The ISME Journal (2008) 2, 171–179

www.nature.com/ismej

ORIGINAL ARTICLE

Enzyme improvement in the absence of structural knowledge: a novel statistical approach

Yoram Barak[1,4], Yuval Nov[2,4], David F Ackerley[1,3] and A Matin[1]
[1]*Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA;* [2]*Department of Statistics, University of Haifa, Haifa, Israel and* [3]*School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand*

Most existing methods for improving protein activity are laborious and costly, as they either require knowledge of protein structure or involve expression and screening of a vast number of protein mutants. We describe here a successful first application of a novel approach, which requires no structural knowledge and is shown to significantly reduce the number of mutants that need to be screened. In the first phase of this study, around 7000 mutants were screened through standard directed evolution, yielding a 230-fold improvement in activity relative to the wild type. Using sequence analysis and site-directed mutagenesis, an additional single mutant was then produced, with 500-fold improved activity. In the second phase, a novel statistical method for protein improvement was used; building on data from the first phase, only 11 targeted additional mutants were produced through site-directed mutagenesis, and the best among them achieved a >1500-fold improvement in activity over the wild type. Thus, the statistical model underlying the experiment was validated, and its predictions were shown to reduce laboratory labor and resources.
The ISME Journal (2008) **2,** 171–179; doi:10.1038/ismej.2007.100; published online 22 November 2007
Subject Category: microbial engineering
Keywords: protein design; Nov–Wein model; directed evolution; rational design

Introduction

Improving the activity of a protein by manipulating its sequence—a process termed *protein design*—is of great interest in medicine and biotechnology, and has been widely practiced. However, the sequence space is 'more than astronomically' vast (Dennett, 1995; Chatterjee and Yuan, 2006), and it is neither experimentally feasible to test all possible mutants of a protein nor is it necessary, since many of the resulting sequences do not fold into functioning proteins (Arnold, 2006).

One mutagenesis approach, termed *rational design*, uses information about the three-dimensional structure of the protein and its target molecule to identify promising sequence changes. Thus, Grove *et al.* (2003) improved the activity of the *Escherichia coli* nitroreductase, NfsB, for prodrug reduction by targeted changes in amino acids around its active site; several other such structure-based improvements in proteins have been made (Chica *et al.*, 2005). However, deciphering the structure of a protein is expensive, laborious and time consuming, and activity predictions based on structure are limited in their success. Thus, design methods that do not rely on structural knowledge are needed, not only for proteins whose structure is not known but also where structural information is available, since activity may be influenced by amino acids not residing in the active site (Qian and Lutz, 2005; Park *et al.*, 2006).

An alternative to structure-based rational design is *directed evolution*—a selective process that mimics nature, whereby a protein is 'bred' through successive generation of gene libraries; the members of these libraries are randomly mutated and shuffled, and their resulting proteins are then screened for improved activity. Common methods for generating such libraries include error-prone PCR and recombination between homologous regions of related genes (Chen and Arnold, 1993; Stemmer, 1994; Aharoni *et al.*, 2005; Barak *et al.*, 2006a, b). Directed evolution is widely practiced and has produced important results, yet it typically

Correspondence: A Matin, Department of Microbiology and Immunology, Sherman Fairchild Science Building, Stanford University School of Medicine, 299 Campus Drive W, Stanford, CA 94305, USA.
E-mail: a.matin@stanford.edu
[4]These authors contributed equally to this work.
Received 4 June 2007; revised 8 October 2007; accepted 9 October 2007; published online 22 November 2007

necessitates expression, purification and screening of thousands of protein mutants. In addition, directed evolution is a 'blind' process, and it is virtually impossible to mathematically predict its success in improving activity.

A third approach to protein design models the relation between the sequence of a protein mutant and its activity (fitness) as a statistical relationship. That is, one assigns a *distribution* of activity levels for each protein mutant, rather than a single predicted activity, and can thereby specify probabilities for the various activity levels. Among the models that belong to this class are the *NK* model of Kauffman and Levin (1987), the Mount Fuji model of Aita and Husimi (2000) and various regression-like methods (Mee *et al.*, 1997; Lejon *et al.*, 2001). Many variants of the Mount Fuji and the *NK* models are Gaussian, and most of the regression-based methods are implicitly Gaussian, as they assume Gaussian distribution of the errors when computing confidence intervals, *P*-values, etc. The statistical approach to protein design circumvents the need to decipher a protein's structure and promotes identification of promising mutant candidates, thus significantly reducing the number of mutants that need to be screened.

Of special interest is recent work by Fox *et al.* (2007), in which the activity of bacterial halohydrin dehalogenase was significantly improved to meet design criteria in the commercial production of atorvastatin (Lipitor), a cholesterol-lowering drug. The enzyme was optimized through a statistical analysis method termed *protein sequence activity relationship*, combined with directed evolution and rational design.

