Engineering Enzyme Specificity Using Computational Design of a Defined-Sequence Library

Citation: Lippow, Shaun M., Tae Seok Moon, Subhayu Basu, Sang-Hwal Yoon, Xiazhen Li, Brad A. Chapman, Keith Robison, Daša Lipovšek, and Kristala L.J. Prather. "Engineering Enzyme Specificity Using Computational Design of a Defined-Sequence Library." Chemistry & Biology 17, no. 12 (December 2010): 1306–1315. © 2010 Elsevier Ltd.

As Published: http://dx.doi.org/10.1016/j.chembiol.2010.10.012

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/96044

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Engineering Enzyme Specificity Using Computational Design of a Defined-Sequence Library

Shaun M. Lippow,[1,3,4,*] Tae Seok Moon,[2,3,6] Subhayu Basu,[1] Sang-Hwal Yoon,[2] Xiazhen Li,[1] Brad A. Chapman,[1] Keith Robison,[1] Daša Lipovšek,[1,5] and Kristala L.J. Prather[2,*]
[1]Codon Devices, Inc., 99 Erie Street, Cambridge, MA 02139, USA
[2]Synthetic Biology Engineering Research Center (SynBERC) and Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[3]These authors contributed equally to this work
[4]Present address: Silver Creek Pharmaceuticals, 409 Illinois Street, San Francisco, CA 94158, USA
[5]Present address: Adnexus, a Bristol-Myers Squibb R&D Company, 100 Beaver Street, Waltham, MA 02453, USA
[6]Present address: Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA
*Correspondence: slippow@gmail.com (S.M.L.), kljp@mit.edu (K.L.J.P.)
DOI 10.1016/j.chembiol.2010.10.012

SUMMARY

Engineered biosynthetic pathways have the potential to produce high-value molecules from inexpensive feedstocks, but a key limitation is engineering enzymes with high activity and specificity for new reactions. Here, we developed a method for combining structure-based computational protein design with library-based enzyme screening, in which inter-residue correlations favored by the design are encoded into a defined-sequence library. We validated this approach by engineering a glucose 6-oxidase enzyme for use in a proposed pathway to convert D-glucose into D-glucaric acid. The most active variant, identified after only one round of diversification and screening of only 10,000 wells, is approximately 400-fold more active on glucose than is the wild-type enzyme. We anticipate that this strategy will be broadly applicable to the discovery of new enzymes for engineered biological pathways.

INTRODUCTION

The use of biological systems as “microbial chemical factories” has increased significantly since the earliest days of large-scale fermentation to produce natural products such as penicillin. Complementing the traditional tools of metabolic engineering (Bailey, 1991), the emerging field of synthetic biology seeks to streamline the construction of highly productive systems (Dueber et al., 2009). Microbial synthesis of chemicals is even more appealing as one considers that advances in technology now enable researchers to think beyond improving flux through known pathways and to begin to move toward designing entirely new biosynthetic schemes (Prather and Martin, 2008). One recent example resulted in the synthesis of molecules that might serve as novel biofuels (Atsumi et al., 2008; Zhang et al., 2008).

Perhaps the most daunting impediment to more widespread de novo design and assembly of biosynthetic pathways is the lack of natural enzymes with high activity and specificity across a wide range of substrates (Yoshikuni et al., 2008; Zhang et al., 2008), combined with the challenge of engineering enzyme specificity (Fox and Huisman, 2008; Gerlt and Babbitt, 2009). Currently, the two main approaches to enzyme engineering are diversity-based screening of large libraries and computational design of a few variants. Diversity-based approaches, which aim to compensate for an incomplete understanding of sequence-structure-function relationships by accessing large numbers of variants (Leemhuis et al., 2009), are often hindered by the availability of only a low- to medium-throughput screen. On the other hand, even structure-based computation (Damborsky and Brezovsky, 2009) cannot reliably design improved variants due to deficiencies in scoring and search functions applied to active sites and catalysis, despite recent advances (Lippow et al., 2007; Jiang et al., 2008; Rothlisberger et al., 2008; Smith and Kortemme, 2008; Chen et al., 2009; Koder et al., 2009; Murphy et al., 2009). The third approach to enzyme engineering is to take advantage of both methods by designing a library based on computational predictions (Moore and Maranas, 2004; Chica et al., 2005; Kang and Saven, 2007; Lappe et al., 2009; Shivange et al., 2009). Thus, experimental screening offsets the shortcomings of modeling by testing more designed variants (e.g., 10^4 instead of 10^1–10^2). In other words, computational design compensates for limited screening capacity by focusing diversity to sequences that are more likely to be functional compared to random or otherwise less rationally designed variations.

Previous studies to combine computational modeling with library screening have treated amino acid positions independently, neglecting pairwise and higher-order correlations (Voigt et al., 2001; Hayes et al., 2002; Amin et al., 2004; Socolich et al., 2005; Otey et al., 2006; Fox et al., 2007; Liao et al., 2007; Treynor et al., 2007; Barderas et al., 2008). In some cases, information on such correlations is absent, such as in a multiple sequence alignment with too few sequences to detect correlations above noise. In other cases, correlations are ignored because of the difficulty of physically encoding them in a library. In particular, using degenerate codons to create position-specific diversity does not allow the encoding of correlations between positions. Yet structure-based, computational protein design can generate substantial and actionable correlated information (Jiang et al., 2008; Rothlisberger et al., 2008). For

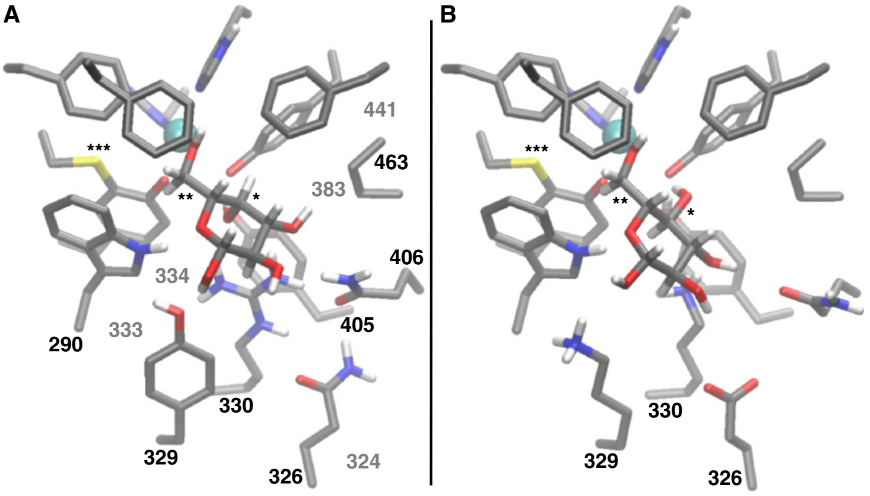

Figure 1. Structural Models of Enzymes with Bound Substrate
The single asterisk marks the C-4 carbon at which the hydroxyl is axial in galactose but equatorial in glucose, the double asterisk marks the C-6 carbon at which the enzyme converts the alcohol to an aldehyde, and the triple asterisk marks the C228-Y272 crosslink. The active-site Cu^{2+} is shown as a cyan ball, and enzyme nonpolar hydrogens are not shown for clarity. (A) Wild-type galactose 6-oxidase with docked galactose. The 12 redesigned positions are labeled, with five second-shell positions (324, 333, 334, 383, and 441) numbered in lighter gray and their side chains not shown for clarity. (B) Model of Des3-2 with docked glucose. Note that the clockwise twist of glucose avoids the potential clash at C-4 with Y495, yet distorts the planarity of the five-membered transition state at C-6. The mutation R330K opens up space below glucose to permit the twist, and the mutations Q326E and Y329K reestablish hydrogen bonds to glucose.

example, two positions buried together in the protein core may accommodate one large and one small amino acid, for two viable sequences, but a position-independent library would encode all four combinatorial sequences. Besides predicting a single low- or lowest-energy protein sequence, computation can provide thousands of low-energy sequences; if a physical library can be focused around such predicted low-energy sequences, it would be expected to contain a higher fraction of functional hits than a random library not based on computational design.

