Applied and Environmental Microbiology, Oct. 2010, p. 6397–6403 Vol. 76, No. 19
0099-2240/10/$12.00 doi:10.1128/AEM.00878-10

Improving Biocatalyst Performance by Integrating Statistical Methods into Protein Engineering[▽]

Moran Brouk,[1] Yuval Nov,[2]* and Ayelet Fishman[1]*

Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel,[1] and Department of Statistics, University of Haifa, Haifa 31905, Israel[2]

Received 11 April 2010/Accepted 4 August 2010

Directed evolution and rational design were used to generate active variants of toluene-4-monooxygenase (T4MO) on 2-phenylethanol (PEA), with the aim of producing hydroxytyrosol, a potent antioxidant. Due to the complexity of the enzymatic system—four proteins encoded by six genes—mutagenesis is labor-intensive and time-consuming. Therefore, the statistical model of Nov and Wein (J. Comput. Biol. 12:247-282) was used to reduce the number of variants produced and evaluated in a lab. From an initial data set of 24 variants, with mutations at nine positions, seven double or triple mutants were identified through statistical analysis. The average activity of these mutants was 4.6-fold higher than the average activity of the initial data set. In an attempt to further improve the enzyme activity to obtain PEA hydroxylation, a second round of statistical analysis was performed. Nine variants were considered, with 3, 4, and 5 point mutations. The average activity of the variants obtained in the second statistical round was 1.6-fold higher than in the first round and 7.3-fold higher than that of the initial data set. The best variant discovered, TmoA I100A E214G D285Q, exhibited an initial oxidation rate of 4.4 ± 0.3 nmol/min/mg protein, which is 190-fold higher than the rate obtained by the wild type. This rate was also 2.6-fold higher than the activity of the wild type on the natural substrate toluene. By considering only 16 preselected mutants (out of ~13,000 possible combinations), a highly active variant was discovered with minimum time and effort.

Enzymes are remarkable biocatalysts that perform numerous chemical reactions. They have evolved in nature to do their task in an efficient and specific way, mostly under aqueous physiological conditions (12). However, the term "biocatalysis" refers to the use of enzymes as process catalysts under artificial conditions, and a major challenge today is to render biocatalysts suitable for the tough reaction conditions of an industrial process (11).

A widely used approach for improving enzyme function is directed evolution, whereby protein sequences are repeatedly selected, mutated, or recombined in a process that mimics natural evolution to produce better and better "generations" of protein variants. Directed evolution has been successfully used in numerous studies, but since it requires generation, purification, and screening of large numbers of variants, it is typically expensive and labor-intensive (28, 35). An alternative to directed evolution is an approach termed rational design, whereby predictions are made as to how mutations in a protein will affect its structure and hence its interaction with the target molecule. Unfortunately, both the sequence-structure and the structure-activity relationships are extremely intricate, and while this approach proved to be fruitful in some cases, its practical use is still limited (15). The rational-design approach also requires knowledge of the three-dimensional structure of the protein, which, unlike with the protein's sequence, is costly and time-consuming to decipher. It has been suggested lately that a combination of both methods may be the best tactic to obtain enzymes with desired activities and selectivities (15, 23).

Yet a third approach for protein improvement, which is less often used, is based on statistical analysis. According to this approach, the activity of any protein variant is viewed as a random quantity, and statistical methods are used to predict from activity data which mutation combinations are likely to improve activity. The statistical approach does not require structural knowledge about the protein at hand and allows one to focus screening efforts on a few promising variants, thus reducing labor, time, and expenses. Earlier statistical models for the sequence-activity relationship include Kauffman's NK model (14) and the "rough Mt. Fuji" model of Aita and Husimi (1). More recently, Fox et al. combined a machine learning technique termed ProSAR with directed evolution and rational design to significantly increase the catalytic function of a halohydrin dehalogenase in the production of the cholesterol-lowering drug atorvastatin (Lipitor) (9, 10). Liao et al. (18) employed eight machine learning algorithms to improve 20-fold the ability of proteinase K to hydrolyze a tetrapeptide substrate.

The statistical model that lies at the center of this work is that of Nov and Wein (22). Unlike the statistical algorithms used in the work of Fox et al. (9, 10) and Liao et al. (18), which are generic methods from the machine learning literature, this model was devised specifically for the protein design problem to capture characteristics of the protein sequence-activity relationship. Barak et al. (3) employed a variation of this model in conjunction with directed evolution to greatly improve the oxidoreductase ChrR in reducing chromate and uranyl.

* Corresponding author. Mailing address for Ayelet Fishman: Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel. Phone: 972-4-829-5898. Fax: 972-4-829-3399. E-mail: afishman@tx.technion.ac.il. Mailing address for Yuval Nov: Department of Statistics, University of Haifa, Haifa 31905, Israel. Phone: 972-4-824-0203. Fax: 972-4-824-0204. E-mail: yuval@stat.haifa.ac.il.

▽ Published ahead of print on 13 August 2010.

