Optimal Scanning of All Single-Point Mutants of a Protein

YUVAL NOV,¹ ALEXANDER FULTON,² and KARL-ERICH JAEGER²

ABSTRACT

In protein engineering, useful information may be gained from systematically generating and screening all single-point mutants of a given protein. We model and analyze an iterative two-stage procedure to generate all these mutants. At each position, L variants are generated in the first stage via saturation mutagenesis and are sequenced. In the second stage, the missing variants (out of the 19 possible single-point substitutions) are produced via site-directed mutagenesis. We study the economic tradeoff associated with varying L , and derive its optimal value given the experimental parameters.

Key words: complete site-saturation mutagenesis; optimal scanning; protein engineering; saturation mutagenesis.

1. INTRODUCTION

TO STUDY THE ACTIVITY, STRUCTURE, OR STABILITY of a protein, scientists often explore libraries that systematically cover single-point mutations at each of the protein's residues. A widely used scanning approach is alanine scanning, whereby each of the protein's wild-type residues is replaced in its turn by alanine, and the resulting variants are screened for the property of interest via an appropriate assay (Morrison and Weiss, 2001). Even more detailed information can be gained from studying all 19 substitutions, rather than only one, at each position of the protein. Such so-called complete site-saturation mutagenesis studies have previously been performed to isolate enantioselective enzymes, for example, a nitrilase (DeSantis et al., 2003) and a lipase (Eggert et al., 2011). A complete site-saturation library of the latter enzyme, the 181-residue-long *Bacillus subtilis* lipase A, is currently analyzed in our laboratory for the contribution of each amino acid residue to a variety of different biochemical properties. Further examples of using this scanning approach toward a sizable subset of a protein's residues include Wu et al. (1998), Chen et al. (1999), and Boakes et al. (2012). The constantly decreasing cost of mutagenesis, sequencing, and screening, as well as the rising of laboratory automation, make this approach increasingly attractive.

One way to generate all single-point mutants of a protein is by a massive use of site-directed mutagenesis, in which each mutant is generated individually in a separate reaction. However, the total cost of such a large number of reactions is currently still prohibitive. A more economical alternative is an iterative two-stage procedure, which is to be repeated at each position: (1) saturation mutagenesis is applied to

¹Department of Statistics, University of Haifa, Haifa, Israel.

²Institute of Molecular Enzyme Technology, Heinrich-Heine-University Duesseldorf and Research Centre Juelich, Juelich, Germany.

construct L variants, which are all sequenced and observed for which of the 19 substitutions are present; this stage will be referred to as *the randomization stage*. (2) Site-directed mutagenesis is used to construct individually all the missing substitutions; this stage will be referred to as *the site-directed mutagenesis stage*.

The main decision variable in this process is L , the number of variants generated and sequenced in the randomization stage. Small L means lower sequencing cost in the randomization stage, yet large L means that more distinct sequences (out of the possible 19) will be generated in the randomization stage, and hence fewer variants will have to be individually generated in the site-directed mutagenesis stage, lowering the cost of this stage. The goal of this work is to establish a methodology for balancing these two conflicting requirements, and thus to optimize the process. Given the various experimental parameters (costs, yield, protein sequence, randomization scheme), one needs to choose optimally L so as to minimize the expected overall cost of the experiment. We emphasize that this work is concerned only with generating the mutant library, and not with the subsequent screening.

2. METHODS

Let M be the number of positions to be scanned. These positions may be all of the protein's positions (as in the lipase A experiment, where $M = 181$), or some predetermined subset thereof. We denote by $A = \{\text{Ala}, \text{Arg}, \dots, \text{Val}\}$ the set of all 20 amino acids. Let c_{primer} be the cost of the required amount of a *single primer* for a single reaction, c_{fixed} be the sum of all other fixed costs associated with a single reaction (mainly labor, dNTPs, polymerase, buffer, template DNA, DpnI, transformation), and c_{seq} be the cost of sequencing a *single variant*. We do not lump together c_{primer} and c_{fixed} , although they determine together the cost of single reaction, because different randomization schemes require different numbers of primers per reaction (see below).

