Protein Engineering vol. 16 no. 6 pp. 451–457, 2003
DOI: 10.1093/protein/gzg057

User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries

Wayne M.Patrick[1], Andrew E.Firth[2] and Jonathan M.Blackburn[1,3]

[1]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA and [2]Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

Wayne M. Patrick and Andrew E. Firth contributed equally to this work.

[3]To whom correspondence should be addressed.
E-mail: jmb50@cam.ac.uk

Directed evolution of proteins depends on the production of molecular diversity by random mutagenesis. While a number of methods have been developed for introducing this diversity, the best ways to sample it are not always clear. Here we used simple statistics to analyse completeness and diversity in randomized libraries generated by oligonucleotide-directed mutagenesis, error-prone polymerase chain reaction (epPCR) and *in vitro* recombination of highly homologous sequences. For oligonucleotide-directed mutagenesis, we derive equations to estimate how complete a given library is expected to be and also to predict the size of library required to give a fixed probability of being 100% complete. We describe the statistical bases for computer programs which estimate the number of distinct variants represented in epPCR and shuffled libraries, dubbed PEDEL and DRIVeR, respectively. These programs allow the user to calculate (rather than guess) the diversity represented in a given library and also provide empirical guidelines for maximizing this diversity. PEDEL and DRIVeR are available at www.bio.cam.ac.uk/~blackburn/stats.html.

Keywords: completeness/diversity/Poisson statistics/randomized libraries

Introduction

In the field of protein engineering, mimicking Darwinian evolution *in vitro* has emerged as a powerful means of generating proteins displaying novel properties and functions. The cornerstone of all directed evolution protocols is the production of molecular diversity by random mutagenesis and a number of methods have been developed to introduce this diversity into protein-encoding genes. The most adaptable and widespread of these are based on the polymerase chain reaction (PCR) and include: oligonucleotide-directed random mutagenesis (Hermes *et al*., 1989), error-prone PCR (epPCR) (Cadwell and Joyce, 1992) and the *in vitro* recombination protocols DNA shuffling (Stemmer, 1994a,b) and staggered extension process (StEP) (Zhao *et al*., 1998). There are many recent examples in which improved proteins have been identified in large libraries of variants generated by one or more of these techniques [reviewed by Brakmann and Taylor (Brakmann, 2001; Taylor *et al*., 2001)].

An assumption underlying all directed evolution experiments is that the amount of molecular diversity theoretically possible is enormous compared with our ability to generate and screen it. Even a small protein of 100 amino acids can be encoded by $4^{300} \approx 10^{181}$ possible DNA sequences, a number vastly larger than the number of atoms in the observable Universe ($\sim 10^{80}$), let alone the biggest protein-encoding libraries accessible in the laboratory [10^{12}–10^{15} using *in vitro* selection methods such as mRNA display (Roberts and Ja, 1999)]. Increasingly it is acknowledged that quantitative, predictive models for the processes underlying randomized library construction will be useful in targeting and interpreting that diversity which we are able to generate experimentally [reviewed by Voigt *et al.* (Voigt *et al*., 2001a)]. Recent studies have included *in silico* modelling of epPCR and the generation of crossovers in DNA shuffling (Moore and Maranas, 2000; Moore *et al*., 2001) and the construction of computational 'pre-screens' both to identify the regions of proteins most likely to yield beneficial mutations on randomization (Voigt *et al*., 2001b) and also to predict the fragments or schemas of proteins able to be recombined with minimal disruption of overall three-dimensional structure (Voigt *et al*., 2002).

While these analyses hint at the insights to be gained from a quantitative approach to directed evolution, they are too complex to be generally applicable for the laboratory researcher. Moreover, the number of mutations or crossovers required, or even optimal, to effect a given functional change remains elusive. In this paper, we argue that the likelihood of finding a variant with improved properties in a given library is maximized when that library is maximally diverse. We used simple statistics to derive a series of widely-applicable equations and computer algorithms for estimating the number of unique sequence variants in libraries constructed by randomized oligonucleotide mutagenesis, epPCR and *in vitro* recombination. Generally, applying these algorithms provides mathematical support for the previously empirical guidelines which have evolved for generating randomized libraries in which diversity is maximized and unwanted degeneracy is minimized, although some new strategies for library construction also become apparent.

Materials and methods

GLUE, PEDEL and DRIVeR are a suite of programs for calculating library statistics. They have been written in Fortran 77 and are available to be downloaded from www.bio.cam.ac.uk/~blackburn/stats.html. Supplementary information at this URL includes comprehensive program notes and a PDF file describing the mathematics underlying the programs in full detail.

