doi:10.1016/j.jmb.2005.05.023 J. Mol. Biol. (2005) 350, 806–816

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Why High-error-rate Random Mutagenesis Libraries are Enriched in Functional and Improved Proteins

D. Allan Drummond[1], Brent L. Iverson[2], George Georgiou[3] and Frances H. Arnold[4]*

[1]*Program in Computation and Neural Systems, California Institute of Technology Mail Code 210-41, Pasadena CA 91125-4100, USA*

[2]*Department of Chemistry and Biochemistry, The University of Texas at Austin, 1 University Station A5300, Austin, TX 78712, USA*

[3]*Department of Chemical Engineering, The University of Texas at Austin, 1 University Station A5300, Austin, TX 78712, USA*

[4]*Division of Chemistry and Chemical Engineering California Institute of Technology, Mail Code 210-41 Pasadena, CA 91125-4100 USA*

**Corresponding author*

The fraction of proteins that retain wild-type function after mutation has long been observed to decline exponentially as the average number of mutations per gene increases. Recently, several groups have used error-prone polymerase chain reactions (PCR) to generate libraries with 15 to 30 mutations per gene, on average, and have reported that orders of magnitude more proteins retain function than would be expected from the low-mutation-rate trend. Proteins with improved or novel function were isolated disproportionately from these high-error-rate libraries, leading to claims that high mutation rates unlock regions of sequence space that are enriched in positively coupled mutations. Here, we show experimentally that error-prone PCR produces a broader non-Poisson distribution of mutations consistent with a detailed model of PCR. As error rates increase, this distribution leads directly to the observed excesses in functional clones. We then show that while very low mutation rates result in many functional sequences, only a small number are unique. By contrast, very high mutation rates produce mostly unique sequences, but few retain function. Thus an optimal mutation rate exists that balances uniqueness and retention of function. Overall, high-error-rate mutagenesis libraries are enriched in improved sequences because they contain more unique, functional clones. Our findings demonstrate how optimal error-prone PCR mutation rates may be calculated, and indicate that "optimal" rates depend on both the protein and the mutagenesis protocol.

Keywords: mutagenesis; PCR; optimal mutation rate; Poisson distribution; neutrality

Introduction

Laboratory evolution has been used to improve protein properties by mimicking natural evolution's stepwise exploration of sequence space[1], steadily improving protein activity or thermostability through repeated rounds of low-frequency mutation and selection. Because the fraction of proteins retaining function appears to decline exponentially with increasing numbers of amino acid substitutions,[2–5] low mutation rates seek to create mutational diversity without destroying activity so that improved clones can be found.[6]

Recently, several groups reported construction of mutant libraries using high-mutation-rate error-prone polymerase chain reactions (EP-PCR) to probe distant regions of sequence space for an antibody fragment (up to an average $\langle m_{nt} \rangle = 22.5$ nucleotide mutations per gene),[3,7] hen egg lysozyme (up to $\langle m_{nt} \rangle = 15.25$),[8] and TEM-1 β-lactamase (up to $\langle m_{nt} \rangle = 27.2$).[9] Where both high and low error rates were assessed, the exponential trend in loss of function established for low $\langle m_{nt} \rangle$ was violated spectacularly at the highest rates, with orders of magnitude more functional clones isolated than would be expected.[3,7,8] Two studies reported improved or novel function more often in these high-mutation-rate libraries,[3,9] leading to suggestions that low mutational pressure may not be optimal,[3,9] and that hypermutagenesis can, without an exponentially increasing cost in inactivated sequences, explore multiple interacting mutations

Abbreviations used: EP-PCR, error-prone polymerase chain reactions; scFv, single-chain Fv.

E-mail address of the corresponding author: frances@cheme.caltech.edu

inaccessible to low-error-rate mutagenesis.[9] These putative interactions could involve synergistic interactions to increase function directly, or combinations in which one or a few mutations increase function at the cost of folding or structural stability, the negative effects of which are suppressed by additional compensatory stabilizing mutations elsewhere in the protein.

The degree to which mutations interact, and thus mutational effects deviate from independence, is known as epistasis. Independent mutational effects imply an exponential decline in fraction functional with mutational distance, so the results of the studies mentioned above suggest that mutations interact epistatically, on average. Such a finding is of fundamental interest in evolutionary biology,[10,11] and is potentially decisive in answering the major open question "Why is there sex?"[12] Moreover, the discovery of reservoirs of positively interacting mutations would fundamentally change strategies for *in vitro* enzyme engineering by evolutionary methods.[9] Therefore, a careful analysis of these results is imperative.

Quantitative analysis of high-frequency mutagenesis results often assumes a Poisson distribution of mutations in EP-PCR, an idea introduced by Shafikhani *et al.*[4] This group's careful study on *Bacillus lentus* subtilisin found an accurately reproducible exponential decline in fraction functional in all libraries where functional proteins were found, up to $\langle m_{nt}\rangle = 15$, contrary to the upward trend reported later.

To examine the mutational distribution generated by high-error-rate EP-PCR, we constructed two large libraries of single-chain Fv (scFv) antibody mutants. The wild-type scFv antibody fragment derived from the 26-10 monoclonal antibody[13] binds digoxigenin with high affinity, and has been expressed as a fusion to the *Escherichia coli* outer membrane protein Lpp-OmpA′, allowing detection of mutants binding fluorescent dye-conjugated digoxigenin by fluorescence-activated cell sorting (FACS).[3] Libraries were assayed for mutant retention of wild-type affinity for digoxigenin (briefly, retention of function). These libraries were constructed and assayed exactly as in a previous study by two of the present authors,[3] making the results of both studies directly comparable. We were able to determine how the mutational statistics relate to PCR experimental parameters and to retention of function.

We show that mutations introduced by EP-PCR at high error rates do not follow the Poisson distribution, but rather a previously proposed distribution derived from a model of the actual PCR process.[14] We derive the expected fraction of functional mutants based on this more realistic model, and show that many reported experimental mutation data follow the predictions of this model. We then introduce a simple measure of optimality to evaluate optimal mutation rates for improvement of protein function. Our results show that the trends observed in earlier studies do not constitute evidence for positive epistasis.

