User-Friendly Algorithms for Estimating Completeness and Diversity in Randomized Protein-Encoding Libraries

(Mathematical Notes)

Andrew E. Firth[1] and Wayne M. Patrick[2]

[1] *Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand*
[2] *Center for Fundamental and Applied Molecular Evolution, Emory University, Atlanta, Georgia 30322, USA*

Present addresses: (Dec 2007)
[1]BioSciences Institute (5th floor), University College Cork, Cork, Ireland
[2]Institute of Molecular Biosciences, Massey University, Albany Campus, Auckland, New Zealand

References:

Wayne M. Patrick, Andrew E. Firth and Jonathan M. Blackburn, 2003, User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries, *Protein Engineering*, 16, 451–457.

Andrew E. Firth and Wayne M. Patrick, Statistics of protein library construction, *Bioinformatics*, 21, 3314–3315.

Web-server: `http://guinevere.otago.ac.nz/stats.html`

Notation: The following mathematical notation has been used:

$\sum_{i=1}^{n} A_i$	Sum up the values A_i with i ranging from 1 to n.
$\prod_{i=1}^{n} A_i$	Multiply the values A_i with i ranging from 1 to n.
n!	Factorial – i.e. $\prod_{i=1}^{n} i = 1 \times 2 \times 3 \times ... \times (n-1) \times n$. ($0! = 1$, by definition.)
$\left(\begin{array}{c} n \\ r \end{array}\right)$	Combinatorial ${}^nC_r$ – the number of ways of choosing r distinct objects out of a total of n distinct objects. Equal to $\frac{n!}{r!(n-r)!}$.

$\exp x$ and $\ln x$ are the natural (base e) exponential and logarithmic functions.

1 Equiprobable variants

Problem: Given a library $\mathcal{L}$ of size L sequences, where each sequence is chosen at random from a set $\mathcal{V}$ of size V of *equiprobable* variants, we wish to calculate the expected number of distinct sequences in $\mathcal{L}$.

Let v_i be one of the V possible variants. Since the variants are equiprobable, the mean number of occurrences of v_i in $\mathcal{L}$ is $\lambda = \frac{L}{V}$. For $\lambda \ll L$ (i.e. $V \gg 1$), the actual number of occurrences of v_i in $\mathcal{L}$ is essentially independent of the number of occurrences of any other v_j, $j \neq i$, so is well-approximated by the Poisson distribution

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ... \tag{1}$$

where $P(x)$ denotes the probability that v_i occurs exactly x times in $\mathcal{L}$. The probability that v_i occurs at least once in $\mathcal{L}$ is $1 - P(0) = 1 - e^{-\lambda} = 1 - e^{-L/V}$. Hence the expected number of distinct variants in the library is $C \approx V(1 - e^{-L/V})$ and the fractional completeness is $F = \frac{C}{V} \approx 1 - e^{-L/V}$. If we wish to generate a library that will have an expected completeness of, for example, 95% then we must solve $F = 0.95 \Rightarrow 1 - e^{-L/V} = 0.95 \Rightarrow L = -V \ln 0.05 \approx 3V$, i.e. we require a library size three times the number of possible variants.

If however we wish to generate a library that has a 95% probability of being 100% complete then we must do a little more work. The probability that *every* variant v_i is represented in $\mathcal{L}$ is $P_c \equiv P(\text{complete}) \approx \prod_i (1 - e^{-L/V}) = (1 - e^{-L/V})^V$. Solving for L gives

$$\begin{array}{rcl} L & = & -V \ln\left(1 - \exp\left(\frac{\ln P_c}{V}\right)\right) \\ & \approx & -V \ln\left(1 - \left(1 + \frac{\ln P_c}{V}\right)\right) \\ & = & -V \ln\left(-\frac{\ln P_c}{V}\right), \end{array} \quad (2)$$

where the approximation holds provided $V \gg -\ln P_c$ (i.e. for any P_c that is not approximately 0)[1]. Since one is generally interested in P_c values of order 90%-100% (and certainly $> 1\%$) this condition is generally true. For an example, if $P_c = 95\%$ and $V = 1000000$, then we obtain $L \approx 17 \times 10^6$. I.e. in order to generate a library which has a 95% probability of being *complete*, we require $L = 17 \times 10^6$.

These two distinct problems – a library that has an expected completeness of 95% and a library that has a 95% probability of being 100% complete – have sometimes been confused in the literature.

`GLUE` is a simple `C++` programme for calculating the expected completeness of a given library, or the library size required to obtain a desired completeness. It is available from the web address given on page 1.

2 Error-prone PCR

Problem: Given a library $\mathcal{L}$ of size L comprising variants of a sequence of N nucleotides, into which random point mutations have been introduced, we wish to calculate the expected number of distinct sequences in $\mathcal{L}$.

