of an Enzyme Family. *Nat. Chem. Biol.* **2014**, *10*, 42−49.

DOI: 10.1021/acscatal.7b03523
ACS Catal. 2018, 8, 2402−2412