

Improving catalytic function by ProSAR-driven enzyme evolution

Richard J Fox^{1,3}, S Christopher Davis^{1,3}, Emily C Mundorff^{1,3}, Lisa M Newman¹, Vesna Gavrilovic¹, Steven K Ma¹, Loleta M Chung¹, Charlene Ching¹, Sarena Tam¹, Sheela Muley¹, John Grate¹, John Gruber¹, John C Whitman¹, Roger A Sheldon² & Gjalb W Huisman¹

We describe a directed evolution approach that should find broad application in generating enzymes that meet predefined process-design criteria. It augments recombination-based directed evolution by incorporating a strategy for statistical analysis of protein sequence activity relationships (ProSAR). This combination facilitates mutation-oriented enzyme optimization by permitting the capture of additional information contained in the sequence-activity data. The method thus enables identification of beneficial mutations even in variants with reduced function. We use this hybrid approach to evolve a bacterial halohydrin dehalogenase that improves the volumetric productivity of a cyanation process ~4,000-fold. This improvement was required to meet the practical design criteria for a commercially relevant biocatalytic process involved in the synthesis of a cholesterol-lowering drug, atorvastatin (Lipitor), and was obtained by variants that had at least 35 mutations.

Although interest in the use of enzymes as biocatalysts for chemical applications is increasing^{1,2}, the performance of natural enzymes is rarely adequate for commercially viable processes. Many enzyme properties, such as specific activity, stability, chemo- and enantioselectivity, susceptibility to substrate and/or product inhibition and sensitivity to rapidly changing conditions, are difficult to optimize using rational design. Despite advances in protein engineering, there is a continuing need for more efficient and effective methods to improve multiple biochemical characteristics of an enzyme simultaneously^{3,4}.

Directed evolution is a powerful tool for protein optimization. The most efficient methods combine multiple rounds of diversity generation and gene recombination with functional screening to identify improved variants. The iterative nature of this approach results in stepwise improvements in overall function, yielding substantial improvements in desired enzyme properties⁵⁻⁹.

Directed evolution can be performed in a variety of ways¹⁰, which are distinguished, for example, by the approach to generating libraries from available diversity. The field of *in vitro* recombination-based directed evolution⁹ has focused on alternative methods for generating gene libraries rather than fundamentally altering the efficiency of the evolutionary process^{7,10}, and despite advances, there remains a need to make the process of increasing enzyme function more efficient^{3,4}.

Our approach involves focusing the selective pressure on the mutations themselves rather than on the mutated gene. Quantitative structure-activity relationships (QSAR) have been used extensively in small-molecule¹¹ and peptide optimization¹² and have been proposed to be useful in protein engineering^{13,14}. The statistical modeling efforts are motivated by the desire to establish causal relationships between the structures of interacting molecules (e.g., small-molecule descriptors and

amino acid sequences) and measurable properties of scientific or commercial interest (e.g., binding affinity and catalytic activity). Such models can then be interrogated to make decisions about how to modify a molecule's structure to achieve improvements in desired properties. By formalizing the decision making processes about which mutations to include in combinatorial libraries, the ProSAR algorithm is an extension of traditional SAR-based approaches to molecular optimization.

This study applies the concepts of QSAR to the problem of enzyme engineering. A multivariate protein optimization strategy based on protein sequence activity relationships (ProSAR)^{15,16} was used to develop a halohydrin dehalogenase (HHDH) with the potential for use in the manufacture of ethyl (*R*)-4-cyano-3-hydroxybutyrate (HN) (Fig. 1), the regulated starting material for the production of the cholesterol-lowering drug atorvastatin (Lipitor). The specifications for the chemical and enantiopurity of HN are tightly controlled, since the hydroxyl present in HN is used to define the second stereocenter in atorvastatin and high chemical purity is essential for successful downstream chemistry.

We projected that an economically viable process could be achieved if the following design criteria were met: complete conversion (100%) of at least 100 g per liter substrate, a volumetric productivity of >20 g product per liter per hour per gram of biocatalyst, a simple HN isolation procedure to recover high quality product in high yield, and a simple enzyme formulation process that obviates the need for extensive enzyme purification.

We first expressed an *Agrobacterium radiobacter* HHDH in *Escherichia coli* and showed that it catalyzed the conversion of ethyl (*S*)-4-chloro-3-hydroxybutyrate (ECHB) to HN at neutral pH (Fig. 1), thereby avoiding the side reactions associated with chemical

¹Codexis, Inc., 200 Penobscot Drive, Redwood City, California 94063, USA. ²Delft University of Technology, Department of Organic Chemistry, Julianalaan 136, 2628 BL Delft, Netherlands. ³These authors contributed equally to this work. Correspondence should be addressed to G.W.H. (gjalt.huisman@codexis.com).

Received 31 October 2006; accepted 17 January 2007; published online 18 February 2007; doi:10.1038/nbt1286

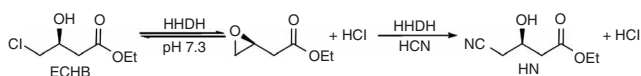


Figure 1 HHDH interconverts halohydrins and epoxides¹⁸ and can accept alternative nucleophiles¹⁷. Chemical cyanation is typically performed at elevated temperatures (80 °C) and pH 9. HHDH catalyzes the single-vessel enzymatic conversion of ethyl (S)-4-chloro-3-hydroxybutyrate (ECHB) to ethyl (R)-4-cyano-3-hydroxybutyrate (HN) in the presence of cyanide at pH 7 where no chemical reaction is observed. The HCl produced in the reaction causes the pH to drop, although both the pH and the cyanide concentration can be maintained by running the reaction in a pHstat with NaCN as the base.

cyanation^{17–19}. HHDH catalyzed the desired transformation with a volumetric productivity of 6×10^{-3} g product per liter per hour per gram of catalyst, indicating that substantial enzyme improvements were needed to enable its commercial application. Traditional hit-shuffling^{6,9,20} (where the best variants were shuffled together) and the multivariate approach were then used in parallel to create variants that were screened for improved function. The apparent advantages of the ProSAR-driven strategy, as described in this report, led us to abandon the traditional hit-based approach after 15 rounds. An additional three rounds of evolution using the ProSAR-driven strategy produced enzymes with the desired characteristics for commercial use.

