Chapter 3

Active Site Engineering of P450 BM3 for Small Alkane Hydroxylation

A. Abstract

To compare the functional richness of mutagenesis libraries generated by error-prone PCR, site-saturation mutagenesis, combinatorial active site saturation with a reduced set of amino acids and structure-based computational library design, seventeen mutagenesis libraries of cytochrome P450 BM3 were designed and constructed. Each library was evaluated for the fraction of variants that had acquired activity for demethylation of dimethyl ether and selected variants were also characterized for propane and ethane hydroxylation. Among these libraries, the ones generated by combinatorial active site saturation with a reduced set of amino acids displayed both a higher fraction of functional variants and variants with higher activity than both an error-prone PCR library with a similar mutation rate (2.1 mutation/protein) and site-saturation mutagenesis libraries targeting the same three residues. The most effective library design for generating variants for both dimethyl ether demethylation and small alkane hydroxylation was the CRAM algorithm developed and described here. While none of the isolated variants of this study achieved the level of specialization for propane hydroxylation previously obtained through multiple rounds of mutagenesis and selection, the levels of activity achieved by these variants show that jumps in sequence space from a specialized enzyme to generalist variants with desired functions are possible through various semi-rational mutagenesis approaches.

B. Introduction

Over the past decades, directed protein evolution has become a versatile tool for both the engineering of protein properties to meet industrial demands (*1 – 2*) and the exploration of structure-function relationships of biocatalysts (*3 – 4*). Using iterative cycles of sequence diversification and functional selection, enzyme variants have been reported with a variety of improved protein functions such as binding, enantioselectivity, thermostability, and altered substrate specificity (*5 – 8*). Recent advances in computational modeling (*9*), combined with the increased availability of structural and sequence information, have resulted in an expansion in the number of mutagenesis techniques and methods employed in directed protein evolution, such as SCOPE (*10*), CASTing (*11*), ISM (*12*), ISOR (*13*), and other structure-based computational library designs (*14 – 16*).

These semi-rational mutagenesis approaches aim to generate functionally enriched libraries by targeting mutations to specific regions of a protein such as an enzyme's active site or a protein-ligand interface determined to be important by structural or sequence analysis (*17*). While the viability of all these methods has been demonstrated by successful examples of their implementation, there have been very few attempts at comparing them with more traditional mutagenesis techniques such as error-prone PCR (EP-PCR) (*18*) or site-saturation mutagenesis (*19*). Part of the difficulty in comparing mutagenesis approaches is the inherent stochastic element of the directed evolution experiment. Since the outcome of such experiments relies on the specific choice of variants selected as parent for the subsequent round of evolution, repeating the same directed evolution experiment can lead to different sequence solutions. Therefore, comparing only the best variants generated through different mutagenesis techniques provides only anecdotal evidence for a method's efficacy. A more informative comparison of mutagenesis techniques would be to evaluate the range of acquired activities and fraction of functional variants generated by each method in a single round of mutagenesis and screening for a defined function, using an identical starting point.

Here, we evaluate four mutagenesis approaches, (1) random mutagenesis by EP-PCR, (2) site-saturation mutagenesis (SSM), (3) Combinatorial Active Site saturation Test with a reduced set of amino acids (reduced CASTing) (*20*), and (4) two-structure-based computational library design approaches (*16*), for their ability to generate cytochrome P450 BM3 (BM3) (*21*) variants with activity for demethylation of dimethyl ether (DME) and hydroxylation of propane and ethane. BM3 is a self-sufficient fusion protein composed of a P450 monooxygenase and an NADPH diflavin reductase that hydroxylates C_{12}-C_{20} fatty acids as its preferred substrates (*22*) and does not have any detectable activity on these three substrates. Previous efforts in our lab generated variant $P450_{PMO}$ (PMO) (*23*) through 16 rounds of mutagenesis with 23 mutations with activity on all three substrates. In addition, PMO accepts propane as its preferred alkane substrate. The evolutionary strategy of enhancing the promiscuous alkane hydroxylation activity of BM3 and subsequent variants used to obtain PMO mimicked a natural evolution pathway and demonstrated that these functions can be acquired upon iterative rounds of mutagenesis and screening. The existing evolutionary lineage from BM3 to PMO allows us to compare the variants generated by these different mutagenesis approaches to determine the degree of specialization that can be obtained through semi-rational library design.

C. Results

C.1. Library design and composition

We designed and constructed 17 mutagenesis libraries of BM3 using EP-PCR (1), SSM (10), reduced CASTing (4) and structure-based computational methods (2), with library compositions listed in Table 3.1 and as described in Chapter 8.D. The residues targeted for SSM, reduced CASTing, and computationally-guided libraries were determined using the crystal structure of the BM3 heme domain bound with N-palmitoyl glycine, PDB:1JPZ (*24*). We identified ten residues (A74, L75, V78, A82, F87, L181, A184, L188, A328, and A330) as mutagenesis targets for both SSM and structure-based computational library design. These residues fall within various substrate recognition sites (SRS) identified for class II P450s (*25*), see Figure 3.1, and were selected over adjacent candidates because their side chains are oriented directly toward the active site. In addition, six of these ten residues (A74, V78, A82, A184, L188, and A328) were previously mutated in the evolutionary path from BM3 to PMO (*23*) and thus were known sites of beneficial mutations.

For the reduced CASTing libraries, mutations were targeted to three residues, V78, A82, and A328, as they were previously found to shift BM3's substrate specificity toward smaller substrates (*26 – 27*). The allowed amino acid cassette was restricted to L, I, M, V, F, A, and W with the use of degenerate codons following the intuition that introducing amino acids with large hydrophobic side chains should improve activity for smaller substrates. Four libraries were constructed, three of which mutated two of the three residues pairwise, and one library which mutated all three targets together. This library will be referred to as the three-site reduced CASTing library.

Instead of constructing a structural model of the BM3 transition state with small alkanes, we elected to use computational tools to find sequences that would maintain the substrate-bound conformation in the absence of substrate for the structure-based computational library design. This choice was made because a structure of the BM3 substrate-enzyme complex in a reactive conformation is not available (*28 – 29*), and the small alkane substrates lack functional groups that would aid the computational design in stabilizing potential transition states. The two designed libraries, C^{orbit} and CRAM, mutated the same ten residues as the site-saturation libraries, but only allowed for two possible amino acids at each position as determined by each algorithm (see Table 3.1). The C^{orbit} algorithm, which has previously been successful in creating diverse libraries of green fluorescent protein (*16*), models protein stability as a surrogate for protein function. In this approach, the selected mutations appeared more frequently in sequences predicted to support the desired enzyme fold. The second approach, which we termed the CRAM algorithm, aggressively packed the active site by computationally determining the largest tolerated amino acid substitutions at each of the ten target positions.