Here, we developed a method to assemble high-fidelity libraries of defined amino acid sequences that include both predicted favorable residues at specific positions and favorable inter-residue correlations. The main advantages of our defined-sequence libraries are that: (1) only amino acid mutations predicted from computational design are included; and (2) within regions close enough in primary sequence to encode on a single oligonucleotide (up to 24-position regions in the present work), only designed amino acid combinations are included. We used this approach to engineer the specificity of galactose 6-oxidase to generate a novel glucose 6-oxidase enzyme for use in a proposed biosynthetic pathway for D-glucaric acid.

D-Glucaric acid is a so-called "value-added chemical from biomass" (Aden et al., 2004) that has been studied for uses ranging from human therapeutics to materials synthesis (Walaszek et al., 1996; Singh and Gupta, 2007). Although D-glucaric acid is a natural product, the elucidated mammalian route is complex and would be difficult to reconstruct in a microbial host. We recently synthesized D-glucaric acid in *Escherichia coli* using a pathway that utilized heterologous enzymes from three disparate sources operating on their natural substrates, but titers remain in the low grams-per-liter range (Moon et al., 2009). As an alternative, we designed de novo a synthetic pathway for production of D-glucaric acid (see Figure S1 available online). The first step of this pathway requires oxidation of the glucose C-6 hydroxyl; unfortunately, no glucose 6-oxidase has yet been found in nature. Arnold and coworkers (Sun et al., 2002) identified a variant of galactose 6-oxidase (M-RQW) with activity on glucose, but improvements are needed for use in microbial synthesis.

Transformation of galactose 6-oxidase into glucose 6-oxidase is challenging because the wild-type enzyme exhibits greater than 10^6-fold specificity for galactose over glucose, because enzyme activity can only be screened in a medium-throughput plate-based assay, and because there is no crystal structure of the enzyme-substrate complex. These challenges also make it a problem well suited for the efficient combination of computational design and library screening. Here, we designed a library based on structure-based computational modeling, assembled and screened the library, and characterized novel enzymes with improved activity and specificity for glucose.

RESULTS

Computational Modeling and Library Design

Galactose was docked into the active site of galactose 6-oxidase (Figure 1A). The docked galactose makes several hydrogen bonds to enzyme side chains without requiring protein conformational change, and its C-6 hydroxyl is near the Cu^{2+} and Y272 as required for catalysis (Wachter and Branchaud, 1996). Twelve active-site and second-shell positions involved in substrate recognition were simultaneously redesigned around glucose, while preserving as wild-type all remaining positions, including five active-site side chains crucial for the conversion of alcohol to aldehyde (Figure 1A). The enzyme redesign protocol was run 20,000 independent times, generating 2,379 unique sequences with low predicted energy of the enzyme-glucose complex. These designs were narrowed down based on criteria including predicted favorable enzyme-glucose binding interaction, enzyme stability, and electrostatic components of these energies. Mutations that contributed to predicted improvement of enzyme stability, but not to glucose recognition, were reverted to wild-type.

The 12 designed positions were grouped into four assembly regions based on proximity in primary sequence (Table 1 and Figure 2). Within each region, we generated library diversity by mixing synthetic oligonucleotides that encoded complete sequences that had been scored as favorable by computational design. This approach maintained all correlations among the

Table 1. The Designed Library and Combinatorial Library

Designed positions	290	324	326	329	330	333	334	383	405	406	441	463	Diversity
Wild-type	W	D	Q	Y	R	N	H	C	Y	Q	F	P	
Assembly region[a]	I	II						III			IV		
Designed amino acid combinations	F	D	E	K	K	N	H	C	G	T	F	A	
	W	D	E	R	K	N	H	C	H	Q	F	D	
		D	K	Y	E	N	H	C	S	Q	F	E	
		D	N	R	H	N	H	C	S	R	F	P	
		D	Q	K	Q	N	H	C	Y	E	F	S	
		D	Q	R	K	N	H	C	Y	K	H	D	
		D	R	E	S	N	H	C	Y	N	H	P	
		D	R	R	N	N	H	C	Y	Q	R	S	
		D	R	Y	K	N	H	C	Y	R			
		D	S	R	K	N	H	C	Y	T			
		G	R	Y	I	N	H	N	Y	Q			
		S	Q	R	D	R	E	N	Y	T			
		S	R	R	S	N	H	S	Y	Q			
		Plus 35 more, totaling 48 (Table S1)						S	Y	T			
Designed library: number of combinations	2	48						14			8		10,752
Number of different amino acids by position	2	3	7	6	11	2	2	3	4	6	3	5	$1.2 \cdot 10^{7}$
Combinatorial library: best degenerate codon[b]	TKS	RRC	VRS	NRS	NDS	MRC	SAS	WRC	NRC	VVS	YDC	BMW	
Number of different amino acids or stop codon	4	4	9	13	18	4	4	4	8	12	6	9	$1.1 \cdot 10^{10}$
Total number of codons	4	4	12	16	24	4	4	4	8	18	6	12	$4.9 \cdot 10^{10}$

[a] Positions encoded on a single oligonucleotide are indicated as Assembly regions I–IV.
[b] R = A,G; Y = C,T; M = A,C; K = G,T; S = C,G; W = A,T; B = C,G,T; D = A,G,T; H = A,C,T; V = A,C,G; N = A,C,G,T.

design positions in that oligopeptide stretch and excluded all non-designed amino acid combinations. For positions encoded by separate oligonucleotides, favorable inter-residue correlations were not encoded in the library, but preferences for specific side chains at specific positions were encoded. The number of sequences selected for inclusion in each synthesized region was chosen so that the overall diversity of the designed library (2 × 48 × 14 × 8 = 10,752), generated by joining the four regions combinatorially, approximately matched the experimental screening capacity (10^{4}). Although the combinatorial step resulted in some full-length sequences that were not designed computationally, the encoded single-position preferences and the encoded inter-residue correlations still greatly focused the diversity as compared to more than 10^{7} possible combinations of all amino acids appearing in at least one design. A second library was designed using best-case degenerate codons to capture single-position diversity, without consideration of pairwise or higher-order correlations from the computational design. From this we created a control, combinatorial library of theoretical diversity greater than 10^{10} (details in Table S2).