RESULTS

Optimization strategy

Our multivariate optimization strategy is an iterative process of diversity generation and screening followed by statistical analysis through linear regression on training sets derived from one or more combinatorial libraries per round (Fig. 2). At the end of each round, the best variant from one of the libraries was selected to serve as a template for programming diversity into the next round. This strategy is based on the creation of ProSAR models that can be used to infer the contributions of mutational effects on protein function. Within a given training set consisting of one or more combinatorial libraries, statistical learning was achieved by formulating an equation that correlates mutations with protein function in the following manner:

$$y = c_{1a}x_{1a} + c_{1b}x_{1b} + c_{2a}x_{2a} + c_{2b}x_{2b} + \dots + c_{ja}x_{ja} + c_{jb}x_{jb} + \dots \quad (1)$$

where y is the predicted function (activity) of the protein sequence, c_{ja} is the regression coefficient corresponding to the mutational effect of having residue choice a present at variable position j , and x_{ja} is a variable indicating the presence ($x_{ja} = 1$) or absence ($x_{ja} = 0$) of residue a at position j . In general, we assumed that the mutational effects are mostly additive^{21,22} and that only linear terms corresponding to each mutation's independent effect on function appear in equation (1). When needed, nonlinear terms can be added to capture

putatively important interactions between mutations. However, they were not required in this study¹⁵. The statistical analysis used sequence-activity data from the training set to adjust regression coefficients so that differences between predicted and measured function values are minimized. Because the system of equations may be underdetermined (more unknowns than equations, that is, more mutations than sequence-activity measurements), partial least-squares (PLS) regression was used to perform the linear regression²³. PLS regression has been a very popular method in QSAR applications^{12,24} because it is computationally efficient and helps prevent the kind of overtraining in underdetermined systems that leads to poor predictive capability. Standard multiple linear regression cannot be used in such situations and some form of regularization or capacity control must be employed to prevent overfitting²⁵.

From an initial pool of mutations found through various diversity-generation methods, mutations were tested in semisynthetic²⁰ combinatorial libraries, with a sample of the library screened for activity and sequenced. We generally designed and constructed one to four libraries per round from a pool of 10–50 mutations. Each library generally consisted of 10–30 programmed mutations with an average incorporation rate of $\sim 50\%$ per mutation and, following binomial theory, a s.d. of 1.6–2.7 mutations per sequence. The sequence-activity data were used to build statistical models that assigned to each mutation a regression coefficient corresponding to that mutation's impact on activity¹⁶. The multivariate optimization strategy then evaluated the regression coefficients to decide whether mutations should be retained, discarded or tested again in new combinatorial libraries.

The more often a mutation was observed, the higher was the confidence of its predicted effect on activity. As the quality of the statistical modeling was a function of variables such as assay noise, mutational context and the amount of sequence-activity data available, mutations were retested in new libraries when we were uncertain about their impact on activity (Fig. 2). As mutations were discarded from further analysis, new mutations were added to the remaining pool.

The diversity pool consisted of both recycled diversity coming from ProSAR analysis and new diversity generated. The pool size (10–50) represents a tradeoff between diversity generation and optimization. Creating larger pools of diversity would come at the expense of being

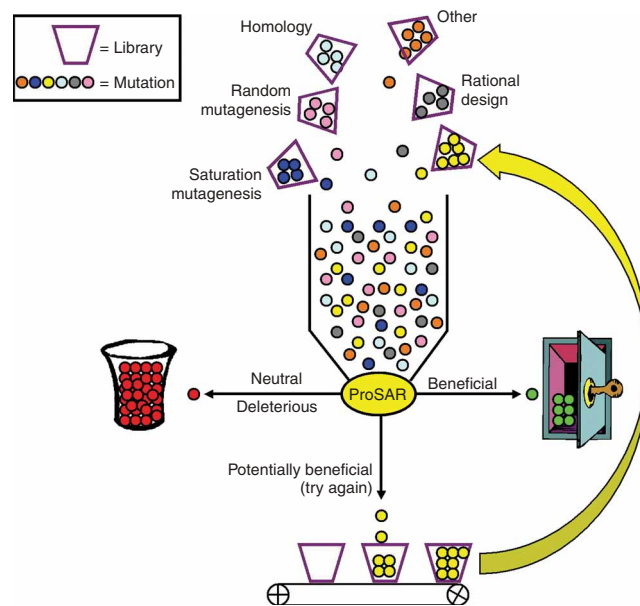


Figure 2 Multivariate optimization of enzymes. At any point, ~ 50 mutations (variables) are evaluated in the combinatorial libraries (in the hopper). The best variants for the desired activity, and a fraction of less improved variants, are sequenced and analyzed as described. After ProSAR analysis, individual mutations are parsed into four classes: 'Beneficial', which are fixed into the population by retention in the next round parental enzyme, 'Potentially beneficial', which are sent back into the hopper for retesting, 'Deleterious', which are discarded and 'Neutral', which have little or no effect on protein function and are discarded. The amount of the diversity under investigation is maintained by addition of additional diversity discovered, for example, through rational design, homologous sequences, saturation or PCR mutagenic libraries or other evolution programs.

Table 1 Library types and number screened

Library type	Tier 1	Tier 2
Hit shuffling	42,710	6,821
Homology	20,390	2,082
Random mutagenesis	58,546	4,524
ProSAR	250,421	30,350
Rational design	49,962	5,973
Site saturation mutagenesis	97,016	16,404
Total	519,045	66,154

The total number of variants screened in each assay tier over the course of the evolution is given for the different library types. The homology, random mutagenesis, rational design and site saturation mutagenesis libraries were used to create diversity. The ProSAR libraries were used to sift through existing diversity as well as generate diversity in the form of random mutations.

unable to efficiently sift through that diversity, whereas smaller pools would lack the diversity that drives the directed evolution process.