Results

Distribution of mutations generated by EP-PCR

The probability $Pr(\mathrm{f})$ that an EP-PCR-amplified sequence retains function can be obtained as follows. Sun modeled EP-PCR by assuming n thermal cycles during which DNA strands are duplicated with probability λ, the PCR efficiency (assumed constant, realistic for large amounts of starting template[15,16]), resulting in $d = n\lambda$ DNA doublings and an average of $\langle m_{nt}\rangle$ nucleotide mutations per sequence.[14] The mutational distribution under these assumptions can be written,[14] with $x = (\langle m_{nt}\rangle(1+\lambda))/(n\lambda)$, as:

$$Pr(m_{nt}) = (1+\lambda)^{-n}\sum_{k=0}^{n}\binom{n}{k}\lambda^{k}\frac{(kx)^{m_{nt}}e^{-kx}}{m_{nt}!}, \quad (1)$$

which has mean $\langle m_{nt}\rangle$ and variance:

$$\sigma^2_{m_{nt}} = \langle m_{nt}\rangle + \frac{\langle m_{nt}\rangle^2}{n\lambda} = \langle m_{nt}\rangle\left(1+\frac{\langle m_{nt}\rangle}{d}\right)$$

At large $\langle m_{nt}\rangle$, small n or low λ, all of which broaden the variance, deviation from the Poisson assumption that the variance is equal to the mean $\langle m_{nt}\rangle$ can be profound. We call equation (1) the PCR distribution.

Results of mutagenesis

To examine the mutational distribution generated by high-error-rate EP-PCR, for which the Poisson-based and PCR-based models make distinct predictions, we generated two libraries (A and B) of scFv antibody clones using similar mutagenic conditions. We assayed both libraries for retention of wild-type-like binding to digoxigenin (retention of function) and sequenced 45+ naïve clones from each library.

Poisson-distributed mutations will have equal mean and variance, while PCR-distributed mutations will always have a variance larger than the mean. Figure 1 shows the distribution of nucleotide mutations observed in library A (46 sequences) and library B (45 sequences); summary statistics are shown in Table 1, and mutational spectra are reported in Table 2.

While visual inspection of the mutation histograms overlaid with the theoretical distributions cannot distinguish between the two models, the relevant statistics are stark and favor the PCR distribution while rejecting the Poisson distribution. For library A, $\langle m_{nt}\rangle = 15.8$ and $\sigma^2_{m_{nt}} = 26.3$; for library B, $\langle m_{nt}\rangle = 19.8$ and $\sigma^2_{m_{nt}} = 36.1$ (Table 1). The probability of measuring variances at least this large given an underlying Poisson distribution with the observed mean is $P < 0.005$ for library A and $P < 0.001$ for library B; the joint probability of

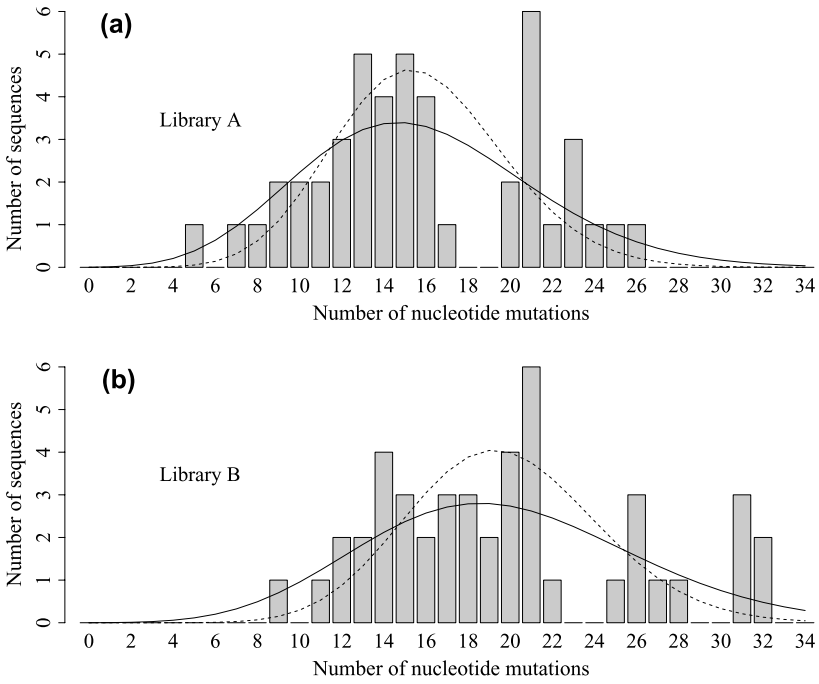

Figure 1. Mutational distributions for two high-error-rate scFv antibody libraries compared with Poisson and PCR distributions. (a) Library A, 46 sequences. (b) Library B, 45 sequences. The corresponding PCR distributions with the same means (see Table 1) (continuous line, $n=30$ cycles and efficiency $\lambda=0.6$) and Poisson distribution (broken line) are shown for comparison. For these histograms, the Poisson distribution may be rejected in favor of the PCR distribution (see the text).

observing two libraries with variances this high is $P<10^{-5}$. With a PCR efficiency of $\lambda=0.6$ (18 doublings), the PCR distribution yields expected variances of 29.6 (library A) and 41.4 (library B), consistent with the observed values.

Using a likelihood ratio test on the mutational samples (see Materials and Methods), we reject the Poisson distribution in favor of the PCR distribution with two additional degrees of freedom (n and λ) for library A ($\chi^2=7.39$, $P<0.025$) and for library B ($\chi^2=8.63$, $P<0.025$). (Using two additional degrees of freedom is conservative, since n is fixed in each experiment.) Thus, the PCR distribution (equation (1)) better describes the data than the previously assumed Poisson model.

Retention of protein function after mutation

What is the effect of the non-Poisson mutational distribution on the fraction of clones in a library that retain function? We assume the probability an individual protein will retain function after m_{aa} amino acid substitutions declines exponentially according to $Pr(f|m_{aa})=\nu^{m_{aa}}$, where ν can be

Table 1. scFv antibody mutational results and corresponding predictions for PCR and Poisson-distributed mutations

Library	Sequenced	$\langle m_{nt}\rangle$	$\sigma^2_{m_{nt}}$ ($P(\sigma^2_{m_{nt}})$ if Poisson)	PCR $\sigma^2_{m_{nt}}$ [a]	Poisson $\sigma^2_{m_{nt}}$
A	46	15.8 ± 0.8	26.3 ($P<0.005$)	29.6	15.8
B	45	19.8 ± 0.9	36.1 ($P<0.001$)	41.4	19.8

[a] Assumed efficiency $\lambda=0.6$ (18 DNA doublings).

