CHEMISTRY
A EUROPEAN JOURNAL

DOI: 10.1002/chem.201103811

Enhancing the Efficiency of Directed Evolution in Focused Enzyme Libraries by the Adaptive Substituent Reordering Algorithm

Xiaojiang Feng,[a] Joaquin Sanchis,[b] Manfred T. Reetz,[b] and Herschel Rabitz*[a]

Abstract: Directed evolution is a broadly successful strategy for protein engineering in the quest to enhance the stereoselectivity, activity, and thermostability of enzymes. To increase the efficiency of directed evolution based on iterative saturation mutagenesis, the adaptive substituent reordering algorithm (ASRA) is introduced here as an alternative to traditional quantitative structure–activity relationship (QSAR) methods for identifying potential protein mutants with desired properties from minimal sampling of focused libraries. The operation of ASRA depends on identifying the underlying regularity of the protein property landscape, allowing it to make predictions without explicit knowledge of the structure–property relationships. In a proof-of-principle study, ASRA identified all or most of the best enantioselective mutants among the synthesized epoxide hydrolase from *Aspergillus niger*, in the absence of peptide seeds with high *E*-values. ASRA even revealed a laboratory error from irregularities of the reordered *E*-value landscape alone.

Keywords: directed evolution · mutagenesis · optimization · protein engineering · structure–activity relationships

Introduction

Directed evolution[1–8] has emerged as a flexible and very successful method to engineer catalytic properties of enzymes, such as thermostability,[9–12] stability in hostile organic solvents,[13,14] and enantioselectivity.[15–17] The directed evolution procedure integrates gene mutagenesis, expression, and screening (or selection) of libraries of enzyme mutants, consisting typically of 10^3 to 10^6 transformants (clones). The most commonly used mutagenesis methods are the error-prone polymerase chain reaction (epPCR) and saturation mutagenesis, as well as recombinant procedures such as DNA shuffling or variations thereof. The vast majority of the clones produced by any of these methods prove to be non-functional, with the frequency of improved variants (hits) in a given library depending upon the choice of the gene mutagenesis method and how it is applied in each evolutionary step. The screening effort remains the bottleneck in the overall process of directed evolution.[18] For this reason recent research has focused on the development of advanced laboratory techniques and strategies for efficiently probing the vast protein sequence space.[19–24] In particular, the use of reduced amino acid alphabets as specified by the respective codon degeneracy has proven to be effective in raising the quality of focused libraries generated by saturation mutagenesis, for example, NDT codon degeneracy encoding 12 amino acids as building blocks.[17,23] The use of bioinformatics is yet another approach,[25,26] optionally in combination with reduced amino acid alphabets.[25]

In the latter endeavor various computational approaches have also been introduced,[27–38] including techniques such as RCA,[39] SIRCH,[40] SCHEMA,[41,42] FamClash,[43] IPRO,[44] HotSpot Wizard,[45] and ProSAR.[46–48] Most of these methods, with the exception of ProSAR, require a certain degree of knowledge about the structures of the target protein and/or the interaction with its substrate. Inspired by the use of quantitative structure–activity relationships (QSAR) in traditional drug design, the ProSAR (protein sequence activity relationship) algorithm is a statistical approach, based on multivariable least squares regression, to model protein sequence–function relationships.[46–48] The first step of ProSAR utilizes a set of sequence–function data to classify individual mutations as beneficial, neutral, or deleterious. The second step exploits this information to design subsequent libraries characterized by more beneficial mutations and less deleterious ones.[46–48]

Herein, we present an alternative approach referred to as the adaptive substitutent reordering algorithm (ASRA),[49–51] which was previously employed for the discovery and property optimization of small molecules. In this work, ASRA is further developed and applied for property optimization in focused protein libraries. For an enzyme library with N substitution positions (i.e., positions in the amino acid sequence where mutations occur) and S_i substitents (i.e., different types of amino acids; normally $S_i = 20$ for proteins) on the

[a] Dr. X. Feng,[+] Dr. H. Rabitz
Department of Chemistry, Princeton University
Princeton, New Jersey 08544 (USA)
E-mail: hrabitz@princeton.edu

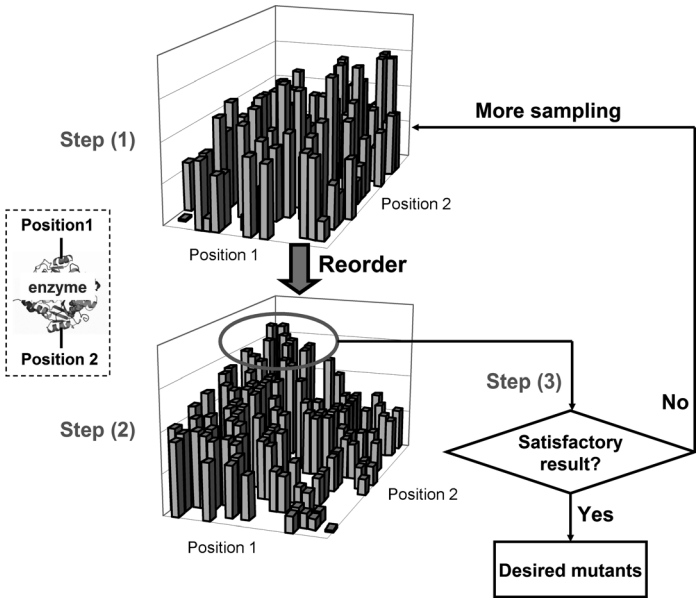
[b] Dr. J. Sanchis,[+] Dr. M. T. Reetz
Max-Planck-Institut für Kohlenforschung
Kaiser-Wilhelm-Platz 1, 45470 Mülheim/Ruhr (Germany)

[+] These authors contributed equally to this work.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/chem.201103811.

i-th position ($i=1,2,\ldots,N$), ASRA views the property y of a protein in the library as a function of an unknown form with N independent variables $y=f(X_1,\ldots,X_i,\ldots,X_N)$, in which $X_i \in [1,Si]$ is a distinct integer assigned to each substituent on the *i*-th position. Consequently, each protein in the library is uniquely associated with an integer vector $\boldsymbol{X}=\{X_1,\ldots,X_i,\ldots,X_N\}$, and the collection of all proteins in the library and their corresponding property values form a discrete N-dimensional property landscape.

