Diversification of Catalytic Function in a Synthetic Family of Chimeric Cytochrome P450s

Marco Landwehr,[1,3,4] Martina Carbone,[1,4] Christopher R. Otey,[2] Yougen Li,[1] and Frances H. Arnold[1,2,*]
[1] Division of Chemistry and Chemical Engineering
[2] Biochemistry and Molecular Biophysics
California Institute of Technology, Mail Code 210-41, Pasadena, CA 91125, USA
[3] Present address: Gevo, Inc., 133 N. Altadena Drive, Suite 310, Pasadena, CA 91107, USA.
[4] These authors contributed equally to this work.
*Correspondence: frances@cheme.caltech.edu
DOI 10.1016/j.chembiol.2007.01.009

SUMMARY

We report initial characterization of a synthetic family of more than 3000 cytochrome P450s made by SCHEMA recombination of 3 bacterial CYP102s. A total of 16 heme domains and their holoenzyme fusions with each of the 3 parental reductase domains were tested for activity on 11 different substrates. The results show that the chimeric enzymes have acquired significant functional diversity, including the ability to accept substrates not accepted by the parent enzymes. K-means clustering analysis of the activity data allowed the enzymes to be classified into five distinct groups based on substrate specificity. The substrates can also be grouped such that one can be a "surrogate" for others in the group. Fusion of a functional chimeric heme domain with a parental reductase domain always reconstituted a functional holoenzyme, indicating that key interdomain interactions are conserved upon reductase swapping.

INTRODUCTION

Enzymes with altered activities and specificities can be generated in the laboratory by processes that mimic mechanisms of natural evolution. Directed evolution combining recombination and random point mutation (e.g., DNA shuffling) is effective in generating both genotypic and phenotypic novelty [1–7]. Although recombination can make many mutations with relatively little structural disruption [8], we do not know the degree of functional diversity that is accessible to a process that only explores combinations of mutations already accepted during natural evolution.

We recently reported construction of a synthetic family of more than 3000 properly folded cytochrome P450 heme domains [9]. Assembled by structure-guided recombination of the heme domains of CYP102A1 from *Bacillus megaterium* (A1) and its homologs CYP102A2 (A2) and CYP102A3 (A3) that exhibit ~65% amino acid identity,

the chimeric proteins differ from the parent sequences by 72 out of 463–466 amino acids on average. Our current goal is to understand how this sequence diversification relates to diversification of function. Initial studies [9, 10] demonstrated that recombination, in the absence of point mutations, can generate functional features outside the range exhibited by the parental P450s. For example, a chimeric heme domain significantly more thermostable than any of the parents was identified (T_{50} = 62°C versus 55°C for the most stable parent) [9]; subsequent analysis of more than 200 chimeric heme domains identified many thermostable proteins (Y.L. et al., unpublished data). Our previous study of selected chimeras of the A1 and A2 heme domains showed that chimeragenesis could also generate activities not exhibited by the parents [10], as has also been reported for recombination of mammalian P450s [11, 12].

The biological functions of cytochrome P450s include key roles in drug metabolism, breakdown of xenobiotics, and steroid and secondary-metabolite biosynthesis [13]; members of the P450 superfamily catalyze hydroxylation and demethylation reactions on a vast array of substrates [14]. Enzymes from the synthetic P450 family could be useful catalysts for synthesis of biologically active compounds if they have acquired the ability to accept substrates not accepted by the parent enzymes (which are all fatty acid hydroxylases). Identifying particular desired products, however, usually requires protein purification and HPLC and/or MS analysis, methods that are cumbersome when testing hundreds of biocatalysts. Thus, in addition to exploring the range of catalytic activities in the chimeric P450 family, a second goal of the current study is to determine to what extent "surrogate" substrates can be used to identify likely catalyst candidates for a particular reaction in a high-throughput screening mode. Can substrates be grouped in such a way that activity toward one member of a group can be used to predict activity toward another?

Enzymes of the CYP102 family are comprised of a reductase domain and a heme domain connected by a flexible linker [15, 16]. With a single amino acid substitution (F87A in A1 and F88A in A2 and A3), the heme domains can function alone as peroxygenases, catalyzing oxygen insertion in the presence of hydrogen peroxide [17].

The synthetic CYP102A family was constructed from parental sequences containing this mutation; all of the chimeric proteins can therefore potentially function as peroxygenases. We are also interested in their ability to be reconstituted into functional monooxygenases, utilizing NADPH and molecular oxygen for catalysis, by fusion to a reductase domain. The reductase domain of CYP102A1 (R1) spans ~585 amino acids and encodes an ~20 amino acid linker and the binding domains for the FMN, FAD, and NADPH cofactors [16]. The reductases from CYP102A2 and CYP102A3 (R2 and R3, respectively,) share 52%–55% sequence identity with R1 and are comparable in size, and the only notable difference is a linker region that is extended in R2 by seven amino acids [18]. Because the chimeric heme domains comprise sequences from three different parents, it is not obvious that fusion to wild-type reductase will generate a catalytically active holoenzyme, nor is it clear which reductase, R1, R2, or R3, should be used. For this initial characterization we therefore selected a set of 14 chimeric heme domains, reconstituted them with all 3 parental reductase domains, and determined peroxygenase and monooxygenase activities on 11 substrates. These activities have been analyzed to (1) assess the functional diversity of the chimeric enzymes, (2) determine whether substrates fall into groups for the purposes of predicting activities, and (3) compare the activities and specificities of the chimeric peroxygenases with those of their reconstituted monooxygenases.

RESULTS

Cloning and Expression of P450 Heme Domains and Holoenzymes

A total of 17 heme domains, including the 3 parent heme domains, were chosen for holoenzyme construction by fusion to a wild-type CYP102A reductase domain. For each heme domain, 4 proteins were examined—the heme domain and its fusion to each of the 3 reductase domains—for a total of 68 constructs. Heme domains contain the first 463 amino acids for A1 and the first 466 amino acids for A2 and A3. The reductase domains start at amino acid E464 for R1, K467 for R2, and D467 for R3 and encode the linker region of the corresponding reductase. A3 and its fusions with R1 and R2 expressed very poorly, yielding only a very small amount of protein after purification, and were therefore not analyzed further.

The chimeric sequences are reported in terms of the parent from which each of the eight sequence blocks is inherited (Table S1; see the Supplemental Data available with this article online). A total of 12 of the 14 chimeras were selected because they displayed relatively high activities on substrates in preliminary studies (data not shown). Chimera 23132233 was chosen because it displayed low peroxygenase activity, while 22312333 was selected because it is more thermostable than any of the parents (T_{50} = 62°C) [9]. For the constructs studied here, the reductase identity is indicated as the ninth sequence element; R0 refers to no reductase (i.e., heme domain peroxygenase).