Table 2. Mutational spectra for libraries

	Library A (33,396 bp sequenced)		Library B (32,670 bp sequenced)	
Type	Number	Fraction	Number	Fraction
A→T, T→A	172	0.24	106	0.12
A→C, T→G	7	0.01	7	0.01
A→G, T→C	336	0.46	202	0.23
G→A, C→T	188	0.26	529	0.60
G→C, C→G	11	0.02	28	0.03
G→T, C→A	11	0.02	17	0.02
Total mutations	725		889	
Non-synonymous	501	0.69	634	0.71
Termination	19	0.03	44	0.05

In each gene, 726 nucleotides were sequenced. Sequences containing frameshift events, which occurred at a very low level (<5%), were discarded.

interpreted as the average fraction of functional one-mutant neighbors on the protein-sequence-space network.[10,17] This assumption is consistent with experimental results obtained without using PCR,[2] and with theoretical considerations.[5] This model assumes no average epistasis.

The probability a nucleotide mutation produces a non-synonymous change is assumed to be binomial, with parameter p_{ns}, corresponding to the assumption that mutations hit distinct codons. This assumption and the value $p_{ns}=0.7$ appear realistic (the precise parameter value will vary somewhat according to the codon composition of a gene).[3] In the following analysis, non-synonymous changes include insertions, deletions, mutations to stop codons, and mutations that change the encoded amino acid: $p_{ns}=p_{ins}+p_{del}+p_{stop}+p_{aa}$. The first three types of changes are assumed to truncate and inactivate the encoded protein; we assume they constitute a fraction $p_{tr}=p_{ins}+p_{del}+p_{stop}\approx 0.05-0.07$ of mutations (see Supplementary Data of Drummond *et al.*[18]) and use the value $p_{tr}=0.06$ for our calculations. The probability that a non-synonymous mutation does not truncate the encoded protein (and thus changes only the encoded amino acid) is $(1-p_{tr}/p_{ns})$. The probability a sequence with m_{nt} nucleotide mutations retains function includes all these effects and is therefore:

$$\begin{aligned} Pr(\mathrm{f}|m_{nt}) &= \sum_{m_{ns}=0}^{m_{nt}} Pr(m_{ns}|m_{nt})Pr(\text{non trunc.}|m_{ns}) \\ &\quad \times Pr(\mathrm{f}|m_{ns}\text{ amino acid changes}) \\ &= \sum_{m_{ns}=0}^{m_{nt}} \binom{m_{nt}}{m_{ns}} p_{ns}^{m_{ns}}(1-p_{ns})^{m_{nt}-m_{ns}} \times (1-p_{tr}/p_{ns})^{m_{ns}} \\ &\quad \times \nu^{m_{ns}} = (1-(1-\nu(1-p_{tr}/p_{ns}))p_{ns})^{m_{nt}} \end{aligned} \tag{2}$$

Under the assumption of Poisson-distributed mutations, Shafikhani *et al.* showed that, if a fraction q_i of nucleotide mutations inactivate a protein, the fraction functional declines exponentially as $e^{-\langle m_{nt}\rangle q_i}$.[4] Because $q_i=(1-\nu(1-p_{tr}/p_{ns}))p_{ns}$, we expect $Pr(\mathrm{f})=e^{-\langle m_{nt}\rangle(1-\nu(1-p_{tr}/p_{ns}))p_{ns}}$ in a Poisson-distributed library. This exponential decline became the experimental expectation for subsequent groups, leading to surprise when functional mutants were later found in great excess at high average mutation rates. By combining equations (1) and (2), and assuming gene length $L\to\infty$, a mild assumption when $\langle m_{nt}\rangle \ll L$, we find the probability a sequence from the library will retain function is:

$$\begin{aligned} Pr(\mathrm{f}) &= \sum_{m_{nt}=0}^{\infty} Pr(\mathrm{f}|m_{nt})Pr(m_{nt}) \\ &= \left(\frac{1+\lambda\exp\left(-\frac{\langle m_{nt}\rangle(1+\lambda)}{n\lambda}(1-\nu(1-p_{tr}/p_{ns}))p_{ns}\right)}{1+\lambda}\right)^{n} \end{aligned} \tag{3}$$

Equation (3) makes several predictions. In the limit of many thermal cycles n, all else equal, the original expectation $Pr(\mathrm{f})=e^{-\langle m_{nt}\rangle(1-\nu(1-p_{tr}/p_{ns}))p_{ns}}$ (above) is recovered. If the number of thermal cycles n is proportional to $\langle m_{nt}\rangle$, following the protocol of Shafikhani *et al.*, then $Pr(\mathrm{f})$ should be a perfect exponential in $\langle m_{nt}\rangle$, which is precisely what this group reports. However, if n is fixed as in other studies,[3,8,9] then $Pr(\mathrm{f})$ curves upward relative to an exponential decline as $\langle m_{nt}\rangle$ increases. PCR efficiency λ decreases with increasing $\langle m_{nt}\rangle$,[19] which increases the expected curvature. In other words, there will be more functional sequences than predicted by the exponential decline.

Using the previously reported scFv antibody data for low $\langle m_{nt}\rangle$,[3] where the Poisson assumption is not unreasonable, and the reported value $q_i=0.6$, we can estimate $\nu\approx 0.2$ for the antibody binding task. For the subtilisin data,[4] we similarly use the