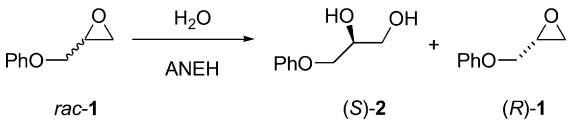
The ASRA formulation above is simple and can represent a broad variety of molecular libraries. However, the setup is not amenable to traditional QSAR methods for molecular discovery. This conclusion follows because 1) there is no a priori way of assigning integers to the substituents (e.g., should alanine be assigned 1 or 9 on a substitution position?), and 2) even when the substituents can be assigned to "proper" values, the form of the function f will remain unknown. Consequently, explicit QSAR functions cannot be constructed from sampling a subset of the target library space for property predictions. ASRA operates in three iterative steps to address these problems and enable efficient property prediction (Scheme 1): 1) Synthesize a small subset of the protein library with simultaneous random mutations on all N substitution positions and then measure the target property value y for each mutant. 2) Based on the data from the sampled mutants, find the optimal substituent ordering (i.e., the best integer assignment X_i* to each amino acid on each substitution position *i*) such that the property landscape is as regular/smooth as possible collectively across all N positions (this step is also called reordering). 3) Based on the observed structure of the reordered property landscape, determine the best set of new mutant(s) to be sampled in the next round of directed evolution experiments. This closed-loop procedure continues until a sufficient number of mutants with desired property value are discovered. Details of the ASRA steps are described in the Materials and Methods section.



Scheme 1. The general steps of ASRA operation: As an example, two substitution positions are selected as mutation targets and each amino acid on each position is assigned a random distinct integer between 1 and 20. The total number of possible mutants is 400. Step 1: randomly synthesize a small subset of the 400 mutants containing substitutions on both positions and measure the target property value for each of them. The initial property landscape is irregular due to the random integer assignments and offers no predictive power. Step 2: identify the optimal integer assignment for each amino acid on each position such that the property landscape is as regular/smooth as possible. Note that when an amino acid on position 1 (or position 2) moves (meaning that its integer assignment is changed from one value to another), all 20 amino acids on the other position will move along with it to keep the indexing consistent. Step 3: based on the geometric features of the reordered property landscape, predict where the best mutants are located (e.g., the circle should be a desired area given the monotonic landscape geometry), and if necessary, synthesize these mutants in the next round of experiment (return to Step 1).

Step 2 is the key step for ASRA operation. Once a regular/smooth property landscape is identified through reordering, protein mutants with more desirable property values can be estimated by a simple observation of the landscape geometry (described in Scheme 1 and later in the article) or an interpolation using empirical fitting functions.[49,51] Unlike traditional QSAR methods as used in ProSAR, Step 2 of ASRA does not depend on assumptions of linearity, additivity, or any explicit form of structure–activity relationship f; it only requires that an underlying regular structure–property relationship (hence a regular property landscape) exists across the library of proteins (i.e., protein biochemistry is sufficiently regular).[52,53] In fact, all existing computational approaches for molecular discovery and/or property estimation, as well as systematic laboratory-driven searches, rely in some fashion on the regularity assumption. The key difference between ASRA and traditional QSAR methods is that ASRA directly exploits the underlying regularity in the structure–property relationship function without determining or assuming its explicit form. As a result, this property of ASRA makes it a generic, "model-free" method applicable to a wide range of situations.

ASRA can in principle be employed together with various laboratory methods in directed evolution. Here we integrate ASRA with the iterative saturation mutagenesis (ISM) method. ISM has recently been proposed and implemented experimentally as a means to accelerate the process of directed evolution, with enhancing enantioselectivity[17,54] or increasing thermostability[55,56] of enzymes being the foci of interest. Appropriate sites, each of which can contain one or more substitution positions, in the enzyme of interest are chosen for saturation mutagenesis. Following library formation and screening, the genes of the hits are used as templates for saturation mutagenesis at the other sites. The choice of the mutation sites depends upon the nature of the catalytic property to be improved. In the case of enantioselectivity, the combinatorial active-site saturation test (CAST)[55,57] is applied, in which all mutation sites around the binding pocket of the enzyme are considered. ISM in the embodiment of CASTing was first utilized in an effort to increase the enantioselectivity of the hydrolytic kinetic resolution of *rac*-**1** (Scheme 2) catalyzed by the epoxide hy-



Scheme 2. Enantioselective ring opening in a racemic mixture of glycidyl phenyl ether (*rac*-**1**) catalyzed by *Aspergillus niger* epoxide hydrolase.

drolase from *Aspergillus niger* (ANEH).[58] Five rounds of iterative CASTing increased the selectivity factor in favor of (*S*)-**2** from $E=4.6$ (wild-type, WT) to $E=115$ (mutant LW202). This remarkable increase was achieved by screening 20000 transformants (clones),[58] which happens to be the same number screened in an earlier study based on epPCR which led to an *E*-value of only 11.[59]

The purpose of the present study is to evaluate the capability of ASRA for predicting desired mutants from limited protein sampling and guiding directed evolution in focused libraries, with the above ANEH-catalyzed reaction serving as the experimental platform. We will demonstrate that ASRA has notable predictive power, making it a useful tool in this type of protein engineering.

Materials and Methods

Laboratory procedures: A racemic mixture of **1** was purchased from Across (Geel, Belgium). Luria Burtani broth (LB) and LB agar were obtained from Invitrogen (Karlsruhe, Germany). Carbenicillin was acquired from Gerbu (Gaiberg, Germany). Methanol and acetonitrile (HPLC grade, LiChrosolv) were purchased from Merck KGaA (Darmstadt, Germany). Distilled water was further purified with a Milli-Q deionization unit (Millipore, Bedford, MA, USA). Sodium dihydrogen phosphate (anhydrous) was obtained from Fluka (Fluka Chemie, Buchs SG, Switzerland). Anhydrous di-sodium hydrogen phosphate was acquired from AppliChem (AppliChem GmbH, Darmstadt, Germany). KOD hot start DNA polymerase, dNTPs and buffer were purchased from Novagen (San Diego, USA). DPN I was obtained from New England Biolabs (NEB, Frankfurt, Germany). GeneRuler 1 kb DNA ladder was acquired from Fermentas (St. Leon-Rot, Germany). Primers were obtained from Invitrogen.

An improved PCR method for the creation of saturation mutagenesis libraries, based on the use of a common reverse non-mutagenic primer, was used to generate all the mutants.[60] LW202 pQEEH plasmid was used as a template (10 ng L^{-1}, 1 μL). For every mutant, desired mutations were both inserted using a forward primer (5′-GGTTCATTTGAAC***XXX***TGC***YYY***ATGAGTGCTC-3′ in which XXX is the mutation at position 215 and YYY the mutation at 217; 2.5 μM, 1 μL), and a non-mutagenic primer (SP2, 3′-CTCGCTCTGCTAATCCTGTTACCAGTGG-5′, 2.5 μM, 1 μL) was used as the reverse one for all cases. Amplification was achieved by using KOD polymerase (1 μL). The temperature cycles were: 1×(94 °C, 3 min), 5×(94 °C, 30 s; 53 °C, 1 min; 72 °C, 5 min), 20×(94 °C, 30 s; 72 °C, 7 min), and 1×(72 °C, 16 min). PCR mixtures were digested twice with DPN I (1 μL, 37 °C, 1 h) and transformed in homemade chemical competent cells (DH5α, 50 μL). After the heat shock (42 °C, 60 s) the cells were grown in SOC (0.2 mL, 37 °C, 1 h, 1400 rpm (mixing stroke 3 mm)). The suspension (15 μL) was streaked out in LB-agar supplemented with carbenicillin (100 μg mL^{-1}) and incubated overnight (37 °C). Plasmid purification was performed by MEDIGENOMIX (Martinsried, Gemany). QER standard primer was used by Medigenomix to read the plasmidic DNA region of interest.