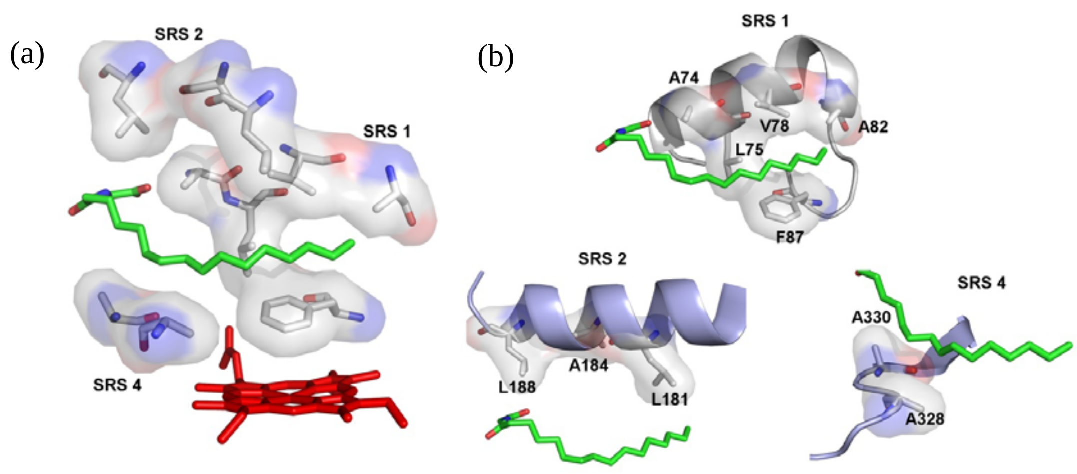


Figure 3.1: (a) The ten residues targeted for site-saturation mutagenesis and structure-based computational library design chosen based on their proximity to the bound N-palmitoyl glycine substrate in the 1JPZ structure of BM3 (*24*); (b) Close-up of SRS 1, 2, and 4 (*25*), heme shown in red. N-palmitoyl glycine is shown in green.

Table 3.1: Library designs and properties

Amino acid	EP-PCR	Site-saturation libraries					Reduced CASTing libraries				Computationally designed libraries	
		V78	A82	A328	A330	Remaining residues	V78/A82	V78/A328	A82/A328	V78/A82/A328	CRAM	Corbit
A74	-	-	-	-	-	-	-	-	-	-	LW	AV
L75	-	-	-	-	-	-	-	-	-	-	LF	LF
V78	-	ALL[a]	-	-	-	-	LIVMFAW	LIVMFAW	-	LIVMFAW	IF	VL
A82	-	-	ALL	-	-	-	LIVMFAW	-	LIVMFAW	LIVMFAW	VL	AS
F87	-	-	-	-	-	-	-	-	-	-	FA	FA
L181	-	-	-	-	-	-	-	-	-	-	LW	LF
A184	-	-	-	-	-	-	-	-	-	-	AV	AT
L188	-	-	-	-	-	-	-	-	-	-	LW	LW
A328	-	-	-	ALL	-	-	-	LIVMFAW	LIVMFAW	LIVMFAW	VF	AF
A330	-	-	-	-	ALL	-	-	-	-	-	LW	AV
$\langle M_{AA} \rangle$[b]	2.1	0.9	0.9	0.9	0.9	0.9	1.7	1.7	1.7	2.6	7.5	5
Library size (number of unique sequences)	1,166[c]	32	32	32	32	192	49	49	49	343	1,024	1,024
Clones screened	1,408	91	91	91	91	546	176	176	176	1,056	2,548	2,548
Library coverage (%)[d]	N/A	94	94	94	94	94	97	97	97	95	92	92
Fraction of folded variants[e]	0.52	0.9	0.96	0.62	0.96	0.95	0.91	0.95	0.97	0.94	0.75	0.84
Fraction of active variants[f]	0.36	0.26	0.21	0.36	0.12	0.01	0.22	0.71	0.74	0.54	0.34	0.31
Average DME activity	0.03	0.01	0.02	0.04	0.01	0.00	0.09	0.24	0.28	0.17	0.19	0.12
Standard deviation of DME activity	0.05	0.03	0.05	0.06	0.05	0.00	0.09	0.23	0.23	0.18	0.38	0.21

[a] All: 20 amino acids as encoded by the NNK codon used in the site-saturation mutagenesis. [b] $\langle M_{AA} \rangle$: average mutation rate. [c] Estimation for the number of unique sequences sampled from 1,408 clones of the EP-PCR library was determined using PEDEL (*30*).[d] Library coverage was determined using GLUE (*30*). [e] Fraction of folded variants were determined by CO-binding spectroscopy, corrected for stop codon presence.
[f] Fraction of variants active for DME (of all variants) were determined in cell-free extract, corrected for background Purpald ® oxidation.

C.2. Library characterization for DME demethylation and protein folding

All 17 libraries were characterized for both DME demethylation activity and protein folding using high throughput assays; the results are summarized in Table 3.1. DME demethylation activity was quantified colorimetrically through the use of a dye, Purpald®, which reacts with the formaldehyde product of the P450 reaction to form a purple adduct with a UV/Vis peak maximum at 550 nm (*27*). The protein folding of each variant was determined by CO-binding difference spectra of the cell-free extract (*31*).

We screened 1,408 variants from a *Taq* polymerase generated EP-PCR library with an error rate of 2.1 amino acid substitutions/protein, which corresponds to sampling ~ 1,166 unique sequences, as calculated by the PEDEL algorithm for estimating the diversity of EP-PCR libraries (*30*). We found 52% of the variants to be folded (CO binding difference > 0.01) and 36% of the variants active for DME (Abs 550 > 0.13, corresponding to background Purpald® oxidation). The ten site-saturation libraries constructed with NNK codons, which encode for all 20 unique single-mutants, were screened to 94% library coverage (91 clones). Nine of the ten libraries contained a high fraction of folded variants, > 90%, with the library at A328 containing only 62% of folded variants. Variants that acquired DME activity were found in libraries targeting residues V78 (15%), A82 (9%), A328 (34%), and A330 (20%).

Given the high mutational tolerance of these active site residues for protein folding, it was unsurprising to find that the reduced CASTing libraries also contained a high fraction of folded variants, > 91%. The pairwise libraries, each having 49 unique members, were screened to 97% library coverage (176 clones), and the three-site library with 343 unique members was screened to 95% library coverage (1,056 clones). The fraction of functional variants varied from 22% for the library mutating V78/A82 to 71% and 74% for libraries mutating V78/A328 and

A82/A328, respectively. The three-site reduced CASTing library had only 54% of variants active for DME demethylation.

Finally, the two structure-based computationally designed libraries C^{orbit} and CRAM, each containing 1,024 unique members, were screened to 92% library coverage (2,548 clones), with 84% and 75% of folded variants, respectively. The fraction of variants with DME demethylation activity was similar between these two libraries, with 34% of CRAM variants and 31% of C^{orbit} variants being functional.

While all four mutagenesis strategies were able to generate variants with DME demethylation activity, the distribution of activity levels varied. The library profiles, i.e., activities of variants plotted in ranked order, for all libraries are shown in Figure 3.2. For simplicity and easier library comparisons, the variants from all ten SSM libraries were grouped and treated as single libraries for this analysis. Likewise, the three pairwise reduced CASTing libraries were also grouped.

