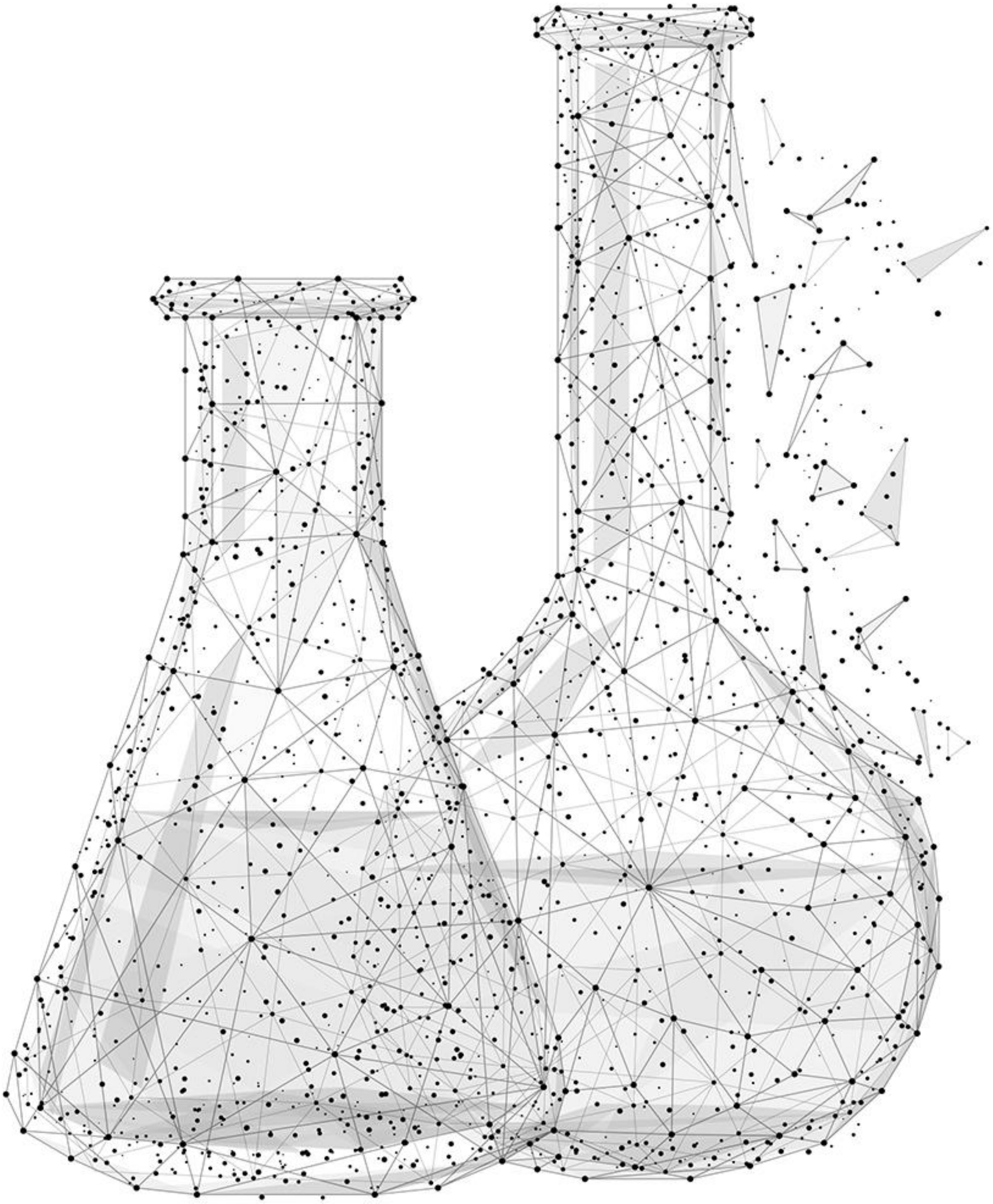


BIOTECHNOLOGY and BIOENGINEERING

Characterizing and Predicting Carboxylic Acid Reductase Activity for Diversifying Bioaldehyde Production

Matthew Moura, Dante Pertusi, Stephen Lenzini, Namita Bhan, Linda J. Broadbelt, Keith E.J. Tyo

Department of Chemical and Biological Engineering, Northwestern University, Evanston 60208, Illinois; telephone: 847-491-5351; fax: 847-491-3728; e-mail: broadbelt@northwestern.edu; telephone: 847-868-0319; fax: 847-491-4089; e-mail: k-tyo@northwestern.edu

ABSTRACT: Chemicals with aldehyde moieties are useful in the synthesis of polymerization reagents, pharmaceuticals, pesticides, flavors, and fragrances because of their high reactivity. However, chemical synthesis of aldehydes from carboxylic acids has unfavorable thermodynamics and limited specificity. Enzymatically catalyzed reductive bioaldehyde synthesis is an attractive route that overcomes unfavorable thermodynamics by ATP hydrolysis in ambient, aqueous conditions. Carboxylic acid reductases (Cars) are particularly attractive, as only one enzyme is required. We sought to increase the knowledge base of permitted substrates for four Cars. Additionally, the Lys2 enzyme family was found to be mechanistically the same as Cars and two isozymes were also tested. Our results show that Cars prefer molecules where the carboxylic acid is the only polar/charged group. Using this data and other published data, we develop a support vector classifier (SVC) for predicting Car reactivity and make predictions on all carboxylic acid metabolites in iAF1260 and Model SEED.

Biotechnol. Bioeng. 2016;113: 944–952.

KEYWORDS: carboxylic acid reductase; reductive aldehyde synthesis; enzyme promiscuity; support vector machines

Introduction

Enzymes have seen wide use as industrial catalysts for stereoisomer purification (Breuer et al., 2004), selective oxidation (Patel et al., 2003), reductive amination (Zhang et al., 2010), and other applications (Patel, 2011). In some cases, these advancements rely on substrate promiscuity, the capability for enzymes to act against non-native substrates (Hult and Berglund, 2007; Khersonsky and Tawfik, 2010; Nobeli et al., 2009). Thus far, these exploits of substrate promiscuity have relied on classes of enzymes with well-documented and highly promiscuous reactions, particularly alcohol dehydrogenases and amino-transferases. To expand the portfolio of biochemical reactions available for chemical production, we will need to expand our knowledge of enzymatic promiscuity in other enzyme classes, as this will reveal new biochemical reactions that can be implemented.

Aldehydes are a class of molecules with industrially relevant characteristics not often explored in the context of biochemistry. This moiety is more reactive than related alcohol and carboxylic acid structures and carries utility for polymerization reagents, reactive precursors for pharmaceuticals and pesticides, and as flavoring or scent additives (Kohlpaintner et al., 2013). Aldehydes are also a necessary intermediate in the conversion of fatty acids to alcohols and alkanes for enhancing biofuel production (Akhtar et al., 2013; Rodriguez and Atsumi, 2014).