All the mutants were drawn from the −80 °C glycerol stock and placed into three 96-deep-well plates containing LB media (0.9 mL), supplemented with carbenicillin (100 μg mL^{-1}). After 12 h, the latter precultures (2 mL, 0.66 mL from each plate) were inoculated in fresh LB (18 mL) containing carbenicillin and incubated until OD_{600} was around 3. Vials (8 mL) were prepared with the second generation bacteria culture (0.818 mL). After the addition of sodium phosphate buffer (20 mM, pH 7.2, 5.5 mL) and a solution of *rac*-**1** in acetonitrile (12.5 mg mL^{-1}, 0.5 mL) and incubation (200 rpm (orbital motion: 25 mm, linear motion 12.5 mm), 30 °C), samples were withdrawn after a definite time (60 min for the first experiment; optimized time for the second one, see Table SI4 in the Supporting Information) and centrifuged (2300 rcf), 20 min). The achieved enantioselectivity in the reaction was analyzed by HPLC in a Chiralcel OD-R HPLC chiral column from Daicel (Essex, UK) eluting with methanol/water (7:3). The retention times were: (*R*)-**2**, $t_R=8.6$ min; (*S*)-**2**, $t_R=9.8$ min; (*R*)-**1**, $t_R=20.9$ min; (*S*)-**1**, $t_R=24.4$ min. *E*-values were calculated as recommended in Faber's work[61] according to the following equation:

$$E=\frac{\ln[(1-ee_s)/(1+ee_s/ee_p)]}{\ln[(1+ee_s)/(1+ee_s/ee_p)]}=\frac{(k_{cat}/k_M)_{fast}}{(k_{cat}/k_M)_{slow}}$$

Most of the experiments were repeated at least three times (see the Supporting Information for the raw data).

Computational analysis procedures: The enzyme library in this study was constructed with two substitution positions ($N=2$) with each amino acid on each position assigned a random distinct integer X_i ($i=1,2$) between 1 and 20. Consequently, each protein in the library is uniquely defined by a length-two vector $\{X_1, X_2\}$ and *E*-value. Following Step 1 as the synthesis and property measurement of the random mutants, Step 2 of ASRA serves to determine the optimal integer assignment $\{X_1^*, X_2^*\}$ for each amino acid on each substitution position that produces the most regular property landscape. To achieve this goal, a quantitative measure *Q* for property landscape regularity needs to be defined, and appropriate optimization algorithms are needed to find the optimal substituent ordering(s) that minimize *Q* (small *Q* values correspond to more regular landscapes by definition). Several different *Q* measures and global optimization algorithms were previously implemented for small molecule libraries.[49–51] In this study, all of them produced satisfactory results. However, going beyond this proof-of-principle system, the computational cost for these algorithms can increase rapidly when the library dimensionality *N* is high (the total number of substituent orderings is $\prod_{i=1,N} S_i!$ in which $S_i=20$ for proteins), which can pose a much more serious problem for protein libraries than for the usually lower-dimensional small molecule libraries. Here we describe a new algorithm which is simple to operate, computationally inexpensive, and easily scalable to higher *N* values.

For the first substitution position of this protein library, the new algorithm computes a score Q_m for the *m*-th amino acid ($m=1,2,\ldots,20$) by the following equation:

$$Q_m=\sum_{m'\neq m}\sum_{n=1}^{20}(a_{mn}-a_{m'n})\frac{1}{1+\omega\sigma_{mn}}\frac{1}{1+\omega\sigma_{m'n}}$$

in which a_{mn} and $a_{m'n}$ are the observed property values (*E*-values in this case) of two variants with position 2 indexed by integer *n* and position 1 indexed by *m* and *m*′, respectively. σ_{mn} and $\sigma_{m'n}$ are the relative standard deviation of a_{mn} and $a_{m'n}$, respectively, and *w* is a weight factor. This simple expression of Q_m provides a global measure of the influence of amino acid *m* on the property value relative to the other amino acids on position 1, averaged over all the mutations on position 2. When Q_m is calculated for all 20 amino acids on position 1, the optimal amino acid ordering on this position is determined by a simple and computationally efficient sorting of the twenty Q_m values. The optimal ordering on position 2 (and additional positions when $N>2$) is calculated in the same way. The computational cost scales linearly with respect to *N*. With this algorithm, even though Q_m is calculated individually for each position, the cooperative/epistatic interactions among different positions are still accounted for (they are averaged over all the other positions because all *N* positions are sampled simultaneously for each mutant). Assumptions of linearity or additivity are not needed for the underlying landscape. When

necessary, the cooperative interactions can also be explicitly expressed with recently developed high dimensional model representation (HDMR) methods.[62–64]

After optimal reordering of the property landscape, Step 3 of ASRA makes prediction of the previously unsampled amino acid mutants that are most likely to give desired property values. This step completes one round of ASRA operation, which then returns to Step 1, when necessary, for further protein synthesis and property measurement. As shown in Scheme 1, a visual inspection of the reordered landscape will usually suffice in Step 3 for identifying the location of new mutations in cases for which two or perhaps three sites are involved. Since Step 3 serves to identify regions with potentially better mutants, this operation is a pattern recognition or interpolation problem that is amenable as well to automation with suitable software. The efficacy of even visual inspection is supported by the results of this proof-of-principle study. We have also developed appropriate interpolation techniques for reliable property prediction when N is large.[49,51] Regardless of this practical difference, successful implementation of Step 3 for both strategies depends critically on the regularity of the reordered landscape, which is the core of ASRA and also distinguishes it from traditional QSAR techniques.

Results and Discussion

Experimental platform: As delineated above, we aimed to test the viability of using the reordering algorithm ASRA to discover ANEH enzymes with optimal enantioselectivity in the selective ring opening of a racemic mixture of glycidyl phenyl ether (**1**). The goal was to see if a limited set of experiments (enzyme syntheses in the form of generated mutants and property measurements as the respective E-values) would suffice to reliably explore a defined portion of the protein sequence space. Accordingly, we studied the ability of ASRA to predict the enantioselectivity of ANEH-mutants as catalysts in the hydrolytic kinetic resolution of the model epoxide *rac*-**1** (Scheme 2). Within the context of this study, reliable prediction of enantioselectivity means the identification of those hits/regions with high E-values, instead of providing explicit E-value predictions.