Figure 3.2 (a) shows that both the EP-PCR library and the combined SSM libraries generated variants with DME activities up to 0.5 A_{550nm}, after correcting for background Purpald oxidation, with library averages of 0.026 $\pm$ 0.050 and 0.010 $\pm$ 0.039, respectively. These averages reflect the overall low functional richness of these libraries with a majority of both variant populations being inactive. In contrast, the variants of the reduced CASTing libraries exhibit DME activities up to 0.97 A_{550nm}, with library averages of 0.12 $\pm$ 0.19 and 0.17 $\pm$ 0.18 for the pairwise and three-site reduced CASTing libraries, respectively. A comparison of the library profiles of SSM libraries at these three residues, the pairwise, and three-site reduced CASTing libraries, Figure 3.2 (b), shows an increase in number of active variants with higher mutation rate and library size. However, the range of obtained DME activities only increased for the pairwise reduced CASTing libraries compared to SSM libraries. The library profiles of the C^{orbit} and CRAM libraries show even higher DME activity: A_{550nm} of 2.6 was reached. However, as only ~ 30% of the variants are functional, the library averages of 0.18 $\pm$ 0.38 and 0.12 $\pm$ 0.21 are similar to the reduced CASTing libraries.

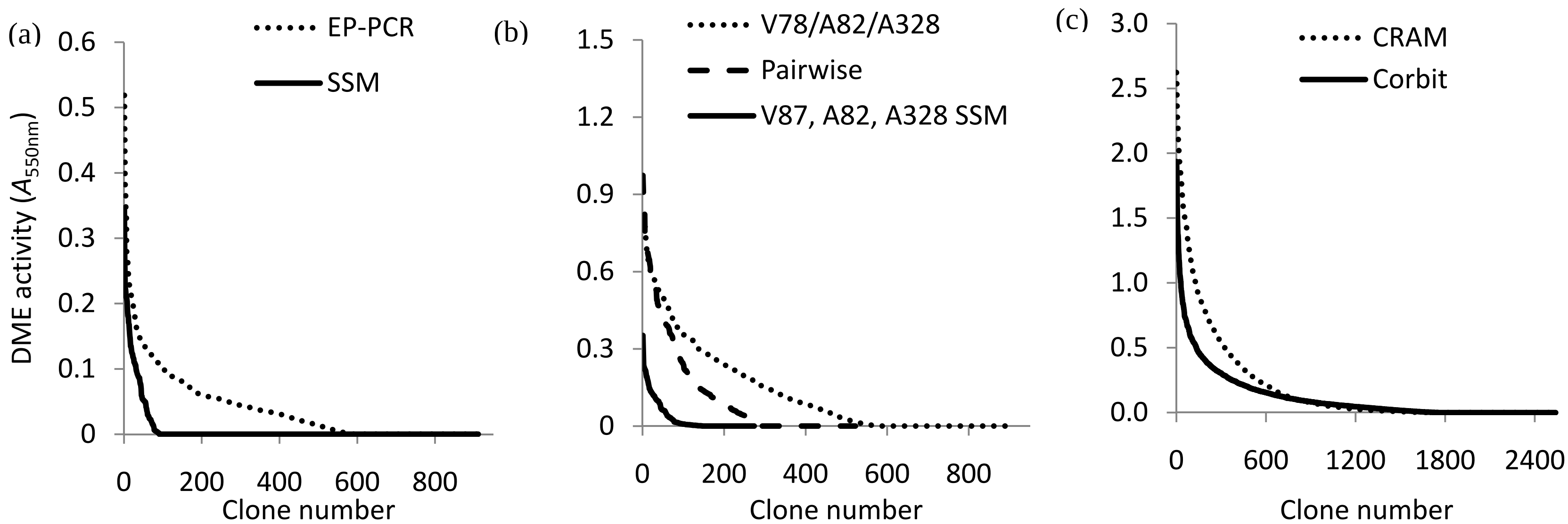


Figure 3.2: Profile of DME activity obtained by mutagenesis libraries with the activity of each variant plotted in ranked order. (a) Variants from the ten site saturation libraries and the top 910 variants of the EP-PCR library. (b) Variants from the site saturation, pairwise, and the complete reduced CASTing libraries targeting residues V78, A82, and A328. (c) Variants from C^{orbit} and CRAM libraries

C.3. Propane and ethane hydroxylation

After the DME demethylation screen was repeated for selected variants from each library in at least duplicate, top-performing variants were purified and characterized for propane and ethane hydroxylation activity (see Appendix A for complete sequence and activity information). Figure 3.3 (a) shows the histogram of propane and ethane turnover number (TON) of variants isolated from these libraries. From the ten SSM libraries, twelve variants were identified supporting propane TON ranging from 120 to 2,200. Mutations at V78 (T, C, S) and A82 (E,Q) located in the B' helix of SRS1 yielded variants with moderate propane activity, 120 to 370 TON. More active variants, > 1,000 propane TON, were obtained with mutations at residues A328 (I, P, L, V) and A330 (L, P, V), which are located in the loop between the J and K helices. The best single active-site variant, A328V, supports 2,200 propane TON with a product formation rate of 7.1 min^{-1} and 8.1% coupling of cofactor consumption.

Six variants were identified from the *Taq* EP-PCR library, supporting 130 to 3,300 propane TON. The best variant, 4F9 (F162L) supports 3,300 propane TON with a product formation rate of 19 min^{-1} and 15% coupling of cofactor consumption. The F162L mutation occurs in the linker between the E and F helices, located outside of the active site. While this residue was not mutated in variants of the PMO lineage, several residues in the adjacent F-helix were mutated, which may suggest the importance of this region for altering substrate specificity.

From the reduced CASTing libraries, nine variants were identified supporting 380 to 4,200 propane TON with only four of the nine variants containing mutations at all three targeted residues. The best variant, WT-A82L-A328V, supports 4,200 propane TON with a product formation rate of 40 min^{-1} and 44% coupling of cofactor consumption. In addition, two of these nine variants, WT-A82L-A328L and WT-A82L-A328V, were also able to hydroxylate ethane, supporting 140 and 200 TON, respectively.

As far more variants from the CRAM and C^{orbit} libraries exhibited high DME demethylation activity compared to the other libraries, we selected the 88 most active variants from each library and screened them for propane and ethane hydroxylation directly as cell-free extracts using the assay outlined in Chapter 5.1. From this screen, 37 variants supporting at least 2,000 propane TON and 100 ethane TON as crude extracts were purified and characterized. All 37 variants were found to support at least 3,500 propane TON with 16 of the variants supporting at least 300 ethane TON as purified enzymes. A much higher number of active variants for propane and ethane hydroxylation was found in the CRAM library, 25 and 13, respectively, compared to the C^{orbit} library, which only produced twelve variants with activity on propane and three variants with ethane activity. The most active CRAM variant was E32 with mutations A74W, V78I, A82L, A184V, L188W, A328F, and A330W. E32 supported 16,800 propane TON and 1,200 ethane TON. The most active C^{orbit} variant, OD2, with mutations A74V, L181F, and A328F, supported 11,600 propane TON and 660 ethane TON. The coupling of cofactor consumption with propanol formation was also determined for a selection of these variants. Most variants exhibited coupling ranging from 36% – 52%, with the best variant, E31, having 68% coupling.