TABLE 1. Primers used for site-directed mutagenesis of the *tmoA* gene in T4MO for generating the different mutants suggested by the statistical model

Primer	Nucleotide sequence[a]
T4MObefEcoRI Front	5′-CCATGATTACGCCAAGCGCG-3′
T4MOAB Rear	5′-TCCATGCTCTTCACTGTTGAC-3′
T4MO_E214G-Front	5′-CAGATGCCGCAG**G**GAGCAGGTGACTAC-3′
T4MO_E214G-Rear	5′-GCAAACGTGTAGTCACCTGCT**C**CTGCGGC-3′
T4MO_D285Q_Front	5′-GGATTACTACACGCCGTTGGAG**CAG**CGCAGCCAG-3′
T4MO_D285Q_Rear	5′-AACTCCTTGAATGACTGGCTGCG**CTG**CTCCAACGG-3′
T4MO_D285I_Front	5′-GGATTACTACACGCCGTTGGAG**ATC**CGCAGCCAG-3′
T4MO_D285I_Rear	5′-AACTCCTTGAATGACTGGCTGCG**GAT**CTCCAACGG-3′
T4MO_I100S_Front	5′-CACTTTGAAATCCCATTACGGCGCC**TCC**GCAGTTGG-3′
T4MO_I100S_Rear	5′-GCTGCATATTCACCAACTGC**GGA**GGCGCCGTAATGG-3′
T4MO_S395C_Front	5′-GTCTCTCCAGAAACCTTGCCC**TGC**GTGTGCAAC-3′
T4MO_S395C_Rear	5′-GCGGTATCTGGCTCATGTTGCACAC**GCA**GGGCAAGG-3′
T4MO_D48E_Front	5′-CGCTGGAAAAATGGGAAAGCTAT**GAA**GAGCCC-3′
T4MO_D48E_Rear	5′-CCGGATAGGATGTCTTATAGGGCTC**TTC**ATAGC-3′

[a] Positions subjected to mutagenesis are underlined, and bases modified to obtain the desired point mutation are indicated in bold.

In this work, we combine all three approaches—directed evolution, rational design, and statistical methods—to improve the capacity of toluene 4-monooxygenase (T4MO) to produce an important antioxidant, hydroxytyrosol (6). This phenol, which is naturally present in olives and olive oil, has the highest free radical scavenging capacity and has been shown to be beneficial in preventing various diseases, such as diabetes, atherosclerosis, and cancer (13, 19, 34). Developing a biotechnological process for the synthesis of this antioxidant is of interest to the food and cosmetics industries.

T4MO from *Pseudomonas mendocina* KR1 is a soluble four-component enzyme belonging to the toluene monooxygenase family. T4MO is composed of six genes, designated *tmoABCDEF*, which are essential for the efficient catalysis and high regiospecificity of the enzyme. Genes *tmoA*, *tmoB*, and *tmoE* encode the α, β, and γ subunits, respectively, that comprise the $(\alpha\beta\gamma)_2$ quaternary structure of the 212-kDa hydroxylase component. The α hydroxylase subunit contains the catalytically active diiron center (2, 8). The *tmoC*, *tmoD*, and *tmoF* genes encode the 12.5-kDa Rieske-type [2Fe-2S] ferredoxin, the 11.6-kDa effector protein, and the 36-kDa NADH oxidoreductase, respectively (2, 16, 21).

Previously, a 35-fold improvement in T4MO activity on 2-phenylethanol (PEA) for the synthesis of hydroxytyrosol was reported for the TmoA I100A mutant (5). The goal of the present work was to further generate a better T4MO variant for the production of hydroxytyrosol. As the cloning steps associated with producing double and triple mutants of this enzyme are very laborious and time-consuming (e.g., QuikChange mutagenesis cannot be applied due to the large plasmid involved [9 kb], and a three-step PCR is needed for each mutant [5, 6]), the integration of the Nov and Wein statistical model was evaluated.

MATERIALS AND METHODS

Bacterial strains and growth conditions. *Escherichia coli* TG1 (*supE hsd*Δ5 *thi* Δ(*lac-proAB*) F′ [*traD36 proAB*$^+$ *lacI*^{q} *lacZ*ΔM15]) with plasmid constructs was routinely cultivated at 37°C in Luria-Bertani (LB) medium (27) supplemented with 100 μg/ml kanamycin to maintain the plasmid. To express the toluene monooxygenase genes stably and constitutively from the same promoter, the expression vector pBS(Kan)T4MO (henceforth T4MO) was constructed as described earlier (31). All experiments were conducted by diluting overnight cells to an optical density at 600 nm (OD_{600}) of 0.1 and growing them to an OD of 1.3. The exponentially grown cells were centrifuged at 8,000 × *g* for 10 min at 25°C in a Sigma 4K15 centrifuge (Sigma, Osterode, Germany) and resuspended in potassium phosphate buffer (100 mM, pH 7.0).

Protein analysis and molecular techniques. Protein samples of cells grown with 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) were analyzed on standard 12% Laemmli discontinuous sodium dodecyl sulfate (SDS)-polyacrylamide gels (27). T4MO variants comprising the initial data set (first phase) along with those comprising the second phase were analyzed by SDS-PAGE to ascertain that the increase in activity is the result of the desired mutation, rather than of unexpected changes in expression level. The bands representing the different enzyme subunits of the mutants and wild type (WT) had similar intensities on the gel. As the cell growth and the biotransformation conditions were identical for the WT and all mutants (including those reported in previous papers), the changes in activity appear to arise from the specific mutations and not from different expression levels.

Plasmid DNA was isolated using a minikit (Qiagen, CA), and DNA fragments were collected using the E-Gel CloneWell system, comprised of 0.8% SYBR Safe gels and the E-Gel iBase as the electrophoresis unit (Invitrogen, Carlsbad, California). Transformation of *E. coli* cells with plasmid DNA was performed via electroporation using a Micro-Pulser instrument (Bio-Rad, CA) with the program Ec2 (2.5 kV, 1 pulse for a 0.2-cm cuvette).