Synthesis of DNA Libraries

The designed and combinatorial libraries were assembled using the PCR-based strategy shown in Figure 2. A 643 bp region of the galactose 6-oxidase gene that spans the 12 designed

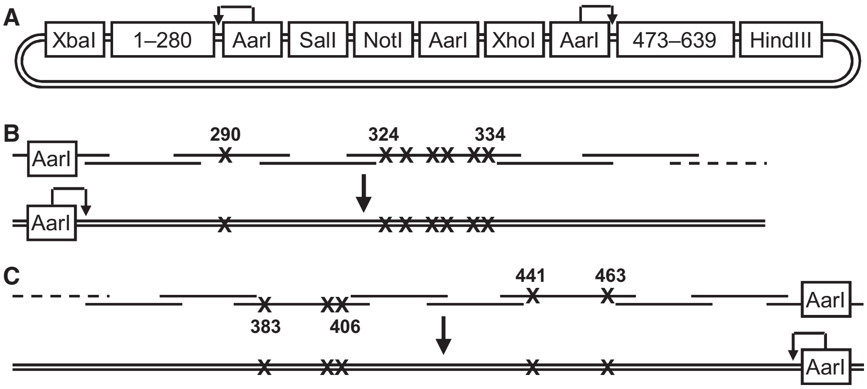

Figure 2. Diagram of Library Construction

(A) The synthesized base construct. Non-designed positions 1–280 and 473–639 (amino acid numbering) are separated by a stuffer, which is excised by the type IIS restriction endonuclease AarI and replaced by the cloned-in library. (B and C) The designed region, 281–472, is arranged into a series of 18 partially overlapping oligonucleotides for assembly by PCR. The oligonucleotides are grouped into two initial steps, shown in (B) and (C), with overlap in the dashed segments for subsequent assembly into the 643 bp full-length library. The resulting library is flanked by AarI sites for scar-free cloning into the base construct. The 12 designed positions are grouped onto four of the oligonucleotide segments, with the first and last designed positions on each segment labeled on the figure.

Table 2. Experimental Screening of the Designed and Combinatorial Libraries

	Number Screened	Number of Hits	Overall Hit Rate (%)
Combinatorial library	11,266	1 (weak)	0.009
Designed library, round 1	10,603	402	3.8
Designed library, round 2	402	Weak: 186	
		Medium: 155	
		Strong: 61	0.57
Designed library, round 3	61	Strongest: 11	0.10

positions was broken into 18 overlapping segments, four of which contained all the variable positions. For the designed library, each defined sequence was encoded by a separate oligonucleotide. Prior to assembly, all oligonucleotides were annealed to their reverse complement and error corrected using the DNA-mismatch-binding protein MutS (Carr et al., 2004; Lipovšek et al., 2009). After construction and cloning into the expression vector, randomly picked clones were sequenced to determine assembly fidelity. In the designed library: 40% of the encoded amino acid sequences conformed to the design exactly; 18% contained at least one non-designed combination of designed amino acids (“crossover”); 8.8% contained a substitution (a non-designed amino acid) at a design position; 9.6% contained a substitution elsewhere in the gene; 21% contained an insertion, deletion, or stop codon; and 2.8% lacked the insert (Table S3). Within each assembly region, the equimolarly mixed defined sequences did not show a significant bias (Figure S2). In the combinatorial library: 27% conformed to the design; 7.0% had a substitution at a non-design position; 65% contained an insertion, deletion, or substitution to stop codon; and 0.8% lacked insert (Table S4).

Library Screening

Over 10,000 clones from each library were screened for glucose 6-oxidase activity in a plate-based, colorimetric assay (Figure S3). Any library member displaying a signal comparable to or better than the M-RQW positive control (Sun et al., 2002) was considered a hit. The combinatorial library yielded one weak hit, whereas the designed library produced 402 hits of equal or better signal strength (Table 2). Subsequent rounds of screening at increased stringency narrowed the 402 hits first to 61 strong hits, and finally to 11 strongest hits. All 402 designed-library hits were sequenced: 70% encoded amino acid sequences exactly as designed; 12% contained a crossover; 10% had a substitution at a designed position; 2.2% had a substitution at a non-designed position; and 0.5% had a deletion (Table S5). Of the 61 strong hits, only two contained a crossover, and none had a substitution, insertion, or deletion. The 11 most active clones isolated in the final round of screening all conform to the theoretical library design, both in the identity and in the combinations of amino acid residues (Tables 1 and 3; Figure S1).

The sequence patterns that emerged with progression of designed-library screening are shown in Figure 3. As stringency increased, specific design positions were clearly enriched for particular amino acids. In the top 11 hits, five of the 12 design positions converged to the wild-type amino acid, two positions were predominantly wild-type, four positions were predominantly mutated (Q326, Y329, Q406, and P463), and one position converged to a single mutation (R330K).

Characterization of Improved Glucose 6-Oxidase Variants

The wild-type galactose 6-oxidase, the previously published variant M-RQW, and the designed-library top hits (“Des3-*x*,” where *x* is a clone number between 1 and 12) were expressed and purified, and specific activities were measured on both

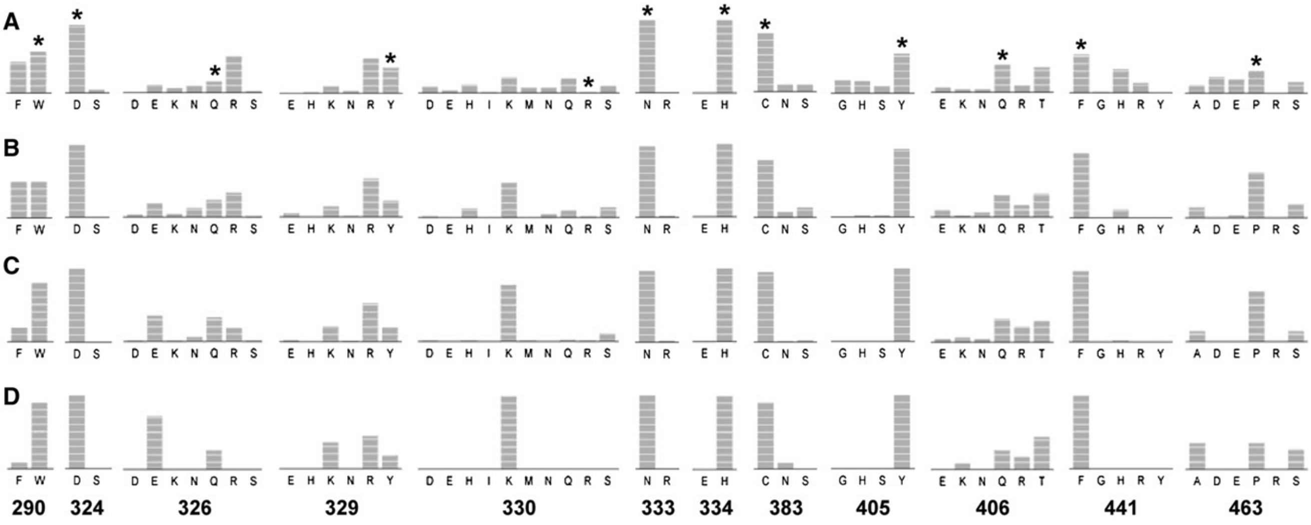

Figure 3. Amino Acid Enrichment by Position during Screening of the Designed Library

Each row shows one round of library screening, and bar heights represent the fraction of clones at that round with the indicated amino acid at each of the 12 designed positions. Asterisks denote the wild-type amino acid at each position. (A) Two hundred forty-two randomly picked clones after DNA assembly, prior to screening. (B) Three hundred eighty-two of the 402 hits from screening round one. (C) Fifty-nine of the 61 hits from screening round two. (D) The final 11 hits from screening round three.

