Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biomolecular Engineering 22 (2005) 57–61

Biomolecular Engineering

www.elsevier.com/locate/geneanabioeng

Mathematical expressions useful in the construction, description and evaluation of protein libraries

Allen D. Bosley, Marc Ostermeier *

Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA

Received 29 July 2004; received in revised form 31 October 2004; accepted 1 November 2004

Abstract

The creation of protein libraries by random mutagenesis and cassette mutagenesis has proven to be a successful method of protein engineering. Appropriate statistical analysis is important for the proper construction of these libraries and even more important for the interpretation of data from these libraries. We present simple mathematical expressions useful in the creation and evaluation of such libraries. These equations are useful in estimating the distribution of mutations, the degeneracy of the library and the frequency of a particular clone in the library. In addition, general equations addressing the probability that a particular clone is in a library, the probability that a library is complete, and as the consequences of retransformation of the library on these probabilities are presented.

Keywords: Protein engineering; Directed evolution; Random mutagenesis; Cassette mutagenesis; Genetic libraries

1. Introduction

Fundamental to all methods of directed evolution is the generation of diversity. Two of the simplest and most commonly used methods are random mutagenesis and cassette mutagenesis. The ability to predict parameters such as the frequency of individual clones F_i and the degeneracy D of the library of these methods of library creation is extremely useful. For example, if screening of a library results in no variant with the desired properties, these equations can be used to determine if a sufficient number of library members were screen to have a high probability that all library members were encountered. Also, these equations can give an estimate as to the number of transformants necessary to have a high probability of creating a complete library.

Here we have compiled and developed equations using simple statistics that will be useful in that regard. Our treatment is similar to that of Patrick et al. [1]; however, we focus on equations for determining the degeneracy of the library, the probability of individual clones in individual libraries and the development of general equations for addressing the probability that a clone is present in a library (or in a sampling of a library), the probability that a library is complete and the consequences of propagating a library by retransformation.

2. Results and discussion

2.1. Library size, diversity and degeneracy

2.1.1. Library size

Library size is an often used but ill-defined term. In general, the term is used in a way that is synonymous with the number of transformants, unless the number of transformants greatly exceeds the maximum degeneracy of the library, in which case the library size is equal to this degeneracy. Further complicating the matter is the fact that, for most libraries not all transformants are ‘meaningful’ library members. For example, consider a cassette mutagenesis library constructed by PCR in which four codons were randomized using NNN. The possible degeneracy on the DNA level is 1.78×10^7. On the protein level, the possible

* Corresponding author. Tel.: +1 410 516 7144; fax: +1 410 516 5510.
E-mail address: oster@jhu.edu (M. Ostermeier).

1389-0344/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bioeng.2004.11.002

degeneracy is 1.94 × 10^5. Suppose a library of 4 × 10^5 transformants is created in which 90% of the transformants include a plasmid that receive the insert with the randomized DNA. What is the 'library size?' Rather than giving a library size, it is much more informative to describe the library in terms of number of transformants, the fraction of the transformants that are 'meaningful' library members (as opposed to those members which, for example, did not receive the insert DNA) and the degeneracy of the library.

D	degeneracy of a library (number of distinct sequences)
$D_{\max}$	maximum degeneracy of a library (assuming an infinite number of transformants
$D_{\max,k_m}$	maximum number degeneracy of protein sequences with k_m mutations in a library
F_i	frequency that a particular sequence i is present in a library
H	Hamming distance
k	number of base mutations in a DNA sequence
k_m	number of amino acid mutations in a protein sequence (or number of non-synonymous mutations in a DNA sequence)
L	number of amino acids in a protein sequence
M	number of residues randomized in a random cassette mutagenesis library
n	number of bases in a DNA sequence
P_c	probability that a library is complete
P_i	probability that a particular sequence i is in the library
$P_{i,S}$	probability that a particular sequence i is encountered in a library sample S times
P_k	probability of having k base mutations in a nucleotide sequence
P_{k_m}	probability of having k_m amino acid mutations in a protein sequence
P_{stop}	probability that an internal stop codon is present in a library member
S	number of times a library is sampled
T	number of transformants
T_{k_m}	number of library members that contain k_m mutations
V_H	number of possible variants that are a Hamming distance of H away from the original protein
Greek symbols	
ε	error rate (frequency at which a bases is mutated)
ε_m	non-synonymous error rate (frequency at which an amino acid is mutated)