We report here a successful first empirical application of a novel method belonging to the last mentioned class; the method is based on a statistical model for the sequence–activity relationship proposed by Nov and Wein (hereafter referred to as 'the model'), whose theoretical and mathematical details were published previously (Nov and Wein, 2005). Briefly, this model is *additive*, in the sense that it assumes that after proper transformation of the data, the change in activity caused by a multiple-residue mutation roughly equals the sum of the activity changes caused by the corresponding single-residue mutations; the degree of non-additivity is captured through one of the model's parameters. The model is sparse in parameters, and is mathematically tractable, conveniently allowing one to update the activity distributions of the yet-unexplored mutants from the sequence–activity data of tested mutants. In addition to their sequence–activity relationship model, Nov and Wein suggested an optimization module for selecting promising mutant candidates; a variant of this module was used in this study. The relevant aspects of the model used in this study are presented in the Materials and methods section.

The improvement efforts targeted the *E. coli* enzyme ChrR, an NAD(P)H-dependant oxidoreductase of unknown structure, which has a wide substrate range (Ackerley *et al.*, 2004), including several beneficial activities such as chromate and uranyl (U(VI)) reduction (useful in the bioremediation of these widespread pollutants (Ackerley *et al.*, 2004, 2006; Barak *et al.*, 2006b)) and prodrug reduction (useful in cancer chemotherapy (Barak *et al.*, 2006a)). Improvement in all three activities is reported.

Materials and methods

Strains, plasmids, genes, primers and growth conditions

Supplementary Table 1 lists the strains, plasmids and primers used in this study. The various strains were grown at 37 °C to mid-exponential phase, induced by 0.5 mM isopropyl-β-D-thiogalactoside and incubated overnight for protein production.

DNA techniques

Routine DNA manipulations were performed as described (Sambrook *et al.*, 1989; Barak *et al.*, 2006a, b). Plasmid DNA purification from *E. coli* was carried out by miniprep (Qiagen Inc., Valencia, CA, USA). DNA was sequenced by Sequetech Corporation, CA, USA, using appropriate primers (Supplementary Table 1).

Directed evolution of the chrR *gene for improving chromate reductase activity*

Error-prone PCR was used to introduce random mutations in the *chrR* gene (Barak *et al.*, 2006a, b), using the GeneMorph II Random Mutagenesis kit (Stratagene Corporation, La Jolla, CA, USA). Forward and reverse *chrR* primers (Supplementary Table 1) were used to amplify full-length hybrid products.

The shuffled genes were ligated into the pET28a$^+$ plasmid, and transformed into *E. coli* BL21 (DE3) (Invitrogen Inc., Carlsbad, CA, USA) to allow overexpression. Recombinants were selected on plates containing kanamycin (50 μg ml^{-1}). High-throughput screening of 7000 recombinants was performed by inoculating colonies into individual wells of 96-well microtiter plates, containing 200 μl Luria–Bertani medium and kanamycin. After growth to stationary phase (overnight incubation, final A_{660}, 1–1.5), 20 μl aliquots from each well were used to inoculate a second series of plates, using M9 minimal medium (Sigma Inc., St Louis, MO, USA). Each well received the same initial inoculum. The first set of plates was stored at −80 °C after addition of glycerol. Cells in the second inoculation series were allowed to grow to mid-exponential phase and then exposed to 0.5 mM isopropyl-β-D-thiogalactoside to induce the recombinant gene expression. After overnight incubation, cells were lysed by addition of 30 μl BugBuster (Novagen Inc., San Diego, CA, USA),

incubated for 20 min at room temperature, and centrifuged for 20 min at 3000 *g*. Supernatant (100 µl) was mixed with 10 µl solution of the following composition: 500 µM potassium chromate, 2 mM NADH, 100 mM Tris-HCl (pH 7) and ddH_2O (Barak *et al*., 2006a, b); and chromate reduction was assayed as described below.

The most efficient enzymes for Cr(VI) reductase activity were purified on nickel columns, as previously described (Ackerley *et al*., 2004), using inocula obtained from the frozen plates. Protein concentrations were determined with the Bio-Rad Dc protein assay kit, using bovine serum albumin as a standard.

Site-directed mutagenesis

Appropriate primers (Supplementary Table 1) were used for site-directed mutagenesis. These were designed to create single-codon mutations following the method of Kuipers *et al*. (1991). Verification that the desired mutations had been generated was obtained by sequencing. Proteins encoded by the modified genes were generated as described above.

Cr(VI) assays

Determination of Cr(VI) reduction rates by cell extract preparation and chromate reductase assays were conducted as described previously (Ackerley *et al*., 2004; Park *et al*., 2000). Kinetic measurements of enzyme activity were performed at pH 7 and at 37 °C. Each assay was conducted four times unless otherwise stated.

Assay for prodrug reduction

Reductive prodrugs become strong killing agents of biological cells upon reduction. The capacity of the mutant enzymes to carry out this reduction was determined with minor modifications as previously described (Barak *et al*., 2006a). Briefly, prodrug reduction mixtures contained mitomycin C, CB 1954 (5-aziridinyl-2,4-dinitrobenzamide) or 17-AAG (17-allylamino-17-demethoxygeldanamycin) at a concentration of 15 µM, 10 µg ml^{-1} pure enzyme, 50 µM NADPH and Dulbecco's modified Eagle's medium (Barak *et al*., 2006a) to a final volume of 0.5 ml. Following prodrug reduction for 30 min at 37 °C, 0.5 ml of JC breast cancer cells (~ 0.5–1×10^5) were added and the cells were incubated for additional 24 h. After the latter incubation, 20 µl of the color reagent, CellTiter 96 AqueousOne (Promega Inc., Madison, CA, USA) was added to 100 µl aliquots of the reaction mixture. Following 1 h of further incubation, A_{490} was measured in an ASYS UVM340 reader.