Mutagenesis of TmoA T4MO. Site-directed mutagenesis at the T4MO *tmoA* gene was performed via overlap extension PCR using the primers listed in Table 1 in a way similar to that described previously (5). The appropriate T4MO plasmids containing a site mutation(s) were used as templates to create the double, triple, and tetra mutants. The PCR program consisted of an initial denaturation at 94°C for 2 min, followed by 25 cycles of 94°C for 45 s, 55°C for 45 s, and 72°C for 2.2 min, with a final extension at 72°C for 8 min. Each pair of fragments was combined during the final reassembly PCR in a 1:1 molar ratio using the outer primers T4MObefEcoRI Front and T4MOABRear. The assembling PCR was programmed similarly to the above-described PCR program, with extension at 72°C for 3.15 min instead of 2.2 min. The assembled PCR fragment was ligated into WT T4MO, after the double digestion of both the vector and the insert with EcoRI and AatII, replacing the corresponding fragment in the original plasmid. The resulting plasmid library was introduced by electroporation into *E. coli* TG1 cells. Verification of the mutations was obtained by sequencing.

Error-prone PCR was performed as described earlier in detail for T4MO (6). Saturation mutagenesis at position I100 was described by Feingersch et al. (7). Saturation mutagenesis at position D285 was described by Brouk et al. (5).

Whole-cell enzymatic biotransformations. Whole-cell activity assays were performed as described previously (6, 7), using screw-cap 16-ml glass vials containing 2 ml cells and 0.25 mM substrate (added from a 100 mM stock solution in ethanol). The substrate, 2-phenylethanol (PEA), was purchased from Sigma-Aldrich Chemical Co. (Sigma-Aldrich, Rehovot, Israel). All vials were shaken at 30°C and 600 rpm (Vibramax 100; Heidolph, Nurenberg, Germany), and the reaction was stopped periodically (a vial was removed at each time period) by filtration of the cells. The progress of enzymatic hydroxylation of PEA was measured by reverse-phase high-performance liquid chromatography (HPLC), and the initial transformation rates (calculated as the amount of PEA decreased

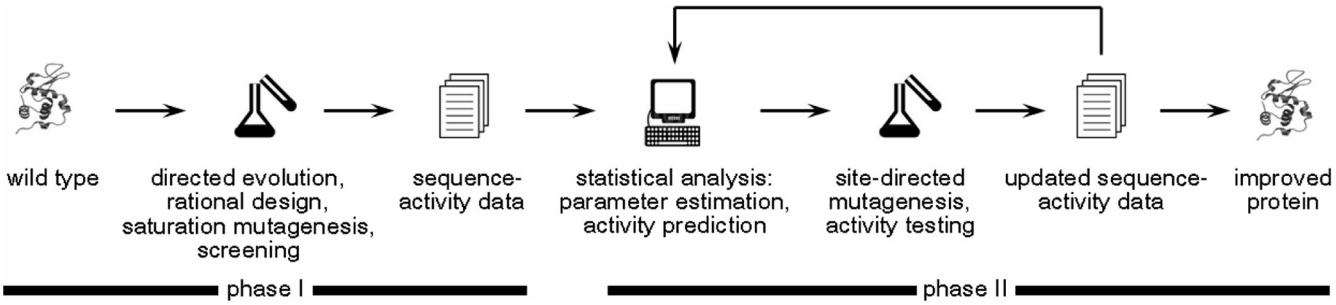

FIG. 1. Flowchart describing the design strategy. The first phase of the experiment consists of traditional protocols—directed evolution, rational design, and saturation mutagenesis—to produce initial genetic diversity. The resulting sequence-activity data are the basis for the second phase, in which several rounds of statistical analysis (two rounds in our study but possibly more, in general) point to a few promising variants with high expected activities; these variants are produced via site-directed mutagenesis and tested for their activity, and the resulting data serve as additional input for the next round.

in the culture supernatant) were determined by sampling at 0.5- to 20-min intervals during the first 2 to 5 h. The specific activity (nmol/min/mg protein) was calculated as the ratio of the initial transformation rate and the total protein content (0.24 mg protein/ml/OD_{600} for T4MO [30]). Activity data reported in this paper are based on at least two independent results.

Analytical methods. The analytical method was the same as we described previously (6). HPLC analysis was performed with an Agilent 1100-series instrument (Agilent Technologies, CA) using an Eclipse XDB-C_{18} column (5 μm, 4.6 by 150 mm; Agilent Technologies, CA) equipped with a photodiode array detector. The isocratic elution was performed with 85% acidic H_2O (0.1% formic acid) and 15% acetonitrile as the mobile phase at a flow rate of 1 ml/min. Compounds were identified by comparison of retention times and UV-visible spectra to those of the appropriate standards.

The sequence-activity statistical model. The statistical model used for data analysis and activity prediction was described in detail previously (22). Briefly, the model captures three characteristics of the protein sequence-activity relationship: dominance in mean of the wild type (*a priori*, mutations are more likely to be deleterious than beneficial), additivity (the activity change from a multiple-position mutation is roughly the sum of the corresponding single-mutation activity changes), and clustering (beneficial mutations tend to cluster at relatively few positions). The model postulates that the activity values (after transformation) of the various T4MO variants may be viewed as a realization of a Gaussian random field, governed by four parameters: m, σ_S^2, σ_R^2, and σ_N^2. With F_s denoting the activity of a sequence, s, the model can be shown to give rise to the following joint distribution:

$$\mathrm{E}(F_s) = d(s) \cdot m$$
$$\mathrm{Var}(F_s) = d(s) \cdot (\sigma_S^2 + \sigma_R^2) + \sigma_N^2$$
$$\mathrm{Cov}(F_s, F_{s'}) = \sigma_S^2 M_1(s, s') + \sigma_R^2 M_2(s, s')$$

where $d(s)$ is the number of positions in which s differs from the wild-type sequence, $M_1(s, s')$ is the number of positions in which both s and s' differ from the wild type, and $M_2(s, s')$ is the number of positions in which s and s' have the same amino acid and this amino acid is different from the corresponding wild-type amino acid.