Table 3. Characterization of the Top Hits from the Designed and Combinatorial Libraries

													Specific Activity ($M^{-1}s^{-1}$)	
	290	324	326	329	330	333	334	383	405	406	441	463	On Galactose	On Glucose
Wild-type	W	D	Q	Y	R	N	H	C	Y	Q	F	P	34,000 ± 5,000	≈0.05[a]
M-RQW	F				K					T			91 ± 8	4.9 ± 0.7
Com1-1[b]	F		N	D	K			Y	N	T	Y	A	Not measured	Not measured
Des3-1[c]			E	R	K					T		A	460 ± 30	11.0 ± 0.6
Des3-2			E	K	K								1,100 ± 300	20. ± 3
Des3-3			E	K	K					T		A	550 ± 70	14.2 ± 1.1
Des3-4					K					T		S	172 ± 2	2.8 ± 0.4
Des3-5				R	K								1,050 ± 150	1.8 ± 0.2
Des3-6			E	R	K					R			870 ± 180	5.8 ± 0.6
Des3-7			E	R	K					K			1,160 ± 40	4.8 ± 0.2
Des3-9			E	R	K							S	210 ± 10	4.4 ± 0.4
Des3-10	F			K	K					R			14.8 ± 0.1	5.6 ± 0.8
Des3-11			E	K	K					T		S	130 ± 30	13.0 ± 1.1
Des3-12					K			N				A	152 ± 6	1.25 ± 0.07

[a] Based on reported 100-fold improvement of M-RQW over wild-type (Sun et al., 2002).

[b] The lone isolate from the control, combinatorial library.

[c] Des3-1 and Des3-8 were found to be identical after sequencing, thus Des3-8 was excluded from the analysis.

galactose and glucose (Table 3). Consistent with previous measurements (Sun et al., 2002), M-RQW exhibits notable activity on glucose and an approximately 400-fold decrease in its activity on galactose, yet it remains 20-fold more active on galactose than on glucose. Among the top hits, four enzymes (Des3-2, Des3-3, Des3-11, and Des3-1) are at least 2-fold more active on glucose than M-RQW, and two enzymes (Des3-10 and Des3-11) are at least 2-fold more specific for glucose than M-RQW. Des3-2 is 4-fold more active than M-RQW and is, to our knowledge, the most active glucose 6-oxidase identified to date. A structural model of Des3-2 with glucose is shown in Figure 1B. The top three variants (Des3-2, Des3-3, and Des3-11) all contain, and are the only characterized enzymes that contain, mutations Q326E, Y329K, and R330K, whereas Des3-1 contains the mutation Y329R. Des3-10, which is as active on glucose as is M-RQW, is 7-fold more specific than M-RQW due to its decreased activity on galactose, and is only 2.5-fold more active on galactose than on glucose.

We found that the enzymes with the W290F mutation (M-RQW and Des3-10) show significant lag times during catalysis (Figure S4). For these mutants, the maximal rate, rather than the initial rate, was used to calculate the specific activity. As a result, the 4-fold improvement that we report for Des3-2 over M-RQW is a conservative estimate. We have investigated this lag phenomenon through a mechanistic model as follow-up work (T.S.M. and K.L.J.P., unpublished data).

DISCUSSION

We engineered the substrate specificity of galactose 6-oxidase for higher activity on glucose by modifying the side chains that bind to and recognize galactose to interact instead with glucose, while preserving the wild-type catalytic side chains critical for the alcohol-to-aldehyde conversion. The design started by docking glucose in a catalytically relevant initial conformation, in which its C-6 hydroxyl completes a five-membered, near-planar transition state with the active-site Cu^{2+} and Y272 oxygen radical (Figure S5) (Wachter and Branchaud, 1996; Whittaker, 2003). Next, 12 potential glucose-recognition residues were mutated in silico, optimizing the stability of the enzyme-glucose complex. The variants with the lowest predicted enzyme-glucose interaction energies were encoded in a defined-sequence library that included many of the predicted favorable combinations of mutations and excluded unfavorable combinations, and used large sets of defined-sequence oligonucleotides to encode such focused diversity. A control library was based on the same computational prediction, but it combined independently diversified single positions, without regard for correlations between positions, and used degenerate nucleotides as the source of diversity.

We found that the defined-sequence library, which incorporated predicted favorable amino acid identities at single positions as well as predicted favorable inter-residue correlations, had an approximately 400-fold higher hit rate than the degenerate nucleotide library, which contained some unfavorable amino acid identities at single positions and no predicted favorable correlational constraints (Table 2). In addition, all the highly active, improved enzymes that were identified in the screen originated in the defined-sequence library (Table 3). The superiority of focused, correlationally informed library diversity was further demonstrated when hit rates were compared between those variants from the defined-sequence library that conformed to the library design and those that did not, due to the imperfect physical library construction method. For example, 18% of enzyme variants in the assembled defined-sequence library (or, more relevantly, 24% of those clones encoding full-length open reading frames) contained a crossover, where combinations of mutations that were not predicted to be favorable appeared in the same clone. In contrast, none of the 11 improved clones isolated after stringent screening contained such a crossover. As the screening became more stringent, the percentage of clones with a crossover dropped monotonically: 24% in the full-length library, 13% in the first-round 402 hits, 3% in the 61 strong hits, and 0% in the top 11 strongest hits. Similarly, 24% of enzyme variants in the assembled defined-sequence library contained at least one mutation that was not predicted to be favorable, but none of the 11 improved clones isolated after stringent screening contained such a random mutation. Together, these observations validate the utility of computationally designed libraries in general and of libraries incorporating information about both single-position preferences and preferred correlations between different positions in an enzyme in particular. Whereas the data described here are not sufficient to estimate the exact contribution of encoded single-position preferences versus encoded preferred inter-residue correlations to the success of our designed library, the observation that the hits screened under the greatest stringency conform to both types of designs suggests that both were selected for, i.e., that both types of design contributed to the high hit rate of the designed library.

The benefit of using this type of designed library can also be illustrated by the efficient sampling of the theoretical sequence space by our screen-limited physical library size, 10^4. Traditional saturation mutagenesis (e.g., using NNK or NNS degenerate codons) would have been limited to diversifying only three positions, generating 32,768 (32^3) DNA sequences and 8,000 (20^3) unique amino acid sequences. For example, the previous work by Arnold and coworkers (Sun et al., 2002), which yielded the enzyme variant M-RQW with weak glucose 6-oxidase activity, required three successive rounds of saturation mutagenesis and screening to isolate single or double-beneficial mutations, which were then combined to create M-RQW. In contrast, our designed, defined-sequence library enabled a more successful search through mutations at 12 enzyme positions simultaneously in just one round of diversification and screening.

Methods to efficiently search multiple positions simultaneously are valuable not only because there are many enzyme positions at which mutations can improve activity but also because a sequential search of only three positions at a time can become trapped in local minima in the sequence-function landscape. In using the best enzyme variant from one round of screening as the input to the next round of searching other positions, mutations identified in earlier screens become locked in, even if they are not as beneficial as other combinations of yet-to-be found mutations. For instance the W290F mutation was isolated early in the previous work and was combined with two mutations identified from a separate screen to generate M-RQW (Sun et al., 2002). However, we find the W290F mutation in only one of our 11 top hits, and our most active variant, Des3-2, retains wild-type W290. Although our approach does not ensure that a global minimum is reached, the ability to screen mutations in combination facilitates sampling of a greater fraction of the energy landscape, possibly arriving at a local minimum representative of a deeper energy well.

The only consensus mutation in our top 11 enzymes is R330K (Table 3), which is also found in the weakly active variant from the combinatorial library (Com1-1) and in M-RQW. After the first round of designed-library screening, 98% of hits had a mutation at R330, with 48% mutated to lysine (Figure 3). According to our models, the R330K mutation replaces an arginine that makes bidentate hydrogen bonds to the C-3 and C-4 hydroxyls of galactose, with a lysine that makes a hydrogen bond to the C-3 hydroxyl of glucose. The mutation allows glucose to dock in a slightly twisted orientation, thus relieving the clash that would have existed between Y495 and its equatorial C-4 hydroxyl, which moved due to the stereochemistry change from galactose (Figure 1). There are two major effects of this twist: the planarity of the five-membered transition state is decreased; and the other enzyme-glucose hydrogen bonds are distorted (Figures 1 and S5). We find that our top enzymes regain substrate recognition through additional mutations. For instance, Y329 is mutated in nine of the 11 top enzymes, either to arginine or lysine, potentially making a new hydrogen bond to the glucose C-1 hydroxyl. The mutation Q326E occurs in eight of these nine variants, potentially forming a salt bridge with the arginine/lysine to stabilize the enzyme. Mutations at Q406 and P463 are predicted to modify hydrogen bonding to the C-2 and C-3 hydroxyls, and although these mutations are not necessary for glucose activity (they remain wild-type in Des3-2), they are more often mutated in variants with improved glucose activity relative to galactose activity.