The best previously evolved mutant LW202 ($E=115$)[58] was selected to be the scaffold for ASRA. Rather than choosing all the sites utilized in the original study (each comprising more than one amino acid position), we decided to simplify the experimental platform by considering only $N=2$ substitution positions, namely Phe215 (position 1) and Asn217 (position 2). All 20 natural amino acids were selected to be mutation targets on each position ($S_1=S_2=20$), with a total of $20^2=400$ possible mutants. A main reason for selecting these two positions is that, from our previous studies,[23,56,65] mutations on these two positions can produce proteins with both very high and very low E-values, providing a desirable system to test the predictive capability of ASRA in large dynamic ranges. The two positions were also selected to be close to each other, to better evaluate ASRA in the presence of potentially strong (synergistic or antagonistic) epistatic interactions.

Figure 1 shows the dimeric structure of mutant LW202 derived from X-ray structural analysis.[65] It is well known that the substrate is bound and activated by hydrogen bonding arising from Tyr251 and Tyr314 with the epoxide oxygen.[66]

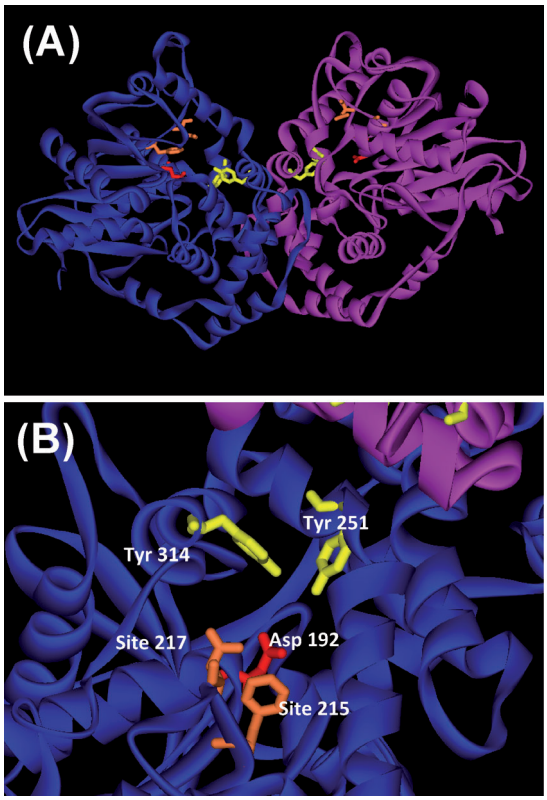


Figure 1. A) Crystal structure of the dimer of LW202.[65] B) Active site of ANEH featuring the catalytically active residues Asp192, Tyr251, and Tyr314 in addition to the two substitution positions 1 (residue 215) and 2 (residue 217).

This is followed by nucleophilic attack by Asp192 in the rate determining step resulting in a covalently bound ester which is rapidly hydrolyzed.[66]

As described earlier, ASRA started with an initial random sampling of the focused library space. A total of 95 random variants with mutations on both positions were prepared by site directed mutagenesis for this purpose (see the Supporting Information, Table SI1). The WT enzyme was already obtained in a previous work.[56,65] With this information ASRA was charged to reveal the general structure of the E-value landscape with respect to the two positions. After performing the standard experimental procedure (60 min of reaction time for all mutants), the E-values for all the reactions were calculated as recommended in Faber's work.[61]

The best application condition requires a conversion higher than 20%, which could not be met for all the mutants during the original screening process.[56,65] Consequently, we performed another set of experiments with the same 95 mutants for which the reaction time was optimized to assure at least 20% conversion (see the Supporting Information, Table SI4). Without synthesizing more mutants, the second set of experiments enabled 1) an evaluation of the influence of prolonged reaction time on E-value and 2) a comparative study of ASRA under the two conditions. Since

the 60-minute reaction condition had been the original screening condition in our ISM experiments, this comparative study allowed for an estimate of the efficacy of the standard screening procedure in discovering the desired mutants.

ASRA reordering with 60-minute reaction data: Using the 60-minute data (see the Supporting Information, Table SI2), Figure 2A shows the landscape of the mean *E*-values for the 95 mutants with a random amino acid ordering (Table 1A). The *E*-value distribution shows a low percentage of variants above a moderate threshold value of $E \geq 40$ (Figure 2D). The landscape resembles the so-called "golf course problem", being flat with several local "good" regions. This suggests that protein discovery in this library can be very difficult compared with problems with smoother distributions.

Table 1. The amino acid orderings on positions 1 and 2 of the ANEH mutants. The integers X_i ($i=1,2$) are the same as the indices along positions 1 and 2, respectively, of the heat maps in the figures. Case A: a random ordering on both positions. Case B: 60-minute reaction time, error weight $w=0$ in cost function Q. Case C: 60-minute reaction time, $w=1.0$. Case D: 60-minute reaction time, $w=0$, the erroneous mutant at position [16,20] in Figure 2B removed from data. Case E: 60-minute reaction time, $w=0$, mutants with $E \geq 50$ excluded from reordering. Case F: optimized reaction time, w=0, the erroneous mutant removed from data. Case G: optimized reaction time, $w=0$, mutants with $E \geq 50$ excluded from reordering.

Case	X_i ($i=1,2$)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	Position 1	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	Position 2	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
B	Position 1	K	R	W	D	E	P	G	V	A	H	I	C	Q	L	T	Y	N	M	S	F
	Position 2	Q	R	E	S	K	A	W	P	H	Y	T	G	I	C	F	M	V	D	N	L
C	Position 1	R	K	W	D	E	P	G	V	A	H	I	C	Q	L	T	M	N	Y	S	F
	Position 2	Q	R	E	W	S	K	A	H	P	Y	T	G	I	C	F	M	V	D	N	L
D	Position 1	K	R	D	E	P	W	V	G	Y	A	H	I	C	Q	L	T	N	M	S	F
	Position 2	Q	K	W	R	P	E	H	S	Y	T	A	G	I	C	F	M	V	D	L	N
E	Position 1	K	P	R	D	E	W	V	Y	A	G	H	I	Q	T	C	L	N	F	S	M
	Position 2	K	Q	R	P	E	S	Y	T	A	G	D	I	H	C	W	F	N	M	V	L
F	Position 1	K	R	P	E	D	W	V	Y	A	C	I	Q	T	L	H	M	G	N	S	F
	Position 2	Q	Y	K	T	E	W	S	R	C	G	P	A	H	I	F	D	V	L	M	N
G	Position 1	K	P	R	E	Y	A	W	V	D	H	I	Q	T	C	F	L	G	M	N	S
	Position 2	K	Q	E	Y	P	T	G	R	S	C	A	D	I	N	F	M	W	H	V	L

Figure 2B shows the regular *E*-value landscape after ASRA reordering (see the Materials and Methods section for details) using the data for all 95 peptides. Despite the unfavorable *E*-value distribution, a reasonably smooth landscape exists with *E*-values gradually decreasing from the lower-right corner to other areas of the landscape. The reordered *E*-value landscape is highly similar when data noise is incorporated in the cost function Q (w=1.0, data not shown). Table 1B and 1C contain the optimal amino acid orderings for both cases.