By the method of Barak et al. (3), the raw activity data were logarithmically transformed, so that F_s was equal to $\log_{10}(V_s/V_{WT})$, where V_s and V_{WT} are the original activity values (in nmol/min/mg protein) of s and the wild-type sequence, respectively.

The model parameters were estimated in a maximum-likelihood (ML) procedure (17), using MATLAB's optimization toolbox. For fitness prediction, the conditional activity distribution of candidate variants, given the observed activity of the variants already screened, was calculated according to the appropriate multivariate normal conditional distribution formulas (32). The parameters used in this calculation were the point ML estimates, except for m, which, in a procedure similar to that in reference 3, was varied across the values $\{-0.1, -0.3, -0.5\}$ to mitigate the bias of the sample.

RESULTS

A two-phase design strategy was employed, as described schematically in Fig. 1.

Phase I: directed evolution and rational design. In phase I of the experiment, prior to application of the statistical model, a broad data set of results was assembled. Table 2 summarizes the influence of mutations in the α-subunit of the T4MO hy-

TABLE 2. Initial data matrix used for the first round of the statistic model to determine the effect of different mutations on the relative activity of T4MO

Approach used for generating the mutants	Mutation(s)	Relative activity of the mutant[a]
	None (WT T4MO)	1
Directed evolution	M37V S46C A171V T351S	1.1
	D48E	4
	M136T F269L W343 (stop codon)	0
	Q243R L306V N446K	0.6
	S395C[b]	8.7
Site-specific saturation mutagenesis	I100A[b]	35
	I100S[b]	33.6
	I100D[b]	24.5
	I100V[b]	21.1
	I100G[b]	35.2
	D285P[c]	3.3
	D285Y[c]	3
	D285C[c]	4
	D285L[c]	5.4
	D285A[c]	2.7
	D285Q[c]	11.7
	D285V	1.7
	D285R	2.4
	D285T	0.9
	D285I[c]	7.4
Site-directed mutagenesis	E214G	11
	I100A E214G	49
	I100G E214G	41
	D285P E214G	11

[a] Relative activity is represented by the initial PEA oxidation rate, normalized to that of WT T4MO. The WT value was considered 1, and the enzyme had an activity of 0.023 nmol/min/mg protein with an initial PEA concentration of 0.25 mM. The cell growth and the biotransformation conditions were identical for the WT and all mutants (including those reported in previous papers). Moreover, the expression levels of the respective proteins were visualized by SDS-PAGE to ascertain similar expression levels.

[b] A variation that was reported previously to result in improved activity by Brouk et al. (6).

[c] A variation that was reported previously to result in improved activity by Brouk et al. (5).

droxylase component on the relative activity of the enzyme, compared to the wild-type (WT) activity. The variants were obtained through rational design and directed evolution, as follows. Directed evolution was performed as described previously (6). Generally, by utilizing error-prone PCR, random mutations were generated, thus creating a diverse library of T4MO variants. Screening a library of over 3,000 mutants using an agar plate assay (6) resulted in the finding of four different variants, whereas following whole-cell biotransformation, only two exhibited over 2-fold improvement in activity (Table 2). Both substitutions, D48E, which improved the enzyme activity 4-fold, and S395C, which improved the enzyme activity 9-fold, are located far from the active site.

The semirational approach of site-specific saturation mutagenesis, which introduces all possible amino acids at a predetermined position in a gene, was utilized for randomizing positions I100 and D285 in T4MO. Residue I100 is positioned at the entrance of the hydrophobic cavity surrounding the diiron binding site and was shown previously to have a major influence on the rate and specificity of the enzyme (5–8, 20, 29, 30). This residue was proposed to act as a gate to the diiron active-site pocket and to influence the substrate alignment in the active-site cavity (5–7). Residue D285 is located at the entrance of the tunnel leading to the active site. The importance of this residue was recently tested using PEA and additional substrates and was shown to influence the activity rate, but not the specificity, of the enzyme (5). It was hypothesized that the polar and bulky residue of Asp 285 might control the substrate entrance and product efflux to/from the active site. Using NNN degenerate primers, two libraries were constructed for the purpose of introducing all possible amino acids at positions I100 and D285 (5, 7). Following screening of both libraries by either the agar plate assay (6) or the one-point biotransformation assay (5), variants with improved activity were sequenced and their activities were evaluated by whole-cell biotransformation, providing a link between the mutations and the enzyme activity (Table 2). As we reported previously, I100A, I100S, I100D, I100V, and I100G variants, which oxidized PEA $\geq$20-fold faster than the WT, were found from examining the T4MO I100 library (6). Moreover, screening the T4MO D285 library resulted in seven additional variants, the D285P, D285A, D285Y, D285L, D285C, D285I, and D285Q variants, which exhibited $>$2-fold improvement in activity compared to the WT (5). The activities of three other D285 variants (D285S, D285R, and D285T variants), which were known by their sequences, were also evaluated and added into the initial data set.