2.1.2. Library degeneracy and diversity

The degeneracy D of a library is the number of different members among the transformants (i.e. the number of independent clones). The degeneracy of a library depends on the number of transformants T, the probability of occurrence of each specific sequence in the library and the maximum degeneracy $D_{\max}$ that could possibly be generated given the method used to create the library (i.e. the number of different members in a library of an infinite number of transformants). In the case where all variants are equally probable, the actual number of occurrences of any variant can be represented by a Poisson distribution and the degeneracy of the library is calculated as follows [1]:

$$D = D_{\max}(1 - e^{-T/D_{\max}}) \tag{1}$$

A distinction must be made between degeneracy and diversity. Degeneracy describes the number of variants in a library, whereas diversity is a qualitative description of how much, on average, two randomly selected library members will differ. This point is emphasized by noting that two libraries can both have 1 million variants, and therefore the same degeneracy, but in one library the average difference between two randomly picked members may be three amino acids, whereas in the other library the average distance between two randomly picked members may be 10 amino acids [2]. Both libraries are equally degenerate but the latter is more diverse.

2.1.3. Number of possible variants of a protein

A fundamental step of directed evolution is the creation of a library of variants of a starting protein(s). An obvious question that arises is how many variants are there of a particular protein. The Hamming distance H is the number of mutational steps it takes to get from one sequence to another. In other words, it is the number of positions that are different between two proteins. As one increases the number of mutations in a protein, the number of possible variants increases exponentially. The number of variants V_H of a protein of amino acid length L that differ by a Hamming distance of H is given by Eq. (2) as follows:

$$V_H = 19^H\left[\frac{L!}{(L-H)!H!}\right] \tag{2}$$

Even for a rather small protein of $L = 150$, the number of variants with a Hamming distance of two is 1.6 × 10^7 and the number of variants with a Hamming distance of three is 1.3 × 10^{11}. Thus, given the limitations of the number of transformants on library size, for most proteins libraries the construction of a complete library of all variants with $H = 3$ is not feasible. We can see how quickly the potential library size grows and how the likelihood that a complete library can be created diminishes.

2.2. Library generation

2.2.1. Random mutagenesis

While it is generally not feasible to make a complete library of all variants with $H = 3, 4$, etc. random mutagenesis allows for the sampling of this sequence space by introducing random point mutations throughout the entire protein. There are many methods for creating such libraries, the most common being error-prone PCR [3]. Regardless of the method, the rate at which these errors are introduced will dictate the distribution of mutations throughout the library. Eq. (3) [3] gives the probability P_k of having k mutations in a sequence length of n bases, where ε is the error rate per position.

$$P_k = \left[\frac{n!}{(n-k)!k!}\right]\varepsilon^k(1-\varepsilon)^{n-k} \tag{3}$$

In the interest of a general treatment that will allow researchers to make useful approximations, the simplification made here is that all mutations on the nucleotide level occur with equal frequency and that all bases occur with equal frequency in a DNA sequence being considered. In truth, this is method specific. Treatments of this subject without these approximations are much more complex [4] and require knowledge of the specific sequence to be mutated and the method of mutation.

In practice, what is most relevant (for protein libraries) is an estimation of the probability of having a non-synonymous mutation P_{k_m} as follows:

$$P_{k_\mathrm{m}} = \left[\frac{L!}{(L-k_\mathrm{m})!k_\mathrm{m}!}\right](\varepsilon_\mathrm{m})^{k_\mathrm{m}}(1-\varepsilon_\mathrm{m})^{L-k_\mathrm{m}} \tag{4}$$