We first focus on the cost associated with a single position out of the M to be scanned. This cost depends on the wild-type amino acid at that position, so for position i , we denote it by $C_i(L)$. We assume that the QuikChange protocol is used, so that the per-position cost of the randomization stage is $2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}$ (recall that the QuikChange protocol requires two primers—one forward and one backward). The cost of the site-directed mutagenesis stage is $K_{i,L}(2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})$, where $K_{i,L}$ is the random variable denoting the number of missing variants that need to be produced individually at position i in the site-directed mutagenesis stage, given that L variants were generated in the randomization stage. The distribution of $K_{i,L}$ depends also on the randomization scheme (see below), but we suppress this dependence in the notation. The total cost of scanning position i is therefore

$$C_i(L) = 2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}} + K_{i,L}(2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})$$

The overall cost of the experiment is then

$$C(L) = \sum_{i=1}^M C_i(L) = M(2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}) + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{i=1}^M K_{i,L} \quad (1)$$

Let n_a be the number of positions in which the wild-type amino acid is a (so that $\sum_{a \in A} n_a = M$), and, with a slight abuse of notation, let $K_{a,L}$ be the random variable denoting the number of missing variants to be produced in the site-directed mutagenesis stage at a position having wild-type amino acid a , given that L variants were generated in the randomization stage. To minimize the overall expected cost of the experiment, one needs to solve the following optimization problem:

$$\min_{L \geq 0} E[C(L)] = \min_{L \geq 0} M(2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}) + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{a \in A} n_a E(K_{a,L}) \quad (2)$$

To analyze this optimization problem, we turn to study the distribution of $K_{a,L}$. Out of the L variants generated in the randomization stage at some given position, let X_a be the number of variants that resulted in amino acid a , and let X_{stp} be the number of variants which resulted in a premature stop codon (so that $\sum_{a \in A} X_a + X_{\text{stp}} = L$). The distribution of X_a is binomial with parameters L and p_a , where p_a is the probability that the randomization in the said position will result in amino acid a (for example, under the popular NNK randomization scheme, since 2 out of the 32 possible codons encode alanine, we have $p_{\text{Ala}} = 2/32$). The

joint distribution of the 21-dimensional random vector $(X_{\text{Ala}}, \dots, X_{\text{Val}}, X_{\text{stp}})$ is multinomial with parameters L and $p_{\text{Ala}}, \dots, p_{\text{Val}}, p_{\text{stp}}$.

Proposition 1. $E(K_{a,L}) = \sum_{b \neq a} (1-p_b)^L$

Proof. Using the notation I_E for the indicator function of the event E , we have that $K_{a,L} = \sum_{b \neq a} I_{\{X_b=0\}}$. Therefore,

$$E(K_{a,L}) = \sum_{b \neq a} E(I_{\{X_b=0\}}) = \sum_{b \neq a} P(X_b=0) = \sum_{b \neq a} (1-p_b)^L. \quad \blacksquare$$

Proposition 2. $\text{Var}(K_{a,L}) = \sum_{b \neq a} (1-p_b)^L + \sum_{b \neq a} \sum_{c \neq a, b} (1-p_b-p_c)^L - \left(\sum_{b \neq a} (1-p_b)^L \right)^2$.

Proof. We first derive the second moment of $K_{a,L}$:

$$\begin{aligned} E(K_{a,L}^2) &= E \left[\left(\sum_{b \neq a} I_{\{X_b=0\}} \right)^2 \right] \\ &= E \left(\sum_{b \neq a} I_{\{X_b=0\}} + \sum_{b \neq a} \sum_{c \neq a, b} I_{\{X_b=0\}} I_{\{X_c=0\}} \right) \\ &= \sum_{b \neq a} P(X_b=0) + \sum_{b \neq a} \sum_{c \neq a, b} P(X_b=0, X_c=0) \\ &= \sum_{b \neq a} (1-p_b)^L + \sum_{b \neq a} \sum_{c \neq a, b} (1-p_b-p_c)^L. \end{aligned}$$

The result now follows from the elementary identity $\text{Var}(X) = E(X^2) - E^2(X)$ and from proposition 1. $\blacksquare$