The strongest candidates for inclusion in subsequent libraries that were not already part of the new backbone (so called 'potentially beneficial' mutations) were those with positive regression coefficients, particularly if the mutations were seen only once or a few times or if their regression coefficients were relatively large compared to other mutations in the library. Mutations were classified as 'beneficial' and therefore fixed in the next library when they were seen enough times to establish confidence they were contributing to improved activity. Beneficial mutations were already present in the new backbone for the next round. If a mutation with a small regression coefficient was seen many times it was considered to be 'neutral', and if it had a large, negative coefficient, it was considered to be 'deleterious'. Neutral and deleterious mutations were discarded, as they were less likely to contribute significantly to improved activity. Coding mutations present in the final variant that did not traverse through intermediate changes at the same position were seen on average in 2.7 libraries before being fixed ($n = 30$; s.d., 1.4; range, 1–6). A detailed description of the mutation selection and decision making used over a round of evolution is described in the **Supplementary Notes** online.

The activity of the HDDH variants obtained by these methods was measured under conditions that simulate the desired process. A three-tier screening process was used to identify improved variants. In brief, library clones exhibiting no enzymatic activity were removed in the first tier using a high-throughput colorimetric plate screen (~14,000 per round). In the second tier, active variants were tested in miniaturized biocatalytic reactions (~1,700 per round), and variants from a selected subset were sequenced. We generally sequenced ~3*N* functionally diverse variants, where *N* is the number of programmed mutations in a library. The sequence-activity data were then subjected to statistical analysis. To generate the highest quality models, we obtained sequence-activity data for variants exhibiting a broad range of activities, ranging from the highest activity to about half the activity of the most active variant identified in the previous round. In the third tier, up to five of the most active biocatalysts were tested in preparative-scale chemical reactions. The gene encoding the best variant emerging from the third tier was then chosen as the sequence on which the next set of semisynthetic libraries was built. Each round of evolution took 3–4 weeks, including library design and construction, screening, sequencing and statistical analysis.

As already mentioned, traditional hit-shuffling was also used during the course of the evolution. Typically, the top five to ten variants in each round were subjected to classical DNA shuffling and the resulting chimeras screened for improved activity. However, the ProSAR libraries were generally screened more deeply (and more overall effort

was expended on them) than the hit-shuffled libraries because they showed the greatest improvements in activity and generally showed the most promise after screening the first five to ten plates in each round (**Table 1**). Thus, the screening effort between hit-shuffled and ProSAR-based libraries was not the same and a direct comparison of the approaches is not possible. Moreover, because the statistical modeling was able to identify beneficial diversity overlooked by the hit-based approach (discussed in more detail below), a controlled comparison of the methods is difficult to conduct; for example, library sizes and sources of diversity could not be kept constant between the methods without altering the essential way in which they operate. Nevertheless, the apparent advantages of the ProSAR-driven strategy led us to abandon the hit-based approach after 15 rounds.

After 18 iterative cycles of ProSAR-driven semisynthetic shuffling and screening, a population of variants was obtained that enabled the desired process. The activities of a sampling of variants under process conditions throughout the evolution are displayed in **Figure 3**. The final product (HN) had 99.5% purity by GC and an enantiomeric excess of >99.9% *R* (no *S*-enantiomer was detected, starting substrate enantiomeric excess > 99.9% *R*). Each round of evolution gave on average a ~1.5-fold improvement in activity over the parent; no single variant showed more than a threefold improvement in a given round.

Mutation analysis

The population of biocatalysts that met the process design criteria consisted of variants that contained at least 35 (from a pool of 47) mutations (see **Supplementary Notes**). These mutations originated from diverse generation methods, illustrating the effectiveness of using different sources of diversity. The mutations were initially identified by random mutagenesis (47%), site saturation mutagenesis (33%), analysis of naturally occurring homologous sequences from libraries (14%) and analysis of the enzyme's tertiary structure²⁶ (6%). The observed mutations were relatively conservative (conservative, 38%; moderately conservative, 47%; moderately radical, 13%; radical, 2%)²⁷, even though many residues were targeted for full randomization and thus not restricted to the generally conservative spectrum of substitutions accessed by single base-pair changes in the genetic code. The relatively conservative nature of the mutations is consistent with the evolutionary adaptation theory that radical changes in phenotype are

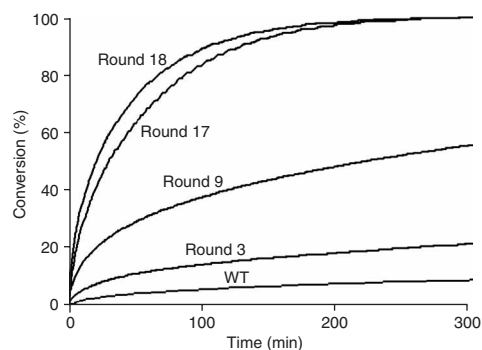


Figure 3 Reaction progress curves for an expression mutant of the wild-type enzyme and representative variants obtained over the course of the evolution at a substrate loading of 130 g/l, substrate/biocatalyst at 100:1 (wt/wt), 40 °C and pH 7.3 (sequences in **Supplementary Notes** online). Although the expression mutant achieves fast conversion over the first few minutes of the reaction, the rate quickly drops, most likely as a result of oxidative instability, thermal instability and/or product inhibition. Based on reactions that completed, the volumetric productivity increased ~4,000-fold.