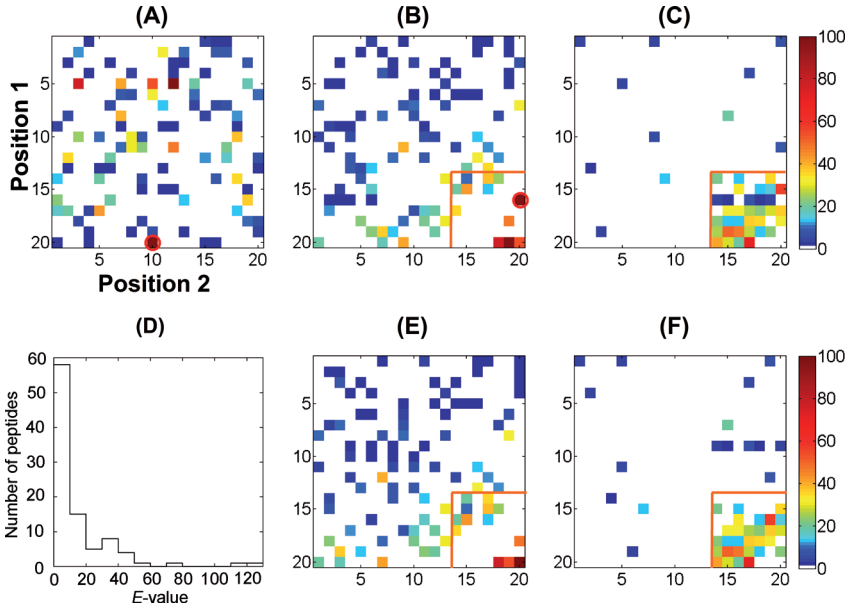


Figure 2. Optimal reordering of the *E*-value landscapes with 60 min reaction time. A) Color heat map for the *E*-value landscape of 95 randomly sampled mutants plotted with a random amino acid ordering (see Table 1A for the integer assignment for each amino acid). Each color square represents one mutant with red indicating a high *E*-value and blue corresponding to a low *E*-value (see color bar on the far right). White squares are unsampled proteins. B) *E*-value landscape of the 95 mutants using the ASRA-identified optimal amino acid ordering (Table 1B). The result predicts that proteins with high *E*-values are most likely located in the lower right corner. The mutant at position 16/20 (circled in red in both A and B) of the reordered landscape turned out to be the same as the mutant at position 20/19; the wrong protein was accidentally placed in this position in the experiment. C) *E*-value landscape for 45 newly sampled mutants, guided by the ordering in B. D) *E*-value distribution for the 95 initial random mutants. E) Reordered *E*-value landscape for the 94 mutants (excluding the erroneous mutant at position 16/20 in B). F) *E*-value landscape for the 45 newly sampled mutants, based on the ordering in E.

Evaluation of ASRA prediction: Based on the reordered *E*-value landscape above, a second set of 45 new mutants were generated to evaluate the reliability of the ASRA predictions (see the Supporting Information, Table SI3). Of the chosen variants, 34 are in the 7×7 box at the most interesting lower-right corner in the reordered landscape (Figure 2B). The other 11 mutants reside randomly in other areas of the landscape for background comparison. When these variants are placed in the *E*-value landscape using the optimal ordering obtained from the 95 random mutants, all those with

$E \geq 40$ are located inside the box (Figure 2C), clearly demonstrating the reliability of ASRA prediction. Note that we could have generated all 400 mutants to test the absolute predictive ability of ASRA. However, the goal of ASRA is to find good regions in a protein library space from a minimal sampling effort. The 45 new mutants synthesized in this work provide a sound evaluation of ASRA's prediction of the good versus the bad regions. In addition, the selection of the new mutants here shows how the predictive Step 3 of ASRA may be performed in practice. Further iterations of ASRA could always be performed on the identified good mutants to find the absolute best member, but that feature is beyond the proof of principle tests in this paper.

Despite the generally high predictive quality, a closer examination of Figure 2B and 2C shows that all but one mutant (circled in red in Figure 2B) in row 16 of position 1 in the reordered landscape have very low *E*-values. Since row 16 overlaps with the selected 7×7 box, this result seems to indicate that the ASRA predictions were unsatisfactory at least for this position. However, a re-sequencing of this "outlier" mutant revealed an experimental error—this mutant was in fact the same as the one at position 20/19 in Figure 2B; the wrong protein was placed at this position. After removing the wrong mutant from the data, the reordered landscape remains smooth (Figure 2E, the ordering is in Table 1D), and all new mutants inside the 7×7 box have relatively high *E*-values (Figure 2F). This experimental accident further illustrates the remarkable capability of ASRA—it can even reveal laboratory errors from irregularities of the reordered landscape alone.

In the above tests, the 94 random samples include several mutants with very high *E*-values. To examine whether ASRA can make reliable predictions without these good seeds, we eliminated all three mutants with $E \geq 50$ from the initial set of 94 and used the rest of the data for ASRA reordering. Despite the differences in the optimal ordering (Table 1D vs. E), the reordered *E*-value landscape remains smooth (Figure 3A), and 5 out of 6 mutants (in the total of 139 variants) with $E \geq 50$ are discovered inside the 7×7 box (Figure 3B) when the rest of the proteins are placed by using the optimal ordering. This observation strongly demonstrates the excellent predictive power of ASRA even in the absence of good seeds, which will be an important property when ASRA is used in directed evolution, for which favorable seeds and/or hits are most likely rare in the initial iterations.

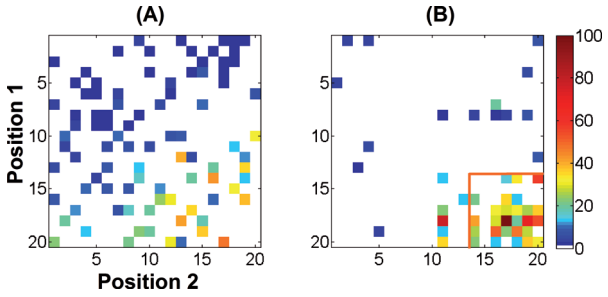


Figure 3. ASRA reordering in the absence of good seeds with the 60-minute reactions. A) ASRA-reordered landscape when all mutants with $E \geq 50$ are excluded from the 94 random samples. B) Location of the rest of the mutants using the amino acid ordering in A.