Finally, the rational-design approach of site-directed mutagenesis was used to introduce a specific substitution at a specific predetermined position (E214G). Position E214 is located in the tunnel entrance, facing residue 285. This position is analogous to position TouA E214 in ToMO, which was found by Vardar et al. (36) to influence the rate, but not regioselectivity, of *o*- and *p*-nitrophenol hydroxylation. It was shown that the size of the residue at position 214 is most likely to be the factor influencing the oxidation rate, leading to 15-fold improvement for *p*-nitrophenol oxidation by the ToMO E214G variant (36). Accordingly, the Glu at position 214 in TmoA T4MO was chosen to be replaced with Gly, which has the smallest side chain. The E214G substitution was introduced into the plasmid of the WT enzyme, as well as into variants that exhibited improved activity on PEA. The E214G variant had an 11-fold improvement in activity, whereas an additive effect was obtained by combining the E214G substitution and other beneficial mutations (Table 2). Thus, the addition of E214G to the I100G variant, which exhibited a 35-fold activity improvement, resulted in a 41-fold increase in activity. Likewise, the E214G I100A double mutant had a 49-fold-increased activity, whereas the I100A variant alone exhibited a 35-fold activity improvement compared with the WT. In contrast, combining the mutation E214G with D285P did not result in an additive effect on activity.

Phase II: statistics-based screening. The initial data set generated in phase I (Table 2) contained activity data for 24 variants, with mutations at 9 positions; the number of substitutions per position ranged between 1 and 10. Clearly, generating all ~13,000 possible mutants spanned by this genetic diversity, sequencing them for verification, and analyzing their activity and specificity by whole-cell biotransformations would be an expensive and time-consuming effort. To alleviate this burden, we used the statistical approach to identify a few promising variants expected to have high activity.

(i) First round. In the first round of phase II, seven variants were identified through statistical analysis. Six of them are the variants with the highest expected activities among all variants with 2 point mutations, and the seventh is the variant with 3 point mutations with the highest expected activity. These seven variants were generated by site-directed mutagenesis, introducing one additional mutation into the *tmoA* gene containing the single or double substitution. The relative activities of the variants were calculated as the initial oxidation rate on PEA (determined via HPLC analysis with an initial PEA concentration of 0.25 mM) normalized to that of WT T4MO, whose activity was considered 1. The variants and their relative activity values are listed in Fig. 2.

Overall, the average activity of the mutants found in the first round of phase II is 4.6-fold higher than the average activity of phase I's mutants ($P = 0.007$). Importantly, the most active variant, TmoA I100A E214G D285Q, exhibited an initial PEA oxidation rate of 4.4 $\pm$ 0.3 nmol/min/mg protein, which is 190-fold higher than the rate obtained by the WT (0.023 $\pm$ 0.001 nmol/min/mg protein) and a 4-fold improvement compared to the rate of 1.1 $\pm$ 0.1 nmol/min/mg protein obtained by the I100A E214G double mutant, the most active variant from phase I (Fig. 2).

(ii) Second round. In an attempt to further improve the enzyme activity toward PEA hydroxylation, a second round of statistical analysis was performed. In this round, the data set used for parameter estimation and activity prediction consisted of both phase I's data (Table 2) and phase II's first-round results. Nine variants were considered: five with 3 point mutations, three with 4 point mutations, and one with 5 point mutations. As in the first round, the variants in each such group were the top variants in their class in terms of conditional expected activity. These variants were generated, and their activities were evaluated using HPLC analysis (see Fig. 2 for the sequences and corresponding relative activity values). Although the average activity of the variants obtained in the second statistical round was 7.3-fold higher than phase I's average activity ($P < 0.0001$) and 1.6-fold higher than phase