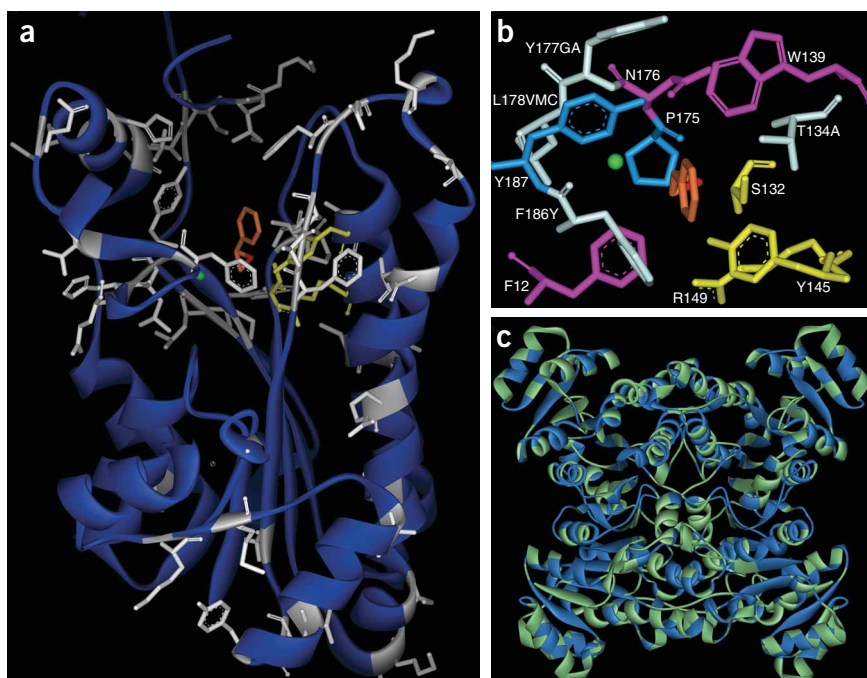


Figure 4 Positions of mutations in the HHDH structure. (a) A monomer of the native HHDH tetramer (pdbid: 1PWZ)²⁶ with (R)-styrene oxide ((R)-SO) shown in orange and Cl⁻ shown in green. Residues that were mutated in the pool of most active variants are shown in white. The catalytic triad S132, Y145, R149 is yellow. (b) A close-up of the active site of HHDH with (R)-SO shown in orange, Cl⁻ in green, the catalytic triad in yellow, unchanged residues not part of the catalytic machinery in blue, mutated residues in the most active population in white, and residues mutated in intermediate active clones in purple. (c) The HHDH tetramer with positions of acceptable mutations shown in green and unchanged positions in blue. All molecular structure figures were drawn with DS Modeling 1.7 (Accelrys).

found at 25 of the 28 positions that are fully conserved between the three haloalcohol dehalogenases, HheA, HheB and HHDH²⁹. Acceptable changes at these fully conserved positions were T7PS, P188S and T189SPC.

We assessed the effectiveness of the ProSAR search strategy by examining all 20 coding

more likely than conservative changes to diverge from an evolutionary optimization path that leads to a specific, improved function²⁸.

Aided by the published wild-type HHDH structure²⁶, we determined that 8 (40%) of the 20 residues proximal to the binding pocket (within 7.5 Å of the bound product in a sphere that encompasses 16% of all the amino acids in the native structure (Fig. 4a)) were mutated in our final population. There are eleven residues that interact with the product or halide ion directly, including two of the catalytic triad residues (S132 and Y145 (Fig. 4b)). Of the other nine interacting residues, the best variants were mutated at four positions: F186Y, T134A, Y177AG and L178CMV. Mutations at three other proximal active-site residues (F12, W139 and N176) had been observed over the course of evolution. F12 could be replaced by I, L or Y with no apparent effect, whereas W139 and N176 were mutated into an apparent salt bridge in two intermediate, active variants W139R/N176E and W139D/N176R. Variants containing the latter mutations had increased function, but better variants were later obtained that did not include these mutations. The remaining two residues in contact with the product, P175 and Y187, remained unchanged throughout the evolution process despite being tested in saturation mutagenesis libraries twice.

HHDH is relatively tolerant to mutations, as alternative residues (those with positive regression coefficients in at least one model) could be introduced at 100 of the 254 amino acids (Fig. 4c)^{26,29}. It is possible that further positions could tolerate such substitutions as we did not quantify substitutability at every position in the protein. Such acceptable diversity was found at 64% percent of the solvent-accessible positions (defined as having >40% of the solvent-accessible surface area, and constituted 15% of the total protein). In contrast, such acceptable diversity was found at only 27% of buried positions (<10% solvent-accessible surface area, 51% of the total protein). The SNDX motif in haloalcohol dehalogenases at residues 78–81, which corresponds to the conserved NNAG motif found in members of the SDR superfamily and is implicated in the proton-relay system²⁶, was unchanged throughout the optimization process, despite the fact that only N79 is conserved in the native haloalcohol dehalogenase homologs HheA and HheB²⁹. Also, no acceptable mutations were

mutations present in the production variant and two silent mutations (believed to be beneficial) that were first observed in a training set of sequence-activity data and became fixed in later rounds (Table 2). The remaining 17 coding mutations in the production variant consisted of those first identified by random mutagenesis (in which case they were not identified in a combinatorial context and thus cannot be viewed as part of the search algorithm) or those that became fixed after their first observation (in which case they were not observed in a subsequent round where their importance could be verified). An additional 13 silent changes in the final variant were not tracked. Analysis of the mutations in subsequent training sets showed that 86% of the mutations were beneficial ($P < 4.3 \times 10^{-4}$; null hypothesis: $\leq 50\%$ beneficial). Among these beneficial mutations, 58% (11 of 19) were not initially identified in any of the most active hit sequences. Such mutations would be overlooked in a given round by blind evolution where only a few of the best variants are taken forward in the next round of evolution. If missed, they would have to be rediscovered through additional mutagenesis. In contrast, only 32% of the beneficial mutations were initially found in the five most active variants for a given round, whereas the remaining 10% were first seen in the ten most active variants. These

Table 2 Hit-oriented optimization methods overlook important mutations in a given round

	First observed in hit (%)	First observed in non-hit (%)	Total (%)
Beneficial	8 (36)	11 (50)	19 (86)
Not Beneficial	1 (5)	2 (9)	3 (14)
Total	9 (41)	13 (59)	22 (100)

For each mutation in the production variant that was initially observed in a training set of sequence-activity data and fixed in a subsequent round ($n = 22$), the rank of the most active sequence carrying the mutation in the initial observation was recorded. Mutations were first observed in either hit sequences (those in the top ten variants greater than the control activity) or mutations were first observed in only non-hit sequences. A mutation was classified as beneficial if its regression coefficient was positive in the next training set in which it was observed; otherwise the mutation was classified as not beneficial.

data demonstrate that ProSAR can help identify beneficial mutations in sequences that are not among the most highly active.