The variable k_m is the number of non-synonymous mutations and ε_m the non-synonymous error rate. This non-synonymous error rate is defined on the amino acid level. Thus, the non-synonymous error rate is found by multiply the error rate at the nucleotide level by three (since ε_m is defined per codon and ε is defined per nucleotide) and also multiplying by the frequency at which a mutation in a codon results in a non-synonymous mutation (Table 1). Thus, $\varepsilon_\mathrm{m} = 2.10\varepsilon$. Eq. (4) ignores the possibility of two mutations occurring in the same codon. This is a very rare event under typical error rates and thus does not appreciably affect the results. Eq. (4) can be used to construct a table detailing the distribution of mutations throughout the library. For example, if average size protein of 300 amino acids ($n = 900$) is subjected to error-prone PCR with a DNA polymerase having error rate of $\varepsilon = 0.003$ per nucleotide the resulting distribution of mutations is shown in Table 2. This table can be used to, along with the number of transformants T, to determine the number of transformants T_{k_m} that contain k_m non-synonymous mutations.

Table 1
Frequency of mutation types as a function of location within a codon

Mutation type	Base position within codon			Overall
	1st	2nd	3rd	
Synonymous	0.068	0.021	0.812	0.300
Non-synonymous	0.932	0.979	0.188	0.700
Missense	0.886	0.958	0.162	0.663
Nonsense	0.046	0.021	0.026	0.036

Table 2
Distribution of mutations in a hypothetical random mutagenesis library of 10^7 transformants constructed from a protein 300 amino acids long using an error rate of $\varepsilon = 0.003$

k or k_m	P_k	P_{k_m}	T_{k_m}	$D_{\mathrm{max},k_\mathrm{m}}$	F_i
0	0.067	0.150	1.501×10^6	1	0.15
1	0.181	0.286	2.856×10^6	1.89×10^3	1.51×10^{-4}
2	0.245	0.271	2.707×10^6	1.78×10^6	1.52×10^{-7}
3	0.221	0.170	1.704×10^6	1.11×10^9	1.53×10^{-10}
4	0.149	0.080	0.802×10^6	5.20×10^{11}	1.54×10^{-13}
5	0.080	0.030	0.301×10^6	1.94×10^{14}	1.55×10^{-16}
6	0.036	0.009	0.093×10^6	6.00×10^{16}	1.56×10^{-19}
7	0.014	0.003	0.025×10^6	1.59×10^{19}	1.57×10^{-22}

$$T_{k_\mathrm{m}} = P_{k_\mathrm{m}}T \tag{5}$$

This can then be compared to the maximum degeneracy $D_{\mathrm{max},k_\mathrm{m}}$ of variants containing k_m mutations accessible using random mutagenesis.

$$D_{\mathrm{max},k_\mathrm{m}} = \left(\frac{L!}{(L-k_\mathrm{m})!k_\mathrm{m}!}\right)6.3^{k_\mathrm{m}} \tag{6}$$

The number 6.3 appears in Eq. (6) because it is the average number of non-synonymous or nonsense mutations that can be made with one nucleotide mutation in a codon. We calculated this value by examining all 27 one-base mutations for each codon (except for nonsense codons) and noting whether a non-synonymous, nonsense or synonymous mutation occurs. The value 6.3 is the sum of all non-synonymous and nonsense mutations divided by the number of codons examined (61).

The degeneracy D of the entire library is found summing the degeneracies of each sub-library (calculated using Eq. (1)) as follows:

$$D = \sum_{k_\mathrm{m}=0}^{\infty} D_{k_\mathrm{m}} = \sum_{k_\mathrm{m}=0}^{\infty} D_{\mathrm{max},k_\mathrm{m}}(1 - \mathrm{e}^{-T_{k_\mathrm{m}}/D_{\mathrm{max},k_\mathrm{m}}}) \tag{7}$$

The degeneracy of the hypothetical library of Table 2 is about 4.3×10^6, or 43% of the number of transformants. The degeneracy in this library is lower than the number of transformants because 15% of the library has no synonymous mutation and members with 1 or 2 non-synonymous mutations appear multiple times among the transformants. One can also calculate the expected average frequency F_i of a particular sequence i with k_m mutations as follows:

$$F_i = \frac{T_{k_\mathrm{m}}/D_{\mathrm{max},k_\mathrm{m}}}{T} \tag{8}$$

When performing error-prone PCR, factors that can alter the error-rate include the relative amounts of dNTP's and the concentration of the template DNA. However, most commonly the error rate is controlled by the $\mathrm{MnCl_2}$ concentration. The error rate can be determined by extensive sequencing of the library or by existing data on the literature on the relationship between PCR conditions and error rate (e.g. Shafikhani et al [5]). In addition, the mutation rate during PCR has been estimated by computational methods [6].