U(VI) determination

For selected mutant enzymes, uranyl reductase activity was also determined. This was carried out as described (Teixeira *et al*., 1999). Briefly, samples were collected after incubation for the specified time. A 120 µl sample was mixed with 130 µl reagent mixture containing 5:1:1:1:5 proportion of complexing solution, TAC (2-(2-thiazolyazo-*p*-cresol)), Triton X-100 (0.15 M), CTAB (*N*-cetyl-*N*,*N*,*N*-trimethyammonium bromide) and triethanolamine buffer (pH 6.5). The method depends on the TAC binding to U(VI), which is aided by Triton and CTAB. After 15 min of color development, the samples were read at $A_{588\,nm}$ using a Micro-Plate Reader (ASYS UVM340).

Table 1 First phase: the effect of sequence changes on chromate reductase activity (V_{max}) of the *E. coli* ChrR protein

Mutants	*Residue substitutions in WT ChrR*	V_{max} *(nmol Cr(VI) reduced per mg protein per min)*
ChrR	None	295 ± 27
ChrR6	V120A, Y128N, T160N, Q175L	8810 ± 611
ChrR7	V120A, Y128N, N154T, T160N, Q184H	23 200 ± 8180
ChrR8	V120A, Y128N, T160N	35 400 ± 2700
ChrR9	V120A, Y128N, G150S	11 100 ± 3450
ChrR10	V120A, Y128N	19 000 ± 2100
ChrR11	V120A	1100 ± 544
ChrR12	A44V, V120A, Y128N	38 300 ± ± 4300
ChrR13	V120A, Y128N, G150S, Q153H	67 500 ± 3950
ChrR14	V120A, Y128N, Q175L	7800 ± 3210
ChrR15	V120A, T160N	52 100 ± 3600
ChrR16	Y128N, T160N	625 ± 235
ChrR17	V120A, Y128N, T160N, Q175H, K187T	20 600 ± 7180
ChrR18	D103G, V120A, T160N	311 ± 86
ChrR19	A120V, T160N, Q175L	252 ± 42
ChrR20	Y128N, T160N, Q175L	9200 ± 545
ChrR21	Y128N	148 600 ± 46 600

Abbreviation: WT, wild type.
Mutants ChrR6–ChrR20 were obtained through directed evolution, and ChrR21 was produced through site-directed mutagenesis.

Computer programs

Sequences were aligned with Clustal W (http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html). Optimization for the maximum likelihood estimation of the model parameters, as well as other scientific programming, was carried out through MATLAB (The MathWorks Inc., Natick, MA, USA).

The model

The model has four parameters: the drift m, which is the expected change in fitness due to introduction of a new, arbitrary mutation (a negative number, as mutations more often decrease than increase fitness); the site variance σ_S^2, which is the variance of the change in expected fitness contribution due to a mutation across sites; the residue variance σ_R^2, which is the variance of the fitness contribution of a specific single-residue mutation within a site; and the non-additivity variance σ_N^2, which captures both the degree of non-additivity and the level of

measurement noise (as in all additive models, these two effects cannot be distinguished from one another). For a thorough presentation of the model, see Nov and Wein (2005). Due to reasons discussed below, a variant of the model, with only three parameters, was used. The fitness F_s of a mutant having sequence s and activity $V_{\max}$ (in nmole substrate converted per milligram protein per minute) was taken to equal $\log_{10}(V_{\max}/v_{\mathrm{wt}})$, where v_{wt} is the $V_{\max}$ value of the wild-type enzyme (which was 295; Table 1). This transformation improved the goodness-of-fit of the data to the model, and set the fitness of the wild type to 0, as required by the model.

The 16 mutant proteins sequenced in the first phase involved mutations in $n=11$ sites (A44, D103, V120, Y128, G150, Q153, N154, T160, Q175, Q184, K187; Table 1). Only one of these sites, Q175, had more than one substituent amino acid—Q175L and Q175H—of which the latter appeared in only one sequence, ChrR17. To improve the numerical stability of the estimation computations, ChrR17 was omitted from the data, so that only 15 sequences were used; otherwise, the parameter σ_{R}^2 would have appeared in only two entries of a 16×16 covariance matrix. It is for this reason that a three-parameter version of the model was used, employing the parameters m, σ_{S}^2 and σ_{N}^2. More specifically, the model is a Gaussian random field $\mathbf{F}=\{F_s\}$, where the index set of $\mathbf{F}$ consists of all $2^{11}=2048$ sequences that may be generated from the genetic diversity of the 15 mutants found in the first phase. The joint distribution of the elements of $\mathbf{F}$ is given by the following equations:

$$\mathrm{E}(F_s) = d(s,\hat{s})m \tag{1}$$

$$\mathrm{Var}(F_s) = d(s,\hat{s})\sigma_S^2 + \sigma_N^2 \tag{2}$$

$$\mathrm{Cov}(F_s, F_{s'}) = M(s,s')\sigma_S^2 \tag{3}$$

where $d(s,\hat{s})$ is the number of sites in which a sequence s differs from the wild-type sequence $\hat{s}$ and $M(s,s')$ is the number of sites in which both sequences s and s' differ from the wild-type sequence. In this three-parameter form (but not in the full four-parameter form), the model is similar to a regression model with random coefficients (sometime called hierarchical regression) without an intercept, in which the predictors are binary variables, indicating the presence or absence of a mutation, their coefficients are $N(m,\sigma_{\mathrm{S}}^2)$ random variables, and the variance of the error terms is σ_{N}^2. As no prior distribution is assumed over the parameters, the model is not Bayesian.

Parameter estimation

The parameters of the model were estimated by the maximum likelihood method. Specifically, m, σ_{S}^2 and σ_{N}^2 were initially estimated to be the maximizers of the likelihood function

$$L(m,\sigma_{\mathrm{S}}^2,\sigma_{\mathrm{N}}^2;\mathbf{F}_1) = \frac{1}{(2\pi)^{r/2}\sqrt{\det(\Sigma_1)}} \times \exp\left\{-\tfrac{1}{2}(\mathbf{F}_1-\mu_1)'\Sigma_1^{-1}(\mathbf{F}_1-\mu_1)\right\}, \tag{4}$$

where $\mathbf{F}_1$ is the log-transformed r-vector ($r=15$) of the 15 $V_{\max}$ values of the mutant proteins (Table 1, excluding the wild type and ChrR17), μ_1 is its mean vector (computed according to Equation (1)) and Σ_1 is its $r\times r$ covariance matrix (computed according to Equations (2) and (3)). The resulting estimates were $\sigma_{\mathrm{S}}^2=0.4861$ and $\sigma_{\mathrm{N}}^2=0.1478$. The estimate of the third parameter, m, was positive, in contrast to the model's assumptions. This finding was expected: the sequences obtained in the first phase were not a random sample from the sequence space, in which *a priori* it is expected that most mutations are deleterious (corresponding to a negative m); rather, these sequences were chosen by the selective directed evolution process due to their improved fitness, and thus carry seriously distorted information about m. Therefore, for fitness prediction purposes (see below), only the two estimated variance parameters σ_{S}^2 and σ_{N}^2 were used, and the value of m was varied, in jumps of size 0.2, across the range -1.5 to -0.1.

For the second application of the model predictions, after the activity information of the first round's five mutants became available, the parameters were re-estimated in a similar way, using $r=15+5=20$ in Equation (4) and appropriately modified $\mathbf{F}_1$, μ_1 and Σ_1. The resulting estimates were $\sigma_{\mathrm{S}}^2=0.4361$ and $\sigma_{\mathrm{N}}^2=0.1961$, very similar to those from the first round.

Fitness prediction

By the additivity of the model, the conditional expected fitness of a sequence s given the data, $\mathrm{E}(F_s\,|\,\mathbf{F}_1)$, is the sum of the conditional expected fitness contributions from each of the 11 mutated sites. The contribution from a site having the wild-type residue is 0 (and hence so is its expected contribution), and that of a site i with a non-wild-type residue is a random variable f_i. The conditional expected value of the vector $\mathbf{f}=(f_1,\ldots,f_n)$ is

$$E(\mathbf{f}|\mathbf{F}_1) = \mathbf{m} + \Sigma_2\,\Sigma_1^{-1}(\mathbf{F}_1-\mu_1), \tag{5}$$

where $\mathbf{m}$ is a constant n-vector, having all of its element equal to m; the matrix Σ_1^{-1} is the inverse of Σ_1 and Σ_2 is an $n\times r$ matrix, having σ_{S}^2 as its (i,j)th entry, if mutant j had a mutation at site i, and 0 otherwise.

In the first round of the second phase, the estimated variance parameters σ_{S}^2 and σ_{N}^2 correspond to a proportion of non-additive variance of $0.1478/(2\times0.4861+0.1478)=0.132$ among double

mutants, which is low enough to allow reliable predictions. As mentioned above, the value of m was not estimated from the data, and was varied from -1.5 to -0.1. For each value of m, the n-vector $E(\mathbf{f}|\mathbf{F}_1)$ was computed according to Equation (5) (see Supplementary Table 2), and the conditional expected fitness values of all possible $n(n-1)/2=55$ double mutants were calculated. Among these, the five double mutants with the highest expected fitness (averaged across all m, and not including sequences already in the data set) were identified, and their sequences are shown in Table 2A (second column). The sequences for the second round were chosen in the same method, with the appropriate changes to r, $\mathbf{F}_1$, μ_1 and Σ_1. Since only triple-residue mutants were considered in this round, the mutants chosen were the top ones, in terms of conditional expected fitness, among all $n(n-1)(n-2)/6=165$ triple mutants.