Activity Assays

To assess the functional diversity of the chimeric P450s, we measured their activities on the 11 substrates shown in Figure 1. Propranolol (PR), tolbutamide (TB), chlorzoxazone (CH), and zoxazolamine (ZX) are drugs that are metabolized by human P450s [3, 19, 20, 21]. 12-p-nitrophenoxydodecanoic acid (PN) is a long-chain fatty acid surrogate; parent A1-R1 holoenzyme and the A1 heme domain (with the F87A mutation) both show high activity on this substrate. Previous work showed that A1 has weak peroxygenase activity on some of the aromatic substrates [10]. Aromatic hydroxylation products of all substrates can be detected quantitatively by using the 4-amino antipyrine assay [22]. PN hydroxylation can be monitored spectrophometrically [23].

Peroxygenase activities of the 16 heme domains (all except A3) were determined by assaying for product formation after a fixed reaction time in 96-well plates (see Experimental Procedures). Similar assays were used to determine monooxygenase activities for each of the fusion proteins. Final enzyme concentrations were fixed to 1 μM in order to reduce large errors associated with low expression and to allow us to compare chimera activities by using absorbance values directly. Protein concentrations were reassayed in a 96-well format and were determined to be 0.88 μM ± 13% (SD/average). All samples were prepared and analyzed in triplicate, and outlier data points were eliminated. Tables S2 and S3 report the averages and standard deviations for each of the assays. More than 85% of the data for each substrate were retained, and more than 95% were retained for 6 of the 11 substrates (Table S4).

Because extinction coefficients are not known for the reaction products, we do not report absolute enzyme activities, nor do we report substrate specificities, which are ratios of enzyme activity on one substrate to activity on another. Our data nonetheless allow us to compare the chimeras with respect to their activities on a given substrate and also to compare their activity profiles and therefore their specificities. Chimeras having a similar profile form the same relative amounts of products from all substrates and are therefore likely to have similar specificities. To better visualize differences among chimeras, the highest average absorbance value for a given substrate was set to 100%, and all other absorbances for the same substrate, but different chimeras, were normalized to this. Figure 2 is a heat plot of the complete data set of normalized absorbances, while Figure S2 shows the substrate-activity profiles in the form of bar plots.

Activities of Parent Enzymes

Figure 3A shows the normalized substrate-activity profiles of the A1 and A2 peroxygenases. Both have relatively low or no activity on any of the substrates except PN, where A1 makes about an order of magnitude more product than does A2. Profiles for the reconstituted parent holoenzymes are shown in Figure 3B. Fusion of A1 and R1 generated an enzyme with profile peaks on ethyl 4-phenylbutyrate (PB) and PN. A1 is in fact the
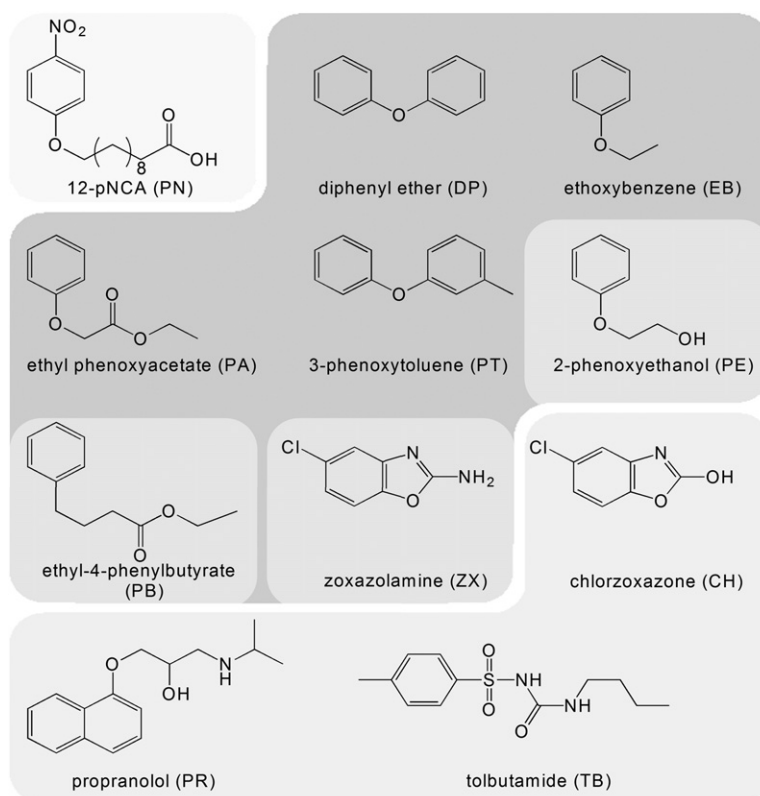
Figure 1. Chemical Structures and Abbreviations

Substrates are grouped according to the pairwise correlations (see text for details). Members of a group are highly correlated; intergroup correlations are low.

second-best-performing enzyme on PB. The A1 peroxygenase activity on this substrate, however, is among the worst, showing that peroxygenase specificity does not necessarily predict that of the monooxygenase. Fusion of A2 to R2 slightly increased activity relative to A2, but it did not alter the profile. The A3-R3 holoenzyme exhibits some activity on the drug-like substrates (PR, TB, CH) as well as PN and PB.

Fusion of the A1 and A2 heme domains to other reductase domains yields holoenzymes that are active on some substrates (Figures 3C and 3D). The A2 fusions have relatively low activities. A1 fusions with R1 and R2, on the other hand, created highly active enzymes with different specificities: the A1-R1 profile has peaks on PN and PB, while that of A1-R2 has peaks on PB, phenoxyethanol (PE), and zoxazolamine (ZX). The A1-R3 fusion is less active on nearly all substrates.

Activities of Chimeras and Identification of Chimera Clusters

The 14 chimeric heme domains generated 56 chimeric peroxygenases and monooxygenases. Nearly all of the chimera fusions outperformed even the best parent holoenzyme, and chimeric peroxygenases consistently outperformed the parent peroxygenases (Figure 2 and Figure S2). The best enzyme for each substrate is listed in Table S5. All of the best enzymes are chimeras. Most of the best enzymes are also holoenzymes: only PE has a peroxygenase as the best catalyst.