DISCUSSION

The ProSAR-driven approach unmask the individual mutations that confer improved function on the biocatalyst. This approach is particularly important in cases where individual mutations do not substantially improve function. Notably, over the course of an extensive diversity-generation campaign (Table 1), we never observed more than a 1.5-fold improvement over the parent for any variant containing a single mutation. Some traditional hit-based directed evolution strategies use a gentle recombination format that is relatively inefficient; had we shuffled only the 'hit' variants, we would have missed many important mutations in a given round and they could have been recovered only through additional mutagenesis efforts (Table 2). Combining ProSAR analysis with library-generation techniques such as semisynthetic shuffling²⁰ enables the power of recombination to be exploited by rapidly sifting through diversity.

Linear forms of the ProSAR models are predicated on the idea that, for local regions of the sequence-function landscape, mutations typically display additive effects. However, the assumption does not hold in all cases and may break down with increased mutational load. Moreover, there is uncertainty about the true effects of a mutation, particularly if it was only observed once or a few times. Thus, it is generally not sufficient to identify individually beneficial mutations and recombine them without first ensuring they will function well with other mutations in a given context or combinatorial library. The ProSAR strategy is designed to strike a balance between mild and more aggressive forms of recombination while moving in a direction that (locally, on a high dimensional sequence-function landscape³⁰) is most likely to improve function.

Several additional observations regarding the ProSAR-driven strategy are worth noting.

1. We theorized originally that the ProSAR strategy would best be applied to determine the quality of the specific set of mutations being tested in a library. However, our semisynthetic libraries contained approximately one random mutation per variant among the various programmed mutations. The ability to identify some of these point mutations as beneficial, even in the context of less active variants, provided a source of diversity that would not have been available without statistical analysis.

2. Traditional *in vitro* evolution strategies that do not attempt to identify the individual mutations contributing to improved function run a higher risk of carrying deleterious mutations into the next round. For example, mutations present in the current set of hits or in hits from previous rounds may be genetic hitchhikers that can reduce function if carried forward into new libraries. Adopting a multivariate, mutation-oriented approach to optimization allows one to work with libraries in parallel while minimizing such risks. The information gained about each mutation is available for use in any future library design, not only in the next round. This allows more hypotheses to be tested in parallel, facilitating the rapid accrual of beneficial diversity in the evolving population and providing concomitant gain in function.

3. Our approach is well suited to multi-objective optimization problems, as it is capable of providing information about the contributions of specific mutations to enhancements of various characteristics (activity, thermostability, enantioselectivity, among others) relevant to the development of efficient biocatalytic processes. Models could be developed for individual objectives and mutations conferring improvements to one or more objectives could be identified separately and recombined in subsequent rounds. Although our work with

HHDH offered several opportunities to selectively engineer different properties (e.g., increased specific activity, increased stability to ECHB and HN, increased thermostability and oxidative stability, increased resistance to changes in ionic strength as well as reduced inhibition by chloride, cyanide and HN), we did not take advantage of the multi-objective capabilities of the ProSAR optimization strategy. We are currently applying such multi-objective strategies using ProSAR for the optimization of other enzymes.

4. Although a relatively high-throughput screening capacity was available, only a modest screening capability was necessary to build statistical models. Variants with an ~4,000-fold improvement in volumetric productivity were obtained after screening ~60,000 variants in a quantitative second-tier assay. A small set of active, functionally diverse variants was all that was required to build models. For example, a library size of 10^9 requires only about 100 sequence-activity measurements ($3N$ sequenced variants for $N = 30$ mutated positions corresponds to a theoretical library size of $2^{30} = 10^9$). This is sufficient as only desirable mutations need to be identified (not the best possible enzyme variant) and is appropriate when combinatorial libraries are constructed in such a way that we expect to see each mutation a number of times in the training set. Conversely, when libraries are generated through random mutagenesis, statistical modeling is not applicable as the majority of mutations are only seen once and making inferences about mutational effects is either trivial (in the case of a single mutation) or not possible (in the case of multiple mutations).

5. Statistical analysis of the sequence-activity data allowed us to better validate and tune the screening methodology, a key component of directed evolution that uses laboratory rather than actual process conditions to measure the function of the biocatalyst variants. Poor correlation between the measured and statistically predicted activities in this analysis, beyond the assay noise, has at least two potential causes. Either the model fails to capture important interactions between the mutations¹⁵, or there is a problem with the assays. Reevaluation of the assays typically uncovered subtle but important problems that were then remedied. In one case we found that the thermal stress placed upon the variants in the miniature high-throughput reactions was less than that of the preparative-scale