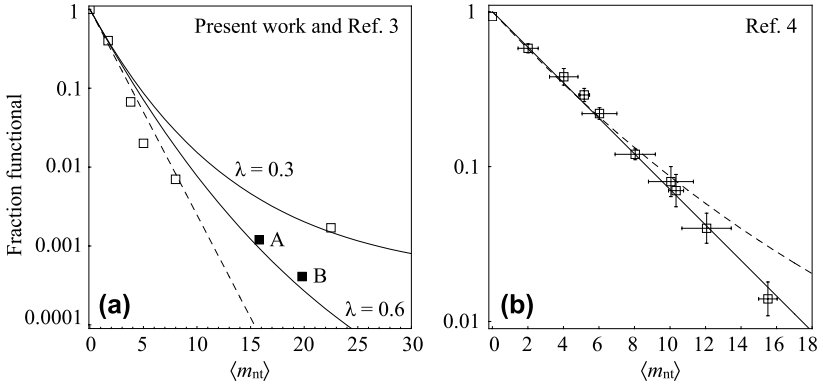

Figure 2. Equation (3) explains previously reported experimental results. (a) Comparison to scFv antibody data from Daugherty *et al.*[3] (□) and present work (■); for conditions, see the footnotes to Table 3. The broken line is the original fit reported,[3] $e^{-\langle m_{nt}\rangle q_i}$ with $q_i=0.6$. The continuous lines show equation (3) for the two libraries reported here (bottom) and for the highest-$\langle m_{nt}\rangle$ library conditions reported previously (top).[3] Changes in line curvature are due entirely to changes in PCR efficiency λ. (b) Comparison to high-$\langle m_{nt}\rangle$ subtilisin data from Shafikhani *et al.*[4] (open squares with standard error bars), which were produced by a multi-round protocol. Conditions (all per-round): $d=n\lambda=10$ DNA doublings, $n=13$ thermal cycles, $\langle m_{nt}\rangle=2.01$ or 5.17 nucleotide mutations per gene. The fractions functional predicted by equation (3) for a multi-round protocol (continuous line) and a single-round protocol (dotted line) show that the theory properly predicts the observed exponential decline in fraction functional.

Table 3. Comparison of retention of wild-type digoxigenin binding for scFv antibody libraries with analytical predictions

$\langle m_{nt} \rangle$	N	Observed functional	Observed % funct.	Predicted % funct.[a] (Poisson)	Predicted % funct.[a] (equation (3))	Predicted U_f
1.7	3×10^5	1.4×10^5	40.0	36.1	38.8	2473
3.8	1×10^6	6.7×10^4	6.7	10.2	12.9	8811
15.8[b]	–	–	0.12	0.0076	0.095	–
19.8[b]	–	–	0.041	0.00069	0.029	–
22.5	6×10^6	1×10^4	0.17	0.00014	0.15	1463

[a] Assumed scFv $\nu = 0.2$ (see the text), efficiency $\lambda = 0.6$ for all but highest-$\langle m_{nt} \rangle$ library, for which we estimate efficiency $\lambda = 0.3$.
[b] Only fractions functional were recorded for these libraries.

reported $q_i = 0.27$ to estimate $\nu \approx 0.65$. With these values for ν, Figure 2 compares the predictions of equation (3) to the observed fractions of functional clones at various library mutation levels $\langle m_{nt} \rangle$ reported by Daugherty *et al.*[3] and in the present work for the scFv antibody fragment (Figure 2(a)) (see also Table 3) and Shafikhani *et al.*[4] for subtilisin (Figure 2(b)). The agreement is quite good and demonstrates that the excess of functional clones can in fact be consistent with an underlying exponential relationship between number of amino acid substitutions and probability of retained wild-type function. To further test our analytical predictions, we simulated single-round EP-PCR using template DNA strands encoding a folded "wild-type" lattice protein. The amplified DNA was translated into lattice proteins, which were scored as functional if they retained the fold and thermostability of the wild-type. We observed excellent agreement with equation (3) (see Supplementary Data).

The reason for deviation from an exponential decline is hinted at in the limit of large average mutation rates, when the exponential part of equation (3) vanishes and $Pr(\mathrm{f})$ approaches a constant, $Pr(\mathrm{f}) \to (1+\lambda)^{-n}$. For a mutationally fragile protein such as the scFv antibody performing the digoxigenin binding task, this can occur at experimentally accessible mutation rates, as can be seen most clearly in the library originally reported[3] and revisited by Georgiou[7]. As the mutation rate increases, the antibody fragment becomes "quite insensitive to mutational load" and $Pr(\mathrm{f})$ flattens out at a value of roughly 0.0018.[7] Most interestingly, this limiting value is a function only of the PCR conditions, and does not depend on the protein at all.

What causes these counterintuitive results? EP-PCR at high frequency generates heavily mutated sequences by a process akin to Xeroxing copies of copies: low-fidelity copies give rise to even lower-fidelity copies, yet a copy, once produced, is not replaced, but remains in the final distribution of copies. During the PCR, the first generation of mutants, amplified directly from the wild-type template gene and carrying few mutations, persists in the mix and continues to reproduce copies with few additional mutations throughout subsequent cycles. The protein products of these less-mutated copies retain function at a greatly elevated rate compared to the average sequence, leading to upward bias in the functional fraction.

Why are improved mutants found more often in high-error-rate libraries?

If statistical effects of the mutagenesis protocol can explain the dramatic deviation from exponential in the fraction of functional sequences without recourse to epistasis, why are high-$\langle m_{nt} \rangle$ libraries enriched in improved clones, despite a smaller number of clones retaining any function? To address this question, we now explore another consequence of PCR's broad mutational distribution.

The effective size of a library is not the number of mutants screened, the number usually reported, but rather the number of unique mutants screened. In a library of 10^6 transformants of the scFv antibody gene (726 bp, 242 amino acid residues) with an average of one mutation per sequence, most of the 2178 possible 1-mutants will occur of the order of 100 times, reducing the effective library size by roughly two orders of magnitude. Most mutagenesis is concerned with protein sequences, where additional losses occur. Truncations due to frameshift mutations or mutations to stop codons eliminate a significant fraction of sequences. With one nucleotide mutation per codon, an average of 5.7 amino acid substitutions (out of a maximum of 19) are accessible due to the conservatism of the genetic code, for a total of $242\times5.7 = 1379$ accessible amino acid sequences with one substitution. (We ignore the effects of synonymous mutations.) Thus 10^6 transformants yield just over 10^3 unique protein sequences, about a 10^3-fold reduction in the effective library size.

We estimate the number of unique sequences in an EP-PCR library in the following way. We derive the distribution of non-synonymous substitutions $Pr(m_{ns})$ after EP-PCR, estimate the number of non-truncated amino acid sequences $N_{m_{ns}}$ with each m_{ns} in a library of a given size, compute the expected number of unique sequences $U_{m_{ns}}$ at each m_{ns} by accounting for recurrence among the $N_{m_{ns}}$ sequences, and then find the expected number of unique sequences U by summing the $U_{m_{ns}}$.