We now show that there exists a discrete set of characteristic substrate-activity profiles to which each chimera can be uniquely assigned. A K-means clustering analysis was applied to the normalized absorbance data to better understand the functional diversity. K-means clustering, a statistical algorithm that partitions data into clusters based on data similarity [24], has been used by Mannervik and coworkers to identify groups of mutants exhibiting similar substrate specificities [25] as well as by others to identify protein fragments (4–7 residues) of similar structure [26] and interacting nucleotide pairs with similar 3D structures [27]. For our analysis, the normalized data were used to ensure that each of the 11 dimensions is given equal weight by the clustering algorithm. The clustering was performed over values of k (number of clusters) ranging from k = 2 to k = 8. The highest silhouette value (see Experimental Procedures) was observed at k = 5.

The cluster composition for k = 5 is depicted in Figure 4. Cluster 1, consisting of chimeras 32312333-R1/R2 and 32313233-R1/R2 (Figure 4B), is characterized by low relative activities on CH, TB, PR, and PN and by high relative activities on all other substrates. In fact, two of these chimeras are the best enzymes on all of the remaining substrates except PB and PE.

Cluster 2 is made up of 22213132-R2, 21313111-R3, and 21313311-R3, which are the most active enzymes on TB, CH, and PR, respectively (Figure 4C). Cluster 2 enzymes are entirely inactive on PN and show low activity on most of the substrates that cluster 1 enzymes accept (PE,
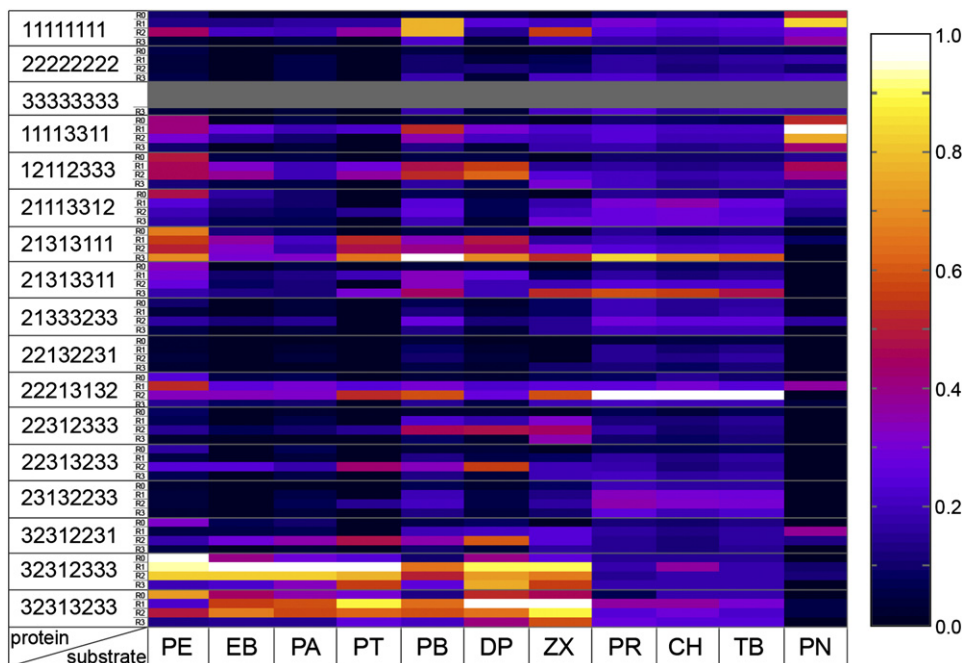
Figure 2. Summary of Normalized Activities for All 56 Enzymes Acting on 11 Substrates
Activities are shown with a color scale (white indicates the highest activity; black indicates the lowest activity); columns represent substrates, and rows represent proteins. Unanalyzed A3, A3-R1, and A3-R2 proteins are shown in gray. Protein rows are ordered by their chimeric sequence first, and then by heme domain (R0) and R1, R2, and R3 fusions.

DP, PA, and EB). Relative activities on the remaining substrates (i.e., PB, ZX, and PT) are moderate (although lower than cluster 1 chimeras). An exception is 21313111-R3, which is the best enzyme for PB and is also fairly good on PE and DP.

Cluster 3 contains chimeras A1-R1/R2, 12112333-R1/R2, 11113311-R1/R2, and 22213132-R1 (Figure 4D). The A1-like sequences are characterized by high relative activity on PN (on which 11113311-R1/R2 and A1-R1 are the three top-ranking enzymes), moderate to high relative activity on PB, and moderate activity on PE.

Cluster 4 contains 21313111-R1/R2, 22313233-R2, 22312333-R2, 32312231-R2, 32312333-R0, 32312333-R3, 32313233-R0, and 32313233-R3 (Figure 4E). This cluster is characterized as having the highest relative

activity on PE, in addition to moderate activities on PT, DP, and ZX. The remaining chimeras appear in a fifth cluster that has relatively low activity on everything except PN and PE (Figure 4F). This cluster contains parental sequences A1-R0, A1-R3, A2-R0, A2-R1/R2/R3, and A3-R3. Native sequences are thus only found in two of the clusters. The remaining clusters (1, 2, and 4) are made up of highly active chimeras that have acquired novel profiles.

The partition created by the clustering algorithm shows that the presence and identity of the reductase can alter the activity profile and thus the specificity of a heme domain sequence. For example, the R1 and R2 fusions of 32312333 and 32313233 appear in cluster 1, whereas their R0 and R3 counterparts are in cluster 4. Sequences 22213132 and 21313111 also behave differently when
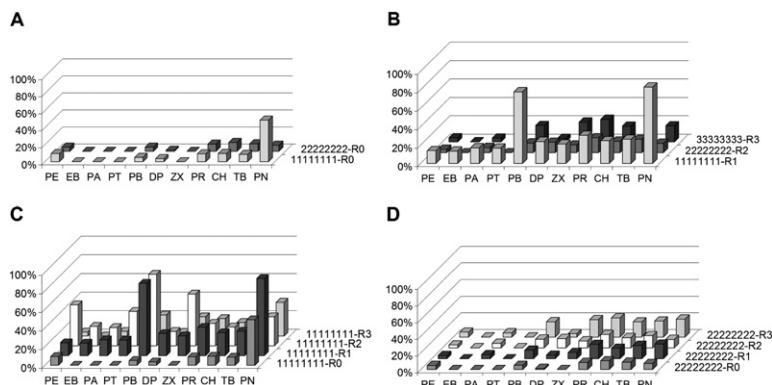