Our enzyme variant Des3-2 exhibits at least 4-fold improved activity on glucose over M-RQW, and approximately 400-fold improved activity over wild-type (Sun et al., 2002). Nevertheless, Des3-2 remains more active on the native substrate, galactose, and its activity on glucose is still 2000-fold lower than the activity of wild-type galactose 6-oxidase on galactose, suggesting that it is still insufficiently active for use in a designed metabolic pathway. We speculate that the desired specificity switch is particularly challenging because the inherent clash between the glucose C-4 hydroxyl and Y495 can only be relieved by mutating or repositioning Y495, which would disrupt catalysis due to its role in stabilizing the active-site Cu^{2+} ion, or by docking glucose in a twisted conformation, which decreases transition state planarity and likely contributes significantly to decreased catalytic efficiency. Thus, additional computational or experimental approaches may be required to achieve further increases in enzyme activity.

SIGNIFICANCE

Combining computational protein design with library-based screening is an attractive protein-engineering strategy that can, in principle, yield the best of both approaches to rapidly generate novel enzymes. Although others have shown that libraries based on computationally designed mutations can be an improvement over site-saturation mutagenesis or random mutations (Voigt et al., 2001; Hayes et al., 2002; Amin et al., 2004; Socolich et al., 2005; Otey et al., 2006; Fox et al., 2007; Liao et al., 2007; Treynor et al., 2007; Barderas et al., 2008), previous methods fail to take full advantage of information from design, generating libraries in a position-independent manner and excluding all correlations between protein positions. We have gone a step further to show that defined-sequence libraries that incorporate some designed amino acid combinations in addition to single-position preferences can further improve hit rates and reduce screening efforts. To our knowledge, our defined-sequence approach is the first method for computational

library design to encode inter-residue correlations of amino acids into the physical library, i.e., to include in the physical library combinations of amino acids that are predicted to be favorable and to exclude combinations that are predicted to be unfavorable. By encoding both single-position preferences and inter-residue correlations between positions close in primary sequence, we were able to engineer an improved glucose 6-oxidase enzyme in just one round of diversification and screening. Our experimental screening effort was at least two orders of magnitude more efficient than the control screening effort, which used focused redundant nucleotide mixes in library construction. This approach to library design and construction should facilitate the engineering of enzymes with new activities and specificities, which we expect will be particularly useful for providing components for creating new biosynthetic pathways.

EXPERIMENTAL PROCEDURES

Computational Modeling and Design

Preparation of Wild-Type Structure

The crystal structure of galactose 6-oxidase, 1GOF (Ito et al., 1991), was prepared for computational modeling in the Schrödinger software suite (Schrödinger, LLC, New York, NY, USA). Acetate ion 703 at the active site was removed, leaving the crystallographic copper ion. Solvent-exposed waters were removed, retaining 73 of 316 crystallographic waters, including eight waters within 8 Å of the copper ion and 13 within 10 Å. Acetate ion 701 and sodium ion 702 were retained, although each are far from the active site. The C228-Y272 covalent crosslink was formed; the side-chain oxygen of Y495 was deprotonated; the copper ion charge was set to +2; and the side-chain oxygen of Y272 was changed to fluorine to approximate an oxygen radical (both are neutral and electronegative).

The Schrödinger Protein Preparation Wizard (version 2007) was used to assign bond orders, add hydrogens, treat metals, detect disulfide bonds, exhaustively search histidine, asparagine, and glutamine side chains for protonation state and terminal-dihedral rotation (due to ambiguity in crystal structures between C, N, and O atoms), and sample water orientations. Next, all water molecules were minimized using MacroModel (version 9.5), and then Prime (version 1.6) was used to refine the side chains of W290, Q326, Y329, R330, and Q406. The RMSD over the side-chain atoms of these five residues was 0.81 Å due to a flip of the terminal dihedral of Q326 and less than 30° rotations in χ_2 and χ_3 of Q406.

Docking Galactose into Active Site of Wild-Type

Schrödinger software was next used to dock a fully flexible galactose into a rigidly held active site of galactose oxidase. First, galactose was generated (in isolation) in a default conformation, and MacroModel mixed torsional/low-frequency-mode conformational search was used to explore different ring structures and torsional angles. One thousand structures were generated, yielding 429 unique conformations. The lowest-energy conformation of galactose was selected, a version of "chair."

Next, Glide (version 4.5) was used to generate a receptor grid for the prepared wild-type structure. The docking site was defined as a box centered at a point in the active site (cartesian coordinates: 37 Å, 8 Å, 17 Å), approximately 3 Å from each of: Cu^{2+}, Y495 deprotonated oxygen, W290 amine hydrogen, Y405 hydroxyl hydrogen, and Y272 oxygen radical (modeled as a fluorine). The box size was set to accommodate ligands with length up to 12 Å.

The standard precision (SP) mode of Glide ligand docking was used. The lowest-energy galactose conformation was used as input, although the ligand was fully flexible during docking, including sampling of ring conformations. Up to 100 poses underwent post-docking minimization with strain-correction terms applied. The pose with the best docking score, "gscore" (−5.47), is also the pose with the best "emodel" score (−41.76); its "einternal" score is 0.69. The docked conformation is as expected: the oxygen of the C-6 hydroxyl is coordinated with the Cu^{2+} at 2.4 Å, and the five atoms of transition state (O radical, Cu^{2+}, C_6O, C_6C, C_6H) are mostly planar (dihedral values: 3.9°, −14.5°, 18.4°, −15.1°, 3.0°).

Docking Glucose into Active Site of a Known Variant

An approximate model of glucose docked in a known glucose 6-oxidase variant was generated to provide an initial conformation of glucose in the active site, which was then varied during enzyme redesign. A model of enzyme variant M-RQWY (US patent 7,220,563), with mutations W290F, Y329R, R330K, and Q406T, was generated with the mutated side chains in default conformations and then Prime side-chain refinement was applied to residues 290, 326, 329, 330, and 406. Receptor grid generation and docking of glucose followed the same procedures as above for galactose. The pose with the best gscore (−4.73), with emodel of −22.76 and einternal of 10.07, does not have C-6 at the active site. Instead, the pose with the best emodel (−33.45) has the native-like expected conformation, with gscore of −4.41 and einternal of 1.35.