We denote the mean number of point mutations per sequence by λ and assume that $\lambda \ll N$. Then the number of point mutations per sequence may be well-described by the Poisson distribution

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ... \quad (4)$$

where $P(x)$ is the probability of there being exactly x mutations in the sequence. The exact mechanics of error-prone PCR may result in some deviation from Poisson statistics and it is left to the reader to judge how serious these deviations may be in her/his particular situation[2]. The number of possible distinct sequences with exactly x mutations is given by

$$V_x = \left(\begin{array}{c} N \\ x \end{array} \right) 3^x = \frac{3^x N!}{x!(N-x)!} \quad (5)$$

since $\left(\begin{smallmatrix} N \\ x \end{smallmatrix} \right)$ 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. Let $\mathcal{L}_x$ denote the sublibrary of $\mathcal{L}$ containing those sequences in $\mathcal{L}$ with exactly x mutations and let L_x denote the expected size of $\mathcal{L}_x$. Then $L_x = P(x) \times L$. Assuming that these sequences are equiprobable, the expected completeness of $\mathcal{L}_x$ is given by $1 - e^{-L_x/V_x}$, provided

[1]The interested reader is referred to Chapter IV.2 of Feller (1968) for a more rigorous approach, which results in

$$P_m(L, V) = \exp(-Ve^{-L/V})\frac{(Ve^{-L/V})^m}{m!}, \quad (3)$$

where $P_m(L, V)$ is the probability that exactly m variants are not present in $\mathcal{L}$. For $m = 0$ (i.e. all variants are present) we have $P_c = \exp(-Ve^{-L/V}) \Rightarrow L = -V \ln(-\frac{\ln P_c}{V})$, as above. Also, since equation 3 is simply the equation for a Poisson distribution with mean $Ve^{-L/V}$, we see that the mean number of variants not present is $Ve^{-L/V}$, so the mean number of variants present is $V - Ve^{-L/V} = V(1 - e^{-L/V})$, also as above.

[2]The programme `PEDEL` (see below) has recently been updated to include the 'PCR distribution' of Sun (1995) as an alternative to the Poisson distribution. This distribution takes into account the number of PCR cycles and the efficiency of each cycle.

$V_x \gg 1$ (see §1). This is generally the case for $x \geq 1$, especially if N is large, but not true for $x = 0$ for which $V_0 = 1$. The expected number of distinct sequences in $\mathcal{L}_x$, $x \neq 0$, is therefore given by

$$\begin{aligned} C_x &\approx V_x(1 - e^{-L_x/V_x}) \\ &= \tfrac{3^x N!}{x!(N-x)!}(1 - \exp(-\tfrac{e^{-\lambda}\lambda^x L(N-x)!}{3^x N!})), \quad x = 1, 2, 3, ... \end{aligned} \tag{6}$$

There is only one possible variant in $\mathcal{L}_0$ (the original unmutated sequence) so if $L_0 \gg 1$ then $C_0 \approx 1$. Since $L_0 = e^{-\lambda}L$ and L is generally large, this will only not be the case when λ is large. In this case the sum of the remaining C_x is very large and the contribution of C_0 is negligible. Hence it suffices to always approximate the first term in $\sum_{x=0}^{\infty} C_x$ by 1. So the total expected number of distinct sequences is

$$\begin{aligned} C &= \textstyle\sum_{x=0}^{\infty} C_x \\ &\approx 1 + \textstyle\sum_{x=1}^{\infty} V_x(1 - e^{-L_x/V_x}) \\ &= 1 + \textstyle\sum_{x=1}^{\infty} \tfrac{3^x N!}{x!(N-x)!}(1 - \exp(-\tfrac{e^{-\lambda}\lambda^x L(N-x)!}{3^x N!})). \end{aligned} \tag{7}$$

At face value the infinite sum may look rather daunting, but it turns out that we can make a few approximations. When L_x is small compared with the total number of possible variants V_x we may expect that almost every member of $\mathcal{L}_x$ will be distinct, in which case $C_x \approx L_x$. This occurs for large x values. Conversely, when L_x is large compared with V_x we may expect $\mathcal{L}_x$ to sample all possible variants in which case $C_x \approx V_x$. This occurs for small x values. We now derive conditions which describe where we can make these approximations and find that in most cases there is only one x value for which we can not use one approximation or the other.

Since $e^{\epsilon} \approx 1 + \epsilon$ for $|\epsilon| \ll 1$ (or $1 - e^{-\epsilon} \approx \epsilon$), in the case where $\frac{L_x}{V_x} \ll 1$, $C_x = V_x(1 - e^{-L_x/V_x})$ reduces to L_x. In fact this approximation is accurate to 5% for $|\epsilon| < 0.1$. We now derive a sufficient condition on x, given L, N and λ, such that $\frac{L_x}{V_x} < 0.1$. Now $\frac{(N-x)!}{N!} = \frac{1}{N(N-1)(N-2)...(N-x+1)} \leq \left(\frac{1}{N-x+1}\right)^x$, so $\frac{L_x}{V_x} = \frac{e^{-\lambda}\lambda^x L}{3^x}\frac{(N-x)!}{N!} \leq Le^{-\lambda}\left(\frac{\lambda}{3(N-x+1)}\right)^x$. For $x \gg \lambda$, $P(x)$ and hence $L_x = P(x) \times L$ become vanishingly small and $\frac{L_x}{V_x} \ll 1$ trivially[3]. Hence, since $\lambda \ll N$, we need only consider $x \ll N$ or, for concreteness, let us assume that $x < \frac{N}{3}$. Then $N - x + 1 > N - x > \frac{2N}{3}$ and $\frac{L_x}{V_x} < Le^{-\lambda}\left(\frac{\lambda}{2N}\right)^x$. Again since $\lambda < N$, this decreases as x increases. Also

$$\begin{aligned} & Le^{-\lambda}\left(\tfrac{\lambda}{2N}\right)^x &= 0.1 \\ \Leftrightarrow\quad & x \ln \tfrac{\lambda}{2N} &= \ln \tfrac{0.1}{Le^{-\lambda}} \\ \Leftrightarrow\quad & x &= \left(\lambda + \ln \tfrac{0.1}{L}\right) / \ln \tfrac{\lambda}{2N} \\ & &\approx \tfrac{\lambda - 2.30 - \ln L}{\ln \lambda - 0.69 - \ln N}. \end{aligned} \tag{8}$$