Figure 3. Substrate-Activity Profiles for Parent Heme Domain Mono- and Peroxygenases
(A) Parent peroxygenases.
(B) Parent holoenzyme monooxygenase profiles.
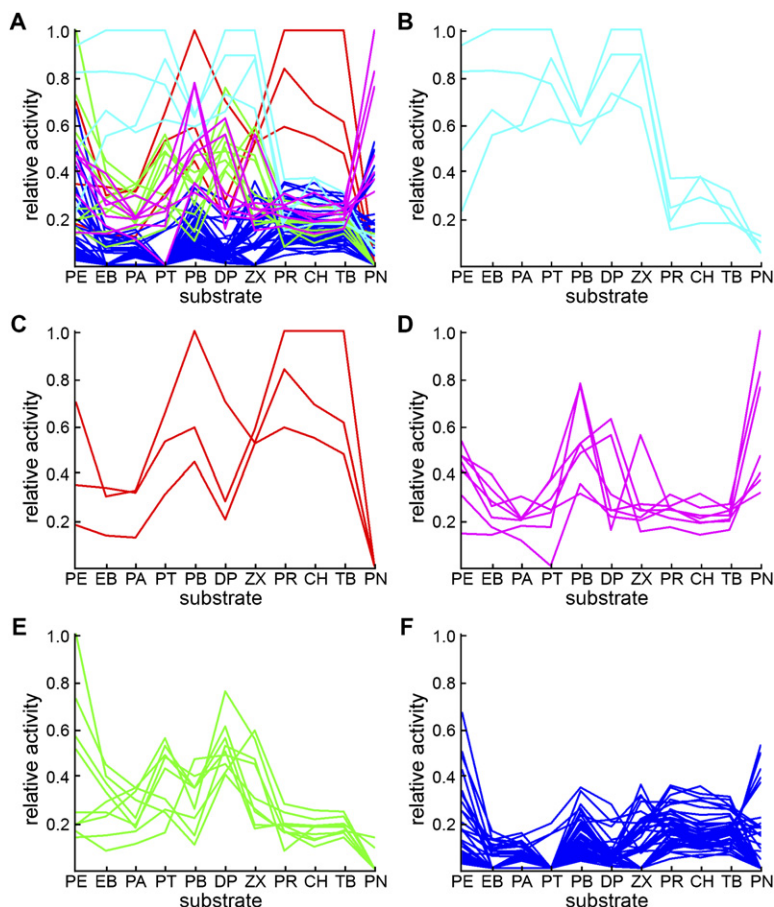(C) The A1 protein set.
(D) The A2 protein set.

Figure 4. K-Means Clustering Analysis Separates Chimeras into Five Clusters
(A–F) All protein-activity profiles are depicted in (A), in which the color identifies the cluster. (B)–(F) show profiles for sequences within each cluster. (B) depicts 32312333-R1/R2 and 32313233-R1/R2. (C) depicts 22213132-R2, 21313111-R3, and 21313311-R3. (D) depicts A1-R1/R2, 12112333-R1/R2, 11113311-R1/R2, and 22213132-R1. (E) depicts 21313111-R1/R2, 22313233-R2, 22312333-R2, 32312231-R2, 32312333-R0, 32312333-R3, 32313233-R0, and 32313233-R3. (F) depicts the remaining sequences.

fused to different reductases. 22213132-R2, for example, displays pronounced peaks on substrates TB, CH, and PR that are not present in the corresponding peroxygenase and R1/R3 profiles (Figure S2E) and is thus the only member with this heme domain sequence appearing in cluster 2. 21313111-R3 and 21313111-R2/R1 have nearly opposite profiles (Figure S2J) and consequently appear in different clusters. Thus, the best choice of reductase depends on both the substrate and the chimera sequence.

Peroxygenase versus Monooxygenase Activities
As shown in Figure 2, each of the 14 chimeric heme domains can be fused to a parental reductase to generate a functional monooxygenase. The resulting monooxygenases are generally more active under these conditions than the corresponding peroxygenases (see Figure S2). The R1 and R2 fusions tend to outperform R3 fusions. While altering reductase identity never completely deactivates the protein, it does affect specificity in some cases. To quantify the differences between the profiles of the four different enzymes that can be made from a given chimera, the pairwise linear coefficients (R^2) of the R0/R1, R0/R2, R0/R3, R1/R2, R1/R3, and R2/R3 profiles were determined for each heme domain sequence (with the exception of A3). The results are shown in Table S1. High correlations represent enzyme pairs with similar specificities.

The results show that peroxygenase and monooxygenase specificities are usually different, that R1/R2 fusions of a chimera are often very similar (five pairs have R^2 values above 0.9), and that the R1 and R2 fusions are less similar to the R3 enzymes.

Identification of Substrate Groups
To understand whether a chimera's activity on one substrate predicts activity on another, the pairwise correlations of the absorbances of all of the possible substrate pairs were determined (Table S5). Mannervik and co-workers used correlations between activities on substrate pairs to identify enzyme variants with novel substrate specificities [7]. Here, we use these correlations instead to identify substrates with similar chimera profiles. This analysis led to the identification of three substrate clusters characterized by high values of the correlation coefficients. Members of different clusters are poorly correlated. DP, PT, PA, and EB all exhibit high correlations with each other (R^2 = 0.71–0.92; see Figure S1A for an example) and were grouped into the core of substrate group A. Group B consists of CH, TB, and PR. The categorization of this group is clearly defined: its members show high correlations with each other (R^2 above 0.9; see Figure S1B for an example), but correlate very poorly with the other substrates (R^2 = 0.01–0.37). PN does not

correlate significantly with any of the other substrates tested (R^2 = 0.00–0.08) and is its own substrate group—termed group C.

ZX, PB, and PE show moderate correlation to members of the group A core (R^2 = 0.56–0.66, 0.39–0.56, and 0.35–0.61, respectively). These substrates are considered loosely associated with group A since they do not belong to any other group due to poor correlation with each other and the remaining substrates.

There exists a correspondence between the chimera clusters and the substrate groups. Group A core substrates have cluster 1 chimeras as their top-performing enzymes, whereas substrates of group B have cluster 2 chimeras as their top-performing enzymes. The top catalysts for group C are three of the cluster 3 chimeras. Members of a substrate group thus share the same best-performing enzymes.

DISCUSSION

SCHEMA Recombination Creates a Family of Functionally Diverse Enzymes

We have begun to characterize the functional diversity in a synthetic P450 family created by structure-guided recombination of bacterial fatty acid hydroxylases. The folded P450s, which make up almost 50% of the 6561 sequences in the SCHEMA library, contain an average of 72 mutations from their closest parent. A large fraction of the folded P450s were shown to be catalytically active [9], but they had been systematically studied on only a single substrate (PN). We therefore selected 11 substrates for this initial characterization of 14 of the active chimeric heme domains and their fusions with each of the 3 parental reductase domains. Although most of the parental enzyme constructs are not very active on the selected substrates, many of the chimeras are significantly more active. In fact, for every single substrate, including one widely used to assay CYP102A1 (PN), the top-performing enzyme is a chimera. Recombining mutations already accepted in natural homologs thus leads to a family of highly active enzymes that accept a broader range of substrates.