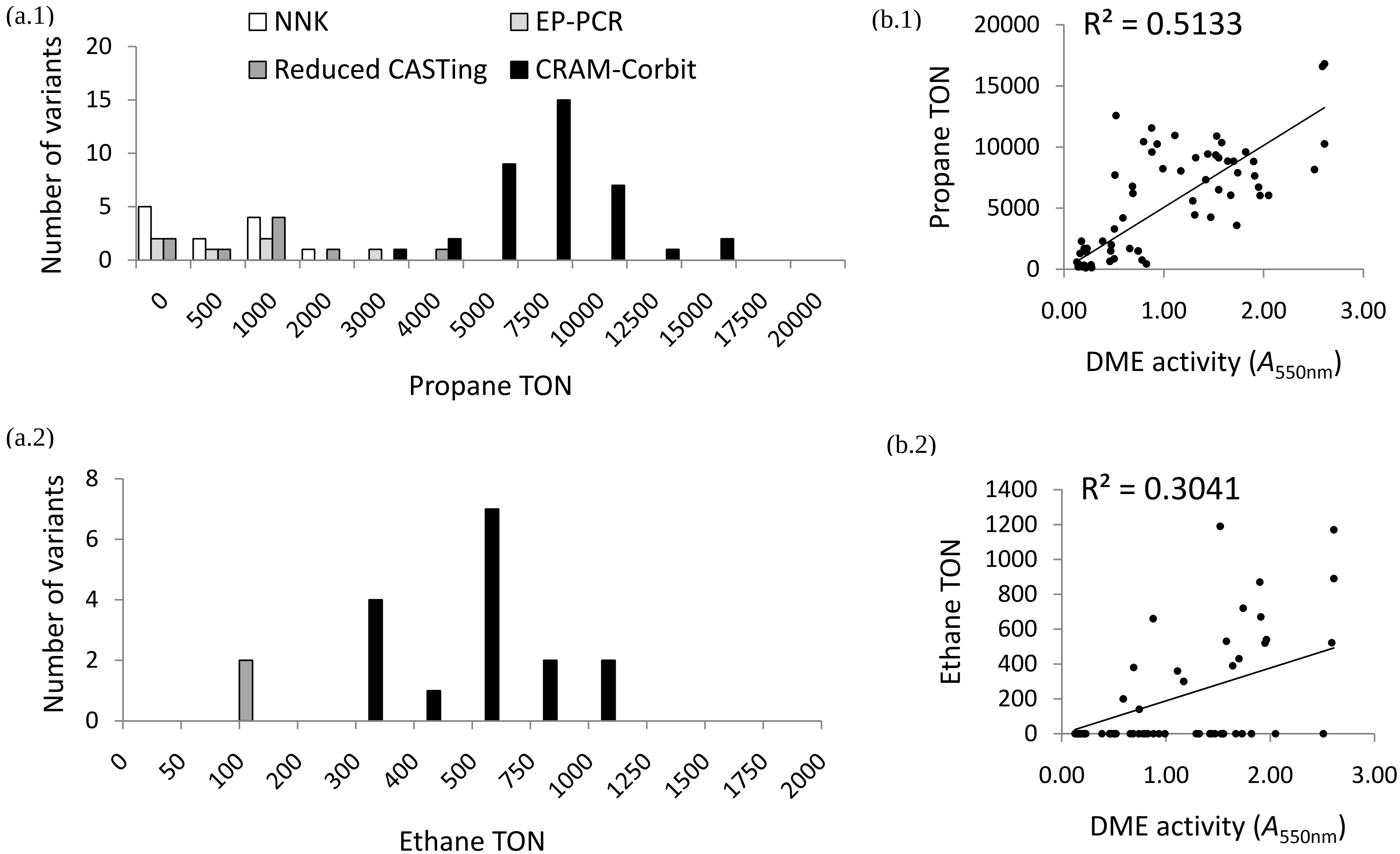
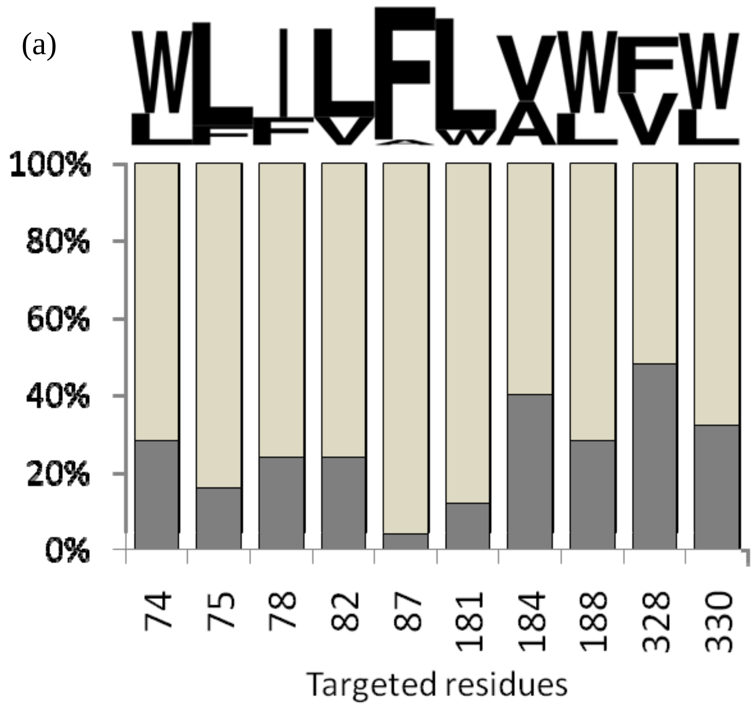


Figure 3.3: (a) Histogram of propane (a.1) and ethane (a.2) hydroxylating variants identified from various libraries. (b) Scatter plots of propane TON vs. DME activity (b.1) and ethane TON vs. DME activity (b.2) for all characterized variants

Figure 3.3 (b) shows the scatter plot of propane and ethane hydroxylation activities of all characterized variants vs. their DME demethylation activity. The data scatter of DME activity vs. propane TON, Figure 3.3 (b.1), appears to be normally distributed with a coefficient of determination (r^2) of 0.51 for linear regression of the data. In comparison, the data scatter of DME activity vs. ethane TON plot, Figure 3.3 (b.2), is not normally distributed, and a strong data bias exists representing variants with DME activity but unable to hydroxylate ethane. An r^2 of 0.30 is obtained for the linear correlation of DME activity and ethane TON. From these plots, we can conclude that both propane and ethane hydroxylation activity is positively correlated with DME demethylation since the p-values for the null hypothesis, i.e., random data scatter, are 1.5×10^{-11} and 1.26×10^{-6}, respectively. In addition, since very few variants with high DME activity were inactive for propane hydroxylation, DME demethylation is a good predictor for propane activity. However, more quantitative conclusions for the differences in predictability of DME demethylation for propane hydroxylation vs. ethane hydroxylation are difficult to determine.