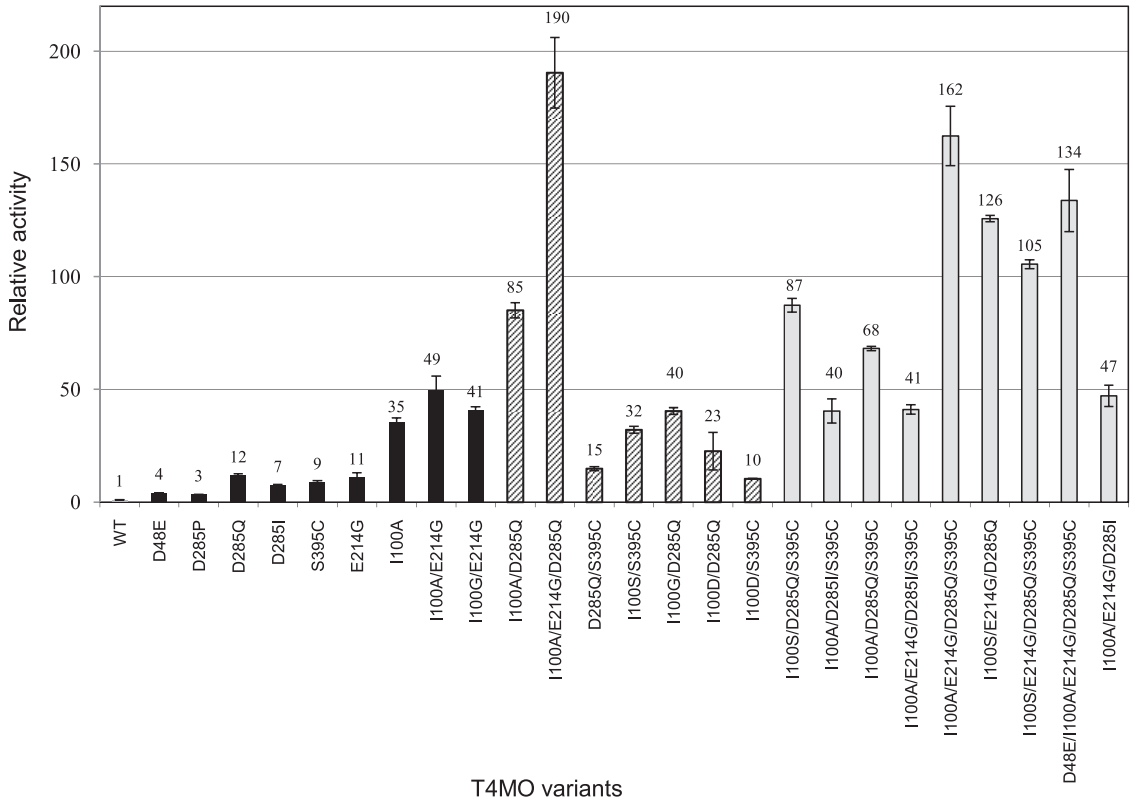

FIG. 2. Evolving T4MO for PEA hydroxylation by the aid of a statistical model. Relative activities on PEA (indicated above each column) of three evolving generations: (i) 10 representative variants out of the 24 T4MO variants comprising the initial data set used for the statistical model (filled bars), (ii) the first-generation mutants obtained by the statistical model (hatched bars); and (iii) the second-generation mutants obtained in the second statistical round (gray bars). Relative activity is presented as the initial PEA oxidation rate (determined via HPLC analysis) normalized to that of WT T4MO. The value for the WT, which had an activity of 0.023 nmol/min/mg protein on an initial PEA concentration of 0.25 mM, is designated 1.

II's first-round average activity ($P = 0.03$), no single variant was found to improve upon the I100A E214G D285Q variant, the best mutant discovered in the first round of phase II.

To further evaluate the statistical model, we also examined variants predicted to have low activity, as a control. Two variants, the M37V D285T and I100D D285V mutants, were predicted (based on the sequence-activity data from phase I and from phase II's first round) to be nonbeneficial combinations. The relative activities of both variants, 10.0 ± 1.0 and 20.71 ± 0.08, respectively, was lower than that of all nine variants produced in the second round (based on exactly the same sequence-activity data); the corresponding average activity was also significantly lower (15.35 versus 90, $P = 0.018$). This result strengthens the validity of the statistical model.

It should be noted that evolving T4MO activity using the statistical model was done with emphasis on the PEA consumption rate. Nevertheless, the product distribution was also measured and considered. It was reported previously that mutations close to the active site (i.e., residue I100) affected both the enzyme activity and product distribution, while other mutations, such as D285, which is located in the tunnel entrance, affected only the activity rate (5, 6). This observation was consistent throughout our work, including with the initial data set and the phase II variants. All of the phase II variants containing a substitution at position I100 were capable of forming hydroxytyrosol, unlike the WT enzyme. Substitutions in other positions did not seem to affect the regiospecificity of the enzyme. Therefore, it can be concluded that by improving PEA oxidation rate and integrating a substitution allowing hydroxytyrosol formation, we succeeded in improving T4MO variants for hydroxytyrosol biosynthesis.

DISCUSSION

Directed evolution is a useful method to improve enzymes in the absence of a structural model, provided that a simple and reliable screen is available. The initial increase in activity is often not high, and subsequent mutagenesis rounds are thus performed on the best variant from the first round of screening to improve further activity or stability. Tracewell and Arnold recently argued that a highly effective directed-evolution strategy is to gradually accumulate single beneficial mutations, either sequentially or by recombination (33). If structural information is available, iterative saturation mutagenesis at a small number of preselected positions is a good option for combining beneficial mutations if the desired property is encoded by changes at the chosen sites (24). However, most methods suggested in the literature for improving enzyme activity have been evaluated with rather simple enzyme systems in which one gene encodes one enzyme (9, 25, 26). In contrast, non-

heme mono- and dioxygenases, which have great potential as industrial catalysts, are multiprotein complexes, encoded by up to six genes (16). Subsequently, the plasmids are large and difficult to engineer. Even though only the hydroxylase subunit is targeted for mutagenesis, the molecular work with a plural gene plasmid of ca. 9 kb is not straightforward. Protocols such as QuikChange, which are routinely used for saturation and site-directed mutagenesis, are not applicable for mutating enzymes such as T4MO. Therefore, when evolving such complex enzymes, the need for methods that reduce the number of variants tailored and evaluated is ever more crucial. The motivation behind this work, of combining statistical analysis with experimental protein engineering, was to greatly reduce time and labor when creating an active variant for hydroxytyrosol synthesis.

A broad data set of mutants with improved activities toward PEA was constructed in phase I through both rational and nonrational protein engineering approaches. Next, in phase II, statistical techniques were employed to identify a few promising variants with increased expected activities.

A triple mutant with 190-fold improvement in activity was found in the first round of phase II. This triple mutant, T4MO I100A E214G D285Q, oxidized PEA 2.6-fold faster than the WT on toluene, the natural substrate of the enzyme (a rate of 1.67 nmol/min/mg protein was reported by Brouk et al. for the same reaction conditions [5]). Interestingly, although the average activity of the variants from phase II's second round increased significantly relative to that of the first round (let alone relative to that of phase I), and although 6 of the 9 variants of the second round were more active than phase I's best variant, none of the variants from phase II's second round surpassed T4MO I100A E214G D285Q in activity. Only a systematic screening of all ~13,000 possible variants spanned from the initial genetic diversity (a Herculean feat for a complex enzyme such as T4MO) could reveal whether phase I's first round indeed exhausted the activity potential of T4MO or whether more-active variants can be discovered. The increasing average activity across the two rounds of phase II, however, indicates that the statistical model is capable of identifying variants with improved activity, while reducing laboratory labor.