With PCR conditions denoted as before and an

average number of nucleotide mutations per sequence $\langle m_{\rm nt}\rangle$, what is the distribution of the number of non-synonymous substitutions per sequence $Pr(m_{\rm ns})$? We assume, as before, that each nucleotide mutation causes a non-synonymous change with probability $p_{\rm ns}$, so we obtain:

$$\begin{aligned} Pr(m_{\rm ns}) &= \sum_{m_{\rm nt}=m_{\rm ns}}^{L} Pr(m_{\rm nt}) \binom{m_{\rm nt}}{m_{\rm ns}} p_{\rm ns}^{m_{\rm ns}} (1-p_{\rm ns})^{m_{\rm nt}-m_{\rm ns}} \\ &= (1+\lambda)^{-n} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \frac{(ky)^{m_{\rm ns}} e^{-ky}}{m_{\rm ns}!} \end{aligned} \tag{4}$$

with:

$$y = \frac{\langle m_{\rm nt}\rangle p_{\rm ns}(1+\lambda)}{n\lambda}$$

That is, the distribution of non-synonymous substitutions $Pr(m_{\rm ns})$ is equivalent, in form, to the distribution of nucleotide mutations $Pr(m_{\rm nt})$, but with an average of $\langle m_{\rm ns}\rangle = \langle m_{\rm nt}\rangle p_{\rm ns}$ substitutions. For simplicity, we will drop the subscript for non-synonymous substitutions and use m.

Of the sequences with m non-synonymous substitutions, some will also be truncated by frameshifts or stop codons. Because we treat all truncations as non-synonymous changes, the fraction of non-truncated sequences with m substitutions is $Pr(\text{non-truncated}|m) = (1-p_{\rm tr}/p_{\rm ns})^m$. Given an EP-PCR library of N transformants, $N_m = N\ Pr(m)\ Pr(\text{non-truncated}|m)$, on average, are non-truncated proteins with m amino acid substitutions.

Of these proteins with m substitutions, how many unique sequences exist? Only one unique sequence has $m=0$. For any m there are, on average,

$$M_m = \binom{L/3}{m} 5.7^m$$

total unique proteins with, at most, one mutation per codon, where L is the length of the gene in nucleotides.

Given $N_{\rm m}$ samples, how many of these $M_{\rm m}$ unique proteins can we expect to find? This is the classic "coupon collector problem"[20] and directly addresses the question of mutant recurrence, since any sample either yields a new, unique protein or one that has been sampled before. The expected number of unique sequences produced by equiprobably sampling $M_{\rm m}$ sequences $N_{\rm m}$ times is:

$$\begin{aligned} U_m &= M_m - M_m(1-1/M_m)^{N_m} \\ &\approx M_m(1-e^{-N_m/M_m}) \end{aligned} \tag{5}$$

For example, to sample 99% of the $M_{\rm m}=1379$ accessible 1-mutants of scFv requires 4.6-fold oversampling ($N_{\rm m}=6350$ samples) on average. Taking 1379 samples, $N_{\rm m}=M_{\rm m}$, on average, yields only 872 unique proteins, or 63% of the total. In practice, for proteins of a few hundred amino acid residues and libraries of a few million transformants, recurrence need be considered only for small values of m ($m<3$), because sequence space becomes large enough to make recurrence extremely unlikely at higher m values so that $U_{\rm m}\approx N_{\rm m}$. The total number of unique sequences in a library is simply the sum over all unique sequences with a specific number of substitutions:

$$U = \sum_{m=0}^{L/3} U_m \tag{6}$$

Figure 3(a) shows the fraction of unique sequences U/N obtained from simulations (see Materials and Methods) in which the scFv gene was mutated according to PCR statistics with the observed frequencies (Table 2, with 3% frameshift rate) or unbiased frequencies (all mutations equally weighted, with 3% frameshift rate). The prediction from equation (6) is also plotted and agrees well. Increasing the mutation rate increases the number of unique sequences because fewer are lost to recurrence. Note that, even at the highest mutation rates, the fraction of unique sequences does not approach 1.0, because sequences truncated by frameshifts and stop codons are not considered unique and accumulate at increasing levels as the mutation rate is increased.

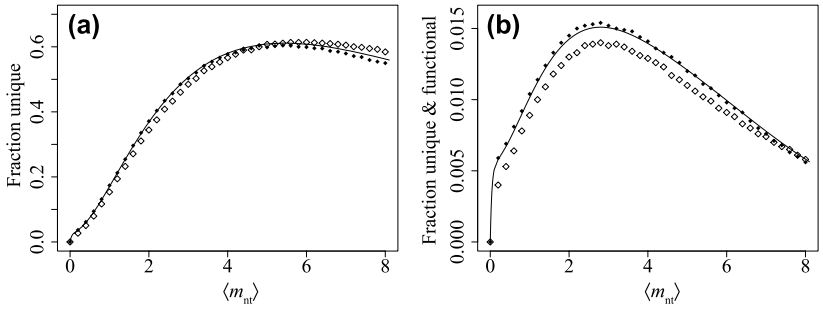

Figure 3. Error-prone PCR error rates strongly influence the fraction of unique and functional sequences. (a) Fraction of unique sequences in a simulated library of $N=50{,}000$ scFv clones ($\nu=0.2$) using the observed mutational spectrum (◇) or an unbiased spectrum (◆). The line is equation (6) (divided by N) evaluated with $n=30$ thermal cycles, efficiency $\lambda=0.6$, $p_{\rm ns}=0.76$ and $p_{\rm tr}=0.07$. (b) Fraction of unique and functional sequences in the same library. The line is equation (7) (divided by N) evaluated using the same parameters. An optimal mutation rate exists that balances uniqueness with retention of function. Mutational biases lower the fraction of unique and functional sequences, but do not alter the optimal mutation rate significantly.

Of greater interest is the expected number of unique sequences in the library that are expected to retain at least wild-type function, because these sequences are a superset of potentially improved sequences. We can estimate the number of unique, functional sequences as:

$$U_f = \sum_{m=0}^{L/3} U_m \nu^m \quad (7)$$

Figure 3(b) shows the fraction of unique, functional sequences U_f/N obtained from the same simulations as in Figure 3(a), with equation (7) plotted for comparison. Biases in mutation frequencies decrease the fraction of unique sequences but preserve the overall form. Results using unbiased frequencies are predicted accurately by our theoretical treatment.

Clearly, low-error-rate libraries suffer from dramatic mutant recurrence, an effect avoided at high error rates. Improved proteins are found often in high-error-rate libraries because these libraries contain more unique functional sequences.

Optimal random mutagenesis

A typical and important goal in protein engineering is to improve an existing protein function, for example by increasing catalytic rate, thermostability, binding affinity, or specificity. While rational engineering has made significant strides, high-throughput screening of large mutant libraries for improved clones is both a dominant strategy to achieve this goal and an area of active research[7].