Enzyme Redesign

Version 2.2.0 of the Rosetta software suite (Rosetta Commons, UW TechTransfer, Seattle, WA, USA) was used to redesign the active-site side chains of galactose 6-oxidase to interact with glucose (Meiler and Baker, 2006). (Updated versions of the software are now available with enhanced capabilities; Kaufmann et al. [2010].) First, the Schrödinger-prepared wild-type structure was modified to conform to Rosetta as follows. The active-site Cu^{2+} was changed to a water oxygen because the metal ion could not be held fixed while the galactose or glucose ligand was docked. The deprotonated Y495 was modeled as a hydroxyl with its hydrogen pointed away from galactose (C-O-H angle = 72°; C-C-O-H dihedral = −90°). The Y272 oxygen radical was modeled as a hydroxyl with its hydrogen pointed away from galactose (C-O-H angle = 72°; C-C-O-H dihedral = 0°). The C228-Y272 crosslink was modeled as a standard-free cysteine and tyrosine, maintaining the same heavy-atom conformations but with overlapping hydrogens. These approximations were acceptable because the clashing atoms and poor bond geometries were held constant during design, maintaining an unfavorable but constant energy from wild-type to enzyme variants, and the parts of the amino acids that faced the active site presented the correct or similar chemical functionalities. In addition, all histidines were changed to one of two neutral tautomers, and the acetate and sodium ion were removed; these changes were acceptable because they occurred far from the active site and did not affect active-site conformations or energies. Finally, galactose and glucose were parameterized with the following atom types: the five ring carbons (C1–C5) = CH1; one hydroxyl carbon (C6) = CH2; six oxygens = OH; five hydroxyl hydrogens = Hpol; and seven nonpolar hydrogens = Hapo.

Galactose 6-oxidase was redesigned to interact with glucose using a protocol that sampled mutations and side-chain conformations at the 12 selected active-site positions and sampled variations in the translation and rotation of the docked glucose. Backbone atoms, non-designed side chains, and the internal conformations of glucose were held fixed. The initial glucose conformation from which perturbations were made was taken from the preliminary docking of glucose into M-RQWY; the protein side-chain conformations were unused. Two designs were executed in parallel, with and without water 39 removed from the input structure. The simulated annealing procedure was modified to anneal to a higher temperature ("lowtemp," which represents kT, was changed in SimAnnealerBase.cc from 0.3 to 30.0 Rosetta energy units) to access a larger space of low-energy results and a greater variety of designed sequences, rather than to aim for the one lowest-energy sequence-structure combination. We found that this modification produced results in which the lowest-found energy was as low as that from an unmodified design, while greatly expanding the diversity of designed sequences in total. In addition, wild-type galactose 6-oxidase was redesigned around galactose, and M-RQWY was redesigned around glucose as controls.

One of the 12 design positions, W290, contributes both to recognition, through a hydrogen bond to the substrate ring oxygen, and to catalysis, due to its ring stacking on residues C228-Y272. The published W290H mutation, which makes a similar hydrogen bond to substrate in our models, slightly improves K_m in galactose 6-oxidase but greatly decreases k_{cat} (Rogers et al., 2007). To avoid a catalytically disruptive mutation at position 290, the designed library was constrained to either the wild-type Trp or to Phe, a mutation predicted to accommodate glucose and previously shown to support catalysis (Sun et al., 2002; Rogers et al., 2007). After preliminary designs with all amino acids (“ALLAA”) at the 12 design positions (and “NATRO” to constrain all other positions), design position W290 was limited to only W or F amino acids for final designs.

The command-line options to Rosetta were: -s input.pdb -design -ligand -fixbb -design_dock -dock_pert 0.2 0.2 5 -resfile resfile -read_all_chains -termini -find_disulf -use_input_sc -ex1 -ex2 -ex3 -ex4 -exOH -extrachi_cutoff 1 -try_both_his_tautomers -no_optH -read_hetero_h2o -use_electrostatic_repulsion -pdbout output -output_pdb_gz -profile -write_interaction_graph -ig_file igfile -ndruns 10000. This produced an 837 MB igfile and designed with 24,402 rotamers.

Each output structure contains intra-protein scores (stability) and protein-ligand interaction scores (binding). Furthermore, stability and binding are each composed of subterms such as van der Waals packing interactions, solvation, and hydrogen bonding. When multiple conformations of the same amino acid sequence were found, the conformation with the most favorable total stability was kept. From the 10,000 outputs from each of the two designs, 1,469 and 1,372 unique sequences were found, respectively, with 2,379 unique sequences overall.

The sequences with lowest total energy were highly mutated and unusual, e.g., exhibiting excess charged residues compared to typical proteins. Additional scores were calculated for each sequence based on physically related terms for comparison to the controls. The following five scores (and their components in parentheses) were defined: Stability$_{total}$ (bk_tot); Stability$_{vdW}$ (fa_atr + fa_rep); Stability$_{elec}$ (hb_srbb + hb_lrbb + fa_sol + fa_h2o_sol + gb_elec + fa_h2o + fa_h2o_hb + hb_sc); Binding$_{vdW}$ (lig_rep + lig_atr); and Binding$_{elec}$ (lig_sol + lig_cou + lig_hb). The output sequences were filtered based on the following five criteria (with the score from the M-RQWY-glucose control shown in parentheses): Stability$_{total}$ < -1132 (-1133.19); Stability$_{vdW}$ < -1609 (-1610.98); Stability$_{elec}$ < 230 (227.28); Binding$_{vdW}$ < -4 (-6.14); and Binding$_{elec}$ < -2 (-3.84).

Modeling Enzyme Variants with Docked Glucose

Models of experimentally characterized enzyme variants with docked glucose were generated using Schrödinger software. Because side-chain and glucose conformations are codependent, but Glide docking requires a fixed receptor conformation, an iterative procedure was applied with side-chain modeling followed by glucose docking and repeated side-chain refinement and docking. Multiple candidate conformations of docked glucose were carried forward in parallel.

In the final round of side-chain modeling and ligand re-docking, six related, candidate conformations of docked glucose, generated from preliminary modeling, were added to each active site, and Prime side-chain prediction was used to refine the following active-site atoms of each enzyme-glucose pair: side chains of positions 194, 290, 326, 329, 330, 383, 406, 407, 441, 463, and 464; waters 2, 39, 49, and 51; and glucose. Select side-chain torsional angles were modified by hand, energies were recalculated with Prime, and conformations with decreased Prime energy were retained. The refined atoms were then minimized in Prime.

Glide receptor grids were generated with a positional constraint (a 2.6 Å sphere centered at the Cu^{2+} ion) to force the C-6 hydroxyl to the catalytic residues during docking. Docking was performed with no ligand-atom scaling and a positional constraint to pin the C-6 hydroxyl oxygen to the Cu^{2+} ion (custom feature: atom 1 of [O][C;H2]); constraint satisfaction was tested only after docking. Outputs from docking were clustered with a modified RMS deviation of 0.05 Å. The best structure for each enzyme variant was selected based on combined best emodel and einternal scores.

Strains, Growth Media, and Plasmids

E. coli strain DH10B [F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*Δ139 Δ(*ara*, *leu*)7697 *gal*U *gal*K λ- *rps*L (Str^R) *nup*G] was used for all molecular biology manipulations and as host for the library screening. BL21 Star (DE3) [F$^-$ *omp*T *hsd*S$_B$ ($r_B^- m_B^-$) *gal dcm rne*131 (DE3)] was used as host for expression of pET21b (Novagen, Madison, WI, USA). Competent cells of both strains were purchased from Invitrogen Corporation (Carlsbad, CA, USA). Cultures were propagated in Luria-Bertani (LB). All molecular biology manipulations were performed according to standard practices (Sambrook and Russell, 2001). The galactose oxidase amino acid sequence described throughout as “wild-type” is the A3.E7 variant, previously evolved for improved stability and expression efficiency in *E. coli* (Sun et al., 2001).

Synthesis of DNA Libraries

The wild-type gene was split into N- and C-terminal constant regions and an internal variable region containing the 12 designed positions (Figure 2). The base construct comprised the constant regions and an internal stuffer with type IIS AarI sites (underlined) to excise the stuffer: [XbaI]-ATG-[codons 1–280]-CATGGCAGGTG-[SalI]-[NotI]-[XhoI]-CACCTGCCATG-[codons 473–639]-TAA-[HindIII]. The base construct was synthesized in pUC19 using Codon Devices’ PCR-assembly methods and error-correction technology (Carr et al., 2004), subcloned into the pTrc99A expression vector, and sequence verified.