Thus for all $x > \frac{\lambda - 2.30 - \ln L}{\ln \lambda - 0.69 - \ln N}$, $\frac{L_x}{V_x} < 0.1$ and $C_x \approx L_x$. We shall denote this threshold by x_u, i.e.

$$x_u = \left(\lambda + \ln \frac{0.1}{L}\right) / \ln \frac{\lambda}{2N} \tag{9}$$

For example, for $L = 10^{12}$, $\lambda = 4$ and $N = 1000$, $x_u = \frac{4 - 2.30 - \ln 10^{12}}{\ln 4 - 0.69 - \ln 1000} \approx 4.2$ (Figure 1).

Note that x_u is mostly dependent on λ. In particular for large λ (e.g. $\lambda > 30$ in the above example) x_u becomes negative (the denominator is always negative since $\lambda < N$) and then $C_x \approx L_x$ for all x (Figure 2). This may be inferred from the $e^{-\lambda}$ factor in $\frac{L_x}{V_x} = \frac{e^{-\lambda}\lambda^x L}{3^x}\frac{(N-x)!}{N!}$ which goes rapidly to 0 with increasing λ. On the other hand, as $\lambda \to 0$, x_u becomes $\frac{-2.30 - \ln L}{\ln \lambda - 0.69 - \ln N}$, where the magnitude of the denominator increases with decreasing λ and again x_u decreases. This may also be inferred from the equation for $\frac{L_x}{V_x}$: as $\lambda \to 0$, $e^{-\lambda} \to 1$ and $\lambda^x \to 0$. In fact the only ways to make x_u larger are to increase L or decrease N – though the condition $\lambda \ll N$ constrains the latter to some extent – for example, for $L = 10^{20}$, $N = 100$ and $\lambda = 1$, $x_u \approx 8.9$. Figure 2 illustrates the dependence of x_u on L, N and λ. It is clear that for reasonable values of L, N and λ, $x_u <\sim 10$.

[3] This can be shown rigorously – but is somewhat involved – and is valid provided N is sufficiently large or λ is sufficiently small (e.g. for $\lambda \leq 0.1N$ and $x \geq \frac{N}{3}$, we require $N >\sim 20$ for $L <\sim 10^8$ and $N >\sim 30$ for $L <\sim 10^{14}$ etc.). These limitations do not, however, apply to the programme `PEDEL` (see below), in which $\frac{L_x}{V_x}$ is calculated directly rather than using x_u and x_l (see below).

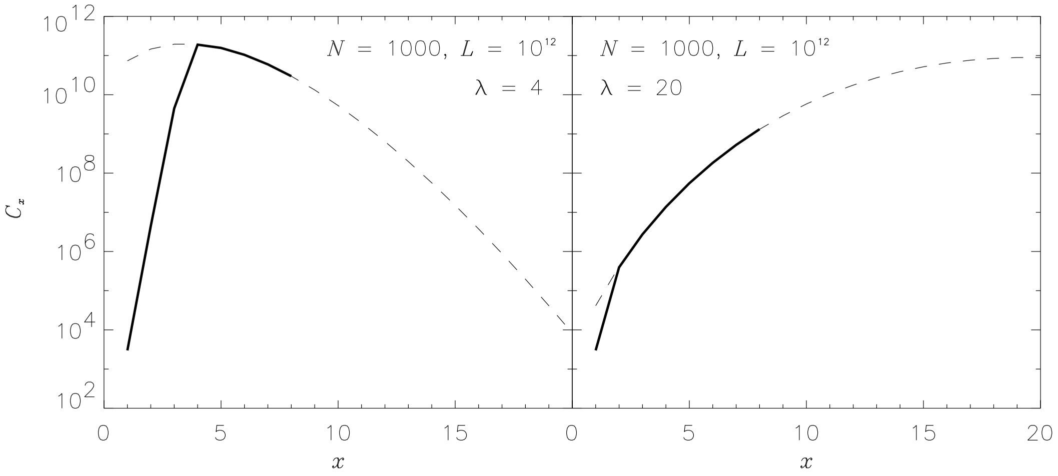

Figure 1: Plot of C_x (solid bold lines) – the expected number of *distinct* sequences with exactly x point mutations, calculated using equation 6, for a library size $L = 10^{12}$, sequence length $N = 1000$, and mean mutation rate $\lambda = 4$ (left) and $\lambda = 20$ (right). Also plotted is L_x (thin dashed lines) – the total expected number of sequences with exactly x point mutations. Using equation 9, $x_u = 4.2$ (left) and 2.2 (right) while equation 16 gives $x_u' = 3.9$ and 1.98 respectively. For $x \geq x_u'$, C_x is very well approximated by L_x. In the case $\lambda = 20$, $C = \sum_{i=0}^{\infty} C_x$ is dominated by the region where the L_x approximation holds.