Given a choice of protein scaffold, a library of fixed size, and no reliable basis for rational engineering, a simple measure of library optimality is the number of unique functional sequences it contains. Figure 3 shows that, given this measure, an optimal mutation rate exists that balances diversity (uniqueness is lost if $\langle m_{nt} \rangle$ is too low) with retained function (functional sequences are rare if $\langle m_{nt} \rangle$ is too high). Mutational biases do not affect the optimal mutation rate significantly.

The optimum depends on the number of transformants sampled, the PCR protocol used, and the wild-type protein being mutated, among other parameters. Figure 4(a) compares predicted optimal mutation rates under identical PCR conditions for the scFv antibody ($\nu \approx 0.2$), depending on whether 10^3 or 10^6 clones are screened. The difference, 1.3 average nucleotide substitutions, corresponds to one amino acid substitution, on average. Figure 4(b) compares predicted optimal mutation rates under identical conditions and with the same wild-type protein, but using 30 thermal cycles (as in the present work) in one case and two cycles (as used by Zaccolo & Gherardi[9]) in the other. A difference of one nucleotide mutation results. Optimal rates depend on protein mutational tolerance as reflected by ν: the more tolerant the protein, the higher the optimal mutation rate (not shown).

Table 3 lists estimates for U_f given the scFv library experimental conditions reported here and previously.[3] Despite the over 200-fold lower observed percentage of functional transformants isolated from the highest-$\langle m_{nt} \rangle$ library relative to the lowest, and the 14-fold fewer functional sequences observed, only 60% fewer unique functional sequences are expected in the highest-$\langle m_{nt} \rangle$ library. Given the experimental parameters of the highest-$\langle m_{nt} \rangle$ library and altering only the mutation rate, the rate $\langle m_{nt} \rangle = 11.0$ is predicted to produce more unique functional sequences (>10,000) than any of the reported libraries. The optimal mutation rate given the highest-$\langle m_{nt} \rangle$ experimental parameters is predicted to be roughly $\langle m_{nt} \rangle = 3.0$, which is predicted to yield >34,000 unique, functional sequences.

Discussion

Laboratory evolution by random mutagenesis remains the most effective strategy for improving

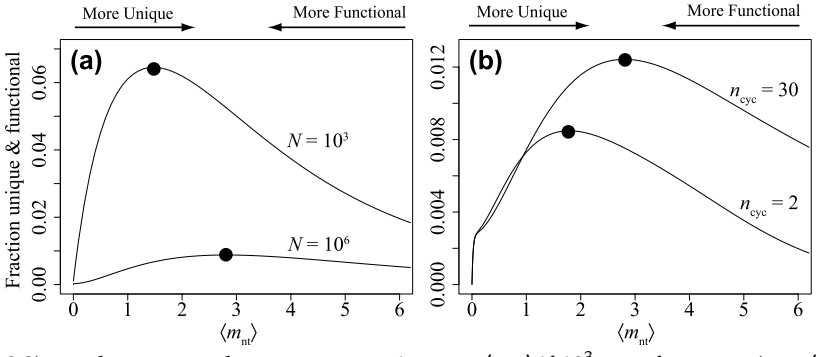

Figure 4. The requirement for uniqueness reduces effective library size and leads to library-dependent and protocol-dependent optimal library mutation rates. (a) Optimal mutation rate (●) depends on library size. Predicted fractions of unique functional sequences given by equation (7) for the same protocol ($n = 30$ thermal cycles with efficiency $\lambda = 0.6$, $p_{ns} = 0.76$ and $p_{tr} = 0.07$) and protein (scFv-like, $\nu = 0.2$) are shown at each average mutation rate $\langle m_{nt} \rangle$ if 10^3 transformants (top, $\langle m_{nt} \rangle_{opt} = 1.5$) or 10^6 transformants (bottom, $\langle m_{nt} \rangle_{opt} = 2.8$) are screened. (b) Optimal mutation rate (●) depends on PCR protocol. Predicted fractions of unique functional sequences given by equation (7) are shown for the same protein (scFv-like, $\nu = 0.2$) and library size (10^5 transformants) using $n = 30$ thermal cycles (top, $\langle m_{nt} \rangle_{opt} = 2.8$) or $n = 2$ thermal cycles (bottom, $\langle m_{nt} \rangle_{opt} = 1.8$). In all cases, recurrence leads to profound loss of uniqueness at low $\langle m_{nt} \rangle$, and the optimal $\langle m_{nt} \rangle$ balances uniqueness and retention of function.

enzyme properties given a choice of scaffold and no reliable basis for rational engineering. The possibility that distant regions of sequence space harbor excesses of functional and, for at least some enzymatic tasks, improved proteins has been advanced several times, with significant experimental evidence to bolster the claims. We have shown that a more accurate model of EP-PCR than used previously, due to Sun,[14] is required to describe adequately the mutational distribution resulting from high-error-rate EP-PCR. This model, in turn, provides straightforward explanations for the previously observed experimental findings: (1) the excess functional proteins observed at high $\langle m_{nt} \rangle$ is predictable using equation (3), is due to low-mutation sequences generated early in the reaction, and is consistent with an exponential decrease in retention of function with amino acid substitution level; and (2) loss of functional sequences at high mutation rates can be balanced by diversity in the form of more unique sequences, improving sampling of sequence space and leading to a higher probability that improved mutants will be found if they exist. We have demonstrated the often-overlooked importance of accounting for recurrence of mutants when estimating how much of sequence space a library covers, extending previous work on modeling effects of mutational bias.[21] With our simple definition of library optimality as maximizing the number of unique, functional proteins, these two observations lead to an optimal mutation rate for EP-PCR, which can be estimated using our analytical results. However, optimal mutation rates are both protocol and protein-dependent. Optimal rates derived for EP-PCR using one set of conditions do not necessarily hold for another set (Figure 4), and are highly unlikely to hold for saturation mutagenesis or site-directed mutagenesis, for which uniqueness is rarely a problem and the distribution of mutation levels in a typical library is tight and easily controllable.