2.2.2. Cassette mutagenesis

Cassette mutagenesis is a method of library creation in which a particular region or regions is targeted for mutagenesis. Generally, the library is created through the use of degenerate oligonucleotides aimed at introducing a predetermined degeneracy into the protein at particular regions. Obviously, one can make libraries in which the diversity at each position varies in any number of different ways. What is covered here is libraries in which the positions are completely, or almost-completely randomized. The simplest approach is to create a library from oligonucleotides in which an equimolar mixture of the four nucleotides is used at each position (NNN libraries). However, using this method there is the possibility of encoding a stop codon within the target sequence. In a library in which M residues are randomized using NNN nucleotides, the probability of a stop codon being present in a sequence is given by Eq. (9) as follows [7]:

$$P_{\text{stop}} = 1 - \left(1 - \frac{3}{64}\right)^{M} \tag{9}$$

To minimize the likelihood of stop codons within the library, a nucleotide mixture of NNB (where B = not A (C or G or T)) can be used. This allows only one possible stop codon of the 48 possible codons while still allowing all possible amino acids to be coded. The probability of a stop codon appearing is reduced to that given by Eq. (10) as follows:

$$P_{\text{stop}} = 1 - \left(1 - \frac{1}{48}\right)^{M} \tag{10}$$

At the DNA level, the maximum degeneracy for an NNN library is $D_{\max} = 4^{n}$ and for a NNB library is $D_{\max} = 4^{2n/3}3^{n/3}$. The degeneracy of the library can be found using Eq. (1), since all DNA sequences can be assumed to be equally probable. On the protein level, the maximum degeneracy of any random cassette mutagenesis library is $D_{\max} = 21^{M}$. However, not all sequences are equally prevalent due to the different number of codons that code for each amino acid and thus Eq. (1) cannot be used to determine the degeneracy of the library. The difference between the most common library member (e.g. variable positions are all arginine, serine or leucine) to the least common member (e.g. a library member in which all variable positions are tryptophan) can be enormous. The difference in frequency increases exponentially with the number of positions varied (6^{M} for NNN libraries and 5^{M} for NNB libraries). This bias may actually be viewed as an advantage since nature has found this distribution of codons to be advantageous for creating proteins.

2.3. Library completeness

2.3.1. What is the probability that a particular sequence is in the library?

The probability P_i of a particular sequence i being in a library is given by Eq. (11) and depends on the number of transformants T and the frequency F_i at which that library member is expected to be present in the library [8].

$$P_i = 1 - (1 - F_i)^{T} \tag{11}$$

F_i is the product of the frequency that i is expected to be present considering the method used to create the library and the frequency of 'meaningful' library members (as opposed to, for example, those not receive the insert).

At the outset of creating a library it is useful to determine the size of library to have in order to have a probability P_i of having a particular sequence i in the library. This can be approximated by Eq. (12) when $(1-F_i)$ is close to one.

$$T = -\frac{\ln(1 - P_i)}{F_i} \tag{12}$$

From this equation one can show that in order to have a 99% probability of having a particular member in the library, the product of T and F_i must be ≥4.6. For a library in which all members are equally frequent, this means that the number of transformants must exceed the maximum degeneracy by a factor of 4.6 in order to have a ≥99% probability that a particular member is present in the library.