The most active variant discovered by the statistical model contains three amino acid substitutions: I100A, E214G, and D285Q. The first position, I100, is located at the entrance of the hydrophobic cavity surrounding the diiron binding site. We have previously reported that replacing the Ile at position 100 with Ala, Ser, or Gly resulted in improved activity on PEA and, more importantly, in the dihydroxylation of PEA to form the desirable hydroxytyrosol (6). Furthermore, both positions E214 and D285 are located at the entrance of the tunnel, leading to the active site and facing one another. It was shown that alternating these positions influences the oxidation rate but not the product distribution (5, 36). Consequently, combining mutations in the active-site entrance which enable hydroxytyrosol formation and accelerate the reaction rate with mutations in the tunnel mouth lead, as shown here, to elevated activity. Generally, introducing additional mutations far from the active site, such as S395C and D48E, resulted in decreased activity (Fig. 2), possibly due to interference with the structure's stability.

Interestingly, even though both D285Q and D285I improved the enzymatic activity in similar ways (12- and 7-fold activity improvements, respectively), all of the combinations which included the D285Q mutation resulted in an activity much higher than those which included the D285I substitution. For example, the I100A E214G D285Q and I100A E214G D285I variants differ only in the amino acid residue located at position 285. While the I100A E214G D285Q triple mutant displayed a synergistic activity and oxidized PEA 190-fold faster than the WT, the addition of the D285I substitution did not increase the activity of the I100A E214G variant at all (Fig. 2).

Mutational additivity. One of the assumptions of the statistical sequence-activity model used in this work is that of mutational additivity, that is, that the combined change in activity due to an introduction of multiple mutations roughly equals the sum of the corresponding individual activity changes. Such additivity was observed in previous studies (4, 37) and is supported by the low estimated fraction of the nonadditivity variance from the total activity variance [$\sigma_N^2/\mathrm{var}(F_S) = 0.117$ for variants with 2 point mutations and 0.081 for variants with 3 point mutations].

When using the raw data (before the logarithmic transformation), deviations from perfect additivity occur in both directions; some mutants exhibit activity which is lower than the sum of the single point mutations constituting them (e.g., combining D285Q and S395C, with relative activities of 12 and 9, respectively, yielded a variant with an activity of 15), while others exhibit synergy (e.g., combining I100A, E214G, and D285Q into a triple mutant) (Fig. 2). When using the transformed data, all deviations are negative; this observation may be partially explained as a regression to the mean effect and may also indicate that a transformation that is "less concave" than the logarithmic value (say, a Box-Cox transformation with a power of <1) may provide a better fit. We pursue these directions in a sequel to this work.

The four-parameter model. To date, the only experimental work done using the statistical approach of Nov and Wein (22) is that of Barak et al. (3). However, due to a certain peculiarity of the data, the model used in the latter study included only three parameters, rather than the four parameters required by the full model. The results presented in this paper support the applicability and usefulness of the four-parameter model.

Generality of the statistical model. This work and the work of Barak et al. (3) demonstrate that the Nov and Wein statistical model is capable of reducing laboratory time and labor when improving a single, desirable enzyme property (PEA oxidation rate here and chromate or uranyl reduction rate in reference 3). The model requires only sequence-activity data, though structural knowledge may be used to select the positions and substitutions generating this data, as in this study but not in reference 3. In a future work, we plan to extend the model so that it can optimize simultaneously two desirable properties (e.g., stability and enzymatic activity), as is often required in industrial settings. We also plan to test whether the model can be used to improve desirable properties of proteins that are not enzymes, for example, the binding strength of an antibody.