More generally, it can be shown that the n th moment of $K_{a,L}$ is

$$E(K_{a,L}^n) = \sum_{\substack{B \subseteq A \setminus \{a\} \\ |B| \leq n}} S(n, |B|) (1-p_B)^L,$$

where $p_B = \sum_{b \in B} p_b$, and $S(n, |B|) = \sum_{k=0}^{|B|} (-1)^{|B|-k} \binom{|B|}{k} k^n$ is the number of surjective functions from a set having n elements onto a set having $|B|$ elements (this number is closely related to the Stirling numbers of the second kind; see Comtet, 1974).

Using proposition 1, the optimization problem in Equation (2) can be now written more explicitly as

$$\min_{L \geq 0} E[C(L)] = \min_{L \geq 0} M(2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}) + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{a \in A} n_a \sum_{b \neq a} (1-p_b)^L \quad (3)$$

From proposition 2, the variance of the total cost of the experiment, $C(L)$, is given by

$$\text{Var}[C(L)] = (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})^2 \sum_{a \in A} n_a \left[\sum_{b \neq a} (1-p_b)^L + \sum_{b \neq a} \sum_{c \neq a, b} (1-p_b-p_c)^L - \left(\sum_{b \neq a} (1-p_b)^L \right)^2 \right]. \quad (4)$$

Proposition 3. *The objective function in equation (3) is convex in L for $L \geq 0$.*

Proof. The objective function is a sum of a linear term and positively weighted exponential terms. All are convex functions, and since the sum of convex functions is convex, the proposition is proved. In more detail, the first derivative of the objective function in Equation (3) is

$$\begin{aligned} \frac{d}{dL} E[C(L)] &= Mc_{\text{seq}} + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{a \in A} n_a \frac{d}{dL} \sum_{b \neq a} (1-p_b)^L \\ &= Mc_{\text{seq}} + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{a \in A} n_a \sum_{b \neq a} (1-p_b)^L \log(1-p_b). \end{aligned}$$

The second derivative is

$$\frac{d^2}{dL^2}E[C(L)] = (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{a \in A} n_a \sum_{b \neq a} (1 - p_b)^L \log^2(1 - p_b).$$

which is positive for each $L \geq 0$, and hence $E[C(L)]$ is convex in L . ■

When viewed as a function of a continuous variable, as we did in the analysis so far, the objective function therefore possesses a unique optimal minimizer L^* , which is either the boundary point $L = 0$, or the solution of the equation $(d/dL)E[C(L)] = 0$. Because of the structure of the objective function $E[C(L)]$, one cannot solve the latter equation analytically. However, numerical methods for finding L^* such as the Newton method are not required, since in practice L is an integer, so one can use, for example, simple binary search across a large enough set of consecutive integers, to find the integer minimizer L^* . Note that, in principle, it may happen that there are two integer minimizers, which will be two consecutive integers flanking the real-valued minimizer of the objective function, when viewed as a continuous function of L .

2.1 Other randomization schemes

The cost function used in Equation (1) corresponds to standard QuikChange randomization, in which a single pair of degenerate primers is used. This randomization approach encompasses several widely used schemes, known as NNN, NNK, NNS, and NNB (Chronopoulou and Labrou, 2011). Recently, two other randomization schemes have been proposed: Kille et al. (2013) presented their so-called “22c-trick,” which uses a mixture of three primer pairs—NDT, VHG, and TGG, at 12:9:1 molar ratio. This randomization scheme is more efficient than each of the four abovementioned single-pair schemes, as it requires smaller libraries to attain the same coverage of sequence space. Similarly, the method of Tang et al. (2012), which uses a mixture of four primer pairs—NDT, VMA, ATG, and TGG, at 12:6:1:1 molar ratio—is even more efficient, at the expense of requiring an additional primer pair. To incorporate either 22c-trick randomization or Tang randomization into the analysis, all that needs to be done is to change the coefficient of the parameter c_{primer} in Equations (3) and (4) from 2 to either 6 or 8, respectively, and to update the probabilities p_a , $a \in A$ (in the 22c-trick randomization, $p_{\text{Leu}} = p_{\text{Val}} = 2/22$ and $p_a = 1/22$ for $a \neq \text{Leu, Val}$; in Tang randomization, $p_a = 1/20$ for all $a \in A$).