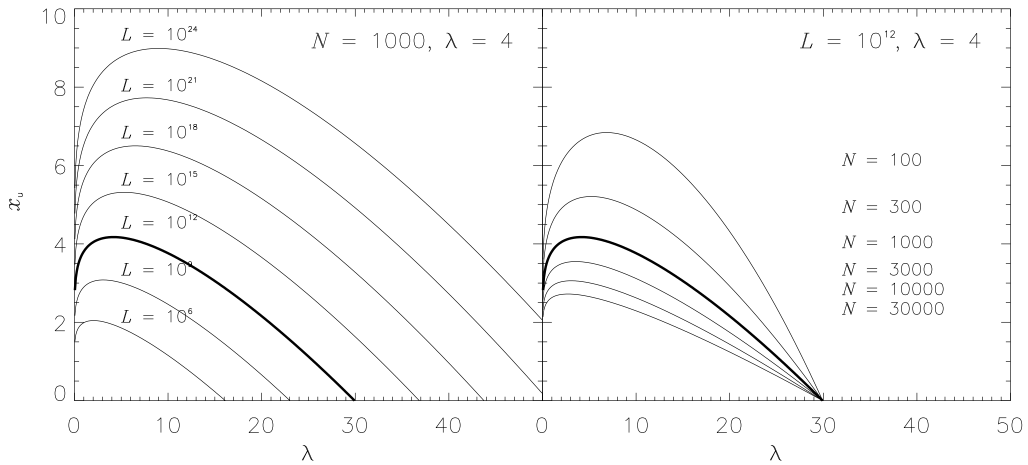

Figure 2: Plot of the threshold value $x_u = (\lambda + \ln \frac{0.1}{L}) / \ln \frac{\lambda}{2N}$ as a function of mutation rate λ for different library sizes L (left) and different sequence lengths N (right). The bold lines correspond to the case $L = 10^{12}$ and $N = 1000$. For large λ, x_u becomes negative; also x_u increases with increasing L and decreasing N, but for all practical library sizes and for $N >\sim 100$, $x_u <\sim 10$.

Conversely, for $\frac{L_x}{V_x}$ sufficiently large, $1 - e^{-L_x/V_x} \approx 1$ and $C_x = V_x(1 - e^{-L_x/V_x})$ reduces to V_x. In fact this approximation is good to 5% for $\frac{L_x}{V_x} > 3$. We now derive a sufficient condition on x, given L, N, and λ, such that $\frac{L_x}{V_x} > 3$. Since $\frac{(N-x)!}{N!} = \frac{1}{N(N-1)(N-2)\ldots(N-x+1)} \geq \left(\frac{1}{N}\right)^x$, we have $\frac{L_x}{V_x} = \frac{e^{-\lambda}\lambda^x L}{3^x}\frac{(N-x)!}{N!} \geq Le^{-\lambda}\left(\frac{\lambda}{3N}\right)^x$, which increases as x decreases. Also

$$\begin{array}{rrcl} & Le^{-\lambda}\left(\frac{\lambda}{3N}\right)^x & = & 3 \\ \Leftrightarrow & x\ln\frac{\lambda}{3N} & = & \ln\frac{3}{Le^{-\lambda}} \\ \Leftrightarrow & x & = & \left(\lambda + \ln\frac{3}{L}\right)/\ln\frac{\lambda}{3N} \\ & & \approx & \frac{\lambda+1.10-\ln L}{\ln\lambda-1.10-\ln N}. \end{array} \tag{10}$$

Thus for all $x < \frac{\lambda+1.10-\ln L}{\ln\lambda-1.10-\ln N}$, $\frac{L_x}{V_x} > 3$ and $C_x \approx V_x$. We shall denote this threshold by x_l, i.e.

$$x_l = \left(\lambda + \ln\frac{3}{L}\right)/\ln\frac{\lambda}{3N} \tag{11}$$

For example, for $L = 10^{12}$, $\lambda = 4$ and $N = 1000$, $x_l = \frac{4+1.10-\ln 10^{12}}{\ln 4-1.10-\ln 1000} \approx 3.4$.

In order to calculate the expected total number C of distinct sequences in $\mathcal{L}$, given by equation 7, we only need to calculate C_x in full (using equation 6) for x values where $0.1 < \frac{L_x}{V_x} < 3$. We now show that in most cases there is only one such x value. Now

$$\frac{V_{x+1}}{V_x} = \frac{3^{x+1}N!}{(x+1)!(N-x-1)!}\frac{x!(N-x)!}{3^x N!} = \frac{3(N-x)}{x+1} \tag{12}$$

and

$$\frac{L_{x+1}}{L_x} = \frac{e^{-\lambda}\lambda^{x+1}L}{(x+1)!}\frac{x!}{e^{-\lambda}\lambda^x L} = \frac{\lambda}{x+1} \tag{13}$$

so

$$\frac{L_{x+1}/V_{x+1}}{L_x/V_x} = \frac{L_{x+1}}{L_x}\frac{V_x}{V_{x+1}} = \frac{\lambda}{x+1}\frac{x+1}{3(N-x)} = \frac{\lambda}{3(N-x)}. \tag{14}$$

For $x < N - \frac{\lambda}{3}$ we have $\frac{\lambda}{3(N-x)} < 1$, in which case $\frac{L_x}{V_x}$ decreases as x increases. Since $\lambda \ll N$ this will be the case for almost all x and in particular for $x \leq x_u$. If $\frac{L_{x+1}/V_{x+1}}{L_x/V_x} < \frac{1}{30}$ between $x = x_u$ and $x = x_l$ then there will be at most one x value with $0.1 < \frac{L_x}{V_x} < 3$. But

$$\frac{L_{x+1}/V_{x+1}}{L_x/V_x} \geq \frac{1}{30} \Rightarrow \lambda \geq \frac{3(N-x)}{30} = \frac{N-x}{10} \approx \frac{N}{10} \tag{15}$$

where in the last approximation we have assumed $x < x_u \ll N$ in general. So for $\lambda < \frac{N}{10}$, at most one C_x value need be calculated in full (and with the initial assumption that $\lambda \ll N$, this will generally be the case).