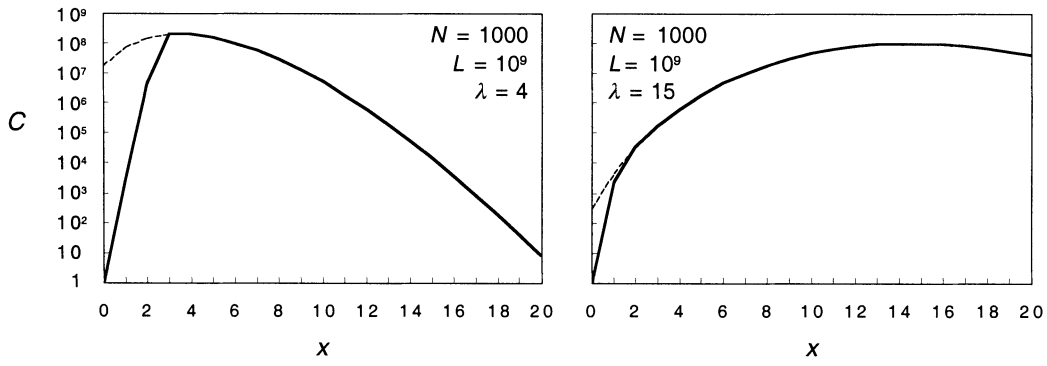

Fig. 1. Plot of the expected number of distinct variants, C_x (solid bold lines), having exactly x point mutations, calculated for an epPCR library of size $L = 10^9$, sequence length $N = 1000$ bp and mean mutation rate per sequence $\lambda = 4$ (left) and $\lambda = 15$ (right). Also plotted over the range $0 < x < 20$ is L_x (thin dashed lines), the total number of sequences in the library expected to contain exactly x mutations. The calculated threshold x values, x_u, are ~3.1 (left) and ~1.6 (right); it can be seen that for $x \geqslant x_u$, C_x is very well approximated by L_x.

Results

Oligonucleotide-directed random mutagenesis

Incorporating randomized codons into one of the primers in a PCR mix allows the generation of molecular diversity at specific locations in a gene. Intuitively, we know that randomizing a greater number of codons reduces the likelihood of sampling all possible random variants. Here we derive simple equations for estimating how many variants a given library will actually contain and how large a library needs to be in order to give a fixed probability (e.g. 95%) that *all* possible sequence variants will be represented.

Consider a library containing a number of clones L, constructed by randomizing M codons or $N = 3M$ base pairs, in which all possible sequence variants v_i are equally probable. Since the variants are equally probable, the mean number of occurrences of any one variant v_i in the library is given by $\lambda = L/V$ (where V is the total number of possible sequence variants). For $\lambda << L$ (e.g. $V > 10$), the actual number of occurrences of any variant v_i is essentially independent of the number of occurrences of any other variant v_j and can therefore be well approximated by the Poisson distribution [see Feller for details (Feller, 1968)]:

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{1}$$

where $P(x)$ denotes the probability that the variant v_i occurs exactly x times in the library. The probability that v_i occurs at least once is $1 - P(0) = 1 - e^{-\lambda} = 1 - e^{-L/V}$. Hence the expected number of distinct variants in the library is

$$C = V(1 - e^{-L/V}) \tag{2}$$

and the fractional completeness of the library is given by $F = C/V = 1 - e^{-L/V}$.

As an example, let us assume that we mutate four codons in a gene using NNS (N = A/C/G/T; S = C/G) codons in the randomization protocol. Because there are 32 possible NNS codons, it follows that there are $V = 32^4 \approx 10^6$ possible sequence variants. If we wish to construct a library expected to contain, say, 95% of all the possible sequence variants, then we must solve $F = 0.95 \Rightarrow 1 - e^{-L/V} = 0.95 \Rightarrow L = -V \ln 0.05 \approx 3.0V$. Three-fold degeneracy, corresponding to 3×10^6 clones in the present example, is therefore required for a library that is expected to contain 95% of the 10^6 possible sequence variants.

A related but distinct problem is to calculate the required size of a library that has a 95% chance of being 100% complete. The probability that *every* variant v_i is represented is $P_c = (1 - e^{-L/V})^V$. Solving for L gives

$$\begin{aligned} L &= -V\ln\left[1 - \exp\left(\frac{\ln P_c}{V}\right)\right] \\ &\approx -V\ln\left[1 - \left(1 + \frac{\ln P_c}{V}\right)\right] \\ &= -V\ln\left(-\frac{\ln P_c}{V}\right) \end{aligned} \tag{3}$$

where the approximation holds provided $V >> -\ln P_c$. Since one is generally interested in P_c values of the order of 90–100% (and certainly >1%), this condition is generally true.

Returning to our example in which four NNS codons are used to generate a randomized library, we can now calculate the size that a library needs to be to have a 95% chance of containing every possible sequence variant. Solving Equation 3 for $P_c = 0.95$ and $V = 10^6$ gives $L \approx 1.7\times10^7$. Hence a much higher level of degeneracy, ~17-fold in this example, is required to be 95% certain of sampling all possible variants. The degree of over-sampling required varies with the number of variants V. For example, if five codons are randomized instead of four, $V = 32^5 \approx 3.4\times10^7$ and solving Equation 3 for $P_c = 0.95$ shows that 20-fold degeneracy is required to be 95% certain that the library will be complete.

While it is reasonably straightforward to calculate manually library completeness C or the probability P_c of having a complete library, using Equations 2 and 3, we have also written a short Fortran 77 program, GLUE, to perform these calculations. GLUE may be downloaded from www.bio.cam.ac.uk/~blackburn/stats.html.