Results

A two-phase strategy for ChrR improvement was employed: a 'blind' directed-evolution approach in the first phase and the model-based predictions to obtain further improvement in the second. In the first phase, ChrR protein mutants were obtained by subjecting the *chrR* gene to three rounds of error-prone PCR. Each round was followed by screening the resulting mutant proteins for chromate reductase activity, using a colorimetric method that provides an approximate indication of the degree of improvement in this activity. Around 6000 mutants were screened. The top 15 mutant proteins were purified and sequenced, and their V_{max} (in nmole Cr(VI) reduced per milligram protein per minute) was measured (mutants ChrR6–ChrR20; Table 1). Eleven of these showed significantly higher V_{max} for this reduction (>25-fold improvement) compared to the wild-type enzyme, the best, ChrR13, showing a V_{max} of 67 500, corresponding to about 230-fold improved activity.

Sequence analysis revealed that the Y128N substitution was common to almost all of the improved mutants isolated in this phase, so an additional mutant, containing the single mutation Y128N, was generated through site-directed mutagenesis. This mutant (ChrR21; Table 1) surpassed the other mutants in chromate reductase activity, exhibiting a 500-fold improvement over the wild type. An additional (fourth) round of directed evolution, using DNA from ChrR6 to ChrR21 as template and screening around 1000 variants, did not yield further improvement.

The second phase of the study consisted of applying the model to the sequence–activity data of Table 1. The parameters of the model were estimated from the entire information of Table 1, and the sequences of the five most promising double mutants (that is, the five mutants that possess the highest conditional expected activity, among those differing from the wild type in two amino acids) were mathematically identified (ChrR22 to ChrR26; Table 2A). These mutant proteins were generated in pure form by site-directed mutagenesis and nickel column purification as described (Barak *et al.*, 2006a), and their V_{max} for chromate reduction was measured (third column of Table 2A). One of these, ChrR23, exhibited a V_{max} of 258 000, corresponding to an 876-fold improvement in activity over the wild type, around fourfold improvement over ChrR13 (the best mutant obtained in four rounds of directed evolution, which necessitated screening of 7000 mutants), and a 1.75-fold improvement over ChrR21 (the best mutant isolated in the first phase). In addition, the average V_{max} of mutants ChrR22–ChrR26 was significantly higher than the average V_{max} of the first-phase mutants ChrR6–ChrR21 (104 000 vs 26 000; $P=0.0084$ in a one-tailed Mann–Whitney test for median comparison).

To further improve the ChrR enzyme, we conducted a second screening round according to the model predictions. The parameters of the model were re-estimated, using the sequence–activity data of both Tables 1 and 2A, and the sequences of the seven most promising triple mutants were identified. One of these mutants could not be generated, but the remaining six were produced as described above, and their chromate reductase V_{max} values were measured (Table 2B). Strikingly, one of these, ChrR30, exhibited 1554-, 6.6- and 3.1-fold improvements over the wild type, ChrR13 and ChrR21, respectively. Thus, by screening just a few mutants, a multifold enhancement was obtained in an enzyme already improved to a large degree. The aggregate average V_{max} of the 11 mutants ChrR22–ChrR32 (117 000) was also significantly higher than that of the first-phase mutants ChrR6–ChrR21 ($P=0.0034$).

Previous results had shown a positive correlation between chromate reductase activity and other useful activities in ChrR mutants (Barak *et al.*, 2006a, b). We therefore examined the activity of three of the most active mutants in chromate reductase—ChrR21, ChrR23 and ChrR30—in two

Table 2A First round of second phase: sequence and V_{max} activity of five mutants predicted by the Nov–Wein model to have improved chromate reductase activity

Mutants	*Amino-acid substitutions from WT ChrR*	V_{max} *(nmol Cr(VI) reduced per mg protein per min)*
ChrR22	A44V, Y128N	38 300 ± 14 500
ChrR23	Y128N, G150S	258 300 ± 21 900
ChrR24	Y128N, Q153H	53 900 ± 17 300
ChrR25	Y128N, N154T	147 200 ± 44 900
ChrR26	Y128N, Q184H	25 000 ± 7200

Abbreviation: WT, wild type.
Predictions were based on the sequence–activity data of Table 1.

additional respects, namely, prodrug and U(VI) reduction. The capacity of the mutants to reduce prodrugs was determined by the efficiency with which they killed cells of the JC breast cancer cell line. Three prodrugs, namely, mitomycin C, CB 1954 and 17-AAG, were used. All three mutants were more potent than the wild-type enzyme in activating each of the drugs, and in causing the drug-mediated killing of the cells (Figure 1). This activity correlated, by and large, with improved chromate reductase activity for each mutant; ChrR30 being the most efficient in this respect.

The three mutants also exhibited improved uranyl reductase activity compared to the wild-type enzyme (Table 3). However, unlike chromate reductase activity, no further improvement in this activity was shown by the other mutants over ChrR21.