To fully evaluate the efficiency and robustness of ASRA, we randomly selected 30, 40, 50, 60, 70, 80, and 90 mutants, respectively, from the first set of 94 (the erroneous mutant was excluded) to determine the best amino acid orderings. Using these orderings, we then calculated, among the available $94+45=139$ mutants, the percentage of "desired" members (defined as mutants with *E*-values higher than a specified threshold) that are located inside the 7×7 box. Figure 4 shows that convergence is reached at around 80

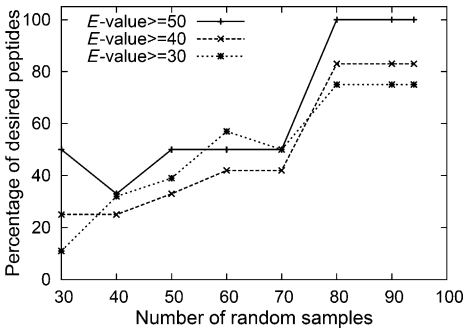


Figure 4. Among all 139 proteins, percentage of desired mutants (above the *E*-value threshold value of 30, 40, or 50) inside the 7×7 box vs. the number of random samples used for reordering.

random mutant samples (i.e., 20% of the library space). When the *E*-value threshold is 40, 10 out of 12 desired mutants are inside the box, corresponding to an approximately 10-fold gain in sampling the box compared with sampling the rest of the library space for finding the hits:

$$\frac{[\text{mutants above threshold in box}]/[\text{all mutants in box}]}{[\text{mutants above threshold outside box}]/[\text{all mutants outside box}]} = \frac{10/45}{2/94} \approx 10$$

This gain is approximately 5-fold when the *E*-value threshold is set to 30 (20 out of 28 desired mutants are inside the box). The lower gain is expected because the percentage of desired mutants increases with the reduced *E*-value threshold, indicating that a larger box should be selected. By contrast, if the *E*-value threshold is set as 50, then all 6 desired mutants are inside the box, corresponding to an infinitely large gain. The selection of the box is a simple means of illustrating the predictive capability of ASRA without knowing/assuming any structure–property relationships of the proteins. More quantitative predictions of precise *E*-values can be performed by applying appropriate interpolation over the reordered landscape and will be a topic of future research.

ASRA reordering with optimized reaction time: *E*-values for the same 139 mutants were also measured using individ-

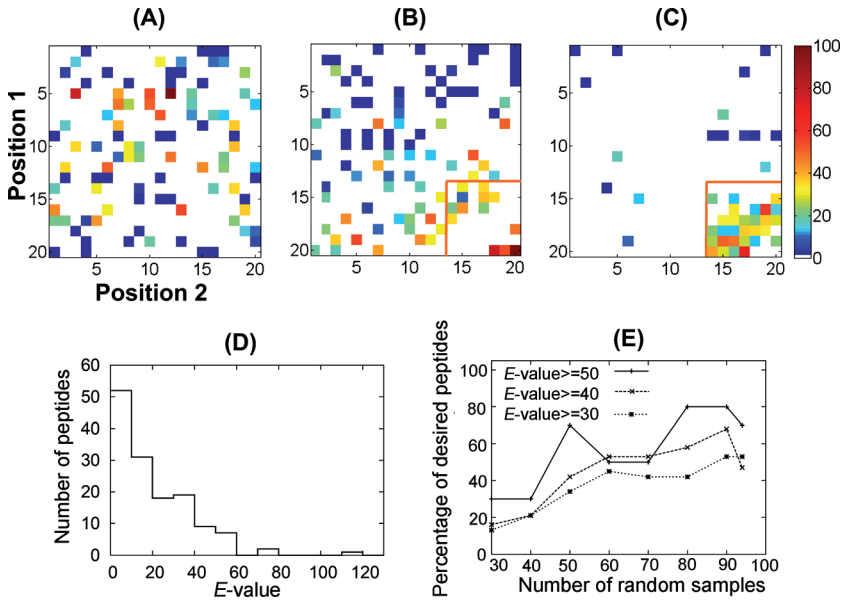


Figure 5. The influence of optimized reaction time on *E*-value and the ASRA predictions. A) Color heat map for the *E*-value landscape (with optimized reaction time) of 94 randomly sampled mutants plotted with a random amino acid ordering (Table 1A). B) *E*-value landscape of the 94 mutants using the optimal amino acid ordering identified from the 60-minute reactions (Table 1D). C) Locations for 45 newly sampled mutants using the same ordering; all mutants with $E \geq 40$ are in the 7×7 box in the lower right corner. D) *E*-value distribution for the 94 initial random mutants. E) Among all 139 proteins, percentage of desired mutants (above the *E*-value threshold value of 30, 40, or 50) inside the 7×7 box vs. the number of random samples used for reordering.

ually optimized reaction times to obtain conversion rates higher than 20% (see the Supporting Information, Tables SI4 and SI5). The *E*-value distribution shows a larger number of mutants with high *E*-values compared with the 60-minute reactions (Figure 2D vs. Figure 5D), reflecting the influence of prolonged reaction times for some proteins. This change also results in a slightly different optimal ordering (Table 1F) and a reordered landscape that decreases more "smoothly" from the lower right corner to the rest of the library space (data not shown). However, using the optimal ordering from the 60-minute reactions (Table 1D and Figure 5B), data from the newly synthesized 45 mutants (see the Supporting Information, Table SI6) again show that all variants with $E \geq 40$ are still located in the 7×7 box (Figure 5C), suggesting that ASRA predictions based on the 60-minute reactions may still be extended to the optimized reaction conditions. This behavior can be beneficial for high-throughput procedures where the reaction conditions, including reaction time, often cannot be easily adjusted for individual reactions. Quantitatively, the gain in efficiency (by sampling inside the 7×7 box) is slightly lower than the 60-minute reactions (Figure 4 vs. Figure 5E), due to the smoother landscape and the increased number of high *E*-value mutants. This does not mean that the former results are better or worse; it only indicates that desired mutants are scattered over a larger area in the whole landscape (and a larger box is needed).

Similar to the 60-minute reactions, we again excluded all seven mutants with $E \geq 50$ from the initial 94-mutant set and applied ASRA to the rest of the proteins. The resultant optimal landscape is similarly smooth (Figure 6A). Placing the rest of the mutants using this ordering (Table 1G) locates 6 out of 10 desired mutants ($E \geq 50$) in the 7×7 box (Figure 6B). Again, the reordered landscape in Figure 6A contains more mutants with high *E*-values than the landscape in Figure 2E, suggesting that the desired mutants should be located in a larger region (e.g., the triangular area in the lower right part of the landscape in Figure 6).

Conclusion

The present study constitutes the first application of ASRA to the directed evolution of an enzyme. Using the experimental platform of hydrolytic kinetic resolution of a chiral epoxide catalyzed by mutants of the epoxide hydrolase from *Aspergillus niger* (ANEH), we have shown that the ASRA technique constitutes a viable procedure for property estimation and guiding directed evolution of enzymes in focused libraries. Based on the reordered *E*-value landscape generated from a small set of random mutants, ASRA gave reliable estimates of the desired mutants with improved enantioselectivity in the model enzyme-catalyzed reaction. ASRA was even able to identify an error in the laboratory data from irregularities in the reordered landscape alone.