While we have shown that there is generally at most one x value with $0.1 < \frac{L_x}{V_x} < 3$, it is not necessarily the case that there is at most one x value with $x_l < x < x_u$ since we used conservative approximations in the derivations of x_u and x_l. Since we have shown (Figure 2) that, for all intents and purposes, $x_u <\sim 10$ for reasonable values of N (e.g. $N >\sim 100$), for x in the vicinity of x_u and x_l, we may approximate $\frac{(N-x)!}{N!} = \frac{1}{N(N-1)(N-2)\ldots(N-x+1)}$ by $\frac{1}{N^x}$, in which case the condition $\frac{L_x}{V_x} = 0.1$ implies $x = (\lambda + \ln\frac{0.1}{L})/\ln\frac{\lambda}{3N}$. So in most cases we may replace x_u in equation 9 with

$$x'_u = \left(\lambda + \ln\frac{0.1}{L}\right)/\ln\frac{\lambda}{3N} \tag{16}$$

and provided $\lambda <\sim \frac{N}{10}$, $x'_u - x_l < 1$.

Using the above, equation 7 reduces to

$$\begin{array}{rcl} C & = & \sum_{x=0}^{\infty} C_x \\ & \approx & 1 + \sum_{x=1}^{s_1} V_x + \sum_{x=s_1+1}^{s_2-1} C_x + \sum_{x=s_2}^{\infty} L_x \\ & = & 1 + \sum_{x=1}^{s_1} V_x + \sum_{x=s_1+1}^{s_2-1} C_x + L - \sum_{x=0}^{s_2-1} L_x \\ & = & 1 + \sum_{x=1}^{s_1} V_x + \sum_{x=s_1+1}^{s_2-1} C_x + L \times (1 - \sum_{x=0}^{s_2-1} P(x)) \end{array} \tag{17}$$

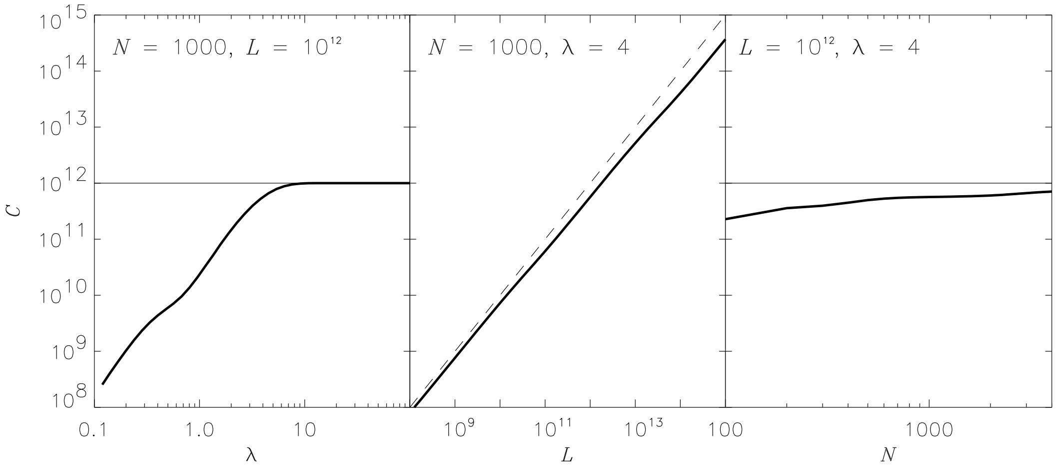

Figure 3: Plot of the total expected number of distinct sequences C (bold lines) as a function of mutation rate λ (left), library size L (centre) and sequence length N (right). **Left:** For $N = 1000$ there are 4^{1000} possible distinct sequences and as λ increases a greater number of these are sampled. The C versus λ plot levels off when C is limited by the total library size L (thin line). **Centre:** Even with only 4 (i.e. λ) mutations there are $\sim 3 \times 10^{12}$ possible sequences while with 8 (i.e. 2λ) mutations there are $\sim 2 \times 10^{23}$ possible sequences. Thus for $\lambda = 4$, even in very large libraries, the degree of redundancy is fairly low and C is of the same order as L. The dashed line plots $C = L$ for reference. **Right:** C changes very little with N for a fixed value of λ (in which case the mean mutation rate *per nucleotide* scales inversely with N). For $N = 100$ there are $\sim 10^{15}$ possible sequences with 8 mutations, so a library of size 10^{12} (thin line) still has a fairly low (factor of ~ 4) degree of redundancy.

where $s_1 = \lfloor x_l \rfloor$ and $s_2 = \lceil x_u \rceil$ are the greatest integer less than x_l and the least integer greater than x_u respectively, V_x is calculated using equation 5, C_x is calculated using equation 6 (if required, i.e. if $s_1 + 1 \not> s_2 - 1$), and $P(x)$ is calculated using equation 4. Note that the last term in the sum can not be ignored – it dominates for larger λ (e.g. see Figure 1).