While the oxidation of alcohols to aldehydes is relatively easy from a biochemical standpoint (many alcohol dehydrogenases demonstrate reversible activity), the reduction of a carboxylic acid to an aldehyde is much more difficult. The reaction is thermodynamically unfavorable and difficult to accomplish through organic chemistry methods (Addis et al., 2011; Bézier et al., 2013). Methods are plagued by a need for high temperatures and pressures, poor tolerance for adjacent functional groups, or difficulties in preventing further reduction to alcohols.

Currently, there have been few specific enzymes found to reduce a carboxylic acid to an aldehyde. Nature typically relies on a two enzyme reaction: first an enzyme thiolates the carboxylate onto a CoA, followed by a second enzyme reducing the molecule to an aldehyde releasing the CoA. Because of the need for two enzymes and available concentrations of the CoA intermediate, this process requires pathway/expression optimization to ensure maximal yields of aldehyde from acid and will suffer losses from substrate diffusion and consumption of intermediates by other enzymatic reactions.

Alternatively, there have been two cases of single enzymes capable of this reduction that have been heterologously expressed and characterized. (Akhtar et al., 2013; Venkitasubramanian et al., 2006).

Conflicts of interest: The authors report no conflicts of interest with this work.
Correspondence to: M. Moura
Contract grant sponsor: Bill and Melinda Gates Foundation, NSF
Contract grant number: CBET-0835800
Contract grant sponsor: Northwestern McCormick School of Engineering
Received 15 June 2015; Revision received 11 October 2015; Accepted 14 October 2015
Accepted manuscript online 19 October 2015;
Article first published online 10 November 2015 in Wiley Online Library
(http://onlinelibrary.wiley.com/doi/10.1002/bit.25860/abstract).
DOI 10.1002/bit.25860

Prior work suggests these enzymes, commonly called carboxylic acid reductases (Cars), can take a wide range of substrates. (Venkitasubramanian et al., 2006, 2007, 2008) A recent review showed that the enzymatic capabilities for this specific transformation are large, but evidence has predominantly been shown with whole-cell catalysis thus far which would not distinguish between single and dual enzyme systems. (Napora-Wijata et al., 2014) A more precisely defined substrate profile for known single-enzyme catalysts would have relevance for metabolic engineering applications.

The mechanism is similar to the two-enzyme system previously mentioned, but with two catalytic domains on a single enzyme. As can be seen in Figure 1A, the two domains interact via a phosphopantetheine (PPT) post-translational modification (Venkitasubramanian et al., 2007). The proposed mechanism is: (i) the adenylation domain forms an unstable acid-AMP adenylate which (ii) reacts with the PPT domain. The PPT domain prevents diffusion of the metabolite away from the Car, and (iii) carries the metabolite to the reduction domain, where (iv) NADPH is used to reductively cleave and release the aldehyde. A literature search for similar enzymes found the Lys2 family of enzymes to use the same catalytic mechanism despite not sharing the "Car" nomenclature used for the two confirmed enzymes described as Cars in prior literature. (Ehmann et al., 1999).

Because of the utility of aldehydes, we sought to better characterize the possible substrates that can be acted on by different Cars. We tested four Cars, as well as two Lys2 enzymes, against a set of substrates covering a range of chemical characteristics. Initial rates were captured through absorbance monitoring of NADPH oxidation, while GC/MS methods were developed and optimized for the reaction conditions to detect trace-level catalysis. This data was used to predict additional substrates from large metabolite databases using machine learning methods.

Materials and Methods

Identifying Enzymatic Candidates

The amino acid sequence for *Nocardia iowensis* Car (NICar) was used to find homologues using BlastP (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The search identified several enzymes, and we selected enzymes from *Nocardia brasiliensis* (NBCar) and *Mycobacterium smegmatis* (MSCar) for further study. Because all enzymes require phosphopantetheination, the *N. iowensis* PPTase sequence was similarly used to find PPTase sequences in the respective organisms for NBCar and MSCar. The fourth Car tested, from *Mycobacterium marinum* (MMCar), was graciously

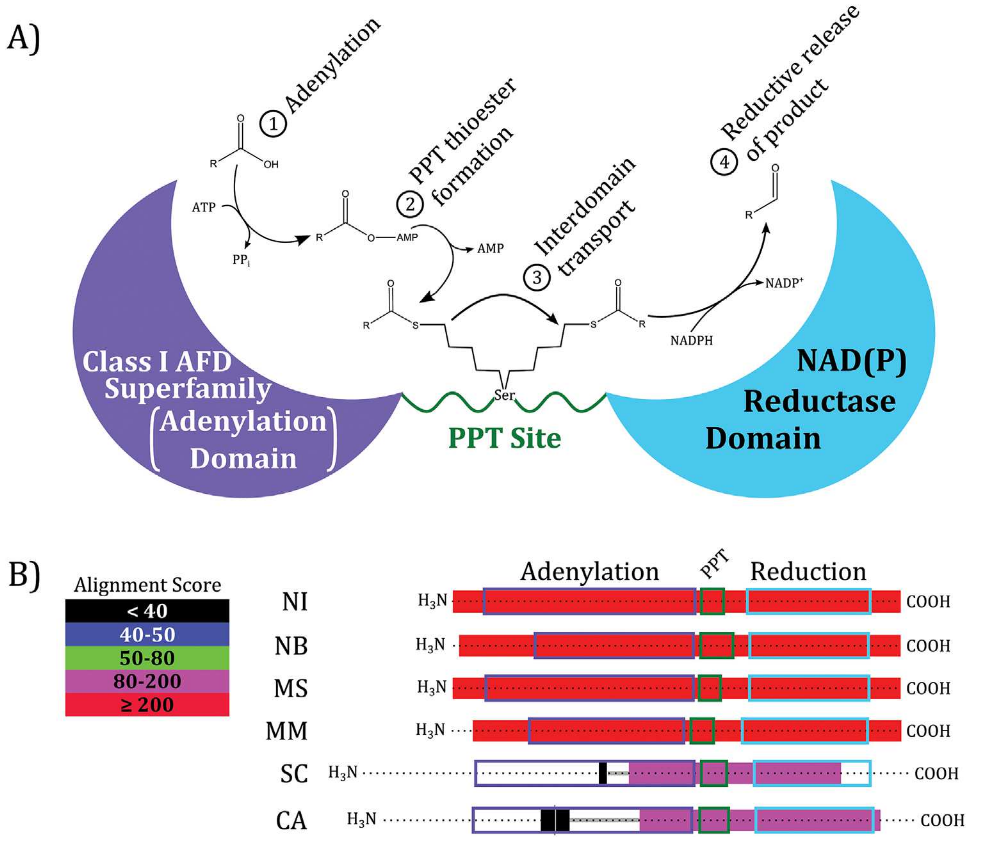


Figure 1. (**A**) Proposed catalytic mechanism of Car, Lys2 enzymes. (**B**) Sequence alignment of six carboxylic acid reducing enzymes. Using NICar (NI) as the query sequence, the remaining five were aligned in the following order moving downwards: NBCar (NB), MSCar (MS), MMCar (MM), SCLys2 (SC), and CALys2 (CA). Filled in boxes correspond to alignment scores, and hollow boxes correspond to the matching colored domains in part A.

provided as an intact construct, with an active PPTase, by the Jones lab at Imperial College London. Lys2 enzymes were selected from *Saccharomyces cerevisiae* (SCLys2) and *Candida albicans* (CALys2), and corresponding Lys5 PPTase sequences were used from the respective organisms.