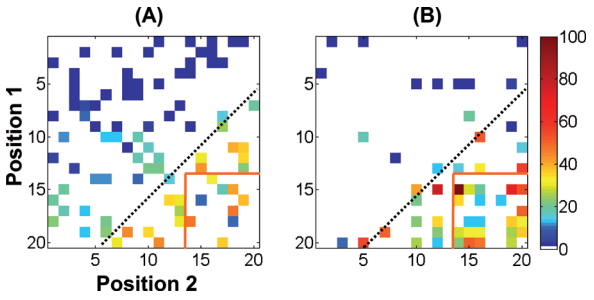


Figure 6. ASRA reordering in the absence of good seeds with optimized reaction times. A) ASRA-reordered landscape when all mutants with $E \geq 50$ are excluded from the 94 random samples. B) Location of the rest of the available mutants using the amino acid ordering in A. The triangular area below the dotted line indicates a better prediction (than the original 7×7 box) of desired mutants that reflects the smooth geometry of the reordered landscape.

In contrast to other computational guides used in directed evolution,[1–8,27–40] such as SCHEMA,[41,42] FamClash,[43–48] or ProSAR,[46–48] the application of ASRA does not require assumptions of linearity, additivity, or any functional form of structure–property relationships. The only requirement is global regularity of the underlying property landscape.[52,53] In addition, ASRA does not require the use of molecular descriptors. Thus, knowledge of enzyme structure is not needed; ASRA can be applied as long as the location of each amino acid on each substitution position can be consistently indexed and followed. With the algorithmic development described in this article, the computational cost of ASRA is very low; the main cost lies in performing the experiments. ASRA is also compatible with Pareto optimization techniques for simultaneously optimizing multiple properties of the same protein library.[67] All these attributes make ASRA a generally applicable and operationally attractive method for efficient protein engineering in focused libraries.

Being a property prediction and optimization tool, ASRA does not directly provide structure–property relationships, but such information is contained in the features of the reordered property landscape and the corresponding optimal amino acid orderings. For example, we observed in this study that structurally and electronically very different amino acids can exert positive effects (higher enantioselectivity). In the absence of a detailed mechanistic study,[65] it is difficult to interpret the role of the point mutations in enhancing enantioselectivity solely. However, the relative position of the amino acids on the reordered landscape may provide valuable insight (as a "free byproduct" of ASRA) when detailed QSAR studies are performed.

An important issue regarding the applicability of ASRA is the scaling of the number of samples needed to make reliable predictions with respect to the number of substitution positions. In this proof-of-principle study, mutations were restricted to two positions on the protein scaffold. However, our research in random sampling high dimensional model representation[62–64] (a related property prediction and optimization method also based on random sampling of the variable space) shows that an increasingly lower percentage of the protein library space is expected to provide convergence for ASRA when the number of substitution positions increases. Moreover, as shown in a recent application,[50] the iterative operation of ASRA can further decrease the experimental cost.

Acknowledgements

MTR thanks the Deutsche Forschungsgemeinschaft (Schwerpunkt 1170) and the Fonds der Chemischen Industrie for generous support. HR acknowledges NSF and EPA environmental bioinformatics and computational toxicology center (ebCTC).

[1] K. M. Arndt, K. M. Muller, *Protein Engineering Protocols (Methods in Molecular Biology), Vol. 352c*, Humana Press, Totowa, New Jersey, **2007**.
[2] S. Bershtein, D. S. Tawfik, *Curr. Opin. Chem. Biol.* **2008**, *12*, 151.
[3] E. G. Hibbert, F. Baganz, H. C. Hailes, J. M. Ward, G. J. Lye, J. M. Woodley, P. A. Dalby, *Biomol. Eng.* **2005**, *22*, 11.
[4] C. Jäckel, P. Kast, D. Hilvert, *Annu.Rev. Biophys.* **2008**, *37*, 153.
[5] J. Kaur, R. Sharma, *Crit. Rev. Biotechnol.* **2006**, *26*, 165.
[6] S. Lutz, U. T. Bornscheuer, *Protein Engineering Handbook, Vol. 1 and 2*, Wiley-VCH, Weinheim, **2009**.
[7] M. T. Reetz, *J. Org. Chem.* **2009**, *74*, 5767.
[8] S. B. Rubin-Pitel, H. Zhao, *Comb. Chem. High Throughput Screening* **2006**, *9*, 247.
[9] A. S. Bommarius, J. Broering, *Biocatal. Biotransform.* **2005**, *23*, 125.
[10] V. G. H. Eijsink, S. Gaseidnes, T. V. Borchert, B. v. d. Burg, *Biomol. Eng.* **2005**, *22*, 21.
[11] C. O'Fagain, *Enzyme Microb. Technol.* **2003**, *33*, 137.
[12] P. L. Wintrode, F. H. Arnold, *Adv. Protein Chem.* **2001**, *55*, 161.
[13] F. H. Arnold, *Acc. Chem. Res.* **1998**, *31*, 125.
[14] K. Chen, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5618.
[15] M. T. Reetz, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5716.
[16] M. T. Reetz, in *Asymmetric Organic Synthesis with Enzymes* (Eds.: V. Gotor, I. Alfonso, E. Garcia-Urdiales), Wiley- VCH, Weinheim, **2008**, pp. 21.
[17] M. T. Reetz, *Angew. Chem.* **2011**, *123*, 144; *Angew. Chem. Int. Ed.* **2011**, *50*, 138.
[18] J. L. Reymond, *Enzyme Assays: High-Throughput Sreening, Genetic Selection and Fingerprinting, Vol. XVIII*, Wiley-VCH, Weinheim, **2006**.
[19] J. D. Bloom, M. M. Meyer, P. Meinhold, C. R. Otey, D. MacMillan, F. H. Arnold, *Curr. Opin. Struct. Biol.* **2005**, *15*, 447.
[20] R. J. Fox, G. W. Huisman, *Trends Biotechnol.* **2008**, *26*, 132.
[21] A. Herman, D. S. Tawfik, *Protein Eng. Des. Sel.* **2007**, *20*, 219.
[22] S. Lutz, W. M. Patrick, *Curr. Opin. Biotechnol.* **2004**, *15*, 291.
[23] M. T. Reetz, D. Kahakeaw, R. Lohmer, *ChemBioChem* **2008**, *9*, 1797.
[24] T. S. Wong, D. Roccatano, M. Zacharias, U. J. Schwaneberg, *J. Mol. Biol.* **2006**, *355*, 858.
[25] M. T. Reetz, S. Wu, *Chem. Commun.* **2008**, 5499.
[26] H. Jochens, U. T. Bornscheuer, *ChemBioChem* **2010**, *11*, 1861.
[27] M. Pasupuleti, B. Walse, B. Svensson, M. Malmsten, A. Schmidtchen, *Biochemistry* **2008**, *47*, 9057.
[28] M. Höhne, S. Schaetzle, H. Jochens, K. Robins, U. T. Bornscheuer, *Nat. Chem. Biol.* **2010**, *6*, 807.
[29] K. A. Armstrong, B. Tidor, *Biotechnol. Prog.* **2008**, *24*, 62.
[30] R. A. Chica, N. Doucet, J. N. Pelletier, *Curr. Opin. Biotechnol.* **2005**, *16*, 378.
[31] P. Francois, V. Hakim, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 580.
[32] S. A. Funke, N. Otte, T. Eggert, M. Bocola, K.-E. Jaeger, W. Thiel, *Protein Eng. Des. Sel.* **2005**, *18*, 509.
[33] U. Krauss, T. Eggert, *Biotechniques* **2005**, *39*, 679.
[34] B.-C. Lee, K. Park, D. Kim, *Proteins Struct. Funct. Bioinf.* **2008**, *72*, 863.
[35] J. Liao, M. K. Warmuth, S. Govindarajan, J. E. Ness, R. P. Wang, C. Gustafsson, J. Minshull, *BMC Biotechnol.* **2007**, *7*, 16.
[36] D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, *Nature* **2008**, *453*, 190.
[37] M. J. Volles, P. T. Lansbury, Jr., *Nucleic Acids Res.* **2005**, *33*, 3667.
[38] D. C. Wedge, W. Rowe, D. B. Kell, J. Knowles, *J. Theor. Biol.* **2009**, *257*, 131.
[39] M. C. Saraf, G. L. Moore, C. Maranas, *Protein Eng.* **2003**, *16*, 397.
[40] G. L. Moore, C. D. Maranas, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 5091.
[41] K. Hiraga, F. H. Arnold, *J. Mol. Biol.* **2003**, *330*, 287.
[42] M. M. Meyer, J. J. Silberg, C. A. Voigt, J. B. Endelman, S. L. Mayo, Z.-G. Wang, F. H. Arnold, *Protein Sci.* **2003**, *12*, 1686.