Error-prone PCR

Although most recent examples of directed evolution use epPCR in conjunction with recombination-based strategies such as DNA shuffling, it is still commonly encountered as a means of generating random diversity at any position in a gene. Further, the epPCR protocol has been refined to afford considerable control over the substitution rate. Low rates of

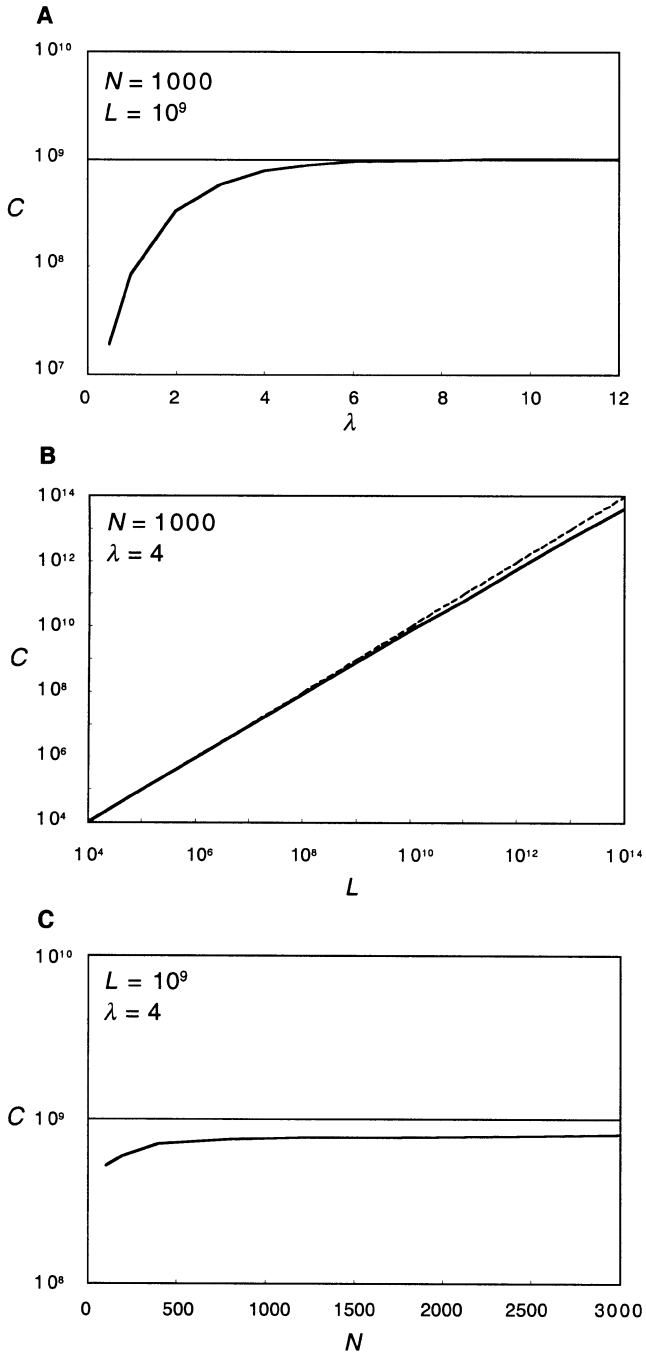

Fig. 2. Plots of the total expected number of distinct variants C (bold lines) in epPCR libraries as a function of (**A**) mutation rate λ, (**B**) library size L and (**C**) sequence length N. (A) For $N = 1000$ there are 4^{1000} possible distinct sequences and as λ increases a greater number of these are sampled. The plot levels off when C is limited by the total library size L (thin line). (B) Even with only four (i.e. λ) mutations there are ~3×10^{12} possible variants whereas with eight mutations there are ~2×10^{23} possible sequences. Thus for $\lambda = 4$, even in very large libraries, the degree of redundancy is low and C is of the same order as L. The dashed line plots $C = L$ for reference. (C) C changes very little with N for a fixed value of λ (in which case the mean mutation rate *per base pair* scales inversely with N).

mutation (typically 2–3 base substitutions per gene, corresponding to ~1 amino acid substitution per sequence per generation) have traditionally been employed, in order to minimize the accumulation of deleterious mutations (Arnold *et al*., 2001). However, using nucleoside analogues to generate a much higher mutational load of 8.2–27.2 mutations per gene in the directed evolution of TEM-1 β-lactamase has also yielded improved variants (Zaccolo and Gherardi, 1999).

By assuming Poisson statistics to analyse the actual compositions of epPCR libraries, here we address the question of how these apparently contradictory approaches can prove equally successful. We describe a simple program, dubbed PEDEL (**P**rogram for **E**stimating **D**iversity in **E**rror-prone PCR **L**ibraries) and available to download from www.bio.cam.ac.uk/~blackburn/stats.html, which estimates the number of unique sequence variants, C, found in a given epPCR library.

The input parameters for PEDEL are the length of the template sequence, N (in base pairs), the size of the randomly-generated library, L, and the mean number of point mutations per sequence, λ (it is assumed that $\lambda \ll N$, e.g. $\lambda < 0.1N$). A value for λ can be obtained experimentally by sequencing a small number of library variants.

The probability that a library variant contains exactly x mutations, given that the mean number of mutations per sequence is λ, can be described by the Poisson distribution $P(x)$ (Equation 1; Cadwell and Joyce, 1992). The number of possible sequences differing from the parental sequence by exactly x mutations is given by

$$V_x = \binom{N}{x}(3^x) = \frac{3^x N!}{x!(N-x)!} \qquad (4)$$

since $\binom{N}{x}$ is the number of ways of choosing x bases to mutate and each of these may be mutated to any of three other bases, giving the 3^x term.

In order to estimate the total number of unique variants, PEDEL divides the library of size L into a series of sub-libraries, each of which contains only variants with exactly x mutations. The expected size of each sub-library is given by $L_x = P(x)L$. Since all sequences with exactly x mutations are equally probable, the completeness statistic C_x derived for cassette mutagenesis (Equation 2) can be applied to each sub-library, substituting V_x (Equation 4) for V.