Plasmid and Strain Construction

All isozymes other than MMCar were codon-optimized with biases blended between *Escherichia coli* and *Saccharomyces cerevisiae* frequencies to allow for potential expression in both hosts. The sequences were also designed with a 6 × N-His sequence. NICar, NBCar, and MSCar were all synthesized by IDT (Coralville, IA). SCLys2 and CALys2 were assembled from LifeTech (Grand Island, NY) GeneArt Strings. The final constructs were cloned with Gibson assembly with the respective PPTase/Lys5 sequences, into a pET21a construct. DH5α *E. coli* was used for plasmid assembly, while BL21 (DE3) was used for protein expression. Gene sequences and primers used for the assembly are provided in Supplementary File 2.

Protein Expression and Purification

All expression strains were grown to an OD of ≈0.6 in TB media at 37°C. Expression was induced with 1 mM IPTG for 4.5 h at temperatures corresponding to growth conditions of the native organisms: 25°C for NICar, NBCar, and CALys2, 30°C for MMCar and SCLys2, and 37°C for MSCar. After induction, cells were spun down, frozen, and stored at −20°C. For purification, cell pellets were thawed for 1 h on ice and then suspended in 4 × lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, 1 mg/mL lysozyme, 1:200 EDTA-free protease inhibitor cocktail (CalBioTech, Spring Valley CA)). Suspensions were sonicated in an ice-water bath at 50 Amps for 6 pulses of 10 s, with 30 s cool-down periods. Soluble fractions were separated by centrifugation at 12,000 *g* for 30 min at 4°C and run through Ni+ spin columns (GE Life Sciences, Pittsburgh PA). Columns were washed three times (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole), and eluted twice (50 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole). Amicon 50 kDa MW cutoff centrifugation filters (Millipore, Billerica, MA) were used for buffer exchange (50 mM Tris–HCl, 1 mM THP, 10 mM $MgCl_2$, 10% glycerol). SDS–PAGE was used to confirm protein expression and purification. Final protein concentrations were measured with the Bradford assay (BioRad, Hercules, CA). For Car and Lys2 studies, benzoic acid and α-amionadipate rates, respectively, were compared to confirm enzymatic consistency across experiments.

To account for discrepancies from differences in enzymatic purity, purified protein solutions were run on an SDS–PAGE. For those values that later showed confirmed activity through GC/MS analysis, initial rates were adjusted for the relative purity levels of the varying purified stocks as described in Supplementary 1B and indicated in Figure 2.

Reaction Buffers

Reaction buffers contained 50 mM Tris, 10 mM $MgCl_2$, 1 mM THP, 10% glycerol, 1 mM ATP, 1 mM NADPH, and 5 mM substrate. The substrates tested are detailed in Table I. Buffers were all pH adjusted to 7.5 with HCl after mixing all components, filtered, and stored at −20°C. All chemicals other than THP (Millipore, Billerica, MA), NADPH (Roche, Basel, Switzerland), and EDBA (ethyl-diamine dibutyric acid) were purchased from Sigma (St. Louis, MO). EDBA was synthesized in the Northwestern Center for Molecular Innovation and Drug Discovery (CMIDD) (Evanston, IL).

NADPH Oxidation Rate Assay

To measure in vitro kinetics, half-area UV transparent 96 well plates were used (Corning, Corning, NY) in a plate reader. Reaction volumes consisted of 150 μL of reaction buffer plus 10 μg of the respective proteins (or equimolar amounts of BSA). NADPH consumption was monitored at $\lambda = 340$ nm overnight in three temperature conditions: 25°C, 30°C, and 37°C. Initial linear rates were estimated using least-squares regression with 95% confidence F-statistic criteria (Supplementary 1A). Two separate conditions were used to control for background NADPH oxidation not resulting from carboxylic acid reduction. To capture any non-enzymatic NADPH oxidation, a BSA control was used. To capture any potential background oxidative acitivy of the Cars, a No Substrate (Sub^-) control was used with no acid substrates in the reaction solution. Rate data was analyzed using the Multiexperiment Viewer (MeV) (Howe et al., 2010), specifically heirarchical clustering (Eisen et al., 1998) data visualization in Figure 2.