By far the best source of variants with activity for propane and ethane hydroxylation is the designed library generated by the CRAM algorithm. The 25 propane hydroxylation variants with mutations at the same ten targeted residues form a concise and convenient data set for sequence analysis. The distribution of amino acids for all 25 variants, shown in Figure 3.4 (a), displays strong biases at seven of the ten targeted positions. Tryptophan appears at residue 74 and 188 in more than 72% of the sequences, likewise, strong preferences exist for L at positions 75 (84%), 82 (76%), and 181 (88%), I at 78 (76%), and F at 87 (96%). Of the remaining targeted positions, a weaker preference existed for W at 330 (68%), and V at 184 (60%), and nearly equal representations of both allowed amino acids were observed at position 328. Further sequence analysis of these variants accounting for their propane TON, Figure 3.4 (c – e), shows a finetuning of amino acid preference and reduction of the sequence space with increased activity. For variants supporting less than 7,500 propane TON, a higher fraction of the less preferred amino acids are found at positions 74, 75, 78, 82, and 188. Proceeding to variants supporting higher propane TON, the occurrences of the less preferred amino acids decrease, culminating in nearly absolute preference for W at position 74, L at positions 75, 82, and 181, I at position 78, and F at position 87 for variants supporting more than 10,000 propane TONs. These results indicate that the screening process was able to find a narrow section of the total allowed sequence space containing the best solutions for propane hydroxylation. Comparing the amino acid preference of CRAM library variants, W74/**L75**/I78/L82/**F87**/**L181,** with the residues found in PMO**,** E74/**L75**/F78/G82/**F87**/**L181**, the positions with a preference for the wild-type amino acid (75, 87, 181) are not mutated in either of the CRAM variants or PMO, while mutations at the other positions (74, 78, 82) differ between the CRAM variants and PMO. The CRAM variants preferred larger hydrophobic residues at these locations, whereas PMO introduced both a charged and a hydrophilic amino acid. Although the choices of amino acids may differ, both sets of mutations may result in a similar constriction of the substrate channel. Due to both the close range of obtained activity and the low number of active variants produced by the other mutagenesis libraries, similar sequence analysis did not yield significant trends.

(a)



(b)

(c)

(d)

(e)

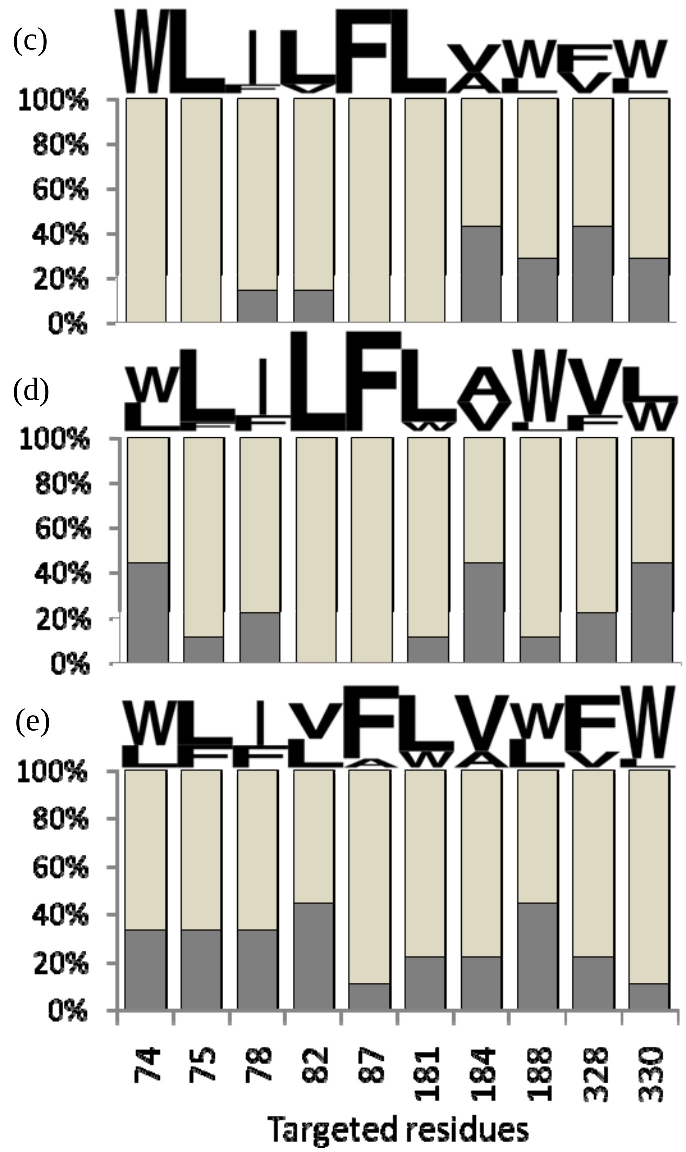


Figure 3.4: Amino acid distribution at each of the ten targeted positions of propane hydroxylating CRAM library variants: (a) all active variants, (b) identity of PMO's amino acid for these residues, (c) variants supporting less than 7,500 TON, (d) variants supporting 7,500 – 10,000 TON, (e) variants supporting > 10,000 TON, sequence logo generated by http://weblogo.berkeley.edu/

D. Discussion

Since variants with activity for DME demethylation and propane hydroxylation were identified from all the mutagenesis approaches we investigated, with as few as one mutation, it appears that finding variants with these two functions was much easier than we anticipated based on previous studies (*26*, *32*). As a substrate, propane shares many similarities with hexane, the smallest known alkane hydroxylated by BM3. They are both hydrophobic with poor water solubility and possess sub-terminal alkane C-H bonds of comparable bond strength as BM3's preferred fatty acid substrates (99–100 kcal/mol). The only difference between propane and other known BM3 substrates is its smaller molecular size, which should result in a lower binding affinity. The major impact of poorly bound substrates on the P450 reaction mechanism is a weaker activation of the catalytic cycle, as the poorly bound substrates cannot displace the distal water-ligand to initiate catalysis (*33*).

For wild-type BM3, the propane binding event needs to be the sole trigger for the activation of catalysis, as the enzyme exhibits low resting state oxidase activity (~ 10 min^{-1}), which indicates that substrate-independent activation of the catalytic cycle occurs rarely. However, variants of BM3 can exhibit much higher resting state oxidase activity, thereby reducing the requirement of propane binding to induce catalysis. In fact, variants generated in the $P450_{PMO}$ lineage and many of the variants found in this study have substrate-free cofactor consumption up to an order of magnitude higher than that of the wild-type enzyme. The existence of this alternative pathway for propane hydroxylation activity, which can be achieved without appreciable propane-induced activation of the catalytic cycle (*34*), could explain the high number of identified propane-hydroxylating variants. The accessibility of this alternative pathway for substrate hydroxylation still requires some degree of substrate binding affinity and should diminish for substrates with lower affinity, such as ethane.