When the sub-library size L_x is small compared with the total number of possible variants V_x in that sub-library, we may expect that almost every member of the sub-library will be distinct, i.e. $C_x \approx L_x$. This is the case when x is large. This approximation is good to within 5% for $L_x/V_x < 0.1$ and in turn allows the calculation of a threshold x value, x_u (see www.bio.cam.ac.uk/~blackburn/stats.html for details), such that for all $x > x_u$ the approximation $C_x \approx L_x$ is acceptable. For example, in a library with a total of $L = 10^9$ members, length $N = 1000$ base pairs and mean mutation rate $\lambda = 4$, one finds that $x_u \approx 3.1$ (Figure 1). PEDEL uses this observation to split the infinite summation

$$C = \sum_{x=0}^{\infty} C_x$$

and hence to calculate the total number of distinct variants in a given epPCR library. Conversely, when x is small enough, L_x will be large compared with V_x and we may expect the sub-library to sample nearly all possible variants, i.e. $C_x \approx V_x$. Indeed, in most scenarios encountered experimentally, there is at most one x value for which one or other of these approximations is not valid. Examples of how C varies as a

Table I. Composition of the epPCR library generated by Saab-Rincón *et al.* (Saab-Rincón *et al.*, 2001), as predicted by PEDEL

x[a]	L_x[b]	V_x[c]	C_x[d]	C_x/V_x (%)[e]
0	1.5×10^{5}	1.0×10^{0}	1.0×10^{0}	100
1	5.3×10^{5}	2.1×10^{3}	2.1×10^{3}	100
2	9.2×10^{5}	2.2×10^{6}	7.6×10^{5}	35
3	1.1×10^{6}	1.5×10^{9}	1.1×10^{6}	0.073
4	9.4×10^{5}	8.0×10^{11}	9.4×10^{5}	1.2×10^{-4}
5	6.6×10^{5}	3.4×10^{14}	6.6×10^{5}	1.9×10^{-7}
6	3.9×10^{5}	1.2×10^{17}	3.9×10^{5}	3.3×10^{-10}
7	1.9×10^{5}	3.5×10^{19}	1.9×10^{5}	5.4×10^{-13}
8	8.4×10^{4}	9.0×10^{21}	8.4×10^{4}	9.3×10^{-16}
9	3.3×10^{4}	2.1×10^{24}	3.3×10^{4}	1.6×10^{-18}
10	1.1×10^{4}	4.3×10^{26}	1.1×10^{4}	2.6×10^{-21}
11	3.7×10^{3}	8.1×10^{28}	3.7×10^{3}	4.6×10^{-24}
12	1.1×10^{3}	1.4×10^{31}	1.1×10^{3}	7.9×10^{-27}
13	2.9×10^{2}	2.2×10^{33}	2.9×10^{2}	1.3×10^{-29}
14	7.2×10^{1}	3.3×10^{35}	7.2×10^{1}	2.2×10^{-32}
15	1.7×10^{1}	4.5×10^{37}	1.7×10^{1}	3.8×10^{-35}
16	3.7×10^{0}	5.8×10^{39}	3.7×10^{0}	6.4×10^{-38}

[a]Number of mutations per sequence.
[b]Expected number of sequences in the library with exactly x mutations.
[c]Number of possible variants with exactly x mutations, calculated using Equation 4.
[d]Expected number of distinct variants in the library with exactly x mutations.
[e]Percentage of all possible variants with exactly x mutations represented in the library.

function of mutation rate λ, library size L and template sequence length N are shown in Figure 2.

To illustrate the use of PEDEL, consider the library of monomeric triose phosphate isomerase (TIM) variants described by Saab-Rincón *et al.* (Saab-Rincón *et al.*, 2001). With a single round of epPCR on the 700 bp gene, the authors constructed a library of 5×10^{6} clones; sequencing 10 of these demonstrated an error rate of 3–4 nucleotide substitutions per gene. Implementing the sub-library algorithm with $\lambda = 3.5$, PEDEL calculates that this library contains ~4.2×10^{6} distinct variants, with the remaining 800 000 clones representing multiple occurrences of some variants. Furthermore, the batch version of the program (named 'stats.batch' and also available at www.bio.cam.ac.uk/~blackburn/stats.html) can be used to display the compositions of individual sub-libraries (Table I) and this shows where the redundancy is found. In particular, 150 000 clones (3% of the total library) are expected to be the parental sequence (i.e. zero mutations), while the 2100 possible sequences with a single nucleotide substitution are over-sampled 250-fold in the $x = 1$ sub-library, which comprises 530 000 clones. Approximately 35% of all possible double-mutant sequences are represented in the library, although sampling of sequences containing a greater number of mutations is sparse owing to the very large numbers of these variants which are possible. Indeed, even the very largest libraries generated *in vitro* would fail to sample all possible TIM sequences containing five mutations.