We have explained several disparate mutagenesis results using only a single parameter unrelated to experimental protocols: ν, the average probability of retaining wild-type function after a random amino acid substitution.[5] It follows that these experiments can be used to measure ν using the analytical tools we have introduced here, with an important caveat. Because multiple mutations per codon, rarely found in EP-PCR even at high mutation rates (though not always[22]), are necessary to experimentally measure ν, such experiments cannot measure this parameter directly but can provide a credible upper bound due to the conservative nature of the genetic code. While ν relates simply to the "structural plasticity" $q_i = (1 - \nu(1 - p_{tr}/p_{ns}))p_{ns}$ proposed by Shafikhani *et al.*,[4] our results show that the emergence of a perfect exponential decline in their experiments likely depended both on a fundamental property of proteins and the particular experimental protocol employed. We also distinguish between genetic mutations that produce truncated protein products, essentially all of which lack function, and those that produce full-length proteins whose structural properties determine whether mutations are tolerated. We believe ν captures the idea of structural plasticity more accurately.

Because optimal mutation rates depend on ν, we can suggest measures that influence ν and that therefore may be used to manipulate the optimal mutation rate. All else being equal, proteins with higher thermodynamic stability (free energy of unfolding) have a higher ν,[5] and tolerate more destabilizing substitutions, suggesting that more stable variants of a protein represent more promising departure points for mutagenesis. If longer proteins are more tolerant of substitutions, as seems plausible, then longer genes will tend to have higher optimal mutation rates. Codon usage may influence ν indirectly, through protein expression; in cases where high protein expression is required for the relevant function, replacement of rare codons with common synonyms may allow higher mutation rates. When the crystal structure of a protein is available, ν can be estimated computationally.[5] We note that the exponential decline in fraction functional holds when many mutations are introduced, as in the present work, but may not always hold for small numbers of mutations.[5]

The intrinsic functional tolerance of a protein to substitutions is only one of many ways in which genetic mutations may affect the fraction of active clones in a library. Biologically relevant or screenable activity may depend on the action of many molecules in an organism, so mutations that hinder expression (e.g. through introduction of non-preferred codons, or in rarer cases by altering mRNA secondary structure) may decrease the fraction of clones scored as active. Disruption of signal sequences may result in improper targeting to cellular locations such as the periplasm or cell membrane. Mutations may destabilize the protein, hindering its folding or exposing it to proteolysis or irreversible misfolding without actually destroying the function of the natively folded molecule. The dominant effect of most random mutagenesis is changes in the primary sequence of a target protein, most of which disrupt native function, and our simple treatment appears to work well under these circumstances.

Our results also illuminate potentially serious methodological flaws in previous studies. For example, the accuracy in measuring average library mutation rate by nucleotide sequencing depends on the variance of the mutational distribution, which at high mutation rates is far broader than that of the Poisson distribution previously assumed. The expected standard error of measurement on a library with $\langle m_{nt} \rangle$ average mutations assessed by sequencing N_{seq} clones is:

$$\sigma_m / \sqrt{N_{seq}} = \sqrt{\langle m_{nt} \rangle (1 + \langle m_{nt} \rangle / n\lambda) / N_{seq}}$$

Zaccolo and Gherardi,[9] for example, report four libraries averaging $\langle m_{nt}\rangle = 8.2$, 19.7, 21.3 and 27.2 mutations per coding region of a 1088 base-pair gene constructed using two, five, ten and 20 thermal cycles with $\langle m_{nt}\rangle$ measured by sequencing at least 2500 base-pairs, effectively $N_{seq} = 2.5$. Even if the true value of $\langle m_{nt}\rangle$ is as measured and perfect PCR efficiency assumed, these measurements have an expected 1σ standard error of 4.3, 6.5, 5.4 and 5.3 mutations per gene, respectively, calling into question the actual levels of hypermutagenesis achieved in these experiments.

The analysis presented here has important consequences for understanding the natural and directed evolution of proteins. Importantly, we have provided a thorough analysis of an apparent manifestation of mutational epistasis. Two issues are often confused: whether mutations interact epistatically on average in individual folded sequences, and whether mutations interact epistatically on average in a library or ensemble that contains both folded and unfolded sequences. Ensemble epistasis is the only measure of interest in studies of the evolutionary persistence of sexual recombination,[12] and of primary interest in deciding which regions of sequence space should be targeted for efficient directed evolution.

If ensemble epistasis existed, as implied by earlier interpretations of the less-than-exponential decline in retention of function with mutational distance discussed here, then individual epistasis would be found on average. Importantly, the reverse is not true. Though folded or improved proteins may display cooperative effects (mutations that are better together than individually), many polypeptides in a random library may carry mutations that are more deleterious together than apart. However, the latter are unlikely to be found by investigators, because such mutants are disproportionately likely to fail to fold, and little if any attention is given to the vast numbers of unfolded proteins in mutant libraries. Confusion arising from the asymmetry between types of epistasis (ensemble epistasis implies individual epistasis, but individual epistasis does not imply ensemble epistasis) may have inspired prior claims that high mutation rates can be used to access reservoirs of cooperative mutations while only a "small proportion" of clones will be lost to disruptive mutations.[9]

As a result of our analysis, several data sets probing high mutation rates can now be seen, despite appearances to the contrary, to provide no evidence for ensemble epistasis, of particular biological interest given the recent discoveries of multiple native error-prone polymerases in bacteria and higher organisms.[23] Meanwhile, recent work providing an explanation for why the fraction of mutant proteins retaining function will decline exponentially suggests that ensemble epistasis is unlikely.[5] We cannot rule out the existence of epistasis; our analysis merely points out one way in which a mutation process can produce results which give the appearance of epistasis when there is none.

Exploration of distant regions of sequence space by random mutation alone appears highly inefficient, reinforcing the role of other search processes such as homologous recombination in creating sequence diversity.[24,25] High-mutation-rate EP-PCR, however, can be used to overcome the "uniqueness sink" that occurs at low mutation rates when using selection or high-throughput screening to assay large numbers of clones. Finally, optimal mutation rates cannot be decoupled from the physical process of mutation, making them dependent on the particular organism or protocol under consideration. There can be no "optimal mutational load for protein engineering," as has been suggested,[22] without specification of the engineering methodology.