Unlike the high number of variants isolated with activity for propane hydroxylation and DME demethylation, far fewer isolated variants exhibited ethane hydroxylation activity. One possible explanation for this result is the poor correlation between DME demethylation and ethane hydroxylation, as shown in Figure 3.3 (b.2). Since DME demethylation was the criterion used to filter variants for ethane hydroxylation characterization, a poor correlation between the activities would result in the elimination of ethane-hydroxylating variants with poor DME demethylation activity. Another potential explanation for this result is that variants with ethane hydroxylation activity are simply rarer in the sequence space that we investigated than variants with activity for propane hydroxylation and DME demethylation. This possibility is quite understandable since ethane is both smaller than propane and lacks the energetically favorable sub-terminal alkane C-H bond that is common to propane and BM3's preferred fatty acid substrates. Therefore, ethane hydroxylation presents challenges to not only the activation of the P450 catalytic cycle due to its smaller size, but also the ability of the P450 to break a ~ 1 kcal/mol stronger C-H bond.

Another finding from these mutagenesis libraries is the extremely high mutational tolerance of the BM3 active site. The high fraction of tolerated mutations observed with BM3 active site residues appears to contradict the general observation that mutations in the core of a protein are on average more destabilizing than mutations of solvent-exposed residues (*35*). Residues in the packed protein core generally have more interactions with neighboring amino acids than solvent-exposed residues. As a result, they have lower site entropy, which has been hypothesized to reflect decreased tolerance for mutation (*36*). However, the BM3 active site residues inherently have a higher degree of flexibility compared to typical core residues, since the active site of BM3 undergoes significant motion between the "closed" substrate-bound state and its "open" resting state (*37*). The active site environment between these two states also differs significantly in terms of solvent accessibility (*33*), which may allow these positions to tolerate polar or even charged amino acid substitutions. Therefore, the flexible nature of the BM3 active site, which is atypical of packed protein core structures, may be responsible for the higher mutations tolerance of these residues.

In comparing the functional richness between the libraries generated by the various mutagenesis methods we investigated, it was surprising to find that the EP-PCR library appears to generate more active variants than the combined efforts of the SSM libraries (see Figure 3.2 (a)). Comparing the functional richness of EP-PCR mutagenesis with SSM is inherently subjective since a poor choice of mutagenesis sites or selection criteria can easily skew the efficacy of the site-saturation libraries. However, since comparable screening effort was required to evaluate the ten site-saturation libraries (910 clones) and the EP-PCR library (1,408 clones), this comparison is reasonable from a practical standpoint. Although the number of clones sampled from these libraries is similar, the actual sequence diversity is quite different: 200 unique variants for the combined NNK libraries vs. 1,166 expected unique variants for the EP-PCR library. This sixfold increase in number of unique sequences could account for the higher number of DME demethylating variants identified in the EP-PCR library. The screening efficacy of the site-saturation libraries can be improved with the use of more efficient codons such as NDT[1] (*20*) or a combination of codons to better match the number of unique nucleotide

[1] NDT codon degeneracy: N = A, T, C, and G; D = A, T, and G

sequences with the number of unique amino acid substitutions of the libraries. However, even with the best possible codon selection, to generate a set of site-saturation libraries with the same number of unique sequences as the EP-PCR library generated in this study would require nearly 60 site-saturation libraries. Clearly, EP-PCR is better than SSM at generating sequence diversity quickly and cheaply, at the cost of not controlling or knowing the sites of mutagenesis. Based on this inherent trade-off, EP-PCR should be superior to SSM in generating diversity for functions that are affected by mutations across protein structure, such as thermostability, solubility, or modulating promiscuous activity. Conversely, for functions that require specific changes in the enzyme's active site, such as altering regio- and enantio-selectivity, the easily generated sequence diversity of the EP-PCR library is wasted, since the majority of the mutations is not created at the necessary locations. Therefore, for these functions, SSM at active site residues has the potential to be more effective.

Comparing functional richness of the site-saturation libraries at V78, A82, and A328 with the reduced CASTing libraries mutating the same residues pairwise, Fig 3.2 (b), shows that the reduced-CASTing libraries are far better in both the range of activities obtained and the number of active variants. In terms of the sequence diversity in this comparison, the pairwise reduced CASTing libraries combine to have 147 unique variants, whereas the three SSM libraries combine to have only 60 possible variants, which should account for some of the differences in the observed functional richness. However, since the mutations found to support DME demethylation and propane hydroxylation at V78 (C, T, S) and A82 (E, Q) were not included in the allowed amino acids of the reduced CASTing libraries (L, I, V, F, M, A, and W), the pairwise reduced CASTing libraries found active variants that would not have been found through recombination of the beneficial point mutations. One obvious question is whether the variants found from the reduced CASTing library are better than those that could have been isolated from either the recombination of the beneficial point mutants identified from site-saturation libraries or iterative rounds of SSM. We cannot answer this question directly, as neither approach was attempted. However, such comparisons would only be anecdotal and cannot determine which method is superior. Ultimately, all three approaches are flawed, as each makes a fallible assumption about the interaction of mutations. In the reduced CASTing library, the initial reduction of the allowed amino acids assumes that better solutions do not exist in the excluded amino acids. Likewise, the strategies of recombining beneficial single mutations or iterative site-saturation assumes that synergistic effects between mutations are minimal, and the best combination of mutations contains mutations found to be beneficial individually. How these assumptions accurately reflect the interaction of mutations for a particular protein will ultimately determine the efficacy of their application.

While the pairwise reduced CASTing libraries displayed both a higher range of obtained activities and a higher number of active variants than SSM libraries mutating the same residues, the range of DME demethylation activity obtained by the three-site reduced CASTing library did not increase compared to that of the pairwise reduced CASTing libraries. In addition, the fraction of functional variants for the three-site reduced CASTing library, 54%, was lower than the pairwise reduced CASTing library involving A328, 71% and 74%. This indicates that with the expanded sequence space, 343 vs. 49 library members, the larger library had more unique active sequences, but a larger fraction of the sequence space was occupied by inactive variants. One explanation for this result is that mutations at V78 and A82, which are located in close proximity, reduce the volume of the active site in the same region. Therefore, adding an additional mutation to the existing mutations at V78 and A328 or A82 and A328, which are already sufficient for function, has only neutral or deleterious effects. Beyond this structure based argument, the reduced benefit of increasing the mutation rate of the CASTing library may be an inherent dilution effect analogous to that observed for random mutagenesis, where increasing the mutation rate beyond 1 – 2 amino acid substitutions results in lower quality libraries (*38*).

The single point mutations at V78, A82, A328, and A330 that resulted in variants with DME demethylation activity and propane hydroxylation activity generally introduced amino acids with bulkier side chains into the active site, which follows the intuition that reducing active site volume would promote activity for a smaller substrate. Since none of the PMO mutations, V78F, A82G, or A328F, were found at these positions, they are not beneficial individually. The presence of proline mutations at A328 and A330 suggests that altering the orientation of the loop containing these residues can result in improved activity in addition to mutations that simply reducing the active site volume. The best mutation, A328V, has been reported to affect fatty acid binding and cause a shift in regioselectivity of the hydroxylation reaction (*39*). A crystal structure of WT-A328V has been solved with N-palmitoyl glycine bound in the active site (PDB: 1ZOA) (*39*). A structure alignment of WT-A328V and wild-type BM3 (Figure 3.5) shows little structural deviation, with an overall RMDS of the α-carbons of only 0.23 Å. The largest deviation between the two structures occurs near the 13^{o} kink in the I-helix, the proposed site of oxygen binding (*40*). However, similar deviations can be observed between different structures in wild-type BM3 in this region, which could reflect the general flexibility rather than the result of this mutation. The methyl group of the valine side chain induces a slight twist in the bound substrate, N-palmitoyl glycine, at the C-5 carbon; otherwise the active site packing is identical.