Chimeras Can Be Clustered by Substrate Specificity

We further showed that the chimeric enzymes exhibit distinct specificities and that they can be partitioned into clusters based on their specificity. One cluster contains parent A1-R1 and all chimeras with A1-like profiles. Another cluster contains low-activity chimeras and includes all remaining parental sequences. The remaining clusters represent highly active chimeras that have acquired new specificities. Members of a cluster are likely to exhibit common structural, physical, or chemical features that account for their similar catalytic properties. If the library is large enough, statistical techniques can be used to determine how sequence elements relate to the observed profiles. In particular, if there are sufficient numbers of chimeras in each cluster, then powerful tools such as logistic regression or machine learning can be used to predict which cluster an untested sequence belongs to [9]. This

type of analysis would enable the prediction of substrate profiles of untested chimeras based on sequence information alone. The functionally diverse enzymes generated by SCHEMA-guided recombination can therefore be used to probe the sequence and structural basis of enzyme specificity. We recently observed the success of such an approach in predicting the thermostabilities of untested chimeras (Y.L. et al., unpublished data). Although the current data set does not contain enough sequences for a comprehensive analysis of sequence-function relationships, anecdotal observations can be used to generate hypotheses for further testing. For example, the chimeras in the library with parent A1 in blocks 1, 3, and 4 are all among the best enzymes for PN. These same enzymes display low relative activity on all of the remaining substrates except for PB. This suggests that having parent A1 sequence at one or more of these blocks improves PN activity and specificity.

Substrates Fall into Groups that Correlate with Chimera Clusters

We were also able to partition the substrates into groups based on the linear correlations of substrate pairs. An enzyme active on one member of a substrate group is therefore likely to be active on another member of the same group. One group consists of the drug-like substrates TB, PR, and CH (Figure 1). Another consists of PT, PA, EB, and DP. If these correlations hold for the larger library of chimeric enzymes, we should be able to predict with reasonable accuracy the relative activities of a chimera on all of the substrates in a group by testing activity on only one. This type of analysis can be expanded to a larger collection of substrates to identify additional groups or additional members of an existing group.

The observed correspondence between the three substrate groups and chimera clusters 1, 2, and 3 illustrates that each group can be associated with a cluster made up of or containing the top-performing enzymes for the substrates in that group. Some degree of correspondence can be expected, given how the partitions were constructed. However, because intragroup correlations are not 1 and intergroup correlations are not 0, the correspondence is not perfect. For this reason, there exist chimeras whose profiles exhibit peaks on only certain members of a group (cluster 4) and others that exhibit peaks on members of different groups (cluster 2 and 3 chimeras). Cluster 4 chimeras have peaks on only certain members of group A and are thus responsible for the lower correlations among group A substrates. Some cluster 2 and cluster 3 chimeras exhibit peaks on PB (on the edge of group A) as well as groups B and C, respectively. In fact, although PB correlates mostly with group A core substrates, it shares its top-performing enzymes with groups B and C and thus displays a hybrid behavior. This is why PB correlates less with group A than core substrates do and why it has higher correlations with group B and group C members than any other substrate not belonging to these groups.

Because chimeras displaying high relative activity have more weight in determining the correlation coefficients,
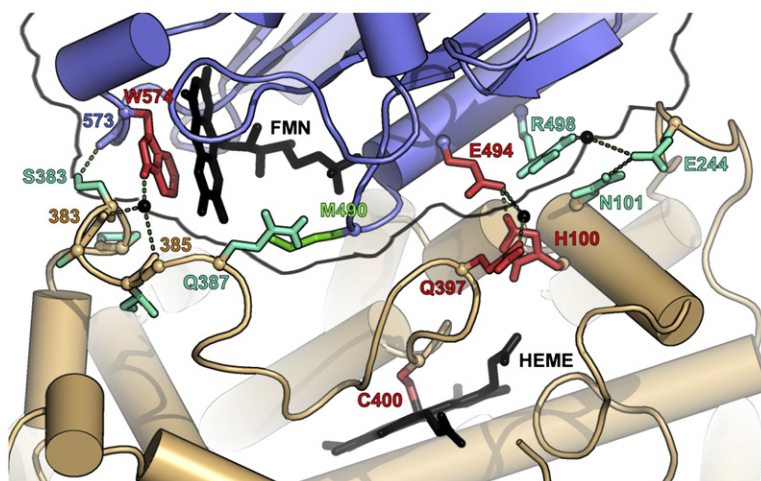
Figure 5. Residue Conservation across the Heme-Reductase Interface

Interface between the FMN (blue backbone) and heme (brown backbone) domains based on the 1BVY structure redrawn according to Sevrioukova et al. [34]. Residue colors indicate the degree of conservation: red, three parents; turquoise, two parents; green, not conserved. Hydrogen bonds are shown as dashed lines. The amino acids correspond to CYP102A1 numbering. PyMOL was used to create this figure [44].

the top-performing enzymes for one member of a substrate group will usually be among the top-performing ones for all of the members of that group. The clearer the definition of the substrate groups, the more likely this is to hold. Given the many important applications of P450s in medicine and biocatalysis, and the lack of high-throughput screens for many compounds of interest, an approach to screening that is based on carefully chosen "surrogate" substrates could significantly enhance our ability to identify useful catalysts. Clearly, any member of a well-defined substrate group can be a surrogate for other members of that group. Further analysis may also help to identify the critical physical, structural, or chemical properties of substrates belonging to a known group. This will make it possible to predict which chimeras will be most active on a new, untested substrate.

Swapping Reductase Domains Consistently Yields Active Monooxygenases and Conserves Key P450-Reductase FMN Domain Interactions

The literature reports multiple cases in which functional P450s have been reconstituted with new reductase domains. In several studies, swapping reductases improved mammalian P450 activity [28–30]. A self-sufficient chimeric mammalian P450 2E1 enzyme was constructed by fusing the 2E1 heme domain to the CYP102A1 reductase [31]. Functional chimeras of CYP102A1 and the flavocytochrome nitric oxide synthase (nNOS) have been generated [32]. Another study reported the functional expression of CYP153A genes by incorporating them into a framework consisting of the N and C termini of homolog CYP153A13a and fusion to the reductase domain of CYP116B2 [33].