At first sight this result may appear counter-intuitive: when the rather high experimentally-derived mutation rate is considered, one might expect that few (if any) library sequences contain 0–1 mutations. However, on statistical analysis, it becomes apparent that not only are the $x = 0$ and $x = 1$ sub-libraries fully sampled, but also that it is in these sub-libraries that overall library diversity is lost. The results described by Zaccolo and Gherardi (Zaccolo and Gherardi, 1999) provide an even more extreme example of the utility of the sub-library approach. While their unselected TEM-1 libraries contained means of between 8.2 and 27.2 mutations per gene, those clones selected on the basis of conferred resistance to cefotaxime showed only 1–11 mutations at the DNA level. That is, by increasing the mutation rate λ they sampled a greater number of sequences with a large number of substitutions, but also continued to sample sequences with many fewer mutations, and it is from these latter sub-libraries that functional variants have been isolated.

Finally, it should be noted that in reality and as detailed by Moore and Maranas (Moore and Maranas, 2000), the distribution of mutations in the final pool of epPCR-generated molecules may deviate from Poisson statistics. The primary cause is the well-documented bias in the types of base substitution introduced by *Taq* DNA polymerase under error-prone conditions (Shafikhani *et al.*, 1997). By making some sequence variants less likely than others, this has the effect of reducing library diversity (i.e. decreasing completeness, C). However, with the recent advent of commercially available polymerases such as Mutazyme (Stratagene, La Jolla, CA), which possesses an opposite bias to *Taq* in its mutational spectrum, unbiased libraries can now be constructed by sequential PCR amplifications with the two polymerases (W.M.Patrick, M.de Lumley and J.M.Blackburn, unpublished data). Given the premise that directed evolution is most likely to identify an improved variant when a library is maximally diverse and that this is the case when all variants in a sub-library are equally probable, we suggest that this dual-polymerase approach should become routine.

In vitro recombination of highly homologous sequences

Methods for recombining genes *in vitro* have proved revolutionary in the field of directed evolution. Here we describe DRIVeR (**D**iversity **R**esulting from ***In vitro*** **R**ecombination), a program for estimating the diversity represented in libraries generated by recombining two highly homologous parental sequences which differ in only a few (e.g. 20) base or amino acid positions. It is available to be downloaded at www.bio.cam.ac.uk/~blackburn/stats.html.

The number of unique sequences in a shuffled DNA library will be critically dependent upon (i) the mean number of crossovers during the recombination process and (ii) the spacing of the varying base pairs, since bases that are closely spaced are less likely to be recombined than those that are far apart in the sequence. In most reported examples of *in vitro* recombination, by either DNA shuffling or StEP PCR, the number of crossovers observed per daughter sequence is small [typically of the order of 1–4 (Zhao *et al.*, 1998; Raillard *et al.*, 2001). Suppose, then, that the mean number of crossovers per daughter sequence is λ and that $\lambda << N$, e.g. $\lambda < 0.1N$ (where N is the sequence length). We assume that the number of crossovers between two consecutive varying positions in a particular daughter sequence follows a Poisson distribution (Equation 1) with mean $(n-1)\lambda/(N-M-1)$, where M is the number of varying base positions and $(n-1)/(N-M-1)$ gives the ratio of the total number of potential crossover points in the sequence (viz. $N-M-1$) to the number of crossover points between the two varying positions ($n-1$). M is incorporated into this expression to account for the observation that, owing to the requirement for a base-paired 3′ clamp in fragment reassembly and extension, crossovers immediately following variable positions are impossible. DRIVeR also therefore sets the probability of a crossover at these positions to zero.