The lack of gross changes in the active site packing between these two structures illustrates that propane hydroxylation activity is obtainable without significant structural deviations.

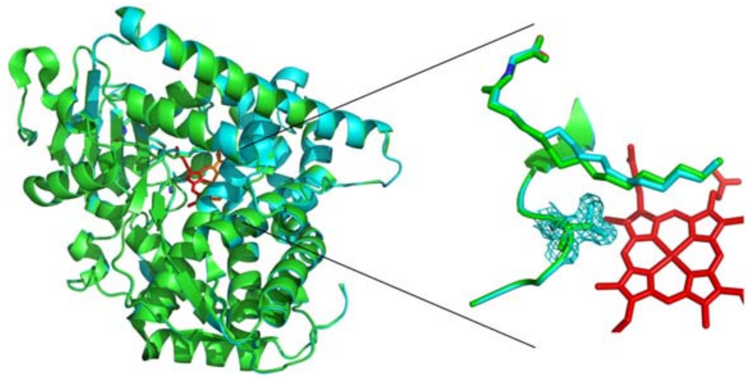


Figure 3.5: Structural alignment of BM3 (1JPZ (*24*)), shown in green, with BM3-A328V (1ZOA (*39*)), shown in cyan, heme shown in red. Close-up of the active site showing a shift of the bound N-palmitoyl glycine due to the presence of V328 side chain electron density

Of the seven mutations found through random mutagenesis, only I260V, located on the I-helix, is near the active site. The mutations F162L and I153V found in the more active variants are clustered in the region between the E and F helices. The remaining mutations E4D, T235M, D232V and Q359R, are at surface-exposed residues, which are typical of mutations found using random mutagenesis (*41*), whose effects are difficult to rationalize.

Across all nine variants isolated from the reduced CASTing libraries, V78L was the only mutation found at position 78, whereas more substitutions were beneficial at A82 (L, W, M, and V) and A328 (L, V, and F). Since the allowed amino acid set excluded glycine, only two PMO mutations, V78F and A328F, could have been found by these variants. Of these two mutations, only A328F was found in the isolated variants. Since this mutation was not found to be beneficial individually, its presence in PMO and these reduced CASTing variants suggests there are synergistic effects between this residue and neighboring amino acids. The effect of the V78L mutation on propane TON was also showed a dependence on the surrounding mutations: when V78L was introduced to WT-A328L or WT-A82L-A328L, the resulting variants lost 48% of the parental propane activity. However, when V78L was mutated in WT-A82W-A328F, the resulting variant had sixfold improved propane activity. This illustrates the ruggedness of active site landscape, in which the effects of mutations are highly dependent on the identity of neighboring amino acids. While the effects of an amino acid substitution are heavily dependent on existing mutations, different amino acid substitutions at a given residue can result in nearly identical activities, for example, mutating A82 to either M and V produced nearly equal increases in propane TON, 3.8-and 3.3-fold in the same parental background of WT-V78L-A328L. These results suggest that reducing the allowed set of amino acids is a good trade-off for mutating more residues simultaneously, as different amino acid substitutions can achieve similar effects, a result which supports the structure-based computationally designed approach we pursued.

The most effective libraries we generated in this investigation for acquiring DME demethylation and propane hydroxylation activity are the CRAM and C^{orbit} structure-based computationally designed libraries. These two libraries mutated all ten targeted active site residues allowing for two possible amino acids at each position. By redesigning the active site in such a global fashion, we obtained variants with propane and ethane activity rivaling those achieved by variants of the $P450_{PMO}$ lineage (*34*). The best variant from the CRAM library, E32, supported 16,800 propane TON and 1,200 ethane TON, which are ~ 50% of PMO's activity on these substrates. The best variant from the C^{orbit} library, OD2, supported 11,600 propane TON and 660 ethane TON, which are 34% and 27% of PMO's activity on these substrates. As a comparison, these activity levels were obtained by variants of $P450_{PMO}$ lineage after 10 – 12 rounds of mutagenesis and screening.

Of the 37 variants we isolated from these two designed libraries, all allowed amino acids were found at least once. While all mutations were represented, there are clear biases in amino acid preference (see Figure 3.4). The consensus sequence of the active CRAM variants, W74, L75, I78, L82, F87, L181, V184, W188, F328, and W330, is actually the sequence of variant E32, the most active variant, which suggests that the screening process was able to find an optimal solution within the allowed sequence space (2^{10}). Of the six positions–75, 78, 87, 181, 184, and 328–where the PMO residue was an allowed choice by the library design, five positions converged on the PMO amino acid. This convergence on mutations found in PMO is not surprising since the PMO active site is a good solution for the selected activities. However, assigning significance to these amino acid preferences in the context of the total possible sequence space of these ten residues (10^{20}) is problematic, as each mutation was compared against only one other amino acid within a limited set of surrounding mutations. Therefore, the best solution obtained from the designed libraries is certainly not the optimal solution for the total sequence space. All we can ascertain from the CRAM library results is that, within the chosen subsection of the sequence space, multiple solutions for propane and ethane hydroxylation exist, and a locally optimal solution is obtainable.

These designed libraries also demonstrate that jumps in sequence space from BM3 to variants with moderate propane hydroxylation activity (~ 10,000 TON) are achievable. None of the obtained variants, however, reached the level of specialization that was previously obtained with $P450_{PMO}$, in either propane TON or coupling of cofactor consumption. In the evolution of BM3 to PMO, the specialization for propane hydroxylation did not occur evenly through the 16 rounds of mutagenesis. In fact, the variants of the lineage can be categorized into three distinct groups by their substrate specificity for linear alkanes as (1) preferring longer chain alkanes, (2) having equal preference for alkanes of chain lengths C_3 – C_{10}, and (3) preferring shorter chains alkanes with the length of propane (*34*). These three groups of variants represent a transition from a specialized fatty acid hydroxylase to generalist P450s with broad alkane substrate acceptance followed by second transition to a specialized propane monooxygenase. This last transition occurs in the final four rounds of mutagenesis where the largest improvements in cofactor coupling (44% to 93%) and propane TON (10,550 to 33,400) occur. In addition, mutations acquired in these final rounds of mutagenesis are located not only in the P450 heme domain but also in the reductase domain. This suggests that mutations outside the active site or even the heme domain in general, may be necessary for functional optimization. The range of obtained propane TON (3,500 – 16,800) and coupling of co-factor consumption (36% – 68%) for variants identified by the designed libraries correspond to those values of the generalist intermediates found preceding the propane specialization phase of the PMO evolution. This suggests semi-rational library design can be an effective strategy to move away from a specialized enzyme toward generalist variants, but functional specialization still requires optimization through several rounds of random mutagenesis and screening.

E. References

1. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., and Hughes, G. J. (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to Sitagliptin manufacture, *Science 329*, 305-309.