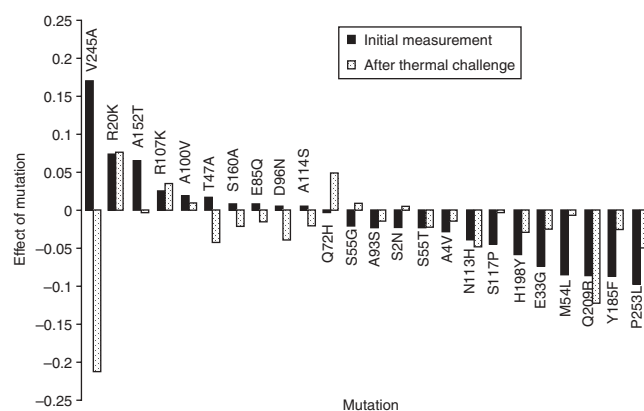


Figure 5 Function contribution by individual mutations under different screening conditions. Sequence-activity models were built for the combinatorial library using function data obtained from the second-tier screen of HHDH activity measurements. The initial measurement was obtained by omission of the 2 h incubation of the lysate at 50 °C before addition of the reaction mixture. Mutations are shown along the x-axis. The predicted impact on function (model regression coefficient) for each mutation is shown on the y-axis.

chemical reactions. Rather than discard the unstable sequences, we retested the sequenced variants under more stringent high-throughput screening conditions. A second statistical model was built, which showed that the very best mutation for increased function under previous high-throughput conditions, V245A, was deleterious under more stringent conditions (Fig. 5). In contrast, Q72H was found to be important for function if thermostability was an operational requirement. Incorporating this information into subsequent optimization led to increased stability and further gains in activity.

These results demonstrate the effectiveness of combining multi-variate optimization theories with genetic recombination and laboratory evolution methodologies for the improvement of biocatalyst function. This generally applicable approach, coupled with rational and random approaches to diversity generation, led to an ~4,000-fold improvement in the performance of HHDH and an industrial process involving the optimized biocatalyst³¹.

METHODS

Materials. All chemicals were purchased from Sigma-Aldrich.

Libraries. Mutational diversity was generated at the outset of the project and throughout the evolution program, as necessary, using random mutagenesis via error-prone PCR and single-gene shuffling³²; semisynthetic shuffling approaches³³, using sequence diversity from the nearest family members²⁹; and information for specific structural regions, such as the active site, the halide-binding pocket and structural loops²⁶. Saturation mutagenesis was also carried out at regions of apparent functional importance, including the randomization of pair-wise positions that were thought to have interactions³⁴. Random mutations present in the ProSAR libraries served as another source of diversity. The number of variants screened for each library type is given in Table 1.

HHDH library construction and expression. The gene for *A. radiobacter* halohydrin dehalogenase (GenBank accession number AF397296) was codon-optimized for expression in *E. coli* based on the amino acid sequence of the halohydrin dehalogenase from *Agrobacterium* sp. The gene was synthesized using 60-mer oligomers and cloned into expression vector pCK110700 under the control of a T5 promoter. The vector was transformed into *E. coli* TOP10 (Invitrogen) from which plasmid DNA was prepared using standard methods. The plasmid DNA was then transformed into *E. coli* BL21 (Stratagene) using standard methods.

HHDH libraries were constructed according to previously described methods^{9,33}, cloned into vector pCK110700, transformed and expressed in *E. coli* BL21 after passage through *E. coli* TOP10.

Oligonucleotides designed to unlink all programmed mutations were spiked into semisynthetic libraries to give ~50% incorporation for single oligonucleotides (lower in the case of multiple oligos at one position). Such libraries contained a random mutation rate of ~1/variant with an s.d. of 1. Unlinking of mutations was accomplished by designing oligos in such a way as to ensure each programmed mutation or combination of mutations was present on oligos within a defined region. In cases where only one position was targeted, either separate oligos were designed for each mutation or a single oligo was designed to have ambiguity at the targeted position corresponding to the different desired codons. Similarly, for two or more mutations near each other, one or more oligos were designed in such a way that all combinations of mutations were made possible by using zero or more ambiguous codons at zero or more positions.

First-tier screen of HHDH activity. On day 1, clones encoding HHDH variants were picked from a QTray (Genetix USA) containing 200 ml Luria Bertani (LB) agar and 1% glucose, 30 µg/ml chloramphenicol (cam) into 384 shallow-well master plates (Nalge Nunc International) containing medium (70 µl per well of 2× YT + 1% glucose, 30 µg/ml cam) using a QBot robot colony picker (Genetix USA) for overnight growth at 30 °C, 250 r.p.m., 2 inch throw and 85% relative humidity (RH). A negative control (*E. coli* BL21 with pCK110700) and a positive control (*E. coli* BL21 with pCK110700 containing the parent HHDH for a particular library) were included. The master well plate cultures were covered with AirPore tape (Qiagen).

On day 2, the master plate cultures were gridded directly from the 384-well plates onto nylon membranes (Pall Biotryne B) and then placed onto a QTray containing 200 ml LB agar + 1% glucose, 30 µg/ml cam. The QTrays were incubated at 30 °C for 8–12 h until growth was detected. Each nylon membrane was transferred to a QTray containing inducing medium (200 ml LB agar (no glucose) + 1 mM isopropyl-β-D-thiogalactoside (IPTG), 30 µg/ml cam), and the QTrays were incubated overnight at 23 °C or room temperature.

On day 3, a low-melt agarose solution (150 ml 10 mM Tris, pH 7.0 with 2.0% low-melt agarose and 0.004% bromocresol purple) that had been prepared the previous day and incubated at 37 °C overnight was amended with ECHB (20 mM) and HN (400 mM), mixed, poured into a QTray, and allowed to solidify. The nylon membrane with the induced colonies was removed from the QTray and inverted onto the assay plate. Conversion of ECHB to the epoxide under these conditions leads to the liberation of HCl, which causes a change in the color of the pH indicator from purple to yellow. Membranes were imaged during the first hour of the reaction through the inverted QTray using the Alpha Imaging ChemStation (Alpha Innotech) with an aperture setting of 4 and a 420 nm (±10 nm) filter. The intensity data for each imaged spot were then normalized to the value of the negative control spots. A normalized value >1 indicated the presence of HHDH activity. Active clones from this screen were further characterized in the second-tier assay.