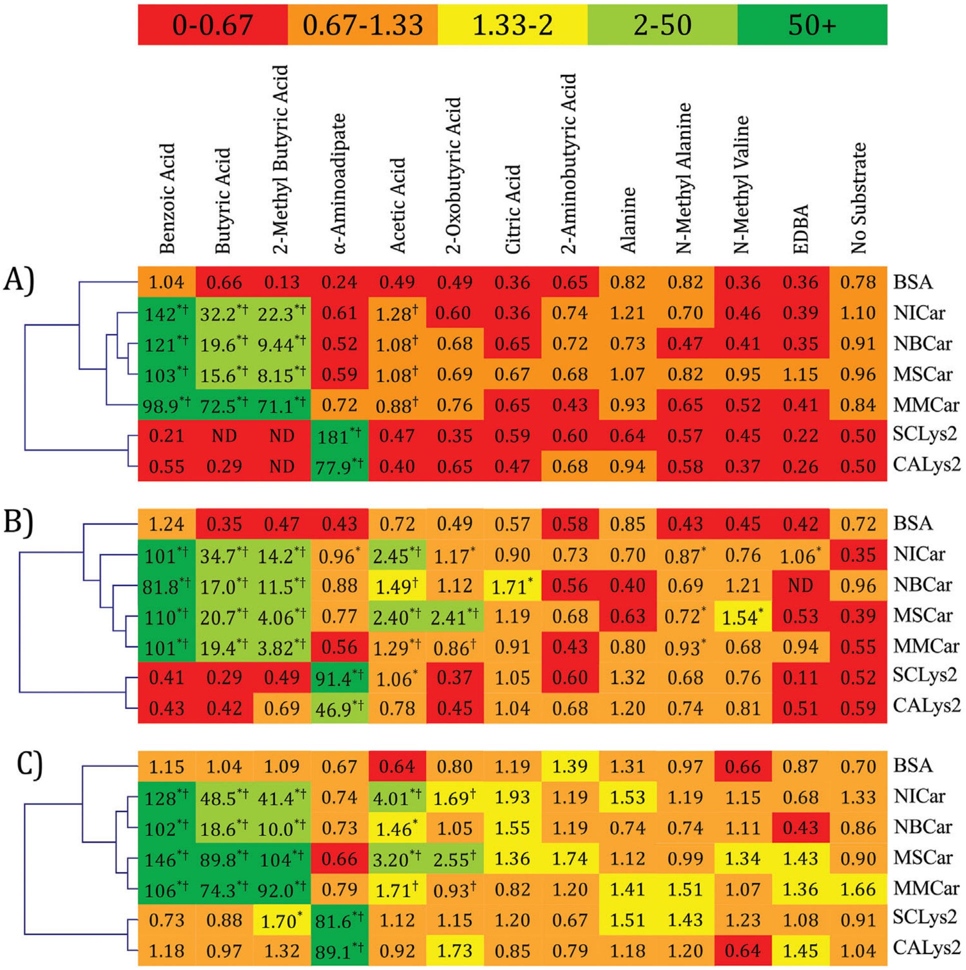
Product Confirmation Assay

Following overnight reactions in sealed microcentrifuge tubes, 0.1 M methoxyamine-HCl was added and allowed to react for 15 min at the assay temperature. Samples were then flash-frozen and stored at −80°C until GC/MS sample preparation.

If the expected product contained no reactive hydrogens (benzoic acid, butyric acid, 2-methylbutyric acid, and 2-oxobutyric acid), the solution was extracted into an equivolume of ethyl acetate and analyzed by GC/MS. The acetic acid reaction was prepared following this procedure, except butyl acetate was used instead of ethyl acetate to extract, because ethyl acetate co-eluted from the GC with the product. For the remaining compounds, *n*-butanol was used for extraction. The butanol phase was separated from the aqueous phase and desiccated over 5–6 h under vacuum at room temperature in a Vacuufuge (Thermo, Waltham, MA). The desiccated remnants were mixed into 10 μL of pyridine, transferred into autosampler vials containing 90 μL of MSTFA + 1% TMCS, and reacted for 45 min at 60°C.

For all samples, gas chromatography was run on an Agilent 7890 GC with an HP-5MS-UI column (Agilent, Santa Clara, CA) with $T_{\text{inlet}} = 250°C$, splitless injections, and a 10°C/min temperature ramp from 80°C to 325°C. Mass spectrometry was conducted in an Agilent 7000 QQQ in scan mode using Positive Chemical Ionization (PCI) with methane as a reagent gas. Products were searched for based on the expected + 1, + 29, and + 41 adducts that are characteristic of Methane-PCI (Agilent, 2014).

Computational methods for the machine learning framework used to predict additional substrates for Car are described in Supplementary 1D.



A)

	Benzoic Acid	Butyric Acid	2-Methyl Butyric Acid	α-Aminoadipate	Acetic Acid	2-Oxobutyric Acid	Citric Acid	2-Aminobutyric Acid	Alanine	N-Methyl Alanine	N-Methyl Valine	EDBA	No Substrate
BSA	1.04	0.66	0.13	0.24	0.49	0.49	0.36	0.65	0.82	0.82	0.36	0.36	0.78
NICar	142*†	32.2*†	22.3*†	0.61	1.28†	0.60	0.36	0.74	1.21	0.70	0.46	0.39	1.10
NBCar	121*†	19.6*†	9.44*†	0.52	1.08†	0.68	0.65	0.72	0.73	0.47	0.41	0.35	0.91
MSCar	103*†	15.6*†	8.15*†	0.59	1.08†	0.69	0.67	0.68	1.07	0.82	0.95	1.15	0.96
MMCar	98.9*†	72.5*†	71.1*†	0.72	0.88†	0.76	0.65	0.43	0.93	0.65	0.52	0.41	0.84
SCLys2	0.21	ND	ND	181*†	0.47	0.35	0.59	0.60	0.64	0.57	0.45	0.22	0.50
CALys2	0.55	0.29	ND	77.9*†	0.40	0.65	0.47	0.68	0.94	0.58	0.37	0.26	0.50

B)

	Benzoic Acid	Butyric Acid	2-Methyl Butyric Acid	α-Aminoadipate	Acetic Acid	2-Oxobutyric Acid	Citric Acid	2-Aminobutyric Acid	Alanine	N-Methyl Alanine	N-Methyl Valine	EDBA	No Substrate
BSA	1.24	0.35	0.47	0.43	0.72	0.49	0.57	0.58	0.85	0.43	0.45	0.42	0.72
NICar	101*†	34.7*†	14.2*†	0.96*	2.45*†	1.17*	0.90	0.73	0.70	0.87*	0.76	1.06*	0.35
NBCar	81.8*†	17.0*†	11.5*†	0.88	1.49†	1.12	1.71*	0.56	0.40	0.69	1.21	ND	0.96
MSCar	110*†	20.7*†	4.06*†	0.77	2.40*†	2.41*†	1.19	0.68	0.63	0.72*	1.54*	0.53	0.39
MMCar	101*†	19.4*†	3.82*†	0.56	1.29*†	0.86†	0.91	0.43	0.80	0.93*	0.68	0.94	0.55
SCLys2	0.41	0.29	0.49	91.4*†	1.06*	0.37	1.05	0.60	1.32	0.68	0.76	0.11	0.52
CALys2	0.43	0.42	0.69	46.9*†	0.78	0.45	1.04	0.68	1.20	0.74	0.81	0.51	0.59

C)

	Benzoic Acid	Butyric Acid	2-Methyl Butyric Acid	α-Aminoadipate	Acetic Acid	2-Oxobutyric Acid	Citric Acid	2-Aminobutyric Acid	Alanine	N-Methyl Alanine	N-Methyl Valine	EDBA	No Substrate
BSA	1.15	1.04	1.09	0.67	0.64	0.80	1.19	1.39	1.31	0.97	0.66	0.87	0.70
NICar	128*†	48.5*†	41.4*†	0.74	4.01*†	1.69†	1.93	1.19	1.53	1.19	1.15	0.68	1.33
NBCar	102*†	18.6*†	10.0*†	0.73	1.46*	1.05	1.55	1.19	0.74	0.74	1.11	0.43	0.86
MSCar	146*†	89.8*†	104*†	0.66	3.20*†	2.55†	1.36	1.74	1.12	0.99	1.34	1.43	0.90
MMCar	106*†	74.3*†	92.0*†	0.79	1.71†	0.93†	0.82	1.20	1.41	1.51	1.07	1.36	1.66
SCLys2	0.73	0.88	1.70*	81.6*†	1.12	1.15	1.20	0.67	1.51	1.43	1.23	1.08	0.91
CALys2	1.18	0.97	1.32	89.1*†	0.92	1.73	0.85	0.79	1.18	1.20	0.64	1.45	1.04

Figure 2. Tabulated initial rates at A: 25°C, B: 30°C, and C: 37°C. Rates are mM_{NADPH} oxidized per hour. Hierarchical clustering is shown on the left of each heat map. *Rates have a $P < 0.05$ (*t*-test) compared to both the BSA (no Car) and no substrate negative controls. †Rates have been modified to account for enzyme purity, based on confirmed products as shown in Table II.