Thus, in order to calculate the total expected number of distinct sequences C in a library of size L, given N and λ, we first use equations 9 (or 16) and 11 to calculate x_u and x_l and then we calculate the sum in equation 17. Examples are shown for various L, N and λ in Figure 3. A simple `C++` programme, dubbed `PEDEL`, to calculate C, P_x, L_x, V_x and C_x, for given λ, N and L, may be found at the web address given on page 1. While x_u and x_l are conceptually useful, `PEDEL` simply calculates C_x using equation 6 until $\frac{L_x}{V_x} < 0.1$ (at $x = s$, say) and then uses $L \times (1 - \sum_{x=0}^{s} P(x))$ for the remaining C_x.

3 Recombination of near-identical sequences

Problem: Given a library $\mathcal{L}$ of L sequences generated by random recombination of two near-identical genes differing at only a small number of known nucleotide (or codon) positions, we wish to calculate the expected number of distinct variants in $\mathcal{L}$.

As an example, Raillard et al. (2001) start with two sequences of 1425 nucleotides differing at 9 positions, giving a total of $2^9 = 512$ possible daughter variants. They screened a library of $L = 1600$ shuffled variants. It is of interest to calculate what proportion of the 512 possible variants we may expect the library to contain. Clearly this will depend upon (a) the mean number of crossovers introduced into the daughter sequences and (b) the spacing of the variable nucleotides, since nucleotides that are closely spaced are less likely to be separated by a crossover than nucleotides that are far apart in the sequence.

Suppose we start with two sequences S and S', each comprising N nucleotides, and suppose that S and S' differ at M nucleotide positions with the varying nucleotides being respectively A_1, A_2,..., A_M and A'_1, A'_2,..., A'_M. For convenience we write $S = A_1A_2...A_M$ and $S' = A'_1A'_2...A'_M$ where we have omitted the intervening nucleotides that do not differ between S and S'. There are 2^M possible daughter variants $D_k = A^k_1A^k_2...A^k_M$ where each $A^k_i = A_i$ or A'_i. It is convenient to relabel the variants, not by the positions of the nucleotides, but by the positions of crossovers. Thus we map each daughter variant D_k onto a binary sequence (sequence of 1's and 0's) $B_k = b_1^k b_2^k ... b_{M-1}^k$ where $b_i^k = 1$ if there is an odd number (1, 3, 5,...) of crossovers between A_i^k and A_{i+1}^k and $b_i^k = 0$ if there is an even number (0, 2, 4,...) of crossovers between A_i^k and A_{i+1}^k. Clearly one crossover between consecutive variable nucleotides produces exactly the same daughter variant as 3, 5, 7,... crossovers, and similarly for an even number of crossovers. As an example, the daughter variant $D_k = A_1 A_2 A_3 A_4' A_5' A_6 A_7 ...$ maps to $B_k = 001010...$. Note that the 'inverse' sequence $D_k' = A_1' A_2' A_3' A_4 A_5 A_6' A_7' ...$, formed from D_k by replacing A_i by A_i' and *vice versa*, maps to the same binary sequence $B_k' = 001010... = B_k$. There are 2^{M-1} such binary sequences, and a one-to-one correspondence between the binary sequences and those daughter variants starting with A_1 and a similar one-to-one correspondence between the binary sequences and those daughter variants starting with A_1'.

Suppose that the mean number of crossovers per daughter sequence is λ and suppose $\lambda \ll N$ (large λ values correspond to small fragments in the reassembly reaction – a situation generally disfavoured by the requirement for annealing of complementary, overlapping fragments). We shall assume that the number of crossovers x in a particular daughter sequence follows a Poisson distribution with mean λ so that

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ... \tag{18}$$

where $P(x)$ is the probability of there being exactly x crossovers in the sequence. Furthermore, we assume that the crossovers in a particular daughter sequence are randomly distributed with the proviso that a crossover cannot occur in the position immediately following a varying nucleotide. By using the Poisson distribution, we are ignoring some of the mechanics of the process (see for example Moore & Maranas 2000; Moore et al. 2001) and so the reader with knowledge of a more accurate probability distribution in his/her own particular situation may wish to modify the equations/results accordingly.

We now calculate the relative probabilities of each binary sequence B_k. As mentioned above, these probabilities will depend on the relative spacing of the variable nucleotides A_i in S. If S has N nucleotides, of which M are varying, then it has $N - M - 1$ potential crossover points. Let n_i be the number of spaces between A_i and A_{i+1} (so that $\sum_{i=0}^{M} n_i = N - 1$, where n_0 and n_M represent the number of spaces before A_1 and after A_M respectively). The probability of there being x crossovers between A_i and A_{i+1} follows a Poisson distribution with mean $\frac{(n_i-1)\lambda}{N-M-1}$. Hence