2. Wohlgemuth, R. (2010) Biocatalysis - key to sustainable industrial chemistry, *Current Opinion in Biotechnology 21*, 713-724.

3. 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., Grate, J., Gruber, J., Whitman, J. C., Sheldon, R. A., and Huisman, G. W. (2007) Improving catalytic function by ProSAR-driven enzyme evolution, *Nat. Biotechnol. 25*, 338-344.

4. Urban, P., Truan, G., and Pornpon, D. (2008) High-throughput enzymology and combinatorial mutagenesis for mining cytochrome P450 functions, *Expert Opin. Drug Metab. Toxicol. 4*, 733-747.

5. Kaur, J., and Sharma, R. (2006) Directed evolution: An approach to engineer enzymes, *Crit. Rev. Biotechnol. 26*, 165-199.

6. Reetz, M. T. (2011) Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions, *Angewandte Chemie-International Edition 50*, 138-174.

7. Tracewell, C. A., and Arnold, F. H. (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time, *Curr. Opin. Chem. Biol. 13*, 3-9.

8. Turner, N. J. (2009) Directed evolution drives the next generation of biocatalysts, *Nature Chemical Biology 5*, 568-574.

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

10. O'Maille, P. E., Bakhtina, M., and Tsai, M. D. (2002) Structure-based combinatorial protein engineering (SCOPE), *Journal of Molecular Biology 321*, 677-691.

11. Reetz, M. T., Wang, L. W., and Bocola, M. (2006) Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space, *Angewandte Chemie-International Edition 45*, 1236-1241.

12. Reetz, M. T., and Carballeira, J. D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes, *Nat. Protoc. 2*, 891-903.

13. Herman, A., and Tawfik, D. S. (2007) Incorporating synthetic oligonucleotides via gene reassembly (ISOR): a versatile tool forgenerating targeted libraries, *Protein Eng. Des. Sel. 20*, 219-226.

14. 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., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro-aldol enzymes, *Science 319*, 1387-1391.

15. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., Clair, J. L. S., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, *Science 329*, 309-313.

16. 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, *Proceedings of the National Academy of Sciences of the United States of America 104*, 48-53.

17. Bershtein, S., and Tawfik, D. S. (2008) Advances in laboratory evolution of enzymes, *Curr. Opin. Chem. Biol. 12*, 151-158.

18. Cadwell, R. C., and Joyce, G. F. (1994) Mutagenic PCR *PCR-Methods Appl. 3*, S136-S140.

19. Kunkel, T. A. (1985) Rapid and efficient site-specific mutagensis without phenotypic selection, *Proceedings of the National Academy of Sciences of the United States of America 82*, 488-492.

20. Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008) Addressing the numbers problem in directed evolution, *Chembiochem 9*, 1797-1804.

21. Fulco, A. J. (1991) P450BM-3 and other inducible bacterial P450 cytochromes - biochemistry and regulation, *Annu. Rev. Pharmacol. Toxicol. 31*, 177-203.

22. Ost, T. W. B., Miles, C. S., Murdoch, J., Cheung, Y. F., Reid, G. A., Chapman, S. K., and Munro, A. W. (2000) Rational re-design of the substrate binding site of flavocytochrome P450BM3, *FEBS Lett. 486*, 173-177.

23. Fasan, R., Chen, M. M., Crook, N. C., and Arnold, F. H. (2007) Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties, *Angewandte Chemie-International Edition 46*, 8414-8418.

24. Haines, D. C., Tomchick, D. R., Machius, M., and Peterson, J. A. (2001) Pivotal role of water in the mechanism of P450BM-3, *Biochemistry 40*, 13456-13465.

25. Pylypenko, O., and Schlichting, I. (2004) Structural aspects of ligand binding to and electron transfer in bacterial and fungal p450s, *Annu. Rev. Biochem. 73*, 991-1018.

26. Glieder, A., Farinas, E. T., and Arnold, F. H. (2002) Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase, *Nat. Biotechnol. 20*, 1135-1139.

27. Peters, M. W., Meinhold, P., Glieder, A., and Arnold, F. H. (2003) Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3, *J. Am. Chem. Soc. 125*, 13442-13450.

28. Jovanovic, T., Farid, R., Friesner, R. A., and McDermott, A. E. (2005) Thermal equilibrium of high- and low-spin forms of cytochrome P450BM-3: Repositioning of the substrate?, *J. Am. Chem. Soc. 127*, 13548-13552.

29. Modi, S., Sutcliffe, M. J., Primrose, W. U., Lian, L. Y., and Roberts, G. C. K. (1996) The catalytic mechanism of cytochrome P450 BM3 involves a 6 angstrom movement of the bound substrate on reduction, *Nat. Struct. Biol. 3*, 414-417.

30. Patrick, W. M., Firth, A. E., and Blackburn, J. M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries, *Protein Eng. 16*, 451-457.

31. Otey, C., and Joern, J. M. (2003) In *Methods in Molecular Biology* (Arnold, F. H., and Georgiou, G., Eds.), Humana Press, Totowa, 141-148.

32. Farinas, E. T., Schwaneberg, U., Glieder, A., and Arnold, F. H. (2001) Directed evolution of a cytochrome P450 monooxygenase for alkane oxidation, *Advanced Synthesis & Catalysis 343*, 601-606.

33. Schlichting, I., Berendzen, J., Chu, K., Stock, A. M., Maves, S. A., Benson, D. E., Sweet, B. M., Ringe, D., Petsko, G. A., and Sligar, S. G. (2000) The catalytic pathway of cytochrome P450cam at atomic resolution, *Science 287*, 1615-1622.

34. Fasan, R., Meharenna, Y. T., Snow, C. D., Poulos, T. L., and Arnold, F. H. (2008) Evolutionary History of a Specialized P450 Propane Monooxygenase, *Journal of Molecular Biology 383*, 1069-1080.

35. Reidhaarolson, J. F., and Sauer, R. T. (1988) Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences, *Science 241*, 53-57.

36. 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, *Proceedings of the National Academy of Sciences of the United States of America 98*, 3778-3783.

37. Arnold, G. E., and Ornstein, R. L. (1997) Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: Cytochrome P450BM-3, *Biophys. J. 73*, 1147-1159.

38. Drummond, D. A., Iverson, B. L., Georgiou, G., and Arnold, F. H. (2005) Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins, *Journal of Molecular Biology 350*, 806-816.

39. Hegde, A., Chen, B., Haines, D.C., Bondlela, M., Mullin, D., Graham, S.E., Tomchick, D.R., Machius, M., Peterson, J.A. Active Site Mutations of P450BM-3 that Dramatically Affect Substrate Binding and Product Formation, (*To be published)*.

40. Ost, T. W. B., Clark, J., Mowat, C. G., Miles, C. S., Walkinshaw, M. D., Reid, G. A., Chapman, S. K., and Daff, S. (2003) Oxygen activation and electron transfer in flavocytochrome P450BM3, *J. Am. Chem. Soc. 125*, 15010-15020.

41. Arnold, F. H. (1998) Design by directed evolution, *Accounts Chem. Res. 31*, 125-131.