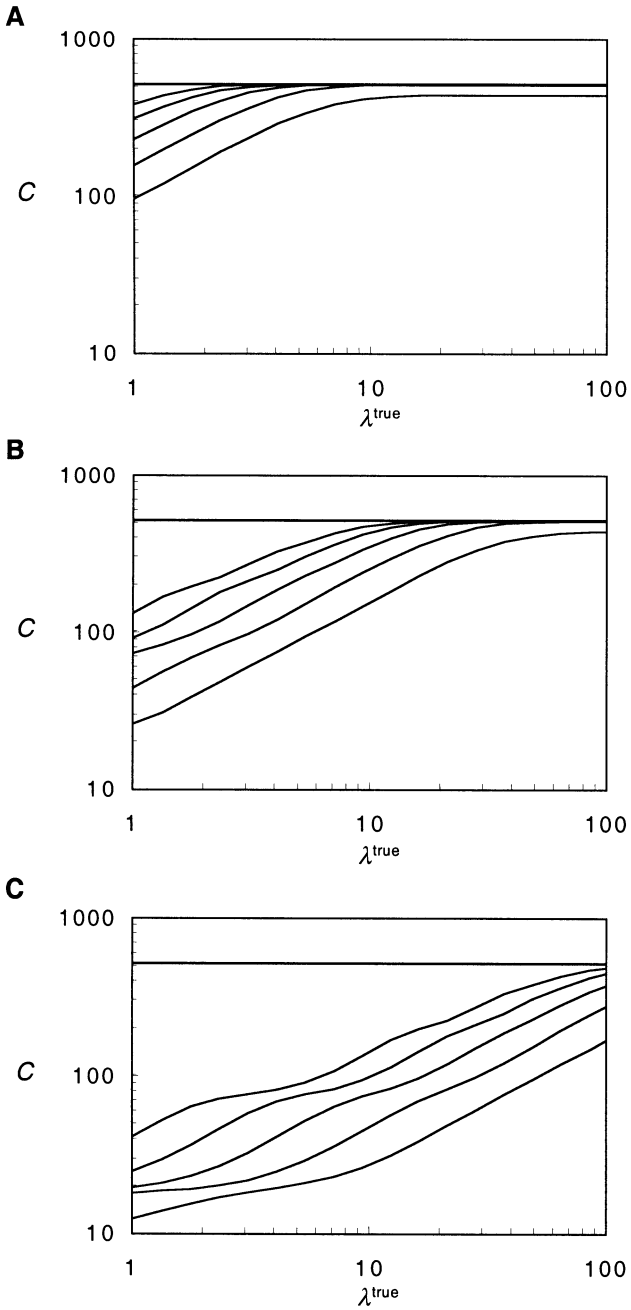

Fig. 3. The expected number C (thin lines) of distinct variants in a shuffled library of size L as a function of λ^{true} for libraries of sizes 1000, 4000, 16 000, 64 000 and 256 000 (listed from lower curve to upper curve). In all cases the specified sequence length is $N = 500$ base pairs and the number of varying base pairs is set at $M = 9$. (**A**) Variable base pairs at positions 50, 100, 150, 200, 250, 300, 350, 400, 450. (**B**) Variable base pairs at positions 100, 110, 120, 130, 140, 150, 160, 170, 180. (**C**) Variable base pairs at positions 100, 102, 104, 106, 108, 110, 112, 114, 116. C increases with increasing L and increasing λ and is greater if the variable base pairs are well spaced along the parent sequences. For sufficiently large L and λ, C levels off at the total number of possible distinct sequences (512 in this case, plotted as a bold line). Note that this is not the case if L is of the same order as the number of possible distinct sequences, e.g. when $L = 1000$ in the examples above (see text for further discussion of this observation).

Note that experimentally, crossovers are only observable if they occur in a region that will produce a distinct daughter sequence. Clearly, one crossover between consecutive varying base pairs produces the same daughter sequence as 3, 5, 7, … crossovers and similarly for any even number of crossovers. Further, any crossovers occurring between one end of the sequence and the first varying base position are also unable to be detected by sequence analysis of daughter variants. Therefore only an 'observed' λ value, λ^{obs}, can be determined from experimental data. However, the statistics used by DRIVeR to estimate the number of distinct sequences in a shuffled DNA library require that the value specified by the user is that of λ^{true}. For a given input value of λ^{true}, DRIVeR calculates the resulting λ^{obs}, allowing a quick, empirical search for the value of λ^{true} based on a λ^{obs} value determined by sequencing a small number of variants from a library. Conversely, experimental uncertainty in estimating λ^{obs} can be assessed by entering a range of possible λ^{true} values into the program and assessing the effects on both λ^{obs} and the overall diversity of the library. Details on these calculations are available in the supplementary material at www.bio.cam.ac.uk/~blackburn/stats.html.

Once a value for λ^{true} in keeping with λ^{obs} has been determined, DRIVeR uses the inputted sequence length N, number of varying base positions M, library size L, λ^{true} and the positions of the varying bases to calculate the relative probability of each possible shuffled variant. This in turn allows the calculation of C, the expected total number of distinct daughter sequences in the library of size L. DRIVeR also tabulates the probabilities of an even or an odd number of crossovers in each interval between varying codons and a further output file contains the probabilities of occurrence of each of the possible 2^M daughter sequences. As one would expect, plotting the expected value of C as a function of L with various combinations of spacing of varying base pairs shows that C increases with library size and is larger if the varying bases are well spaced along the parental sequences (Figure 3).

In most cases, C levels off at the total number of possible shuffled sequences (*i.e.* 512 in Figure 3). Perhaps counter-intuitively, however, this is not the case for the library of size $L = 1000$ in Figure 3(a) and (b), for which the maximum value of C is 440, falling short of a complete sampling of all variants. If λ^{true} is very large, then the varying base pairs will be essentially randomly assigned in each daughter sequence, all possible variants will be essentially equally likely and increasing λ^{true} further will have no effect on C. At this point, C depends only on the library size L and the problem reduces to one analogous to oligonucleotide-directed mutagenesis (in which each variant is also equally probable). Indeed, solving Equation 2 for $L = 1000$ and $V = 512$ gives $C = 439$, in excellent agreement with the maximum completeness predicted by DRIVeR for this library.

In a recent example from the literature, Raillard *et al.* (Raillard *et al.*, 2001) used DNA shuffling to recombine two bacterial triazine hydrolase genes (*atzA* and *triA*, GenBank accession numbers U55933 and AF312304, respectively), which differed at nine of 1425 bases, coding for proteins with nine varying amino acids. There are therefore $2^9 = 512$ possible shuffled daughter variants. The authors screened a library of $L = 1600$ shuffled variants, although they also noted that this probably fell short of a screen of the complete library. However, it was not immediately obvious what fraction of the complete library had actually been sampled here. We therefore specified $N = 1425$, $M = 9$, $L = 1600$ and the positions of the variable bases were 250, 274, 375, 650, 655, 757, 763, 982 and 991. Trial and error show that a λ^{true} of 10 crossovers per sequence corresponds to $\lambda^{obs} \approx 2$, which in turn agrees with the

data obtained experimentally (the authors state that 'every variant sequenced had undergone at least one and as many as four recombination events'). On inputting $\lambda^{true} = 10$, DRIVeR calculates that the screened library of 1600 members is expected to contain ~161 of the 512 possible distinct sequence variants. The biggest factor leading to the low diversity is the close spacing of the variable bases at positions 650 and 655, 757 and 763, and 982 and 991, between each of which recombination events are unlikely to occur. Indeed, even a 1000-fold larger library of size $L = 1.6\times10^6$ would still be incomplete, containing an expected 484 distinct sequences. Further, the close spacing of these variable bases means that the uncertainty in library diversity, C, associated with estimating λ^{obs} is significant in this example: if λ^{obs} was in fact 1.6, the library would be expected to contain 126 variants, while $\lambda^{obs} = 2.4$ gives $C = 221$. This effect is reduced for sequences in which the variable bases are more evenly spaced.