Reconstitution of the chimeric CYP102A heme domains with the three parental reductases generated functional monooxygenases in all cases. Although their specificities were often different (particularly when fused to R3), fusion to a reductase was never detrimental to activity, and swapping the reductase never completely inactivated the enzyme (Figure S2). Subtle changes in the structure and coupling behavior that affect total product formation may account for specificity differences. The fact that the

parental reductase domains are accepted without loss of function, however, suggests that key domain-domain interactions are conserved upon reductase swapping.

Although a complete crystal structure of a CYP102A holoenzyme is not available, a partial CYP102A1 structure (1BVY) includes the interface between the heme and the reductase FMN domains. Only a few direct contacts, including one hydrogen bond, one salt bridge, and several water-mediated contacts, make up this A1-R1 interface [34]. We aligned the parental sequences by using ClustalW [35], and we found that the interactions depicted in the 1BVY crystal structure involve amino acids that are mostly conserved in the parent proteins. Figure 5 displays the interface between the heme and reductase domains of CYP102A1 and highlights the amino acids involved in key interactions. The salt bridge is formed between reductase residue E494 and heme domain residue H100, both of which are conserved in all three parents. Thus, this key interaction would be retained upon reductase swapping that conserves the orientation of the two domains.

The direct hydrogen bond occurs between the reductase backbone carbonyl of N573 and the side chain hydroxyl group of heme domain residue S383. N573 is only conserved in R1 and R2, but because the interaction involves the backbone oxygen, the reductase side of the interface is not affected by changes in the side chain identity. S383 is only conserved in parents A1 and A3. However, the corresponding residue in A2, D385, may also be capable of forming the hydrogen bond. This interaction may therefore be present in all of the chimeras.

There are two water-mediated hydrogen bonds between the hydrogen of the indole nitrogen of reductase residue W574 and the backbone carbonyl of S383 and I385. W574 was earlier shown to be crucial for electron transfer from the FMN to the heme [36] and is conserved in R1, R2, and R3. S383 and I385 are conserved in A1 and A3, but not A2, in which the corresponding residues are D385 and V387, respectively. Because the hydrogen bonds involve the backbone oxygens of these residues, these interactions may be retained upon domain substitution. Also, all possible pairwise interactions that can be

formed at these positions by domain swapping already exist in at least one of the parental sequences and are thus likely not to be destabilizing. Finally, the substitutions that do occur are conservative, replacing a hydrophilic residue with another hydrophilic residue and a hydrophobic residue with another hydrophobic residue. The third water-mediated hydrogen bond between the side chains of reductase residue R498 and heme domain residue E244 (block 5) is conserved in A1-R1 and A2-R2, but not A3-R3, in which the corresponding residues are G501 and V246, respectively. A3-R3 thus cannot form this interaction, nor can any chimera that inherits A3 sequence at block 5 and/or is fused to R3.

In summary, it appears that the direct hydrogen bond, two of the three water-mediated hydrogen bonds, and the salt bridge are all conserved in the chimera-reductase fusions. The third water-mediated hydrogen bond is conserved only in R1/R2 fusions that do not have parent A3 in block 5 (8 out of 17 sequences). Thus, the activities of the reconstituted monooxygenases are consistent with their sequences, the domain-domain interactions identified in the 1BVY structure, and the assumption that the overall structures and orientations are conserved upon reductase swapping. These results demonstrate the highly conservative nature of mutation by recombination of protein domains: as long as key interactions are retained, the remaining sequences can vary extensively.

SIGNIFICANCE

The evolvable cytochrome P450 scaffold has diversified over millions of years of mutation and natural selection to exhibit the myriad activities of the natural enzyme family, of which more than 4500 sequences are known [37]. We constructed a large synthetic P450 family by recombining sequence elements from three bacterial P450s [9]. We have now shown that members of this synthetic family exhibit diverse activities and specificities, including activities toward substrates that are not accepted by the parent P450s and drug-like compounds that are substrates of human P450s. Thus, enzymes in this family have acquired the ability to mimic important reactions in human drug metabolism. The grouping of substrates according to the likelihood that a given chimera will accept them, as has been demonstrated here, will aid in the identification of useful catalysts from this synthetic family by high-throughput screening of substrate "surrogates." We anticipate that these enzymes will be useful for synthesis of drug metabolites [19], which are needed for toxicity testing and drug discovery.

EXPERIMENTAL PROCEDURES

Abbreviations
Parent CYP102A1 heme domain, A1; parent CYP102A2 heme domain, A2; parent CYP102A3 heme domain, A3; standard deviation, SD; background, BG; 2-phenoxyethanol, PE; ethoxybenzene, EB; ethyl phenoxyacetate, PA; 3-phenoxytoluene, PT; ethyl 4-phenylbutyrate, PB; diphenyl ether, DP; zoxazolamine, ZX; propranolol, PR; chlorzox-

azone, CH; tolbutamide, TB; 12-p-nitrophenoxydodecanoic acid, PN, 12-pNCA.

Nomenclature and Construction of Holoenzymes from Chimeric Heme Domains
Details of chimera construction have been reported previously [9]. Sequences are given an eight-digit number, in which each digit indicates the parent from which each of the eight blocks was inherited. The identity of the reductase is indicated by R0 (for no reductase) or by R1, R2, or R3 for the CYP102A1, CYP102A2, or CYP102A3 reductases, respectively.

To construct the holoenzymes, the chimeric heme domains were fused to each of the three wild-type reductase domains after amino acid residue 463 when the last block originates from CYP102A1 and after amino acid 466 for CYP102A2 and CYP102A3. The holoenzymes were constructed by overlap extension PCR [38] and/or ligation and were cloned into the pCWori expression vector [39]. All constructs were confirmed by sequencing.

Protein Expression and Purification
Proteins were expressed in *E. coli* as described previously and were purified by anion exchange on Toyopearl SuperQ-650M from Tosoh [40]. After binding of the proteins, the matrix was washed with a 30 mM NaCl buffer, and proteins were eluted with 150 mM NaCl (all buffers used for purification contained 25 mM phosphate buffer [pH 8.0]). Proteins were rebuffered into 100 mM phosphate buffer and concentrated by using 30,000 MWCO Amicon Ultra centrifugal filter devices (Millipore). Proteins were stored at $-20°C$ in 50% glycerol.

Protein concentration was measured by CO absorption at 450 nm as described [41]. A protein concentration of 1 μM was chosen for the activity assays. Protein concentrations were reassayed in a 96-well format and were determined to be 0.88 μM $\pm$ 13% (SD/average).