Results

Sequence Analysis

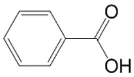
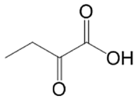
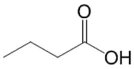
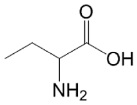
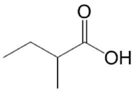
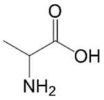
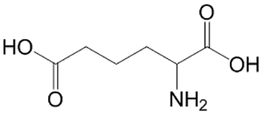
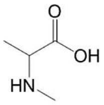
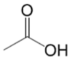
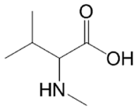
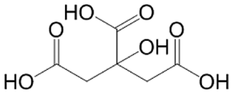
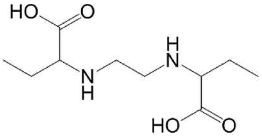
At the beginning of the study, NICar was the only characterized Car, and the protein sequence alignment using BlastP (Altschul et al., 1990) was used to identify additional enzyme candidates to analyze for substrate promiscuity. NBCar and MSCar were chosen for their high identity scores as enzymes of interest, 75% and 64%, respectively. MMCar was reported during the course of this work (Akhtar et al., 2013), and was incorporated into the analysis. Lastly, SCLys2 and CALys2 were also included as comparative examples of enzymes with the same catalytic mechanism, but with different sequences.

Sequence comparison revealed interesting aspects of the enzymes, which can be seen in Figure 1B. All enzymes show a similar domain architecture of a Class I Adenylate Forming Domain (AFDI) connected to a second domain classified as a NAD(P)H reductase sequence with a Rossmann Fold (Rao and Rossmann, 1973) with a PPT attachment site in between. While all Cars aligned well (scores ≥ 200), the Lys2 enzymes were sequentially distinct. The majority of the AFDI domains showed little to no alignment between the two enzyme families, but the remaining portions of the sequences did show good scoring, including predicted "substrate binding residues" in the NCBI's CDD predictions (Marchler-Bauer et al., 2014). Given the observed difference in substrate profiles, this would suggest the AFDI domains control enzyme specificity.

Substrate Profiles

The range of substrates tested against the enzymes composed a wide range of chemical structures and moieties in the α-position to the carboxylic acid moiety, shown in Table I. Given that a substrate preference for aromatic substrates has been seen by others (Venkitasubramanian et al., 2006), a range of moieties with varying different electrophilic properties α-positioned to the active site were tested to broaden the known substrate profile for Cars. To elucidate any temperature effects on the enzymes' activity profiles, substrates were tested at three different physiological temperatures: 25°C,

Table I. Chemical names and structures for compounds tested against all enzymes in this study.

Compound	Structure	Compound	Structure
Benzoic acid		2-oxobutyric acid	
Butyric acid		2-aminobutyric acid	
2-methyl butyric acid		Alanine	
α-aminoadipate		N-methyl alanine	
Acetic acid		N-methyl valine	
Citric acid		EDBA	

EDBA, Ethyldiamine-dibutyric acid.

30°C, and 37°C. It was hypothesized that promiscuous reactions may become more apparent at non-native temperatures to the enzymes' original species, as this has been shown to influence promiscuity (Hult and Berglund, 2007).

The substrates tested can be broken into three activity categories: High, Low, and Non-detected. High activity substates were active across all temperatures, low activity substrates showed thermo-sensitive promiscuity and oxidative rates that were difficult to distinguish from background oxidation of NADPH, and compounds that were unable to be confirmed as active substrates were considered non-detected. Previous studies on NICar and MMCar found that benzoic acid was a high-activity substrate (Akhtar et al., 2013; Venkitasubramanian et al., 2008), and this trend was confirmed both with the same two enzymes as well as the two uncharacterized ones. It was also found that the short chain organic acids demonstrated high initial rates for the Cars. The Lys2 enzymes only showed affinity for their native substrate, α-aminoadipate.

The only consistent low-activity substrate from the NADPH oxidation analysis was acetic acid. This activity only showed significance at the two higher temperatures and was seen

Table II. GC/MS confirmed reactions at all three respective temperatures.

	NICar			NBCar			MSCar			MMCar			SCLys2			CALys2		
	25°C	30°C	37°C	25°C	30°C	37°C	25°C	30°C	37°C	25°C	30°C	37°C	25°C	30°C	37°C	25°C	30°C	37°C
Benzoic acid	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—	—	—	—	—
Butyric acid	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—	—	—	—	—
2-methyl butyric acid	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—	—	—	—	—
α-aminoadipate	—	—	—	—	—	—	—	—	—	—	—	—	✓	✓	✓	✓	✓	✓
Acetic acid	✓[a]	✓	✓	✓[a]	✓[a]	—	✓[a]	✓	✓	✓[a]	✓	✓	—	—	—	—	—	—
Citric acid	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2-oxobutyric acid	—	—	✓[a]	—	—	—	—	✓[a]	✓[a]	—	✓[a]	✓[a]	—	—	—	—	—	—
2-aminobutyric acid	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Alanine	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
N-methyl alanine	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
N-methyl valine	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
EDBA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

[a] Reaction products detected in GC/MS analysis that were not apparent through NADPH oxidation assay.

consistently for the MS and NI Cars at both temperatures. There were additional cases of substrates with low, but significant activities, shown in Figure 2, though none were consistent across multiple temperatures as was observed for acetic acid.

In some cases, temperature had dramatic effects on the kinetics. MMCar showed a preference for broad catalysis outside of 30°C, though its activity towards benzoic acid was consistent across temperatures. Meanwhile MSCar's rates dramatically increased at 37°C for the organic acids as well as for benzoic acid. *Nocardia* enzymes demonstrated activity towards benzoic acid comparable to the *Mycobacterium* enzymes, but showed less preference for the organic acids, especially NBCar; yet NICar had the highest rate for acetic acid. Hierarchical clustering of enzymes based on the rate data was consistent with evolutionarily groupings only at 37°C (Fig. 2). This observation is most likely due to the higher promiscuous activity observed at 37°C. While approaches like this can provide an understanding of the promiscuous capabilities of enzymes (Beigi et al., 2014; Venkitasubramanian et al., 2006), an intensive kinetic analysis will require exploring activities over a range of substrate concentrations.