REFERENCES

1. **Aita, T., and Y. Husimi.** 2000. Adaptive walks by the fittest among finite random mutants on a Mt. Fuji-type fitness landscape. II. Effect of small non-additivity. J. Math. Biol. **41:**207–231.
2. **Bailey, L. J., J. G. McCoy, G. N. Phillips, and B. G. Fox.** 2008. Structural consequences of effector protein complex formation in a diiron hydroxylase. Proc. Natl. Acad. Sci. U. S. A. **105:**19194–19198.
3. **Barak, Y., Y. Nov, D. F. Ackerley, and A. Matin.** 2008. Enzyme improvement in the absence of structural knowledge: a novel statistical approach. ISME J. **2:**171–179.
4. **Benos, P. V., M. L. Bulyk, and G. D. Stormo.** 2002. Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. **30:**4442–4451.
5. **Brouk, M., N. L. Derry, J. Shainsky, Z. Ben-Barak Zelas, Y. Boyko, K. Dabush, and A. Fishman.** 2010. The influence of key residues in the tunnel entrance and the active site on activity and selectivity of toluene-4-monooxygenase. J. Mol. Catal. B Enzym. **66:**72–80.
6. **Brouk, M., and A. Fishman.** 2009. Protein engineering of toluene monooxygenases for synthesis of hydroxytyrosol. Food Chem. **116:**114–121.
7. **Feingersch, R., J. Shainsky, T. K. Wood, and A. Fishman.** 2008. Protein engineering of toluene monooxygenases for synthesis of chiral sulfoxides. Appl. Environ. Microbiol. **74:**1555–1566.
8. **Fishman, A., Y. Tao, W. E. Bentley, and T. K. Wood.** 2004. Protein engineering of toluene 4-monooxygenase of *Pseudomonas mendocina* KR1 for synthesizing 4-nitrocatechol from nitrobenzene. Biotechnol. Bioeng. **87:**779–790.
9. **Fox, R. J., 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, and G. W. Huisman.** 2007. Improving catalytic function by ProSAR-driven enzyme evolution. Nat. Biotechnol. **25:**338–344.
10. **Fox, R. J., and G. W. Huisman.** 2008. Enzyme optimization: moving from blind evolution to statistical exploration of sequence-function space. Trends Biotechnol. **26:**132–138.
11. **Hatti-Kaul, R., U. Tornvall, L. Gustafsson, and P. Borjesson.** 2007. Industrial biotechnology for the production of bio-based chemicals: a cradle-to-grave perspective. Trends Biotechnol. **25:**119–124.
12. **Illanes, A. (ed.).** 2008. Enzyme biocatalysis: principles and applications. Springer, New York, NY.
13. **Jemai, H., A. El Feki, and S. Sayadi.** 2009. Antidiabetic and antioxidant effects of hydroxytyrosol and oleuropein from olive leaves in alloxan-diabetic rats. J. Agric. Food Chem. **57:**8798–8804.
14. **Kauffman, S., and S. Levin.** 1987. Towards a general theory of adaptive walks on rugged landscapes. J. Theor. Biol. **128:**11–45.
15. **Kazlauskas, R. J., and U. T. Bornscheuer.** 2009. Finding better protein engineering strategies. Nat. Chem. Biol. **5:**526–529.
16. **Leahy, J. G., P. J. Batchelor, and S. M. Morcomb.** 2003. Evolution of the soluble diiron monooxygenases. FEMS Microbiol. Rev. **27:**449–479.
17. **Lehmann, E. L., and G. Casella.** 1998. Theory of point estimation. Springer-Verlag, New York, NY.
18. **Liao, J., M. K. Warmuth, S. Govindarajan, J. E. Ness, R. P. Wang, C. Gustafsson, and J. Minshull.** 2007. Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol. **7:**16.
19. **Marrugat, J., M. I. Covas, M. Fit, H. Schroder, E. Miro-Casas, E. Gimeno, M. C. Lopez-Sabater, R. de la Torre, and M. Farre.** 2004. Effects of differing phenolic content in dietary olive oils on lipids and LDL oxidation. Eur. J. Nutr. **43:**140–147.
20. **McClay, K., C. Boss, I. Keresztes, and R. J. Steffan.** 2005. Mutations of toluene-4-monooxygenase that alter regiospecificity of indole oxidation and lead to production of novel indigoid pigments. Appl. Environ. Microbiol. **71:**5476–5483.
21. **Mitchell, K. H., J. M. Studts, and B. G. Fox.** 2002. Combined participation of hydroxylase active site residues and effector protein binding in a *para* to *ortho* modulation of toluene 4-monooxygenase regiospecificity. Biochemistry **41:**3176–3188.
22. **Nov, Y., and L. M. Wein.** 2005. Modeling and analysis of protein design under resource constraints. J. Comput. Biol. **12:**247–282.
23. **Otten, L. G., F. Hollmann, and I. W. Arends.** 2010. Enzyme engineering for enantioselectivity: from trial-and-error to rational design? Trends Biotechnol. **28:**46–54.
24. **Reetz, M. T., and J. D. Carballeira.** 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat. Protoc. **2:**891–903.
25. **Reetz, M. T., D. Kahakeaw, and R. Lohmer.** 2008. Addressing the numbers problem in directed evolution. Chembiochem **9:**1797–1804.
26. **Reetz, M. T., M. Puls, J. D. Carballeira, A. Vogel, K. E. Jaeger, T. Eggert, W. Thiel, M. Bocola, and N. Otte.** 2007. Learning from directed evolution: further lessons from theoretical investigations into cooperative mutations in lipase enantioselectivity. Chembiochem **8:**106–112.
27. **Sambrook, J., and D. W. Russell.** 2001. Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, New York, NY.
28. **Shivange, A. V., J. Marienhagen, H. Mundhada, A. Schenk, and U. Schwaneberg.** 2009. Advances in generating functional diversity for directed protein evolution. Curr. Opin. Chem. Biol. **13:**19–25.
29. **Tao, Y., W. E. Bentley, and T. K. Wood.** 2005. Regiospecific oxidation of naphthalene and fluorene by toluene monooxygenases and engineered toluene 4-monooxygenases of *Pseudomonas mendocina* KR1. Biotechnol. Bioeng. **90:**85–94.
30. **Tao, Y., A. Fishman, W. E. Bentley, and T. K. Wood.** 2004. Altering toluene 4-monooxygenase by active-site engineering for the synthesis of 3-methoxycatechol, methoxyhydroquinone, and methylhydroquinone. J. Bacteriol. **186:**4705–4713.
31. **Tao, Y., A. Fishman, W. E. Bentley, and T. K. Wood.** 2004. Oxidation of benzene to phenol, catechol, and 1,2,3-trihydroxybenzene by toluene 4-monooxygenase of *Pseudomonas mendocina* KR1 and toluene 3-monooxygenase of *Ralstonia pickettii* PKO1. Appl. Environ. Microbiol. **70:**3814–3820.
32. **Tong, Y. L.** 1990. The multivariate normal distribution. Springer-Verlag, New York, NY.
33. **Tracewell, C. A., and F. H. Arnold.** 2009. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr. Opin. Chem. Biol. **13:**3–9.
34. **Tripoli, E., M. Giammanco, G. Tabacchi, D. Di Majo, S. Giammanco, and M. La Guardia.** 2005. The phenolic compounds of olive oil: structure, biological activity and beneficial effects on human health. Nutr. Res. Rev. **18:**98–112.
35. **Turner, N. J.** 2009. Directed evolution drives the next generation of biocatalysts. Nat. Chem. Biol. **5:**567–573.
36. **Vardar, G., and T. K. Wood.** 2005. Alpha-subunit positions methionine 180 and glutamate 214 of *Pseudomonas stutzeri* OX1 toluene-*o*-xylene monooxygenase influence catalysis. J. Bacteriol. **187:**1511–1514.
37. **Voigt, C. A., S. Kauffman, and Z. G. Wang.** 2000. Rational evolutionary design: the theory of in vitro protein evolution. Adv. Protein Chem. **55:**79–160.