Functional Assays
Proteins were assayed for mono- or peroxygenase activities in 96-well plates as described [9, 42]. Heme domains were assayed for peroxygenase activity by using hydrogen peroxide as the oxygen and electron source. Reductase domain fusion proteins were assayed for monooxygenase activity, by using molecular oxygen and NADPH. Reactions were carried out in 100 mM EPPS buffer (pH 8), 1% acetone, 1% DMSO, 1 μM protein in 120 μl volumes. Substrate concentrations depended on their solubility under the assay conditions. Final concentrations were as follows: 2-phenoxyethanol (PE), 100 mM; ethoxybenzene (EB), 50 mM; ethyl phenoxyacetate (PA), 10 mM; 3-phenoxytoluene (PT), 10 mM; ethyl 4-phenylbutyrate (PB), 5 mM; diphenyl ether (DP), 10 mM; zoxazolamine (ZX), 5 mM; propranolol (PR), 4 mM; chlorzoxazone (CH), 5 mM; tolbutamide (TB), 10 mM; 12-p-nitrophenoxydodecanoic acid (PN), 0.25 mM. The reaction was initiated by the addition of NADPH or hydrogen peroxide stock solution (final concentration of 500 μM NADPH or 2 mM hydrogen peroxide) and was mixed briefly. After 2 hr at room temperature, reactions with substrates 1–10 were quenched with 120 μl of 0.1 M NaOH and 4 M urea. A total of 36 μl of 0.6% (w/v) 4-aminoantipyrine (4-AAP) was then added. The 96-well plate reader was zeroed at 500 nm, and 36 μl of 0.6% (w/v) potassium persulfate was added. After 20 min, the absorbance at 500 nm was read [22]. Reactions on PN were monitored directly at 410 nm by the absorption of accumulated 4-nitrophenol. All experiments were performed in triplicate, and the absorption data were averaged.

Data Analysis
The background absorbance (BG) was subtracted from the raw data. BG reactions contained buffer, cofactor, and substrate in the absence of protein sample and were done in triplicate. All absorbance measurements were done once on three separate samples (triplicate sampling). Data points with an SD/average $\geq$20% that did not lie within the average $\pm$1.1*SD were eliminated. 1.1*SD was chosen so that for each substrate at least 85% of the points were retained. This never resulted in the elimination of more than one point from each triplicate set of

measurements. All points with an average absorbance <BG were set to 0, because they are assumed to belong to inactive proteins. The absorbance matrix thus obtained for all 68 proteins on all 11 substrates is displayed in Table S2. The SD/average matrix is displayed in Table S3. SD/average was calculated by ignoring values for inactive enzymes.

Cluster Analysis
K-means clustering is a partitioning method that divides a set of observations into k mutually exclusive clusters. K-means treats each data point as an object with a location in m-dimensional space (m = 11 in this analysis) [24]. It then finds a partition such that members of the same cluster are as close as possible to each other and as far as possible from members of other clusters. For this reason, a measure of the meaningfulness of a partition is given by the silhouette value

$$s = avg\left(\frac{b(i) - a(i)}{\max[a(i), b(i)]}\right),$$

where a(i) is the average distance from point i to all other points in its cluster, and b(i) is the average distance from point i to all points in the closest cluster. It is evident that $-1 \leq s \leq 1$, and that the quality of the clustering increases as s approaches 1 [43]. Distances are measured by the square of the Euclidean distance.

Supplemental Data
Supplemental Data include the pairwise correlations of the activities of corresponding peroxygenase and monooxygenase enzymes reported as R^2 values, the normalized activity matrix and the corresponding standard deviation matrix, the error statistics for each substrate, a list of the most active chimeras on each substrate and the pairwise correlations of the activities on each substrate reported as R^2 values, scatter plots of examples of pairwise substrate correlations and bar plots of the activities of all of the the chimeras and parental enzymes in this study and are available at http://www.chembiol.com/cgi/content/full/14/3/269/DC1/.

REFERENCES

1. Taly, V., Urban, P., Truan, G., and Pompon, D. (2006). A combinatorial approach to substrate discrimination in the P450 CYP1A subfamily. Biochim. Biophys. Acta. *1770*, 446–457.

2. Raillard, S., Krebber, A., Chen, Y., Ness, J.E., Bermudez, E., Trinidad, R., Fullem, R., Davis, C., Welch, M., Seffernick, J., et al. (2001). Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. Chem. Biol. *8*, 891–898.

3. Ernstgard, L., Warholm, M., and Johanson, G. (2004). Robustness of chlorzoxazone as an in vivo measure of cytochrome P450 2E1 activity. Br. J. Clin. Pharmacol. *58*, 190–200.

4. Hansson, L.O., Bolton-Grob, R., Massoud, T., and Mannervik, B. (1999). Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling. J. Mol. Biol. *287*, 265–276.

5. Christians, F.C., Scapozza, L., Crameri, A., Folkers, G., and Stemmer, W.P. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. Nat. Biotechnol. *17*, 259–264.

6. Griswold, K.E., Kawarasaki, Y., Ghoneim, N., Benkovic, S.J., Iverson, B.L., and Georgiou, G. (2005). Evolution of highly active enzymes by homology-independent recombination. Proc. Natl. Acad. Sci. USA *102*, 10082–10087.

7. Broo, K., Larsson, A.K., Jemth, P., and Mannervik, B. (2002). An ensemble of θ class glutathione transferases with novel catalytic properties generated by stochastic recombination of fragments of two mammalian enzymes. J. Mol. Biol. *318*, 59–70.

8. Drummond, D.A., Silberg, J.J., Meyer, M.M., Wilke, C.O., and Arnold, F.H. (2005). On the conservative nature of intragenic recombination. Proc. Natl. Acad. Sci. USA *102*, 5380–5385.

9. 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*, e112.

10. Otey, C.R., Silberg, J.J., Voigt, C.A., Endelman, J.B., Bandara, G., and Arnold, F.H. (2004). Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. Chem. Biol. *11*, 309–318.

11. Brock, B.J., and Waterman, M.R. (2000). The use of random chimeragenesis to study structure/function properties of rat and human P450c17. Arch. Biochem. Biophys. *373*, 401–408.

12. Ramarao, M.K., Straub, P., and Kemper, B. (1995). Identification by in vitro mutagenesis of the interaction of two segments of C2MstC1, a chimera of cytochromes P450 2C2 and P450 2C1. J. Biol. Chem. *270*, 1873–1880.

13. Lewis, D.F.V. (2001). Guide to Cytochromes P450: Structure and Function (London, New York: Taylor & Francis).

14. Mansuy, D. (1998). The great diversity of reactions catalyzed by cytochromes P450. Comp. Biochem. Physiol. C Pharmacol. Toxicol. Endocrinol. *121*, 5–14.