Discussion

It is apparent that random mutagenesis is very efficient at producing molecular diversity; however, the best ways to exploit this diversity are not always clear. In many cases the size of a library that can be produced and screened in the laboratory represents only an infinitesimally small fraction of all possible sequences and furthermore any library will inevitably contain wasted diversity: variants encoding premature stop codons; synonymous DNA mutations; mutations in regions known to disrupt protein structure [predicted by the SCHEMA algorithm of Voigt *et al.* (Voigt *et al*., 2002), for example]; or duplicates of one sequence at the expense of another unsampled sequence. Estimating the true amount of useful diversity represented in a given library therefore becomes a multi-faceted problem, although one for which the solution promises valuable insights into how best to design and screen such a library for that rare improved variant.

Here we have used simple arguments and the approximating assumption of Poisson statistics in order to analyse completeness and therefore degeneracy in libraries generated by three common PCR-based methods. Applying the algorithms described provides laboratory researchers with either an estimate of how many unique sequence variants their library is most likely to contain or conversely a target library size to enable them to sample a specified proportion of all possible sequences.

In the first example we derive simple equations for characterizing the diversity arising from oligonucleotide-directed mutagenesis. In calculating library sizes required for fixed levels of completeness, an important distinction is made between 'a library that is most likely to be 95% complete' (Equation 2) and 'a 95% chance of having a 100% complete library' (Equation 3). As one might expect, we show that libraries containing much higher levels of sequence degeneracy are required to effect the latter situation. The assumption underpinning these equations is that each random variant is equally probable and at the DNA level this will be the case. However, an interesting recent study of M13 phage-displayed peptides (Rodi *et al*., 2002) uncovered biases in the sequences of those peptides which are successfully translated and displayed, suggesting that in this example the underlying biology has affected functional library diversity in a manner not predicted by Poisson statistics. Of the methods commonly employed for screening or selecting from randomized libraries, display on filamentous phage is the one most likely to show such a bias (reflecting the constraints on sequence imposed by periplasmic virus assembly) and this should be considered when implementing GLUE to assess the diversity represented in phage-displayed libraries.

Provided that the number of randomized codons is low, oligonucleotide-directed mutagenesis offers the potential to generate a library which samples all possible sequence variants. In epPCR, however, the ability to generate mutations at any position in a gene ensures that the possible number of variants is enormous and it becomes more informative to analyse how many distinct sequences are actually represented in a given library. The sub-library approach described here and implemented by PEDEL gives insights into the composition of epPCR libraries. For example, if the mutation rate λ is kept low (the case in the majority of published experiments), then a large proportion of the resulting library (as high as 37% when $\lambda = 1$) will contain the parental template sequence, thus significantly reducing the effective size of that library. However, a mid-sized library with a higher mutation rate (e.g. $L = 10^6$, $\lambda = 6$, $N = 1000$) will still contain all possible variants differing from the template sequence by a single nucleotide, whilst reducing the fraction of unmutated template sequence present to <0.25%. In addition, it will also sample many sequences which differ from the template in 2–7 positions. Hence PEDEL allows an informed trade-off to be made between maximizing the number of unique sequence variants sampled and ensuring that those sequences least likely to contain deleterious mutations are well represented.

DNA shuffling and its variants have proven to be extremely flexible methodologies, enabling the inclusion of many different pieces of genetic information in their protocols. Diversity can be introduced by incorporating an epPCR step, shuffling a large number of parental sequences from different species or by incorporating synthetic oligonucleotides to target specific regions for mutagenesis (Minshull and Stemmer, 1999). Here we have analysed a 'simplest case' scenario, in which two highly homologous parental sequences are recombined. In modelling the shuffling reaction as a Poissonian process we have simplified it considerably, since as detailed by Moore *et al*. (Moore *et al*., 2001), crossovers will tend to accumulate in regions of high sequence identity and are better predicted by considering the thermodynamics of the reassembly reaction. However, in examples such as that of the triazine hydrolases *atzA* and *triA* (Raillard *et al*., 2001), where the parental sequences show near 100% sequence identity, it can be inferred that crossover position will be well approximated by a Poisson distribution and that DRIVeR will therefore provide a useful measure of the diversity represented in a given library. A further advantage of DRIVeR is that it can be used in the analysis of libraries generated by both DNA shuffling and StEP PCR, as the critical parameter is the number of crossovers, rather than the size of the reassembled fragments (a concept that is irrelevant in the latter methodology).

A key result from the analysis of recombined libraries using DRIVeR was that diversity is largely dependent on λ^{true} and hence that it is maximized when crossovers are sufficiently frequent that each daughter variant is equally probable. In such a scenario, the number of variants in a given library is best estimated using the formula derived for oligonucleotide-directed mutagenesis (Equation 2). Two modifications to the DNA shuffling protocol that were described recently (Ness *et al*., 2002; Zha *et al*., 2003) are directly predicted by this

observation. Both groups designed overlapping oligonucleotides for assembly in a synthetic DNA shuffling reaction, resulting in the production of functional daughter variants through the recombination of tightly linked diversity. In both cases, *in vitro* recombination was reduced to a problem of equally probable variants, with concomitant success in maximizing the resultant library diversity. A third report (Moore and Maranas, 2002) has recently outlined a method for engineering codon usage to maintain amino acid sequence but to optimize or direct recombination, based on earlier considerations of predicting crossover generation. In combination, these approaches offer the prospect of a more rational approach to increasing diversity and therefore the likelihood of identifying improved variants in recombined libraries.