2.2 Yield

The analysis so far assumed that both the randomization and the site-directed mutagenesis are always successful. In practice this is not always the case, and one determines the *yield* of the experiment as the probability that the intended mutation is inserted successfully into the gene of interest. Let $\alpha \in (0, 1]$ denote the yield of the experiment.

The yield influences the above analysis at both stages of the process. At the randomization stage, the probability p_b that the randomization will result in amino acid b is lowered by a factor α . At the site-directed mutagenesis stage, when generating each of the $K_{i,L}$ missing variants, there is a probability $1 - \alpha$ that sequencing will reveal that the cloning failed, so the mutagenesis will have to be repeated until it is performed successfully. The total cost in Equation (1) therefore changes to

$$C(L) = M(2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}) + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \sum_{i=1}^M \sum_{j=1}^{K_{i,L}} N_{i,j},$$

where $N_{i,j}$ is the number of cloning trials of variant j (out of the $K_{i,L}$ missing ones) until successful cloning has occurred, in the site-directed mutagenesis stage of position i . From this definition, $N_{i,j}$ is geometrically distributed, with probability of success α .

Using Wald’s equation for the mean of a random sum, the objective function of Equation (3) changes to

$$E[C(L)] = M(2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}) + (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}}) \frac{1}{\alpha} \sum_{a \in A} n_a \sum_{b \neq a} (1 - \alpha p_b)^L.$$

The optimal L^* can be still easily found using the approach described above.

Using the well known equation for the variance of a random sum, the variance of the total cost (Equation (4)) changes to

$$\begin{aligned}\text{Var}[C(L)] &= (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})^2 \sum_{i=1}^M \text{Var} \left(\sum_{j=1}^{K_{i,L}} N_{i,j} \right) \\ &= (2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})^2 \sum_{a \in A} n_a [\text{Var}(N_{i,j}) E(K_{a,L}) + E^2(N_{i,j}) \text{Var}(K_{a,L})],\end{aligned}$$

where $E(N_{i,j}) = 1/\alpha$ and $\text{Var}(N_{i,j}) = (1-\alpha)/\alpha^2$ (the mean and variance of a geometric distribution), and

$$\begin{aligned}E(K_{a,L}) &= \sum_{b \neq a} (1 - \alpha p_b)^L \\ \text{Var}(K_{a,L}) &= \sum_{b \neq a} (1 - \alpha p_b)^L + \sum_{b \neq a} \sum_{c \neq a, b} (1 - \alpha p_b - \alpha p_c)^L - \left(\sum_{b \neq a} (1 - \alpha p_b)^L \right)^2.\end{aligned}$$

3. RESULTS

In this section, we present the practical consequences of the mathematical results derived in the Methods section.

The left panel of Figure 1 shows the expected overall cost of the experiment as a function of L , the number of variants generated and sequenced in the randomization stage. The black curve corresponds to the ideal case of 100% yield, and the red curve to 80% yield. In both cases it was assumed that randomization was done using NNK degenerate oligonucleotides, and that the cost parameters are $c_{\text{primer}} = 8$, $c_{\text{fixed}} = 2$, and $c_{\text{seq}} = 2.8$ (these are realistic figures in Euros, at the time of writing this article).

It can be seen that the optimal L in the case of 100% yield is $L^* = 39$, which correspond to an optimal expected overall cost of $E[C(L^*)] = 37,272$. The cost curve varies considerably near its minimum: Sequencing $L = 96$ variants in the randomization stage (an entire microplate) would result in an expected overall cost of 54,012, an increase of 45%. As expected, the optimal overall expected cost increases when the yield drops to 80%, to 45,877, which is attained with $L^* = 49$.