15. Ruettinger, R.T., Wen, L.P., and Fulco, A.J. (1989). Coding nucleotide, 5′ regulatory, and deduced amino-acid sequences of P-450BM-3, a single peptide cytochrome P-3′450-NADPH-P-450 reductase from *Bacillus megaterium*. J. Biol. Chem. *264*, 10987–10995.

16. Ortiz de Montellano, P.R. (1995). Cytochrome P450: Structure, Mechanism, and Biochemistry (New York: Plenum Press).

17. Cirino, P.C., and Arnold, F.H. (2002). Regioselectivity and activity of cytochrome P450BM-3 and mutant F87A in reactions driven by hydrogen peroxide. Adv. Synth. Catal. *344*, 932–937.

18. Gustafsson, M.C., Roitel, O., Marshall, K.R., Noble, M.A., Chapman, S.K., Pessegueiro, A., Fulco, A.J., Cheesman, M.R., von Wachenfeldt, C., and Munro, A.W. (2004). Expression, purification, and characterization of *Bacillus subtilis* cytochromes P450 CYP102A2 and CYP102A3: flavocytochrome homologues of P450 BM3 from *Bacillus megaterium*. Biochemistry *43*, 5474–5487.

19. Otey, C.R., Bandara, B., Lalonde, J., Takahashi, K., and Arnold, F.H. (2006). Preparation of human metabolites of propranolol using laboratory-evolved bacterial cytochromes P450. Biotechnol. Bioeng. *93*, 494–499.

20. Lee, C.R., Pieper, J.A., Frye, R.F., Hinderliter, A.L., Blaisdell, J.A., and Goldstein, J.A. (2003). Tolbutamide, flurbiprofen, and losartan as probes of CYP2C9 activity in humans. J. Clin. Pharmacol. *43*, 84–91.

21. Lasker, J.M., Huang, M.T., and Conney, A.H. (1982). In vitro activation of zoxazolamine metabolism by flavone. Science *216*, 1419–1421.

22. Otey, C.R., and Joern, J.M. (2003). High-throughput screen for aromatic hydroxylation. Methods Mol. Biol. *230*, 141–148.

23. Schwaneberg, U., Schmidt-Dannert, C., Schmitt, J., and Schmid, R.D. (1999). A continuous spectrophotometric assay for P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A. Anal. Biochem. *269*, 359–366.

24. McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematics, Statistics, and Probability. University of California, Berkeley, Berkeley, CA, 281–297.

25. Larsson, A.K., Emren, L.O., Bardsley, W.G., and Mannervik, B. (2004). Directed enzyme evolution guided by multidimensional analysis of substrate-activity space. Protein Eng. Des. Sel. *17*, 49–55.

26. Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. *323*, 297–307.

27. Sykes, M.T., and Levitt, M. (2005). Describing RNA structure by libraries of clustered nucleotide doublets. J. Mol. Biol. *351*, 26–38.

28. Shet, M.S., Fisher, C.W., Holmans, P.L., and Estabrook, R.W. (1993). Human cytochrome P450 3A4: enzymatic properties of a purified recombinant fusion protein containing NADPH-P450 reductase. Proc. Natl. Acad. Sci. USA *90*, 11748–11752.

29. Shet, M.S., Fisher, C.W., Arlotto, M.P., Shackleton, C.H., Holmans, P.L., Martin-Wixtrom, C.A., Saeki, Y., and Estabrook, R.W. (1994). Purification and enzymatic properties of a recombinant fusion protein expressed in *Escherichia coli* containing the domains of bovine P450 17A and rat NADPH-P450 reductase. Arch. Biochem. Biophys. *311*, 402–417.

30. Harlow, G.R., and Halpert, J.R. (1996). Mutagenesis study of Asp-290 in cytochrome P450 2B11 using a fusion protein with rat NADPH-cytochrome P450 reductase. Arch. Biochem. Biophys. *326*, 85–92.

31. Fairhead, M., Giannini, S., Gillam, E.M., and Gilardi, G. (2005). Functional characterisation of an engineered multidomain human P450 2E1 by molecular Lego. J. Biol. Inorg. Chem. *10*, 842–853.

32. Fuziwara, S., Sagami, I., Rozhkova, E., Craig, D., Noble, M.A., Munro, A.W., Chapman, S.K., and Shimizu, T. (2002). Catalytically functional flavocytochrome chimeras of P450 BM3 and nitric oxide synthase. J. Inorg. Biochem. *91*, 515–526.

33. Kubota, M., Nodate, M., Yasumoto-Hirose, M., Uchiyama, T., Kagami, O., Shizuri, Y., and Misawa, N. (2005). Isolation and functional analysis of cytochrome P450 CYP153A genes from various environments. Biosci. Biotechnol. Biochem. *69*, 2421–2430.

34. Sevrioukova, I.F., Li, H., Zhang, H., Peterson, J.A., and Poulos, T.L. (1999). Structure of a cytochrome P450-redox partner electron-transfer complex. Proc. Natl. Acad. Sci. USA *96*, 1863–1868.

35. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. *31*, 3497–3500.

36. Klein, M.L., and Fulco, A.J. (1993). Critical residues involved in FMN binding and catalytic activity in cytochrome P450BM-3. J. Biol. Chem. *268*, 7553–7561.

37. Nelson, D. (2006). Cytochrome P450 Homepage (http://drnelson.utmem.edu/CytochromeP450.html).

38. Higuchi, R., Krummel, B., and Saiki, R.K. (1988). A General method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Res. *16*, 7351–7367.

39. Barnes, H.J., Arlotto, M.P., and Waterman, M.R. (1991). Expression and enzymatic activity of recombinant cytochrome P450 17α-hydroxylase in *Escherichia coli*. Proc. Natl. Acad. Sci. USA *88*, 5597–5601.

40. Schwaneberg, U., Sprauer, A., Schmidt-Dannert, C., and Schmid, R.D. (1999). P450 monooxygenase in biotechnology I: single-step, large-scale purification method for cytochrome P450 BM-3 by anion-exchange chromatography. J. Chromatogr. A. *848*, 149–159.

41. Otey, C.R. (2003). High-throughput carbon monoxide binding assay for cytochromes. In Directed Enzyme Evolution: Screening and Selection Methods, F.H. Arnold and G. Georgiou, eds. (Totowa, NJ: Humana Press), p. 450.

42. Cirino, P.C., and Arnold, F.H. (2003). A self-sufficient peroxide-driven hydroxylation biocatalyst. Angew. Chem. Int. Ed. Engl. *42*, 3299–3301.

43. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

44. DeLano, W.L. (2002). The PyMOL Molecular Graphics System (http://www.pymol.org).