Second-tier screen of HHDH activity. Potentially improved variants (10 µl) were transferred from the master plates into the wells of 96-well shallow plates (each well containing 200 µl LB and 1% glucose, 30 µg/ml cam) for overnight growth at 30 °C, 250 r.p.m., 2 inch throw and 85% RH. The positive controls were picked from the prescreened master well plates.

The next day, 10 µl aliquots of the overnight cultures were subcultured in deep-well microtiter plates (Costar), each well containing 300 µl 2× YT, 100 mM NaH₂PO₄/Na₂HPO₄ pH 7, 1 mM MgSO₄ and 30 µg/ml cam. The plates were incubated at 30 °C, 250 r.p.m., 2 inch throw and 85% RH for 2–4 h, until the cell density reached an OD₆₀₀ of ~0.6. The plates were then induced with 1 mM IPTG and incubated overnight at 30 °C, 250 r.p.m. and 85% RH.

The next day, the plates were centrifuged (3,220g, 10 min, 4 °C) to pellet the cells, and the spent medium was discarded. (The plates may be stored at –80 °C for 1 h to aid in cell breakage.) The pellets in each well were resuspended in 200 µl B-PER lysing solution (Pierce) containing 2.04 M HN and ~200 U/µl DNase (200 µl per well) and incubated at 50 °C with shaking for 2 h. A solution of 50 mM sodium phosphate pH 7.0–7.2 with 1 M HCN, 2 M NaCl and 20 mM ECHB was prepared in a fume hood; 200 µl of the solution were added to the lysed cells in each well. The reaction mixture was designed to mimic the end of the reaction; reaction product (HN) was added to the reaction mixture to screen under product-inhibited conditions. The plates were heat sealed and shaken at room temperature ~22 °C for 120 min. After shaking, the plates were unsealed, and 1 ml of 1 mM thymol (internal standard for extraction efficiency) in ethyl acetate was added to each well. The plates were resealed and shaken vigorously for ~2 min. After phase separation, 150-µl aliquots of the upper layer were transferred into 96-well shallow plates, resealed and stored at –20 °C until analyzed.

The remaining ECHB in the reaction mixture was analyzed by gas chromatography (GC) with an Agilent 19091J-413 HP-5 5% phenyl methyl siloxane column, 30.0 m long × 320 µm ID × 0.25 µm nominal, at a flow rate of 2.6 ml/min, using a program of 1 min at 100 °C, 50 °C/min for 2 min, 2 min hold, 10 min cycle time. The detector conditions were 300 °C, 40 ml/min H₂ and 450 ml/min clean, dry air. Under these conditions, retention times for ECHB, HN and thymol were 2.51 min, 2.92 min and 3.04 min, respectively. Activity was characterized by the quantity of ECHB remaining, normalized to the extraction efficiency (area ECHB/area thymol).

Third-tier screen of HHDH activity. Growth medium (10 l) containing 0.528 g/l (NH₄)₂SO₄, 7.5 g/l of K₂HPO₄·3H₂O, 3.7 g/l of KH₂PO₄, 2 g/l of Tastone-154 yeast extract, 0.05 g/l FeSO₄·7H₂O and 3 ml/l of a trace element solution containing 2 g/l of CaCl₂·2H₂O, 2.2 g/l of ZnSO₄·7H₂O, 0.5 g/l MnSO₄·H₂O, 1 g/l CuSO₄·7H₂O, 0.1 g/l Na₂B₄O₇·10H₂O, and 0.5 g/l EDTA, was prepared at 30 °C. The starter culture, a late-stage exponential culture of *E. coli* BL21 equipped with a plasmid containing the isolated HHDH variant gene of choice, was added to the fermentor, which was then agitated

at 500–1,500 r.p.m. and aerated to maintain a dissolved oxygen level of at least 30% saturation. The pH of the culture was controlled at 7.0 by addition of 20% vol/vol NH_4OH . After the culture reached an OD_{600} of 40, the temperature was reduced to 25 °C, and the expression of haloalcohol dehalogenase was induced by the addition of IPTG to a final concentration of 1 mM. The culture was grown for another 15 h. After expression, the cells were harvested by centrifugation and washed with 10 mM potassium phosphate buffer, pH 7.0. The cell paste was used directly in the downstream recovery process or was stored at –80 °C until use.

The cell paste was washed by suspending one volume wet cell paste in three volumes of 100 mM Tris/sulfate (pH 7.2) followed by centrifugation at 5,000g for 40 min at ambient temperature, ~22 °C. The washed cell paste was suspended in two volumes of 100 mM Tris/sulfate (pH 7.2). The intracellular HHDH was released from the cells by passing the suspension through a homogenizer in two passes using a pressure of 14,000 psig for the first pass and 8,000 psig for the second pass. The cell lysate was allowed to cool to 4 °C between passes through the homogenizer. The lysate was warmed to room temperature, ~22 °C, followed by the addition of 10% wt/vol solution of polyethyleneimine, pH 7.2, (0.6–1.0% wt/vol final concentration) and stirring for 30 min. The homogenate was centrifuged at 5,000–10,000g at ambient temperature, ~22 °C, for 30–60 min. The supernatant was decanted and dispensed in shallow containers, frozen at –20 °C and lyophilized to a powder that was stored at –80 °C.