Over the past 2 decades, technological advances have kept lowering biotechnological costs considerably, and are expected to continue doing so. The right panel of Figure 1 shows how the results change when the sequencing cost c_{seq} drops from 2.8 to 1.5. With 100% yield, the optimal value of L is $L^* = 54$, which

FIG. 1. Expected overall cost as a function of L , the number of variants generated and sequenced in the randomization stage. *Black curve* corresponds to 100% yield and *red* to 80%. *Left:* $c_{\text{primer}} = 8$, $c_{\text{fixed}} = 2$, and $c_{\text{seq}} = 2.8$. *Right:* c_{seq} lowered to 1.5. All curves refer to NNK randomization and to a protein with $M = 181$ randomized positions, as in the *Bacillus subtilis* lipase A experiment.

FIG. 2. *Left:* The optimal expected overall cost, $E[C(L)]$, as a function of the yield. *Right:* The optimal number of variants generated and sequenced in the randomization stage, L^* , as a function of the yield.

correspond to an optimal expected overall cost of 25,814. Under 80% yield, the figures change to $L^* = 68$ and $E[C(L^*)] = 31,522$.

The influence of the yield on the expected overall cost can be seen in more detail in Figure 2. The left panel of the figure depicts the optimal value of the cost function (the minimal point of the cost curve in Figure 1) as a function of the yield. As expected, the resulting curve tends to infinity as the yield approaches zero. The influence of decreasing the yield on L^* is *a priori* less clear, as lower yield affects adversely both stages of the scanning process. The right panel of Figure 2 shows that not only the optimal overall cost, but also L^* , increases as the yield tends to 0.

Although the distribution of $C(L)$ is certainly not normal, it may be well approximated by a normal distribution having the mean and variance of $C(L)$ (Equations (3) and (4) in the Methods section), as shown in Figure 3. The central limit theorem explains the good fit when M , the number of scanned positions, is large ($M = 181$ in the left panel), but the fit remains good also when only few positions are scanned ($M = 10$ in the right panel).

FIG. 3. Histogram of 10,000 simulated values of $C(L)$, the total cost of the experiment, with a density plot of a normal distribution having the mean and variance of $C(L)$. *Left:* $M = 181$, as in the lipase A study. *Right:* $M = 10$. In both cases $L = 40$, and other parameters are as in the left panel of Figure 1.

FIG. 4. Expected overall cost as a function of L , the number of variants generated and sequenced in the randomization stage, for three randomization schemes. *Left:* 100% yield. *Right:* 80% yield. Other parameters are as in the left panel of Figure 1.

All the results presented so far assumed that NNK randomization was used. Figure 4 shows the expected overall cost as a function of L for two additional randomization schemes: the “22c-trick,” and the method of Tang et al. (2012). It can be seen that in both methods, the savings due to the improved efficiency do not compensate for the increase in primer cost.

4. DISCUSSION

The goal of this work is to establish a methodology for determining the optimal resource allocation, when scanning all single-point mutants of a protein. Only a small adjustment is needed if one wishes to scan not all 19 substitutions at each position, but rather, some other predetermined number. One can analyze similarly also a process whereby all two-mutant variants of a protein are to be systematically generated.

The cost function we employ is rather simplistic, as it ignores, for example, the cost of the nonexpendable laboratory equipment. Still, it captures the most essential components of the situation: the tradeoff between increased randomization cost and increased site-directed mutagenesis cost, as a result of varying L , the number of variants generated and sequenced in the randomization stage. Neglecting the integrality constraint, the optimal L is the solution of the equation obtained by equating the derivative of the objective function to zero. This solution lends itself to an intuitive interpretation: When increasing L by a unit, the per-position increase in the randomization cost is the derivative of the expression $2c_{\text{primer}} + c_{\text{fixed}} + Lc_{\text{seq}}$, which is exactly c_{seq} , as one more variant will have to be sequenced. At the same time, the expected decrease in the site-directed mutagenesis cost (a decrease that takes place as fewer variants will have to be generated individually) is the derivative of $(2c_{\text{primer}} + c_{\text{fixed}} + c_{\text{seq}})E(K_{a,L})$. The optimal L is indeed the value that makes these two forces equal to each other, and is therefore found by equating to zero the derivative of their sum, the total expected cost.