2.3.2. What is the probability that a particular sequence was encountered in a sampling of a library?

The probability $P_{i,\ S}$ that a particular sequence i was encountered in a sampling of the library depends on the probability that the library contained the sequence to begin with (P_i) and the probability that it will be encountered in sampling the library S times as given by Eq. (13) as follows:

$$P_{i,S} = P_i(1 - (1 - F_i)^{S}) \tag{13}$$

2.3.3. What is the probability that a library is complete?

It is never correct to state that a library is complete, rather the probability that a library is complete should be stated. This is found by taking the product of the probabilities that each particular library member is present in the library.

$$P_{\mathrm{c}} = \prod_{i=1}^{D} P_i \tag{14}$$

where P_i is evaluated for each member of the library by Eq. (11). If the probability of occurrence of each member of the library is equal, then Eq. (14) reduces to Eq. (15) [9].

$$P_{\mathrm{c}} = [1 - (1 - F)^{T}]^{D_{\max}} \tag{15}$$

For such a library, it is useful to rearrange and simplify Eq. (15), as shown in Eq. (16), in order to calculate the number of transformants needed in order to have a certain probability that the library is complete.

$$T = \frac{\ln(1 - P_c{}^{1/D})}{\ln(1 - F)} \approx \frac{\ln(1 - P_c{}^{1/D})}{-F} \quad (16)$$

For typical degeneracies (10^3 to 10^7), the number of transformants must exceed the degeneracy by a factor on the order of 10–25 in order to have a $\geq 99\%$ probability of having a complete library. However, it should be noted that rarely is the frequency of each member of the library the same and, thus, the simplification of Eqs. (14)–(16) can rarely be made. Even in the case where the Eq. (16) is not mathematically justified, it is still a useful, quick lower estimate of the number of transformants needed to have a certain probability of a complete library. If some library members are more rare than the frequency used to calculate T in Eq. (16), then the T calculated will be underestimated. For libraries created through random mutagenesis, it is easiest to view the resulting library as several sub-libraries differentiated by the number of mutations and use the above equations to determine the probability of completion of each sub-library.

2.3.4. Propagation of a Library

If a library present on a plasmid is to be retransformed into a new host or sub-cloned into a new vector, this must be taken into account when considering the probabilities described above. Basically, the probability must be the product of the probability in the original library and the probability associated with creation of the sub-cloned or retransformed library. The calculation of this product is analogous to the probability associated with sampling a library (Eq. (13)) as the second set of transformants can be viewed as a sampling of the first library.

For example, consider a library of 10^6 transformants in which the maximum degeneracy is 8×10^5. If we make the simplification that all transformants are 'meaningful' library members and that each library member of the degeneracy has an equal probability of occurrence, the probability that a particular sequence is present is 0.714 (from Eq. (11)). If this library were then sub-cloned into a new plasmid resulting in 2×10^6 transformants, the probability that a particular sequence is present would decrease (to $0.714 \times 0.917 = 0.655$) even though the number of transformants in the sub-cloned library was two-fold higher than the original library. It follows also that this sub-cloned library is less diverse than the original library. This emphasizes the point that in order to maintain the number of unique members of the original library upon sub-cloning or retransformation can require a number of transformants that greatly exceeds the number of transformants in the original library.

2.4. Other sources and online programs

Patrick et al. [1] have developed similar mathematical expressions on some of the topics presented here and, in addition, they offer a treatment of libraries created by in vitro recombination. These equations are the basis of computer programs available online at http://www.bio.cam.ac.uk/~blackburn/stats.html.

Acknowledgement

This work was supported by a grant from the National Science Foundation (#BES-0239088).

References

[1] Patrick WM, Firth AE, Blackburn JM. Protein Eng 2003;16:451–7.
[2] Sun F. J Comput Biol 1995;2:63–86.
[3] Caldwell RC, Joyce GF. Mutageneic PCR. In: Dieffenbach C, editor. PCR Primer: A Laboratory Manual. Plainview, NY: Cold Spring Harbor Laboratory Press, 1995.
[4] Moore GL, Maranas CD. J Theor Biol 2000;205:483–503.
[5] Shafikhani S, Siegel RA, Ferrari E, Schellenberger V. Biotechniques 1997;23:304–10.
[6] Wang D, Zhao C, Cheng R, Sun F. J Comput Biol 2000;7:143–58.
[7] Steipe B. Curr Top Microbiol Immunol 1999;243:55–86.
[8] Clarke L. Carbon J Method Enzymol 1979;68:396–408.
[9] Zilsel J, Ma PH, Beatty JT. Gene 1992;120:89–92.