Table 2B Second round of second phase: sequence and V_{max} activity of six additional mutants chosen according to the Nov–Wein model

Mutants	*Amino-acid substitutions from WT ChrR*	V_{max} *(nmol Cr(VI) reduced per mg protein per min)*
ChrR27	Y128N, V120A, Q153H	19 200 ± 2060
ChrR28	Y128N, Q153H, N154T	43 300 ± 9300
ChrR29	Y128N, Q153H, G150S	22 500 ± 2400
ChrR30	Y128N, G150S, N154T	458 300 ± 83 600
ChrR31	Y128N, N154T, V120A	203 000 ± 80 200
ChrR32	Y128N, Q153H, A44V	18 600 ± 5600

Abbreviation: WT, wild type.
Predictions were based on the data of Tables 1 and 2A.

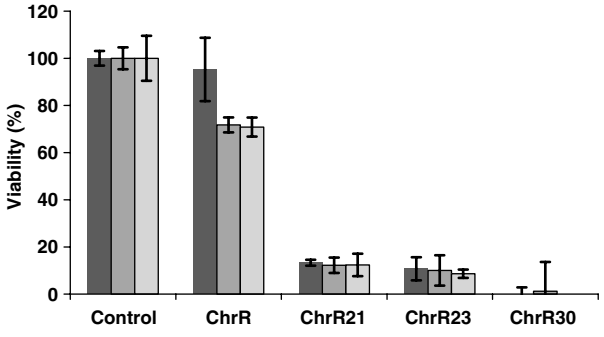

Figure 1 The effect of mitomycin C (■), CB 1954 (5-aziridinyl-2,4-dinitrobenzamide) (■) and 17-AAG (17-allylamino-17-demethoxygeldanamycin) (□) on the killing of JC breast cancer cells in the presence of the wild-type or the evolved enzymes (10 μg ml^{-1}). The concentration of the drugs was 15 μM. The enzymes were incubated with the drug for 30 min (37 °C), followed by the addition of the cells. After 24-h incubation (37 °C), cell viability was determined as described in the Materials and Methods section.

Discussion

Directed evolution has resulted in successful generation of many improved proteins, but this approach is blind, laborious and time consuming. Typically, the improvements achieved in the early rounds of directed evolution are significant, but in later rounds, even when a large number of further mutants is screened, improvements become smaller and less frequent. For example, Minagawa and Hiroki (2000) and Minagawa *et al.* (2007) were able to improve the thermostability of lactate oxidase by 18-fold after screening around 3000 mutants, but had to screen more than 20 000 additional mutants for a twofold further improvement. Since mutations are more often deleterious to protein activity, it has been thought that increased mutational rate was likely to correlate with loss of function, and 1–3 mutation rate per gene was considered desirable (Suzuki *et al.*, 1996; Arnold, 1998). Recently, however this notion has been questioned. Daugherty *et al.* (2000) and Drummond *et al.* (2005) have shown that higher mutation rate libraries (15–30 per gene) have a better probability of generating improved mutant proteins. The mutation rate employed in our experiment was low (1–5 per gene) and therefore the fourth round of directed evolution resulting with no improvement might be explained by 'masking' of deleterious mutations over beneficial ones.

While we could have obtained further improvement in ChrR activity using the directed evolution process by screening a large number of additional mutants (perhaps 20 000 or more) in later rounds, our use of the Nov and Wein model clearly afforded a significant saving in screening effort in these later stages. To provide a perspective: it was necessary to screen around 7000 mutants in four rounds of directed evolution (the last of which yielded no additional improvement) to obtain a 230-fold increase in ChrR activity; in contrast, the model made

Table 3 Uranyl reduction kinetics of selected evolved mutants

Mutants	V_{max} *(nmol U(VI) reduced per mg protein per min)*	K_m *(μM)*	K_{cat} *(s^{-1})*	K_{cat}/K_m *(M^{-1} s^{-1})*
ChrR	213 ± 17	108 ± 49	29 ± 11	$2.7 \times 10^5 \pm 2.3 \times 10^5$
ChrR21	6010 ± 226	228 ± 13	361 ± 24	$1.5 \times 10^6 \pm 1.6 \times 10^5$
ChrR23	4810 ± 462	221 ± 54	333 ± 37	$7 \times 10^5 \pm 7 \times 10^4$
ChrR30	5830 ± 502	237 ± 75	446 ± 32	$1.8 \times 10^6 \pm 1.5 \times 10^5$

it possible to improve the enzyme significantly further (>6-fold improvement over the best mutant obtained by directed evolution and >1500-fold improvement over the wild type) by screening only 11 targeted new mutants. This saving in screening is especially attractive when the screening cost is high compared to the cost of producing site-directed mutants, as required by the model.

Recently, Fox *et al*. (2007) elegantly demonstrated how a statistical model can augment directed evolution to significantly improve the cyanation activity of bacterial halohydrin dehalogenase. Although both studies employed additive statistical models coupled with traditional techniques, their results do not permit easy comparison, since (a) enzyme activity was measured differently in the two studies and (b) it is not known which of the two enzymes is more amenable to optimization. However, Fox *et al*. improved activity by ~4000-fold after screening more than 500 000 mutants in 18 rounds, while in the present work, we achieved ~1500-fold activity improvement after screening ~7000 variants in six rounds. Furthermore, as the structure of the halohydrin dehalogenase enzyme is known, some of the diversity in Fox *et al*. experiments was generated through rational design. In this work, no structural knowledge was used, as the structure of the ChrR enzyme is unknown. Both studies demonstrate the power of statistical modeling in protein design, and both permit beneficial use of information gained from mutants with reduced activity.