The original problem formulation assumed that the same L will be used in the randomization stage for all of the protein’s position. However, since each position is dealt with separately, and since the per-position cost $C_i(L)$ depends on the specific wild-type amino acid at that position, one could in principle find and use an optimal L for each position. By doing so the overall expected cost will indeed decrease, but it can be shown mathematically that the decrease will be of only about 1%, a small difference that will most likely get buried in the experimental noise.

Our analysis shows that the yield has a substantial influence on both the optimal cost and on the optimal L required to attain it. We recommend estimating the yield in a small preliminary experiment, and then use the estimate for determining the optimal L in the full-scale experiment.

In the process studied in this work, the genetic diversity created in the randomization stage was generated via saturation mutagenesis. Other alternatives are random mutagenesis techniques such as error-prone PCR

(Cadwell and Joyce, 1992) and SeSaM (Wong et al., 2004), which have been used successfully in many mutagenesis experiments. However, upon analyzing mathematically these alternatives, it turns out that even under ideal conditions (zero probability for premature stop codons and complete control over the mutation spectrum), the resulting expected overall cost will be considerably higher than that obtained via saturation mutagenesis. The reason for this is that random mutagenesis protocols are whole-protein methods, which invariably produce a large portion of multiple-point mutants, whereas the experimental goal discussed in our work is to generate systematically only single-point mutants. The inevitable additional cost, incurred by generating and sequencing these multiple-point mutants, makes random mutagenesis methods less suitable for our specific goal. Another alternative for creating the desired library would be via a massive use of gene synthesis; however, at the current synthesis prices, this approach will be about an order of magnitude more expensive than the two-stage process studied above.

ACKNOWLEDGMENTS

We thank Ulrich Krauss, group “Molecular Biophotonics,” Institute of Molecular Enzyme Technology, for advice and discussions. AF and KEJ gratefully acknowledge support by the German Research Foundation (DFG) within research training group 1166 “Biocatalysis using Non-Conventional Media—BioNoCo”.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Boakes S, Ayala T, Herman J, et al. 2012. Generation of an actagardine A variant library through saturation mutagenesis. *Appl Microbiol Biotechnol* 95, 1509–1517.
- Cadwell RC, and Joyce GF. 1992. Randomization of genes by PCR mutagenesis. *Genome Res* 2, 28–33.
- Chen G, Dubrawsky I, Mendez P, et al. 1999. In vitro scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site. *Protein Engin* 12, 349–356.
- Chronopoulou EG, and Labrou NE. 2011. Site-saturation mutagenesis: A powerful tool for structure-based design of combinatorial mutation libraries. *Curr Protoc Protein Sci* 26, 26.6.
- Comtet L. 1974. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. D. Reidel Publishing Company.
- DeSantis G, Wong K, Farwell B, et al. 2003. Creation of a productive, highly enantioselective nitrilase through gene site saturation mutagenesis (GSSM). *J Am Chem Soc* 125, 11476–11477.
- Eggert T, Funke SA, Andexer JN, et al. 2011. Evolution of enantioselective *Bacillus subtilis* lipase. *Protein Engineering Handbook, Volume 1 & Volume 2*, pp. 441–451.
- Kille S, Acevedo-Rocha C, Parra L, et al. 2013. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synthet Biol* 2, 83–92.
- Morrison KL, and Weiss GA. 2001. Combinatorial alanine-scanning. *Curr Opin Chem Biol* 5, 302–307.
- Tang L, Gao H, Zhu X, et al. 2012. Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques* 52, 149–158.
- Wong TS, Tee KL, Hauer B, et al. 2004. Sequence saturation mutagenesis (SeSaM): A novel method for directed evolution. *Nuc Acids Res* 32, e26–e26.
- Wu H, Beuerlein G, Nie Y, et al. 1998. Stepwise in vitro affinity maturation of Vitaxin, an $\alpha_v\beta_3$ -specific humanized mAb. *Proc Natl Acad Sci USA* 95, 6037–6042.

Address correspondence to:

Dr. Yuval Nov
Department of Statistics
University of Haifa
Mount Carmel
Haifa 31905
Israel

E-mail: yuval@stat.haifa.ac.il