In a 170-ml flask connected to an automatic titrator, 1.5 g NaCN was dissolved in 50 ml water. The vessel was sealed; the solution was heated to 40 °C and the pH was adjusted to 7 with concentrated H_2SO_4 . The enzyme was added as a lyophilized powder (1–50 g/l, depending upon the activities of the variants), followed by 5 g ECHB. The automatic titrator maintained the pH at 7 by the addition of 4 M NaCN. The progress of the reactions was monitored by recording the cumulative volume of the NaCN solution added.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank David Gray, Alice Wang, Su Chen, John Peterson, Walter Heath, Tim Brandon, Jon Postlethwaite, Anjali Srivastava and Bill Dewhirst for help with bioprocess development and production; Malissa Jefferson and Susan Louie for additional assay support; Patricia Babbitt, Pim Stemmer, Lynne Gilson, Lori Giver, Birthe Borup, Anke Krebber, Stephen DelCardayre and Jonathan Blanding for careful reading of the manuscript and helpful suggestions; and three anonymous reviewers for insightful commentary and critical feedback.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/naturebiotechnology/>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>

- Schmid, A. *et al.* Industrial biocatalysis today and tomorrow. *Nature* **409**, 258–268 (2001).
- Schoemaker, H.E., Mink, D. & Wubbolts, M.G. Dispelling the myths—biocatalysis in industrial synthesis. *Science* **299**, 1694–1697 (2003).
- Hibbert, E.G. & Dalby, P.A. Directed evolution strategies for improved enzymatic performance. *Microb. Cell Fact.* **4**, 29 (2005).
- Tawfik, D.S. Biochemistry. Loop grafting and the origins of enzyme species. *Science* **311**, 475–476 (2006).
- Castle, L.A. *et al.* Discovery and directed evolution of a glyphosate tolerance gene. *Science* **304**, 1151–1154 (2004).
- Cramer, A., Raillard, S.A., Bermudez, E. & Stemmer, W.P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291 (1998).
- Ness, J.E. *et al.* Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat. Biotechnol.* **20**, 1251–1255 (2002).
- Ness, J.E. *et al.* DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **17**, 893–896 (1999).
- Stemmer, W.P.C. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
- Yuan, L., Kurek, I., English, J. & Keenan, R. Laboratory-directed protein evolution. *Microbiol. Mol. Biol. Rev.* **69**, 373–392 (2005).
- Kubinyi, H. QSAR and 3D QSAR in drug design Part1: methodology. *Drug Disc. Today* **2**, 457–467 (1997).
- Hellberg, S., Sjöström, M. & Wold, S. The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure-activity relationship. *Acta Chem. Scand. B* **40**, 135–140 (1986).
- Eroshkin, A.M., Fomin, V.I., Zhilkin, P.A., Ivanisenko, V.A. & Kondrakhin, Y.V. PROANAL version 2: multifunctional program for analysis of multiple protein sequence alignments and studying structure-activity relationships in protein families. *Comp. Appl. Biosci.* **11**, 39–44 (1995).
- Eroshkin, A.M., Zhilkin, P.A. & Fomin, V.I. Algorithm and computer program Pro_Anal for analysis of relationship between structure and activity in a family of proteins or peptides. *Comput. Appl. Biosci.* **9**, 491–497 (1993).
- Fox, R. Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J. Theor. Biol.* **234**, 187–199 (2005).
- Fox, R. *et al.* Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng.* **16**, 589–597 (2003).
- Nakamura, T., Nagasawa, T., Yu, F., Watanabe, I. & Yamada, H. A new catalytic function of haloalcohol hydrogen-halide-lyase, synthesis of beta-hydroxynitriles from epoxides and cyanide. *Biochem. Biophys. Res. Commun.* **180**, 124–130 (1991).
- van den Wijngaard, A.J., Reuvekamp, P.T. & Janssen, D.B. Purification and characterization of haloalcohol dehalogenase from *Arthrobacter* sp. strain AD2. *J. Bacteriol.* **173**, 124–129 (1991).
- Matsuda, H., Shibata, T., Hashimoto, H. & Kitai, M. Method for producing (R)-4-cyano-3-hydroxybutyric acid lower alkyl ester. US patent 5,908,953 (1999).
- Stemmer, W.P. DNA *in vitro* shuffling by random fragmentation and reassembly *in vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* **91**, 10747–10751 (1994).
- Wells, J.A. Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509–8517 (1990).
- Sandberg, W.S. & Terwilliger, T.C. Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc. Natl. Acad. Sci. USA* **90**, 8367–8371 (1993).
- de Jong, S. SIMPLS: an alternative approach to partial least squares regression. *Chemomet. and Intell. Lab. Sys.* **18**, 251–263 (1993).
- Mee, R.P., Auton, T.R. & Morgan, P.J. Design of active analogues of a 15-residue peptide using D-optimal design, QSAR and a combinatorial search algorithm. *J. Pept. Res.* **49**, 89–102 (1997).
- Bennett, K. & Embrechts, M. An Optimization Perspective on Partial Least Squares. *Advances in Learning Theory: Methods, Models and Applications, NATO Science Series III: Computer & Systems Sciences* vol. 190. (eds. Suykens, J., Horvath, G., Basu, S., Micchelli, J. & Vandewalle, J.) 227–250, (IOS Press, Amsterdam, 2003).
- de Jong, R.M. *et al.* Structure and mechanism of a bacterial haloalcohol dehalogenase: a new variation of the short-chain dehydrogenase/reductase fold without an NAD(P)H binding site. *EMBO J.* **22**, 4933–4944 (2003).
- Li, W.H., Wu, C.J. & Luo, C.C. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71 (1984).
- Hartl, D.L. & Taubes, C.H. Towards a theory of evolutionary adaptation. *Genetica* **102–103**, 525–533 (1998).
- Hylckama Vlieg, J.E.T. *et al.* Haloalcohol dehalogenases are structurally and mechanistically related to short-chain dehydrogenases/reductases. *J. Bacteriol.* **183**, 5058–5066 (2001).
- Kauffman, S., *The Origins of Order* (Oxford University Press, New York, 1993).
- Thayer, A. Competitors want to get a piece of Lipitor. *Chem. Eng. News* **84**, 26–27 (2006).
- Zhang, J.H., Dawes, G. & Stemmer, W.P. Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. USA* **94**, 4504–4509 (1997).
- Stutzman-Engwall, K. *et al.* Semi-synthetic DNA shuffling of aveC leads to improved industrial scale production of doramectin by *Streptomyces avermitilis*. *Metab. Eng.* **7**, 27–37 (2005).
- Reetz, M.T., Wang, L.W. & Bocola, M. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. *Angew Chem. Int. Ed. Engl.* **45**, 1236–1241 (2006).