Substrates that showed little to no detectable activity largely all shared amines in the α position to the carboxylic acids. Alanine and 2-aminobutyric acid did not show significant catalysis over all three temperatures. The N-methyl amines, as well EDBA, citric acid, and 2-oxobutyrate, showed some low rates at 30°C, but these activities were not reproduced at 37°C. The native functions of Cars has not yet been identified, however, we can begin to hypothesize function based on the substrate profiles (see Discussion).

GC/MS Confirmation of Catalysis

To confirm the NADPH oxidation was a result of the expected carboxylic acid reduction, chemical analysis of products by GC/MS was used. In addition, the enhanced sensitivity of GC/MS identified additional catalysis not evident in the NADPH oxidation. Confirmed GC/MS detection is summarized in Table II, while chromatograms and spectra can be seen in Supplementary 1C. Benzaldehyde, butyraldehyde, 2-methylbutyraldehyde, acetaldehyde, and α-aminoadipate semialdehyde were all confirmed as methoximed derivatives. No aminated products were detected on the GC, agreeing in large part with the oxidation rate data. Although not detected by NADPH oxidation, a very low signal for a successful reduction of 2-oxobutyrate to 2-oxobutyraldehyde was detected as the expected $+1$ adduct for the di-methoxime product. This product was seen for the MS and MM Cars at 30°C, and for these two as well as NICar at 37°C (though not in negative controls). The higher promiscuous rates of the *Mycobacterium* Cars towards the short-chain organic acids support the higher promiscuity of these enzymes, as well as the validity of this putative reduction.

NICar Support Vector Classifier (SVC)

Because the synthesis of bioaldehydes may have broad application, it was potentially interesting to use the substrate data collected by others and us as training data to predict other NICar substrates. Two other studies profiled additional Car substrates and were used in combination with the data from this study (Li and Rosazza, 1997; Venkitasubramanian et al., 2006). We used an SVC for identifying putative substrates. By using a properly trained SVC, all carboxylic acids from large biochemical databases could be assessed as a potential substrate for NICar. We constructed both unweighted and weighted SVCs to account for the imbalance of aromatic and aliphatic carboxylic acids in the training set. To determine the predictive capability of our SVCs, both were subjected to 100 iterations of both three and five-fold stratified cross validation. Validation accuracy scores ranged from 77% to 78% while F_1 scores, the harmonic mean of the classifier's precision and recall values, ranged from 85% to 86% (see Supplementary 1D for calculations). We then used the SVC trained with the full dataset to classify carboxylic acids in the iAF1260 *E. coli* model (Feist et al., 2007) and Model SEED (Henry et al., 2010). Of the 304 compounds in iAF120, the unweighted SVC categorized 9% as positive, all of which were aromatic. The weighted SVC classified 14% as positive; of these, 48% were aromatic and the balance aliphatic. Of the 3,137 carboxylic acids in Model SEED, 16% were classified as positive by the unweighted classifier, which once again identified only aromatic carboxylic acids. The weighted SVC categorized 22% as positive, with a 50–50 split between aromatic and aliphatic hits. The predictions are included in Supplements 3–6 and can be viewed with MarvinView (ChemAxon) or the OpenBabel software package.

Discussion

Most oxygenated metabolites contain alcohol, keto, or carboxylic acid moieties with the diversity of metabolites with aldehydes being relatively low. The reactivity and associated toxicity of aldehydes likely resulted in evolutionary pressure against aldehydes within metabolism (Kunjapur and Prather, 2015; Langevin et al., 2011). In vivo, aldehydes can often be short lived as there are many enzymes that reduce (alcohol reductases) or oxidize (aldehyde dehydrogenases) aldehydes to alcohols and carboxylic acids, respectively. It has been shown in several studies that the removal of non-essential isozymes of these two classes dramatically increases the amount of aldehydes within the metabolome, demonstrating that there is potential for novel bioaldehyde production routes (Kunjapur et al., 2014; Rodriguez and Atsumi, 2014). From a production standpoint, toxicity and alcohol reductase activity may complicate biosynthesis of aldehydes. That said, Car enzymes have interesting utility and have many promising applications.

In this study, we characterized Cars from *N. brasiliensis*, *N. iowensis*, *M. marinum*, and *M. smegmatis*. Though this work focused on four isozymes, more than 100 proteins were found with identity scores ranging from 60 to 81% to NICar, predominantly from *Mycobacterium* and *Nocardia* species. It is interesting that these two genera are the main source of these enzymes as they are evolutionarily linked and were once even considered the same (McMurray, 1996).

Comparing the sequences of the Car enzymes to the Lys2 enzymes gives some clue towards where the enzyme's substrate preferences are localized. The lack of sequence similarity in the AFDI domain between the Car and Lys2 enzymes (compared to the relative similarity in the rest of the protein) suggests the AFDI domains are responsible for the substrate selectivity. From a metabolic perspective, it makes sense to exert control on the first step of the reaction mechanism to prevent wasting excess ATP within the first domain and the generation of partial products.

Though there are no crystal structures available to analyze, a simple analysis of the AFDI sequences' hydrophobicity shows that the Cars contain significantly more hydrophobic residues than the Lys2 sequences. Using the ExPASy ProtParam (Gasteiger et al., 2005), the averaged Grand Average of Hydropathy (GRAVY) scores, a measure of overall amino acid hydrophobicity, of the Cars and Lys2 AFDI's are -0.033 ± 0.019 vs. -0.301 ± 0.0055, respectively. The more positive value for the Cars may reflect their preference for nonpolar substrates, while the Lys2 enzymes' single substrate requires hydrophilic recognition of the amine and additional carboxylic acid residues.

Functionally, it is clear that the Car and Lys2 enzymes have two distinct substrate profiles. Lys2 showed no promiscuous activity towards the tested non-native substrates which, in the case of SCLys2, is in line with prior studies that showed a restricted substrate profile (Ehmann et al., 1999). Correlational studies have shown that high substrate specificity is more prevalent in essential enzymes than in non-essential ones (Nam et al., 2012). It is possible that the essentiality of Lys2 to many fungal species, as demonstrated through its common use as an auxotrophic marker in minimal media, has selected for high specificity (Xu et al., 2006).

In contrast to the Lys2 enzymes, the Car enzymes clearly show a preference for substrates of a very different character. Cars showed no activity for α-aminoadipate but high turnover for butyric acid, a substructure of α-aminoadipate. The Cars showed a poor (or nonexistent) activity towards substrates with charged residues adjacent to the acid site. Even the two low-activity substrates, acetic acid and 2-oxobutyric acid, contain either no charge or only a negative dipole. The success of 2-methyl butyric acid shows there is some steric allowance in the binding pockets of these enzymes. Our results and others (Akhtar et al., 2013) show that alkyl and aryl-acids are the preferred substrates. Enzymes with slow but non-zero rates are ideal candidates for protein engineering or directed evolution.