The genetic diversity spanned by the directed evolution mutants (Table 1) encompasses more than 3000 possible combinations, among which (after omitting ChrR17) 55 are double mutants and 165 are triple mutants. As diversity increases, the numbers grow exponentially: when one considers 15 mutated positions with two possible mutations in each, there are $>10^7$ possible combinations (420 double mutants, 3640 triple mutants); and with 20 mutated positions and three possible mutations in each, there are $>10^{12}$ possible combinations (1710 double mutants, >30 000 triple mutants). Thus, exhaustive search in a laboratory, even only for double and triple mutants, does not scale well, and systematically producing and screening all of them would be an extensive and highly laborious feat. The predictions of the model allow one to screen instead only a few targeted mutants, and still improve activity.

It is serendipitously possible to identify promising mutants by simply 'gazing' at activity data and detecting beneficial mutations, as was done in the discovery of the single-residue mutant ChrR21 in this work. However, a systematic mathematical approach is needed to identify more complex mutants, such as ChrR30. The model is shown here to be a valuable tool for such situations, as it allows one to rigorously separate the expected contribution to activity from each of the mutations in a data set of mutant activity (such as Table 1), and thus to isolate mutants that otherwise may have been difficult to discover.

All mutants designed in the second phase based on the model predictions are built from mutations generated through directed evolution in the first phase. Thus, in principle, it was possible to obtain the new mutants through additional rounds of directed evolution, without using the statistical model. However, as directed evolution is a blind process governed by chance, it is not clear screening of how many additional mutants would have been required to achieve an improvement comparable to that which the application of the model made possible; it should be kept in mind that the last round of directed evolution yielded no improvement.

The model postulates that mutations are approximately additive. Is this assumption supported by the data? Based on the second, more complete estimate of the parameters, the fraction of the total variance of the fitness of a double mutant that is due to non-additivity (and measurement noise) is $0.1961/(2\times0.4361+0.1961)=0.18$; for triple mutants, the fraction is $0.1961/(3\times0.4361+0.1961)=0.13$. These relatively low numbers indicate that the data are not particularly noisy, there are no strong epistatic effects and that the mutational effects are mostly additive. Two additional points regarding additivity are noteworthy. First, additivity is assumed to apply to the *transformed* activity measurements, rather than to the raw data. For example, ChrR31 is the combination of ChrR11 and ChrR25, and the deviation from perfect additivity in the raw data (202 778 vs $1000+147\,222$) is much greater than that in the transformed data (2.83 vs $0.53+2.70$). Second, as often happens in statistical analysis, even when approximate additivity holds, some considerable exceptions occur; this can be seen in our data when comparing ChrR10, whose transformed activity is 1.81, against ChrR11 combined with ChrR21, whose sum of transformed activities is 3.23.

Producing designed mutants through site-directed mutagenesis, as our approach required, is not always simple, as certain designed mutants are difficult to generate in a laboratory. A potential remedy for this problem is to create in the second phase, after statistically analyzing sequence–activity data from the first phase, combinatorial libraries containing only putatively beneficial mutations. These focused libraries will then be subject to directed evolution, and are more likely to achieve improvement than straightforward directed evolution libraries that do not incorporate statistical analysis in their design. This approach is pursued in a sequel to this work.

One might suggest that our statistical analysis could benefit from adopting a Bayesian approach, where prior distributions are set over the parameters. However, as this work is the first to study

enzyme activity data in light of the model, we could not use informative priors for Bayesian estimation. The proper choice of non-informative priors is under debate among statisticians, especially for parameters of the type appearing in our model, which are not constrained to lie in a known interval. We note, though, that when varying the value of m in our analysis, we effectively used a Bayesian-like approach with a non-informative prior for estimation and prediction.

Work is now in progress to use ChrR30 in improving bacterial bioremediation and prodrug cancer chemotherapy.

Acknowledgements

We are grateful to Drs Bruno Salles, Mike Benoit and Ms Mimi Keyhan for their useful advice and stimulating discussion. We thank Dr Stephen H Thorne for kindly supplying us with freshly made JC breast cancer cells. We also thank three anonymous referees whose insightful comments and suggestions greatly improved this article. This work was supported by Grants DE-FG03-97ER-624940 and DE-FG02-96ER20228 from the Natural and Accelerated Bioremediation Program of US Department of Energy, and Stanford Office of Technology Licensing (1105626-100-WOAAA). YB and DFA were supported, in part, by a Postdoctoral Fellowship from Lady Davis Postdoctoral Fellowship and FRST New Zealand (STAX0101) Fellowship, respectively.

References

Ackerley DF, Barak Y, Lynch SV, Curtin J, Matin A. (2006). Effect of chromate stress on *Escherichia coli* K12. *J Bacteriol* **188**: 3371–3381.