$$\begin{aligned} P(b_i = 0) &= P(x=0) + P(x=2) + P(x=4) + ... \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \sum_{x=0,2,4,...}^{\infty} \frac{1}{x!} \left(\frac{(n_i-1)\lambda}{N-M-1}\right)^x \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \cosh\left(\frac{(n_i-1)\lambda}{N-M-1}\right) \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \left[\frac{e^{(n_i-1)\lambda/(N-M-1)} + e^{-(n_i-1)\lambda/(N-M-1)}}{2}\right] \\ &= \frac{1}{2}\left[1 + e^{-\frac{2(n_i-1)\lambda}{N-M-1}}\right] \end{aligned} \tag{19}$$

where $\cosh\theta$ is the hyperbolic cosine function defined by $\cosh\theta = \frac{e^\theta + e^{-\theta}}{2}$. The third equality follows from the identity $e^\theta = \sum_{x=0}^{\infty} \frac{\theta^x}{x!}$. Similarly $\sinh\theta$ is the hyperbolic sine function defined by $\sinh\theta = \frac{e^\theta - e^{-\theta}}{2} = \sum_{x=1,3,5,...}^{\infty} \frac{\theta^x}{x!}$. So we have also

$$\begin{aligned} P(b_i = 1) &= P(x=1) + P(x=3) + P(x=5) + ... \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \sum_{x=1,3,5,...}^{\infty} \frac{1}{x!} \left(\frac{(n_i-1)\lambda}{N-M-1}\right)^x \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \sinh\left(\frac{(n_i-1)\lambda}{N-M-1}\right) \\ &= e^{-(n_i-1)\lambda/(N-M-1)} \left[\frac{e^{(n_i-1)\lambda/(N-M-1)} - e^{-(n_i-1)\lambda/(N-M-1)}}{2}\right] \\ &= \frac{1}{2}\left[1 - e^{-\frac{2(n_i-1)\lambda}{N-M-1}}\right]. \end{aligned} \tag{20}$$

For $\frac{2(n_i-1)\lambda}{N-M-1} \ll 1$, i.e. for closely spaced variable nucleotides, this approximates to $\frac{(n_i-1)\lambda}{N-M-1}$ and is small.

For the example of Raillard et al. (2001), $M = 9$, so there are eight $P(b_i = 0)$ and eight $P(b_i = 1)$ probabilities to be calculated. All possible products $P(B_k) = \prod_{i=1}^{8} P(b_i^k) = \prod_{i=1}^{8} e^{-(n_i-1)\lambda/(N-M-1)} \mathrm{sch}\left(\frac{(n_i-1)\lambda}{N-M-1}\right) =$

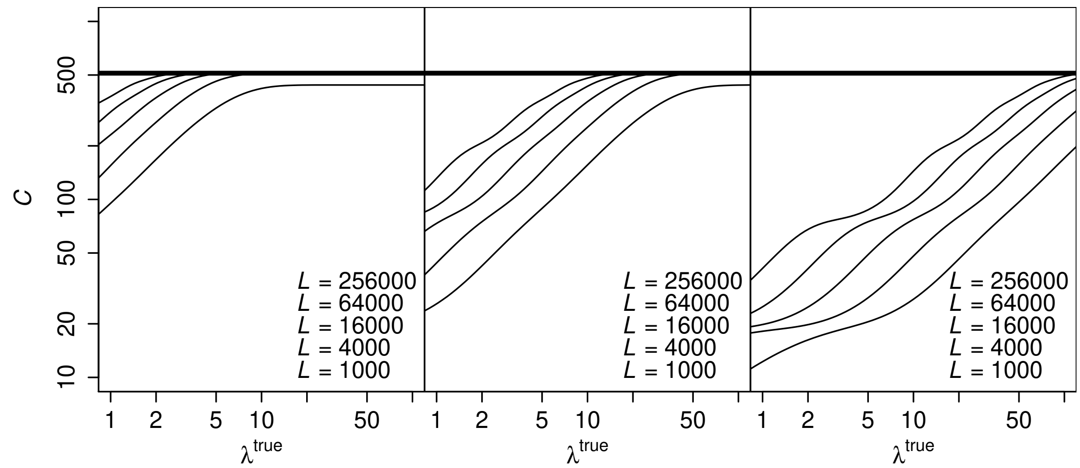

Figure 4: The expected number C (thin lines) of distinct sequences in a library of size L as a function of λ^{true} for libraries of sizes 1000, 4000, 16000, 64000 and 256000 (listed from lower curve to upper curve). In all cases the specified sequence length is $N = 500$ nucleotides and the number of variable nucleotides is fixed at $M = 9$. **Left:** variable nucleotides at positions 50, 100, 150, 200, 250, 300, 350, 400, 450. **Centre:** variable nucleotides at positions 100, 110, 120, 130, 140, 150, 160, 170, 180. **Right:** variable nucleotides 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 nucleotides are well-spaced along the parent sequences. C levels off at the total number of possible distinct sequences (512 in this case, plotted as a bold line) unless constrained by L (e.g. for $L = 1000$, above). For very large λ (especially for well-spaced variable nucleotides), C becomes independent of λ.

$\exp\left(\frac{-\lambda\sum_{i=1}^{8}(n_i-1)}{N-M-1}\right)\prod_{i=1}^{8}\text{sch}\left(\frac{(n_i-1)\lambda}{N-M-1}\right)$, where $\text{sch}\,\theta = \cosh\theta$ if $b_i^k = 0$ and $\text{sch}\,\theta = \sinh\theta$ if $b_i^k = 1$, can then be calculated (there are $2^8 = 256$ possibilities – easily calculable by computer). Each corresponds to the two equiprobable inverse sequences represented by the same binary sequence and, when divided by 2, these 256 products give the relative probabilities $Q_k = \frac{P(B_k)}{2}$ of the 512 possible daughter variants D_k.