Materials and Methods

Library construction, sequencing and functional assay

We constructed two libraries, A and B, from EP-PCR reactions as described.[19] Identical mutagenesis conditions were used for both libraries but produced different mutation levels in each library. In particular, 2.50 mM $MgCl_2$, 0.5 mM $MnCl_2$, 0.35 mM dATP, 0.40 mM dCTP, 0.20 dGTP, and 1.35 mM dCTP were used along with Taq DNA polymerase. The PCR reaction was continued for 30 cycles rather than 16. All other parameters were set and subsequent ligation, transformation and fluorescence-activated cell sorting functional analysis procedures performed as described.[3]

Statistical characterization of mutational distributions

To characterize the sequencing results and relate them to two theoretical distributions (the Poisson distribution:

$$Pr(m;\langle m_{nt}\rangle) = \frac{\langle m_{nt}\rangle^m e^{-\langle m_{nt}\rangle}}{m!}$$

and the PCR distribution, equation (1)), we used the likelihood ratio test, which compares the probabilities of observing a particular mutational sample under competing distributions. A mutational sample, obtained by sequencing, consists of N sequences $i = 1 = N$ having m_i mutations. Given a theoretical mutational distribution $Pr(m)$ which gives the probability of randomly choosing a sequence having m mutations, the likelihood of a sample is $L = \prod_{i=1}^{N} Pr(m_i)$. The likelihood ratio test evaluates the statistic $LR = 2[\ln(L_{Poisson}/L_{PCR})]$ which has approximately a χ^2 distribution. Significance values (P values) can be computed from the likelihood ratio statistic, the χ^2 distribution and a number of degrees of freedom, which in this case is 2, corresponding to the two additional parameters in the PCR distribution, the number of thermal cycles n and the replication efficiency λ.

Simulation

To simulate the EP-PCR process, two approaches were taken. First, we exhaustively simulated the EP-PCR process using genes encoding simplified model proteins

(compact lattice model, 25 residues consisting of any of 20 amino acids) which were then folded and assayed for retention of wild-type structure. Details of this simulation and results are presented in Supplementary Data.

We found that a vastly simpler simulation produced nearly identical results (see Supplementary Data Figure S2) and used this simulation to generate Figure 3. In this simplified simulation, the scFv gene was mutated to produce $N = 50{,}000$ sequences at each $\langle m_{nt} \rangle$ according to the observed mutation frequencies (Table 2, Library A) and the PCR distribution, equation (1), with parameters as indicated in the Figure legend. Each mutated gene was translated into a protein sequence according to the universal genetic code. Truncated proteins, either from stop codons or frameshifts, were discarded. Whether a full-length sequence was functional or not was estimated by counting the number of amino acid substitutions relative to wild-type and designating the protein functional with probability $Pr(\mathrm{f}|m_{aa}) = \nu^{m_{aa}}$. All full-length protein sequences were inserted in a set that retained only unique sequences. Numbers and fractions of unique, functional and jointly unique and functional sequences were then tabulated.

Acknowledgements

We thank G. Chen and R. Loo for creation and screening of the scFv libraries, J. D. Bloom for optimal mutation rate calculations, C. C. Adami for guidance, C. O. Wilke for lattice protein folding code and advice, and Z. -G. Wang for insightful comments on the manuscript. We are grateful for the critical input of two anonymous reviewers. D.A.D. acknowledges NIH National Research Service Award 5 T32 MH19138. This research is supported by Army Research Office Contract DAAD19-03-D-0004.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.05.023

References

1. Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature*, **225**, 563–564.
2. Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E. & Loeb, L. A. (1996). Tolerance of different proteins for amino acid diversity. *Mol. Divers.* **2**, 111–118.
3. Daugherty, P. S., Chen, G., Iverson, B. L. & 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.
4. Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. (1997). Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotechniques*, **23**, 304–310.
5. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA*, **102**, 606–611.
6. Arnold, F. H. (1998). Enzyme engineering reaches the boiling point. *Proc. Natl Acad. Sci. USA*, **95**, 2035–2036.
7. Georgiou, G. (2001). Analysis of large libraries of protein mutants using flow cytometry. *Advan. Protein Chem.* **55**, 293–315.
8. Kunichika, K., Hashimoto, Y. & Imoto, T. (2002). Robustness of hen lysozyme monitored by random mutations. *Protein Eng.* **15**, 805–809.
9. Zaccolo, M. & Gherardi, E. (1999). The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 beta-lactamase. *J. Mol. Biol.* **285**, 775–783.
10. Wilke, C. O. & Adami, C. (2001). Interaction between directional epistasis and average mutational effects. *Proc. Roy. Soc. ser. B*, **268**, 1469–1474.
11. Elena, S. F. & Lenski, R. E. (2001). Epistasis between new mutations and genetic background and a test of genetic canalization. *Evol. Int. J. Org. Evol.* **55**, 1746–1752.
12. Kondrashov, A. S. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature*, **336**, 435–440.
13. Francisco, J. A., Campbell, R., Iverson, B. L. & Georgiou, G. (1993). Production and fluorescence-activated cell sorting of *Escherichia coli* expressing a functional antibody fragment on the external surface. *Proc. Natl Acad. Sci. USA*, **90**, 10444–10448.
14. Sun, F. (1995). The polymerase chain reaction and branching processes. *J. Comput. Biol.* **2**, 63–86.
15. Weiss, G. & von Haeseler, A. (1995). Modeling the polymerase chain reaction. *J. Comput. Biol.* **2**, 49–61.
16. Weiss, G. & von Haeseler, A. (1997). A coalescent approach to the polymerase chain reaction. *Nucl. Acids Res.* **25**, 3082–3087.
17. Bornberg-Bauer, E. & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl Acad. Sci. USA*, **96**, 10689–91064.
18. Drummond, D. A., Silberg, J. J., Wilke, C. O. & Arnold, F. H. (2005). On the conservative nature of intragenic recombination. *Proc. Natl Acad. Sci. USA*, **102**, 5380–5385.
19. Fromant, M., Blanquet, S. & Plateau, P. (1995). Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal. Biochem.* **224**, 347–353.
20. Flajolet, P., Gardy, F. & Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms, and self-organizing search. *Discrete Appl. Mathematics*, **39**, 207–229.
21. Moore, G. L. & Maranas, C. D. (2000). Modeling DNA mutation and recombination for directed evolution experiments. *J. Theoret. Biol.* **205**, 483–503.
22. Zaccolo, M., Williams, D. M., Brown, D. M. & Gherardi, E. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.* **255**, 589–603.
23. Goodman, M. F. (2002). Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu. Rev. Biochem.* **71**, 17–50.
24. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer,

W. P. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288–291.
25. Andrews, T. D. & Gojobori, T. (2004). Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen *Neisseria meningitidis*. *Genetics*, **166**, 25–32.

Edited by J. Karn

(Received 18 February 2005; received in revised form 6 May 2005; accepted 10 May 2005)