The variable, library region of the gene was constructed by PCR-based assembly of 18 partially overlapping oligonucleotides (Tian et al., 2004). Subsets of eight and ten oligonucleotides, or segments, were first assembled and then the two subassemblies were combined. The 12 designed positions were excluded from the overlaps between the 18 segments. The average segment length was 58.9 base pairs, with a minimum of 29 base pairs and maximum of 124 base pairs. The entire assembly was flanked by AarI sites oriented to be excised upon cleavage and generate four-base overhangs compatible with the similarly linearized base construct. Infusion-cloning tags were also included for an alternate, unused cloning strategy. The assembled sequence, GACGGCCAGT-[EcoRI]-CACCTGCCATGAGAC-[codons 281–472]-CCTGGTACGCAGGTGCCTGA-[BamHI]-TCTAGAGTCG, was 643 base pairs in length, becoming 584 base pairs (including overhangs) after AarI cleavage.

For the designed library, pairs of forward and reverse-complementary oligonucleotides (2 + 48 + 14 + 8 = 72 total pairs, which create 10,752 combinations) were mixed to form duplexes, and 10 μM of each duplex was error corrected by incubating with MutS (6 μg/μl) at 60°C for 20 min and flowing through a nitrocellulose filter (Carr et al., 2004; Lipovšek et al., 2009). Separate oligonucleotides encoding each designed amino acid combination were independently duplexed and filtered and then mixed equimolar. For the combinatorial library, a single oligonucleotide was used for each of the four variable segments. At each designed position a degenerate codon was chosen to encode all of the designed amino acids while minimizing the number of stop codons and non-designed amino acids and avoiding all rare codons for *E. coli* expression. Each library was inserted into the base construct by restriction cloning and transformed into DH10B.

Sequence Validation of Libraries

Clones from the combinatorial and designed libraries, 288 each, were sequenced with forward and reverse primers reading from the base construct into the inserted variable region. Sequencing traces were called with Phred (Ewing et al., 1998) and aligned using NCBI BLAST to the wild-type sequence and separately to each construction oligonucleotide. The forward and reverse reads of each clone were merged with the BLAST alignments, adding the Phred quality values if the two reads agreed and using the higher-scoring value if they did not. Positions with Phred score determined in this matter of 15 or greater were considered “high-quality” positions; clones with all positions of “high-quality” were subjected to further analysis. BLAST alignments to the construction oligonucleotides were used to identify nucleotide substitutions in the variable regions, and alignments to the wild-type sequence were used to identify nucleotide substitutions in the constant regions.

For the designed library, nucleotide substitutions at designed positions within a single construction oligonucleotide were classified as crossover if a perfect match for the clone sequence could be found by the in silico recombination of any pair of construction oligonucleotides across any 12 shared nucleotides. This method of determining crossover is conservative in that it does not allow for multiple recombinations or for recombination plus mutation in the same construction oligonucleotide.

Library Screening and Characterization of Top Hits

Detailed experimental methods are described in the Supplemental Experimental Procedures. Library screening and activity assays were based on a colorimetric method that detects hydrogen peroxide generation during catalysis (Baron et al., 1994). Single colonies were inoculated into 96-well deep-well microplates (2 ml well volume, VWR) containing 1 ml LB and 0.1 mg/ml carbenicillin. Cell lysates were mixed with horseradish peroxidase (Sigma-Aldrich), ABTS (2,2′-azino-bis[3-ethylbenzothiazoline-6-sulfonic acid]; Sigma-Aldrich), and glucose (Mallinckrodt Baker), and absorbance was monitored at 405 nm. Hits were identified after 24, 6, or 0.5 hr of incubation at room temperature for the first, second, and third rounds of screening, respectively.

The ten unique top hits (Des3-8, a duplicate of Des3-1, was excluded), the next-best hit (Des3-5), wild-type, and M-RQW were subcloned into the 6× His-tag-containing pET21b and transformed into BL21 Star (DE3). Proteins were purified using the ProBond Purification System (Invitrogen) and each ran as a single band with the expected molecular weight (70 kDa). Protein concentrations were measured using sandwich ELISA against the T7 and hexahistidine tags. The rate-substrate-concentration curves do not exhibit typical saturation behavior up to the practical solubility limit of glucose (0.8 M) and galactose (0.4 M), implying that the K_m values are on the order of molar and are impractical to measure. Instead, the k_{cat}/K_m ratio was determined from the slope of the specific activity-substrate-concentration curve, assuming that K_m is much greater than the substrate concentration (glucose: 4–300 mM; galactose: 2–70 mM).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and five tables and can be found with this article online at doi:10.1016/j.chembiol.2010.10.012.

ACKNOWLEDGMENTS

We thank Kathy Galle for help in coordinating the research; Paulo Gouveia, Christina Lee, Gina Prophete, and Farhana Syed for colony-picking and sequencing support; and Ashagari Habtegabriel, Roland Telson, and Ben Cerio for the synthesis of the base construct. We thank Bracken King for help in analyzing the coupling between design positions. We thank undergraduate researchers Joseph Roy-Mayhew, Theresa Dinio, and Amanda Lanza for assistance with library screening. We are also grateful to Professor Frances Arnold (California Institute of Technology) for the generous donation of the galactose oxidase mutants A3.E7 and M-RQW.

Work performed at Codon Devices, Inc., was funded by the Codon Devices, Inc., research budget. Work performed at MIT was supported by grants from the Office of Naval Research Young Investigator Program (Grant No. N000140510656) and the National Science Foundation through the Synthetic Biology Engineering Research Center (Grant No. EEC-0540879), as well as with start-up funds provided by MIT to K.L.J.P.

Received: March 21, 2010
Revised: October 19, 2010
Accepted: October 19, 2010
Published: December 21, 2010

REFERENCES

Aden, A., Bozell, J., Holladay, J., White, J., and Manheim, A. (2004). Top Value-Added Chemicals from Biomass. Volume I: Results of Screening for Potential Candidates from Sugars and Synthesis Gas. T. Werpy and G. Petersen, eds. (Golden, CO: Pacific Northwest National Laboratory and National Renewable Energy Laboratory).

Amin, N., Liu, A.D., Ramer, S., Aehle, W., Meijer, D., Metin, M., Wong, S., Gualfetti, P., and Schellenberger, V. (2004). Construction of stabilized proteins by combinatorial consensus mutagenesis. Protein Eng. Des. Sel. *17*, 787–793.

Atsumi, S., Hanai, T., and Liao, J.C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature *451*, 86–89.

Bailey, J.E. (1991). Toward a science of metabolic engineering. Science *252*, 1668–1675.

Barderas, R., Desmet, J., Timmerman, P., Meloen, R., and Casal, J.I. (2008). Affinity maturation of antibodies assisted by in silico modeling. Proc. Natl. Acad. Sci. USA *105*, 9029–9034.

Baron, A.J., Stevens, C., Wilmot, C., Seneviratne, K.D., Blakeley, V., Dooley, D.M., Phillips, S.E.V., Knowles, P.F., and McPherson, M.J. (1994). Structure and mechanism of galactose oxidase. The free radical site. J. Biol. Chem. *269*, 25095–25105.

Carr, P.A., Park, J.S., Lee, Y.J., Yu, T., Zhang, S.G., and Jacobson, J.M. (2004). Protein-mediated error correction for de novo DNA synthesis. Nucleic Acids Res. *32*, e162.