Given a particular daughter variant D_k, the probability that a given sequence in the library is not D_k is $1 - Q_k$ so the probability that D_k does not occur at all in a library of size L is $(1 - Q_k)^L$. Hence the probability that D_k does occur in the library is $1 - (1 - Q_k)^L$ and the expected number of sequences in the library is $\sum_{k=1}^{512} 1 - (1 - Q_k)^L$ (or, alternatively and as in §1, $\approx \sum_{k=1}^{512} 1 - \exp(-LQ_k)$.)

In general then, for two parent sequences of length N that differ at M positions, a mean crossover rate of λ and a library of size L, the mean expected number of distinct daughter variants in the library will be

$$\begin{aligned} C &= \sum_k 1 - (1 - Q_k)^L \\ &= 2\sum_{b_1=0}^{1}\sum_{b_2=0}^{1}\cdots\sum_{b_{M-1}=0}^{1}\left[1 - \left(1 - \frac{1}{2}\exp\left(\frac{-\lambda\sum_{i=1}^{M-1}(n_i-1)}{N-M-1}\right)\prod_{i=1}^{M-1}\text{sch}\left(\frac{(n_i-1)\lambda}{N-M-1}\right)\right)^L\right] \end{aligned} \tag{21}$$

where $\text{sch}\,\theta = \cosh\theta$ if $b_i = 0$ and $\text{sch}\,\theta = \sinh\theta$ if $b_i = 1$. C increases with library size and is larger if the variable nucleotides are well-spaced rather than being clustered together (Figure 4).

Up until now we have assumed that $\lambda \equiv \lambda^{\text{true}}$ is known. It is apparent however that many crossovers may not in fact be observable – e.g. two crossovers between two adjacent variable nucleotides *looks* like no crossover at all. Similarly, a crossover at either end of the sequence – not containing any variable nucleotide – is also not observable. In fact the observed crossovers are exactly equivalent to the presence of a '1' in the binary sequence B_k corresponding to a given daughter sequence. Hence the mean number of observed crossovers per sequence is

$$\lambda^{\text{obs}} = \sum_k P(B_k)\rho(B_k) \tag{22}$$

where $\rho(B_k)$ is the number of '1's in the binary sequence B_k. Now $\sum_k P(B_k)\rho(B_k)$ is a sum of the form $\sum_j a_j \exp(b_j \lambda^{\rm true})$ for some values a_j, b_j, which in general is not analytically soluble for $\lambda^{\rm true}$ in terms of $\lambda^{\rm obs}$. However, by trying various values of $\lambda^{\rm true}$ and calculating $\lambda^{\rm obs}$, one may find by trial and error the value $\lambda^{\rm true}$ that reproduces a given $\lambda^{\rm obs}$.

Note that it is impossible to have $\lambda^{\rm obs} > \frac{M-1}{2}$: if $\lambda^{\rm true}$ is very large, then the variable nucleotides will be essentially randomly assigned in each daughter sequence and all possible daughters will be essentially equally likely – giving a 50 per cent probability of an observable crossover between any two variable nucleotides (equation 20, as $\lambda \to \infty$). Since we assume that crossovers cannot occur immediately following a variable nucleotide, if two variable nucleotides are adjacent in the parent sequence, then they will remain linked in all daughter sequences and the number of possible daughter sequences will be reduced accordingly. Once $\lambda^{\rm true}$ is large enough so that all possible daughter sequences are essentially equally likely, increasing $\lambda^{\rm true}$ further will have no effect on the mean number of distinct sequences C in the library. At this point C depends only on L, the library size. The problem essentially reduces to the problem of equiprobable variants of §1.

A simple `C++` programme to calculate C, $P(b_i = 0)$, $P(b_i = 1)$, $P(B_k)$, $\lambda^{\rm obs}$ and $\lambda^{\rm true}$, given N, M, L, either $\lambda^{\rm obs}$ or $\lambda^{\rm true}$, and the positions of the variable nucleotides, may be found at the web address given on page 1. For most intents and purposes the programme may be used equally well on sequences of nucleotides or sequences of codons. The only proviso is that if codons are treated as the basic unit, then one tacitly ignores that crossovers occurring *within* variable codons may or may not lead to an observable crossover, and may in fact lead to new amino acids not present in either parental sequence. In the Raillard et al. (2001) case with $N = 1425$, $M = 9$, $L = 1600$, $\lambda^{\rm obs} \sim 2$ and the positions of the variable nucleotides being 250, 274, 375, 650, 655, 757, 763, 982, 991, `DRIVeR` estimates that $\lambda^{\rm true} \sim 10$, and that the library is expected to contain ~161 distinct variants out of a total of 512 possible variants. The biggest factor leading to the low diversity is the close spacing of the variable nucleotides 650 & 655, 757 & 763, 982 & 991, between which crossovers are unlikely to occur.

References

Feller W., 1968, An Introduction to Probability Theory and its Applications, John Wiley, New York

Moore G. L., Maranas C. D., 2000, *J. Theor. Biol.*, 205, 483

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

Raillard S. et al., 2001, *Chem. Biol.*, 8, 891

Sun F., 1995, *J. Comput. Biol.*, 2, 63