Ackerley DF, Gonzalez CF, Park CH, Blake R, Keyhan M, Matin A. (2004). Chromate reducing properties of soluble flavoproteins from *Pseudomonas putida* and *Escherichia coli*. *Appl Environ Microbiol* **70**: 873–882.

Aharoni A, Gaidukov L, Khersonsky O, Gould McQS, Roodveldt C, Tawfik DS. (2005). The 'evolvability' of promiscuous protein functions. *Nat Gen* **37**: 73–76.

Aita A, Husimi Y. (2000). Adaptive walks by the fittest among finite random mutants on a Mt. Fuji-type fitness landscape. *J Math Biol* **41**: 207–231.

Arnold FH. (1998). Enzyme engineering reaches the boiling point. *Proc Natl Acad Sci USA* **95**: 2035–2036.

Arnold FH. (2006). Fancy footwork in the sequence space shuffle. *Nat Biotech* **24**: 328–330.

Barak Y, Ackerley DF, Dodge CJ, Lal B, Cheng A, Francis AJ *et al.* (2006b). Analysis of novel soluble Cr(VI) and U(VI) reductases and generation of improved enzymes using directed evolution. *Appl Environ Microbiol* **72**: 7074–7082.

Barak Y, Thorne SH, Ackerley DF, Lynch SV, Contag CH, Matin A. (2006a). New enzyme for reductive cancer chemotherapy (YieF) and its improvement by directed evolution. *Mol Cancer Ther* **5**: 97–103.

Chatterjee R, Yuan L. (2006). Directed evolution of metabolic pathways. *Trends Biotech* **24**: 28–38.

Chen K, Arnold FH. (1993). Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of Subtilisin E for catalysis in dimethylformamide. *Proc Natl Acad Sci USA* **90**: 5618–5622.

Chica RA, Doucet N, Pelletier JN. (2005). Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr Opin Biotechnol* **6**: 378–384.

Daugherty PS, Chen G, Iverson BL, Georgiou G. (2000). Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc Natl Acad Sci USA* **97**: 2029–2034.

Dennett DC. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster Inc.: New York, NY.

Drummond DA, Iverson BL, Georgiou G, Arnold FH. (2005). Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J Mol Biol* **350**: 806–816.

Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK *et al.* (2007). Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotech* **25**: 338–344.

Grove JI, Lovering AL, Guise C, Race PR, Wrighton CJ, White SA *et al.* (2003). Generation of *Escherichia coli* nitroreductase mutants conferring improved cell sensitization to the prodrug CB1954. *Cancer Res* **63**: 5532–5537.

Kauffman SA, Levin S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**: 11–45.

Kuipers OP, Boot HJ, de-Vos WM. (1991). Improved site-directed mutagenesis method using PCR. *Nucleic Acids Res* **19**: 4558.

Lejon T, Strom MB, Svendsen JS. (2001). Antibiotic activity of pentadecapeptides modeled from amino acid descriptors. *J Pept Sci* **7**: 74–81.

Mee RP, Burton TR, Morgan PJ. (1997). Design of active analogues of a 15-residue peptide using D-optimal design, QSAR and a combinatorial search algorithm. *J Pept Res* **49**: 89–102.

Minagawa H, Hiroki K. (2000). Effect of double mutation on thermostability of lactate oxidase. *Biotechnol Lett* **22**: 1131–1133.

Minagawa H, Yoshida Y, Kenmochi N, Furuichi M, Shimada J, Kaneko H. (2007). Improving the thermal stability of lactate oxidase by directed evolution. *Cell Mol Life Sci* **64**: 77–81.

Nov Y, Wein LM. (2005). Modeling and analysis of protein design under resource constraints. *J Comput Biol* **12**: 247–282.

Park C-H, Keyhan M, Wielinga B, Fendorf S, Matin A. (2000). Purification to homogeneity and charcterization of a novel *Pseudomonas putida* chromate reductase. *Appl Environ Microbiol* **66**: 1788–1795.

Park H-S, Nam SH, Lee JK, Yoon CN, Mannervik B, Benkovic SJ *et al.* (2006). Design and evolution of new catalytic activity with an existing protein scaffold. *Science* **311**: 535–538.

Qian Z, Lutz SJ. (2005). Improving the catalytic activity of *Candida antarctica* lipase B by circular permutation. *Am Chem Soc* **127**: 13466–13467.

Sambrook J, Fritsch EF, Maniatis T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbour Laboratory Press: Cold Spring Harbor, NY.

Stemmer WP. (1994). DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for

molecular evolution. *Proc Natl Acad Sci USA* **91**: 10747–10751.

Suzuki FC, Christians B, Kim A, Skandalis MEB, Loeb LA. (1996). Tolerance of different proteins for amino acid diversity. *Mol Divers* **2**: 111–118.

Teixeira LSG, Costa ACS, Ferreira SLCM, Freitas LM, Carvalho S. (1999). Spectrophotometric determination of uranium using 2-(2-thiazolylazo)-*p*-cresol (TAC) in the presence of surfactants. *J Braz Chem Soc* **10**: 519–522.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)