The native function or pathway for Cars in vivo is not currently known, but clues can be gleaned from the substrate preferences. *Mycobacterium* and *Nocardia* species contain a range of very distinctive and diverse lipids, such as mycolic, phenolic, and trehalose-containing glycolipids (Marrakchi et al., 2014; Neyrolles and Guilhot, 2011). Several of these have been shown to contribute to virulence in the well-studied *Mycobacterium tuberculosis* (*Mtb*). Based on the confluence of Cars in these two genera, and the enzymes' preferences for organic and aromatic acids (both of which are present in the lipid profiles of *Mtb*), it is possible that these enzymes play an as-yet unknown function in non-standard lipid metabolism. A TBCar homologue was found through BlastP that had no assigned function on the Tuberculist database (Lew et al., 2011). The homologue, named FadD9, is not essential, though the PPTase homologue (PptT) that likely activates FadD9 is essential for both growth and virulence, suggesting that these enzymes may play an as-yet unknown role in infection.

For future work seeking to exploit this biochemical transformation, some Cars did demonstrate higher promiscuity than others. *Mycobacterium* isozymes showed the highest rates for the short-chain organic acids, and comparable rates to the *Nocardia* enzymes for benzoic acid. If high turnover is desired, *Mycobacteria* may then contain better candidate enzymes. NBCar was the poorest performer, as its rates were the lowest, and low-turnover substrates found for the other Cars were not detected by GC/MS for this enzyme.

We substantially extend the usefulness of this data for different bioaldehyde synthesis applications by constructing a machine learning classifier that predicts substrates likely to be reduced by NICar. To our knowledge, this is the first use of machine learning cheminformatics that does not rely on enzyme sequence or structure to propose targets of an enzyme. We borrow an approach used in pharmacology to predict drug/P450 interactions and apply it to predicting metabolic promiscuity in the cytoplasm. Using training data collected by ourselves and others, we constructed classifiers for NICar that correctly predicted compounds with observable activity with NICar in vitro that was not in the classifier's training set. The preference for aromatic carboxylic acids suggested by the unweighted SVC (Supplementary 3–4) is likely due to either a true preference of NICar for aromatic carboxylic acids or an imbalance in representation in the training set between aromatic and aliphatic acids. For this reason, we give more credence to the results of the weighted SVC (Supplementary File 5–6). Encouragingly, the weighted SVC predicted activity by NICar on fatty acids of varying lengths, a result which has been shown elsewhere (Sheppard et al., 2014). The SVC also predicted vanillic acid as a substrate, a result that has been previously demonstrated (Venkitasubramanian et al., 2007). Other putative substrates of interest for NICar are involved in the biosynthesis of various amino acids and ubiquinone. The broadness of the substrate promiscuity on iAF1260 compounds by NICar suggests it may be effective at diverting flux from existing pathways for metabolic engineering applications. Given the broad alcohol dehydrogenase activity present in the cytoplasm, it is likely that in an in vivo environment these aldehydes would be further reduced into alcohols.

Overall, we sought to expand the available enzymatic options to reduce carboxylic acids to aldehydes. Through expressing and characterizing four Cars, we found substrate preferences common across the four homologs for non-polar organic acids and made further predictions about likely substrates in vivo based on machine learning techniques. Further, through comparing the activity profiles and sequences to the mechanistically related Lys2 enzymes, it is clear that Cars have a distinct reaction profile, and that this specificity is potentially a result of the adenylation domain enzyme sequence. The adenylation domain would be an interesting candidate for future engineering strategies. We also applied machine learning methods to classify the Car activity profiles against native metabolites.

Methods to predict enzymatic promiscuity, like those used here, will be necessary for expanding available biochemistries for the engineering of new pathways. Recent advances have made it feasible to design and implement pathways with non-natural biochemistry (Atsumi et al., 2008; Bhan et al., 2013; Campodonico et al., 2014). These pathways rely on enzymatic promiscuity towards non-native substrates, which can be difficult to discover if the promiscuous reaction is slow. The use of computational strategies to hypothesize likely enzyme-substrate pairs and highly sensitive analytics like in this study can be broadly used for nearly any enzyme-substrate pairing. High-sensitivity studies on the bioproduction of other rare moieties will undoubtedly change what types of molecules can be made through metabolic means.

The authors would like thank several Northwestern core facilities for their assistance: The Integrated Molecular Structure Education and Research Center (IMSERC) for GC/MS resources and technical assistance, the Center for Molecular Innovation and Drug Discovery (CMIDD)'s ChemCore facility for custom synthesis, the Recombinant Protein Production Core (rPPC) for plate reader access and protein expression guidance, and the Genomics Core for sequencing services. We also would like to thank the Jones (Imperial College London, UK) and Turner (University of Manchester, UK) labs for providing the MMCar construct. Funding was provided by the Bill and Melinda Gates Foundation, NSF CBET50835800, and the Northwestern McCormick School of Engineering.

References

Addis D, Das S, Junge K, Beller M. 2011. Selective reduction of carboxylic acid derivatives by catalytic hydrosilylation. Angew Chem Int Ed Engl 50:6004–6011. http://www.ncbi.nlm.nih.gov/pubmed/21648027

Agilent. 2014. Personal communication.

Akhtar MK, Turner NJ, Jones PR. 2013. Carboxylic acid reductase is a versatile enzyme for the conversion of fatty acids into fuels and chemical commodities. Proc Natl Acad Sci USA 110:87–92. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3538209&tool=pmcentrez&rendertype=abstract

Altschul S, Gish W, Miller W. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. http://www.cmu.edu/bio/education/courses/03510/LectureNotes/Altschul1990.pdf

Atsumi S, Hanai T, Liao JC. 2008. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 451:86–89. http://www.ncbi.nlm.nih.gov/pubmed/18172501

Beigi M, Waltzer S, Zarei M, Müller M. 2014. New Stetter reactions catalyzed by thiamine diphosphate dependent MenD from *E. coli.* J Biotechnol 191:64–68. http://linkinghub.elsevier.com/retrieve/pii/S0168165614007962

Bézier D, Park S, Brookhart M. 2013. Selective reduction of carboxylic acids to aldehydes catalyzed by B(C6F5) 3. Org Lett 15:496–499. http://linkinghub.elsevier.com/retrieve/pii/S0168165614007962

Bhan N, Xu P, Koffas MAG. 2013. Pathway and protein engineering approaches to produce novel and commodity small molecules. Curr Opin Biotechnol 24:1137–1143. http://www.sciencedirect.com/science/article/pii/S0958166913000335

Breuer M, Ditrich K, Habicher T, Hauer B, Kesseler M, Stürmer R, Zelinski T. 2004. Industrial methods for the production of optically active intermediates. Angew Chem Int Ed Engl 43:788–824. http://www.ncbi.nlm.nih.gov/pubmed/14767950

Campodonico MA, Andrews BA, Asenjo JA, Palsson BO, Feist AM. 2014. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. Metab Eng 25:140–158. http://www.sciencedirect.com/science/article/pii/S1096717614001001

Ehmann DE, Gehring a M, Walsh CT. 1999. Lysine biosynthesis in Saccharomyces cerevisiae: Mechanism of alpha-aminoadipate reductase (Lys2) involves posttranslational phosphopantetheinylation by Lys5. Biochemistry 38:6171–6177. http://www.ncbi.nlm.nih.gov/pubmed/10320345

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868. http://www.pnas.org/content/95/25/14863.full

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ.. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1911197&tool=pmcentrez&rendertype=abstract

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. Proteomics protocols handbook. pp: Humana Press. p 571–607. http://web.expasy.org/docs/expasy_tools05.pdf

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28:977–982.

Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. 2010. MeV: MultiExperiment viewer. Biomed Informatics Cancer Res 267–277. http://link.springer.com/10.1007/978-1-4419-5714-6

Hult K, Berglund P. 2007. Enzyme promiscuity: Mechanism and applications. Trends Biotechnol 25:231–238. http://www.ncbi.nlm.nih.gov/pubmed/17379338

Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: A mechanistic and evolutionary perspective. Annu Rev Biochem 79:471–505. http://www.ncbi.nlm.nih.gov/pubmed/20235827

Kohlpaintner C, Schulte M, Falbe J, Lappe P, Weber J, Frey G. 2013. Aldehydes, Aliphatic. Ullmann's Encycl. Ind Chem 1–31. http://www.ncbi.nlm.nih.gov/pubmed/20235827

Kunjapur AM, Prather KLJ. 2015. Microbial engineering for aldehyde synthesis. Appl Environ Microbiol 81:1892–1901. http://aem.asm.org/content/81/6/1892.short

Kunjapur AM, Tarasova Y, Prather KLJ. 2014. Synthesis and accumulation of aromatic aldehydes in an engineered strain of *Escherichia coli.* J Am Chem Soc 136:11644–11654.

Langevin F, Crossan GP, Rosado I V, Arends MJ, Patel KJ. 2011. Fancd2 counteracts the toxic effects of naturally produced aldehydes in mice. Nature 475:53–59. http://www.nature.com/nature/journal/v475/n7354/pdf/nature10192.pdf

Lew JM, Kapopoulou A, Jones LM, Cole ST. 2011. TubercuList—10 years after. Tuberculosis (Edinb) 91:1–7. http://www.sciencedirect.com/science/article/pii/S1472979210001113

Li T, Rosazza JPN. 1997. Purification, characterization, and properties of an aryl aldehyde oxidoreductase from Nocardia sp. strain NRRL 5646. J bacteriol 179:3482–3487.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2014. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. http://nar.oxfordjournals.org/content/43/D1/D222

Marrakchi H, Lanéelle MA, Daffé M. 2014. Mycolic acids: Structures, biosynthesis, and beyond. Chem Biol 21:67–85.

McMurray DN. 1996. Mycobacteria and nocardia. In: Baron S, editor. Medical microbiology. 4th edn. Galveston (TX): University of Texas Medical Branch at Galveston. Chapter 33.

Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO. 2012. Network context and selection in the evolution to enzyme specificity. Science 337:1101–1104. http://www.sciencemag.org/content/337/6098/1101.short

Napora-Wijata K, Strohmeier GA, Winkler M. 2014. Biocatalytic reduction of carboxylic acids. Biotechnol J 9:822–843. http://www.ncbi.nlm.nih.gov/pubmed/24737783

Neyrolles O, Guilhot C. 2011. Recent advances in deciphering the contribution of Mycobacterium tuberculosis lipids to pathogenesis. Tuberculosis 91:187–195. http://www.sciencedirect.com/science/article/pii/S1472979211000163

Nobeli I, Favia AD, Thornton JM. 2009. Protein promiscuity and its implications for biotechnology. Nat Biotechnol 27:157–167. http://www.ncbi.nlm.nih.gov/pubmed/19204698

Patel RN. 2011. Biocatalysis: Synthesis of key intermediates for development of pharmaceuticals. ACS Catal 1:1056–1074.

Patel RN, Chu L, Mueller R. 2003. Diastereoselective microbial reduction of (S)-[3-chloro-2-oxo-1-(phenylmethyl) propyl]carbamic acid, 1,1-dimethylethyl ester. Tetrahedron: Asymmetry 14:3105–3109. http://www.sciencedirect.com/science/article/pii/S095741660300658X

Rao ST, Rossmann MG. 1973. Comparison of super-secondary structures in proteins. J Mol Biol 76:241–256. http://www.sciencedirect.com/science/article/pii/0022283673903884

Rodriguez GM, Atsumi S. 2014. Toward aldehyde and alkane production by removing aldehyde reductase activity in *Escherichia coli.* Metab Eng 25:227–237. http://www.sciencedirect.com/science/article/pii/S1096717614001037

Sheppard MJ, Kunjapur AM, Wenck SJ, Prather KLJ. 2014. Retro-biosynthetic screening of a modular pathway design achieves selective route for microbial synthesis of 4-methyl-pentanol. Nat Commun 5:5031. http://www.nature.com/ncomms/2014/140924/ncomms6031/full/ncomms6031.html?WT.ec_id=NCOMMS-20141001

Venkitasubramanian P, Daniels L, Das S, Lamm AS, Rosazza JPN. 2008. Aldehyde oxidoreductase as a biocatalyst: Reductions of vanillic acid. Enzyme Microb Technol 42:130–137. http://www.ncbi.nlm.nih.gov/pubmed/22578862

Venkitasubramanian P, Daniels L, Rosazza JPN. 2006. Biocatalytic reduction of carboxylic acids: Mechanism and applications. In: Patel RN, editor. Biocatal pharmaceutical biotechnology industries. CRC Press. p 425–440.

Venkitasubramanian P, Daniels L, Rosazza JPN. 2007. Reduction of carboxylic acids by Nocardia aldehyde oxidoreductase requires a phosphopantetheinylated enzyme. J Biol Chem 282:478–485. http://www.jbc.org/content/282/1/478.short

Xu H, Andi B, Qian J, West AH, Cook PF. 2006. The alpha-aminoadipate pathway for lysine biosynthesis in fungi. Cell Biochem Biophys 46:43–64. http://www.ncbi.nlm.nih.gov/pubmed/16943623

Zhang K, Li H, Cho KM, Liao JC. 2010. Expanding metabolism for total biosynthesis of the nonnatural amino acid L-homoalanine. Proc Natl Acad Sci USA 107:6234–6239. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2852006&tool=pmcentrez&rendertype=abstract

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.