In addition to addressing the question of how much molecular diversity is present in a randomized library, the statistics and algorithms described here provide evidence for a number of empirical guidelines for library construction. In general, of course, the larger the library, the more diverse it is expected to be, although in the case of oligonucleotide-directed mutagenesis GLUE provides insight into just how large is large enough to sample a required proportion of variants. In epPCR and DNA shuffling, we have demonstrated that maximizing library diversity requires consideration of both library size, L, and mean mutation rate or crossover frequency, λ. Contrary to the prevailing dogma for construction of libraries by epPCR, we have shown that an elevated mutation rate can be useful for increasing the absolute number of variants sampled, in addition to reducing wasteful over-sampling of the unmutated template and single point mutants. Likewise, in DNA shuffling a very large mean number of crossovers per sequence (corresponding to small fragments in the reassembly reaction), such that every daughter variant is essentially equally probable, ensures that the maximum number of shuffled variants will be represented. This is especially important for ensuring recombination between two or more varying bases that are closely spaced in the parent sequences. In combination, we believe that the descriptive and predictive aspects of GLUE, PEDEL and DRIVeR will prove useful and generally applicable in guiding the construction and interrogation of randomized protein-encoding libraries.

Acknowledgements

The authors thank Mark Liddament for helpful discussions and critical reading of the manuscript. W.M.P. gratefully acknowledges financial support from the Cambridge Commonwealth Trust's Prince of Wales Scholarship and an Overseas Research Student Award. A.E.F. acknowledges financial support from Trinity College, Cambridge, and use of facilities at the University of Otago. J.M.B. thanks the Royal Society for a University Research Fellowship.

References

Arnold,F.H., Wintrode,P.L., Miyazaki,K. and Gershenson,A. (2001) *Trends Biochem. Sci.*, **26**, 100–106.

Brakmann,S. (2001) *ChemBioChem*, **2**, 865–871.

Cadwell,R.C. and Joyce,G.F. (1992) *PCR Methods Appl.*, **2**, 28–33.

Feller,W. (1968) *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

Hermes,J.D., Parekh,S.M., Blacklow,S.C., Köster,H. and Knowles,J.R. (1989) *Gene*, **84**, 143–151.

Minshull,J. and Stemmer,W.P.C. (1999) *Curr. Opin. Chem. Biol.*, **3**, 284–290.

Moore,G.L. and Maranas,C.D. (2000) *J. Theor. Biol.*, **205**, 483–503.

Moore,G.L. and Maranas,C.D. (2002) *Nucleic Acids Res.*, **30**, 2407–2416.

Moore,G.L., Maranas,C.D., Lutz,S. and Benkovic,S.J. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 3226–3231.

Ness,J.E., Kim,S., Gottman,A., Pak,R., Krebber,A., Borchert,T.V., Govindarajan,S., Mundorff,E.C. and Minshull,J. (2002) *Nat. Biotechnol.*, **20**, 1251–1255.

Raillard,S. *et al.* (2001) *Chem. Biol.*, **8**, 891–898.

Roberts,R.W. and Ja,W.J. (1999) *Curr. Opin. Struct. Biol.*, **9**, 521–529.

Rodi,D.J., Soares,A.S. and Makowski,L. (2002) *J. Mol. Biol.*, **322**, 1039–1052.

Saab-Rincón,G., Juárez,V.R., Osuna,J., Sánchez,F. and Soberón,X. (2001) *Protein Eng.*, **14**, 149–155.

Shafikhani,S., Siegel,R.A., Ferrari,E. and Schellenberger,V. (1997) *Biotechniques*, **23**, 304–310.

Stemmer,W.P.C. (1994a) *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.

Stemmer,W.P.C. (1994b) *Nature*, **370**, 389–391.

Taylor,S.V., Kast,P. and Hilvert,D. (2001) *Angew. Chem., Int. Ed. Engl.*, **40**, 3310–3335.

Voigt,C.A., Kauffman,S. and Wang,Z.-G. (2001a) *Adv. Protein Chem.*, **55**, 79–160.

Voigt,C.A., Mayo,S.L., Arnold,F.H. and Wang,Z.-G. (2001b) *Proc. Natl Acad. Sci. USA*, **98**, 3778–3783.

Voigt,C.A., Martinez,C., Wang,Z.-G., Mayo,S.L. and Arnold,F.H. (2002) *Nat. Struct. Biol.*, **9**, 553–558.

Zaccolo,M. and Gherardi,E. (1999) *J. Mol. Biol.*, **285**, 775–783.

Zha,D., Eipper,A. and Reetz,M.T. (2003) *ChemBioChem*, **4**, 34–39.

Zhao,H., Giver,L., Shao,Z., Affholter,J.A. and Arnold,F.H. (1998) *Nat. Biotechnol.*, **16**, 258–261.

Received November 6, 2002; revised May 6, 2003; accepted May 20, 2003