Chen, C.Y., Georgiev, I., Anderson, A.C., and Donald, B.R. (2009). Computational structure-based redesign of enzyme activity. Proc. Natl. Acad. Sci. USA *106*, 3764–3769.

Chica, R.A., Doucet, N., and Pelletier, J.N. (2005). Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr. Opin. Biotechnol. *16*, 378–384.

Damborsky, J., and Brezovsky, J. (2009). Computational tools for designing and engineering biocatalysts. Curr. Opin. Chem. Biol. *13*, 26–34.

Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J., and Keasling, J.D. (2009). Synthetic protein scaffolds provide modular control over metabolic flux. Nat. Biotechnol. *27*, 753–759.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

Fox, R.J., and Huisman, G.W. (2008). Enzyme optimization: moving from blind evolution to statistical exploration of sequence-function space. Trends Biotechnol. *26*, 132–138.

Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S., et al. (2007). Improving catalytic function by ProSAR-driven enzyme evolution. Nat. Biotechnol. *25*, 338–344.

Gerlt, J.A., and Babbitt, P.C. (2009). Enzyme (re)design: lessons from natural evolution and computation. Curr. Opin. Chem. Biol. *13*, 10–18.

Hayes, R.J., Bentzien, J., Ary, M.L., Hwang, M.Y., Jacinto, J.M., Vielmetter, J., Kundu, A., and Dahiyat, B.I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. Proc. Natl. Acad. Sci. USA *99*, 15926–15931.

Ito, N., Phillips, S.E., Stevens, C., Ogel, Z.B., McPherson, M.J., Keen, J.N., Yadav, K.D., and Knowles, P.F. (1991). Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. Nature *350*, 87–90.

Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F., et al. (2008). De novo computational design of retro-aldol enzymes. Science *319*, 1387–1391.

Kang, S.G., and Saven, J.G. (2007). Computational protein design: structure, function and combinatorial diversity. Curr. Opin. Chem. Biol. *11*, 329–334.

Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H., and Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry *49*, 2987–2998.

Koder, R.L., Anderson, J.L.R., Solomon, L.A., Reddy, K.S., Moser, C.C., and Dutton, P.L. (2009). Design and engineering of an O_2 transport protein. Nature *458*, 305–309.

Lappe, M., Bagler, G., Filippis, I., Stehr, H., Duarte, J.M., and Sathyapriya, R. (2009). Designing evolvable libraries using multi-body potentials. Curr. Opin. Biotechnol. *20*, 437–446.

Leemhuis, H., Kelly, R.M., and Dijkhuizen, L. (2009). Directed evolution of enzymes: library screening strategies. IUBMB Life *61*, 222–228.

Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C., and Minshull, J. (2007). Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol. *7*, 16.

Lipovšek, D., Mena, M., Lippow, S.M., Basu, S., and Baynes, B.M. (2009). Library construction for protein engineering. In Protein Engineering and Design, S.J. Park and J.R. Cochran, eds. (Boca Raton, FL: CRC Press), pp. 83–108.

Lippow, S.M., Wittrup, K.D., and Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. Nat. Biotechnol. *25*, 1171–1176.

Meiler, J., and Baker, D. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins *65*, 538–548.

Moon, T.S., Yoon, S.H., Lanza, A.M., Roy-Mayhew, J.D., and Prather, K.L.J. (2009). Production of glucaric acid from a synthetic pathway in recombinant *Escherichia coli*. Appl. Environ. Microbiol. *75*, 589–595.

Moore, G.L., and Maranas, C.D. (2004). Computational challenges in combinatorial library design for protein engineering. AIChE J. *50*, 262–272.

Murphy, P.M., Bolduc, J.M., Gallaher, J.L., Stoddard, B.L., and Baker, D. (2009). Alteration of enzyme specificity by computational loop remodeling and design. Proc. Natl. Acad. Sci. USA *106*, 9215–9220.

Otey, C.R., Landwehr, M., Endelman, J.B., Hiraga, K., Bloom, J.D., and Arnold, F.H. (2006). Structure-guided recombination creates an artificial family of cytochromes P450. PLoS Biol. *4*, 789–798.

Prather, K.L.J., and Martin, C.H. (2008). De novo biosynthetic pathways: rational design of microbial chemical factories. Curr. Opin. Biotechnol. *19*, 468–474.

Rogers, M.S., Tyler, E.M., Akyumani, N., Kurtis, C.R., Spooner, R.K., Deacon, S.E., Tamber, S., Firbank, S.J., Mahmoud, K., Knowles, P.F., et al. (2007). The stacking tryptophan of galactose oxidase: a second-coordination sphere residue that has profound effects on tyrosyl radical behavior and enzyme catalysis. Biochemistry *46*, 4606–4618.

Rothlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. Nature *453*, 190–195.

Sambrook, J., and Russell, D.W. (2001). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Shivange, A.V., Marienhagen, J., Mundhada, H., Schenk, A., and Schwaneberg, U. (2009). Advances in generating functional diversity for directed protein evolution. Curr. Opin. Chem. Biol. *13*, 19–25.

Singh, J., and Gupta, K.P. (2007). Induction of apoptosis by calcium D-glucarate in 7,12-dimethyl benz [a] anthracene-exposed mouse skin. J. Environ. Pathol. Toxicol. Oncol. *26*, 63–73.

Smith, C.A., and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. *380*, 742–756.

Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H., and Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature *437*, 512–518.

Sun, L., Petrounia, I.P., Yagasaki, M., Bandara, G., and Arnold, F.H. (2001). Expression and stabilization of galactose oxidase in *Escherichia coli* by directed evolution. Protein Eng. *14*, 699–704.

Sun, L., Bulter, T., Alcalde, M., Petrounia, I.P., and Arnold, F.H. (2002). Modification of galactose oxidase to introduce glucose 6-oxidase activity. ChemBioChem *3*, 781–783.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. (2004). Accurate multiplex gene synthesis from programmable DNA microchips. Nature *432*, 1050–1054.

Treynor, T.P., Vizcarra, C.L., Nedelcu, D., and Mayo, S.L. (2007). Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. Proc. Natl. Acad. Sci. USA *104*, 48–53.

Voigt, C.A., Mayo, S.L., Arnold, F.H., and Wang, Z.G. (2001). Computational method to reduce the search space for directed protein evolution. Proc. Natl. Acad. Sci. USA *98*, 3778–3783.

Wachter, R.M., and Branchaud, B.P. (1996). Molecular modeling studies on oxidation of hexopyranoses by galactose oxidase. An active site topology apparently designed to catalyze radical reactions, either concerted or stepwise. J. Am. Chem. Soc. *118*, 2782–2789.

Walaszek, Z., Szemraj, J., Hanausek, M., Adams, A.K., and Sherman, U. (1996). D-glucaric acid content of various fruits and vegetables and cholesterol-lowering effects of dietary D-glucarate in the rat. Nutr. Res. *16*, 673–681.

Whittaker, J.W. (2003). Free radical catalysis by galactose oxidase. Chem. Rev. *103*, 2347–2363.

Yoshikuni, Y., Dietrich, J.A., Nowroozi, F.F., Babbitt, P.C., and Keasling, J.D. (2008). Redesigning enzymes based on adaptive evolution for optimal function in synthetic metabolic pathways. Chem. Biol. *15*, 607–618.

Zhang, K., Sawaya, M.R., Eisenberg, D.S., and Liao, J.C. (2008). Expanding metabolism for biosynthesis of nonnatural alcohols. Proc. Natl. Acad. Sci. USA *105*, 20653–20658.