[43] M. C. Saraf, A. R. Horswill, J. S. Benkovic, C. D. Maranas, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4142.
[44] M. C. Saraf, G. L. Moore, N. M. Goodey, v. Y. Cao, S. J. Benkovic, C. D. Maranas, *Biophys. J.* **2006**, *90*, 4167.
[45] A. Pavelka, E. Chovancova, J. Damborsky, *Nucleic Acids Res.* **2009**, *37*, W376.
[46] R. Fox, A. Roy, S. Govindarajan, J. Minshull, C. Gustafsson, J. T. Jones, R. Emig, *Protein Eng.* **2003**, *16*, 589.
[47] R. J. Fox, *J. Theor. Biol.* **2005**, *234*, 187.
[48] R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon, G. W. Huisman, *Nat. Biotechnol.* **2007**, *25*, 338.
[49] F. Liang, X. Feng, M. Lowry, H. Rabitz, *J. Phys. Chem. B* **2005**, *109*, 5842.
[50] S. R. McAllister, X. Feng, P. A. DiMaggio, Jr., C. A. Floudas, J. D. Rabinowitz, H. Rabitz, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5967.
[51] N. Shenvi, J. Geremia, H. Rabitz, *J. Phys. Chem. A* **2003**, *107*, 2066.
[52] K. W. Moore, A. Pechen, X.-J. Feng, J. Dominy, V. Beltrani, H. Rabitz, *Chem. Sci.* **2011**, *2*, 417.
[53] K. W. Moore, A. Pechen, X.-J. Feng, J. Dominy, V. J. Beltrani, H. Rabitz, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048.
[54] M. T. Reetz, D. Kahakeaw, J. Sanchis, *J. Mol. Biosyst.* **2009**, *5*, 115.
[55] M. T. Reetz, J. D. Carballeira, *Nat. Protoc.* **2007**, *2*, 891.
[56] M. T. Reetz, J. D. Carballeira, A. Vogel, *Angew. Chem.* **2006**, *118*, 7909; *Angew. Chem. Int. Ed.* **2006**, *45*, 7745.
[57] M. T. Reetz, M. Boccola, J. D. Carballeira, D. Zha, A. Vogel, *Angew. Chem.* **2005**, *117*, 4264; *Angew. Chem. Int. Ed.* **2005**, *44*, 4192.
[58] M. T. Reetz, L.-W. Wang, M. Boccola, *Angew. Chem.* **2006**, *118*, 1258; *Angew. Chem. Int. Ed.* **2006**, *45*, 1236.
[59] M. T. Reetz, C. Torre, A. Eipper, R. Lohmer, M. Hermes, B. Burnner, A. Maichele, M. boccola, M. Arand, A. Cronin, Y. Genzel, A. Archelas, R. Furstoss, *Org. Lett.* **2004**, *6*, 177.
[60] J. Sanchis, L. Fernandez, J. D. Carballeira, J. Drone, Y. Gumulya, H. Hobenreich, D. Kahakeaw, S. Kille, R. Lohmer, J.-P. Peyralans, J. Podtetenieff, S. Prasad, P. Soni, A. Taglieber, S. Wu, F. E. Zilly, M. T. Reetz, *Appl. Microbiol. Biotechnol.* **2008**, *81*, 387.
[61] K. Faber, *Biotransformations in Organic Chemistry*, 5th ed. Springer, Berlin, **2004**.
[62] G. Li, C. Rosenthal, H. Rabitz, *J. Phys. Chem. A* **2001**, *105*, 7765.
[63] H. Rabitz, O. Alis, *J. Math. Chem.* **1999**, *25*, 197.
[64] H. Rabitz, O. Alis, in *Sensitivity Analysis* (Eds.: A. Saltelli, K. Chan, M. Scott), John Wiley & Sons, **2000**, pp. 199.
[65] M. T. Reetz, M. Bocola, L. W. Wang, J. Sanchis, A. Cronin, M. Arand, J. Zou, A. Archelas, A. L. Bottalla, A. Naworyta, S. L. Mowbray, *J. Am. Chem. Soc.* **2009**, *131*, 7334.
[66] J. Zou, B. M. Hallberg, T. Bergfors, F. Oesch, M. Arand, S. L. Mowbray, T. A. Jones, *Structure* **2000**, *8*, 111.
[67] Y. Collette, P. Siarry, *Multiobjective Optimization: Principles and Case Studies*, Springer, **2004**.

Received: December 5, 2011
Published online: March 20, 2012

Please note: Minor changes have been made to this manuscript since its publication in *Chemistry—A European Journal* Early View. The Editor.

