BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 15 2005, pages 3314–3315
doi:10.1093/bioinformatics/bti516

Sequence analysis

Statistics of protein library construction

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

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

Received on April 14, 2005; revised on May 21, 2005; accepted on May 23, 2005
Advance Access publication June 2, 2005

ABSTRACT

Summary: We have investigated the statistics associated with constructing and sampling large protein-encoding libraries. Using fairly simple statistics we have written algorithms for estimating the diversity in libraries generated by the most commonly used protocols, including error-prone PCR, DNA shuffling, StEP PCR, oligonucleotide-directed randomization, MAX randomization, synthetic shuffling, DHR, ADO and SISDC.

Availability: Web interface and C++ source code available at http://guinevere.otago.ac.nz/stats.html

Contact: aef@sanger.otago.ac.nz

Supplementary information: Complete mathematical notes, model assumptions and justification, users' guide and worked examples at above website.

INTRODUCTION

Directed evolution is a powerful strategy for generating new proteins with desirable properties. Central to the technique is the generation of large sequence libraries. There are a number of methods for generating molecular diversity in these libraries (reviewed by Lutz and Patrick, 2004). However, to maximize the chances of finding a desired and rare improved variant, it is important to understand the statistics of library construction.

Previously, we introduced a suite of algorithms for calculating library statistics for a variety of protocols. Since then, the equations and programs have been used a number of times (e.g. Hughes *et al*., 2005). However, the programs were a little unwieldy and required compiling by the user. In this short paper we present an improved and easy-to-use web interface, which will return a variety of library statistics and graphics for user-defined library sizes, mutation rates, sequence lengths, etc. These statistics may be used to direct experimental design (e.g. to determine what library size is required to sample a given amount of diversity, or to optimize the mutation rate to maximize diversity) and to interpret results (e.g. by estimating how many distinct sequences are represented in a given library).

We note that more detailed models of some of the processes involved in library construction have been published (reviewed by Moore and Maranas, 2004). However, these models are not generally accessible to most laboratory researchers, can be CPU-intensive, and are less widely applicable than the generic tools that we present here.

The web interface is available at http://guinevere.otago.ac.nz/stats.html. Users are referred to our original paper (Patrick *et al*., 2003) for experimental details, usage examples and a few caveats. Users interested in the mathematics behind the programs are invited to read the mathematical notes on our website.

In the remainder of this short paper, we introduce the three main programs, `GLUE`, `PEDEL` and `DRIVeR`, and list situations in which they may be useful.

EQUALLY PROBABLE VARIANTS

The simplest program, `GLUE`, is broadly applicable to any protocol where all possible variants are equally likely to occur in the library. Examples include oligonucleotide-directed randomization, MAX randomization, synthetic shuffling, DHR, ADO and SISDC.

Given the total number of possible variants, `GLUE` may be used to calculate (1) the expected number of distinct variants represented in a given library, (2) the library size required to sample a given fraction of the variants or (3) the library size required to have a given probability of sampling all possible variants. For example, if there are 1 million possible variants (e.g. an oligonucleotide-directed randomization involving four NNK codons allows $32^4 = 1\,048\,576$ variants), `GLUE` shows that a library of ∼3 million transformants will be ∼95% complete, while a library of ∼17 million transformants has a ∼95% probability of being 100% complete.

ERROR-PRONE PCR (epPCR)

In this protocol, random base substitutions are introduced into a parent sequence. 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.

The program `PEDEL` can be used to calculate the expected number of distinct variants present in a library, given the library size, mean substitution rate and parent sequence length. On the web page, the user may produce plots of the expected number of distinct daughter sequences as a function of library size and substitution rate. The user can also produce statistics and plots for the total number of variants with exactly x mutations, the expected size of the sub-library comprising those sequences with exactly x mutations, the completeness of each sub-library, and the redundancy of each sub-library.

For example, given a library of 10^7 clones, a parent sequence length of 600 nt, and a mean substitution rate of 2 bases per daughter sequence, `PEDEL` calculates that the library is expected to contain $\sim 4.5 \times 10^6$ distinct sequences. These comprise ∼1.3, ∼1.8, ∼0.9, ∼0.4 and ∼0.1 million distinct sequences with, respectively, exactly 2, 3, 4, 5 and 6 mutations, together with the parent sequence, the

*To whom correspondence should be addressed.

1800 distinct sequences with exactly 1 mutation, and $\sim 4.5 \times 10^4$ sequences with >6 mutations. The rest of the 10^7 clones break down into $\sim$1.4, $\sim$2.7 and $\sim$1.4 million redundant sequences with, respectively, exactly 0, 1 and 2 mutations.

PEDEL uses a generic Poisson model of sequence mutations. All base substitutions are assumed equally likely. In reality, polymerases favour some substitutions over others. This will reduce the number of distinct sequences compared with the PEDEL predictions. The effect is limited by the fact that, for low substitution rates, the library tends to saturate all possible variants while, for high substitution rates, there are so many possible variants that, even with substitution bias, nearly every library member is distinct (Patrick *et al*., 2003). Note that, by using sequential PCR amplifications with two different polymerases with opposite substitution biases, it is possible to produce unbiased libraries.

Another possible source of bias results from the uneven representation of mutations introduced early and late in the epPCR process. However, in practice one might use 10^9 identical template sequences, amplify them to perhaps 10^{15} product molecules in the epPCR, but usually only end up with a library of $<10^7$ variants after ligation and transformation. Under such conditions, amplification bias would have a typical frequency of 1 in 10^9, and would be undetectable in the final library of 10^7. In addition, different parent molecules may be copied a different number of times, but empirically the end result is a library with a Poisson distribution of mutations (Cadwell and Joyce, 1992).

DNA SHUFFLING AND StEP PCR

The program DRIVeR is applicable to libraries generated by recombining two highly homologous parent sequences differing at only a few (e.g. $\leq$20) nucleotide or amino acid positions. Examples include DNA shuffling and StEP PCR. (Note that GLUE should be used for synthetic shuffling, ADO, DHR and SISDC since, in these protocols, all daughter sequences are equally likely.)

The number of distinct daughter sequences depends on library size, the mean crossover rate and the spacing of the positions that vary between the two parent sequences. Closely spaced variable positions tend to remain linked in daughter sequences, resulting in reduced library diversity.

Given the library size, parent sequence length, mean crossover rate and the positions of the variable nucleotides (or amino acids), DRIVeR calculates the expected number of distinct daughter sequences in the library. On the web page, the user may also produce plots of the expected number of distinct daughter sequences as a function of library size and crossover rate. For example, for a sequence length of 1425 nt, nine variable nucleotides at positions 250, 274, 375, 650, 655, 757, 763, 982 and 991, a library of size 1600, and a mean crossover rate of 10 crossovers per sequence, the expected number of distinct sequences in the library is 161 (out of 512 possible variants).

DRIVeR uses a generic Poisson model for crossover positions. The parent sequences are assumed to be highly homologous. For parent sequences that are homologous at the amino acid level but divergent at the nucleotide level, crossovers preferentially occur in regions with greater nucleotide sequence similarity. This bias is not reflected in the DRIVeR model which, nevertheless, provides a useful upper bound on library diversity.

ACKNOWLEDGEMENTS

This work was funded in part by the New Zealand Foundation for Research, Science and Technology (grant number UOOX0304).

Conflict of Interest: none declared.

REFERENCES

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

Hughes,M.D. *et al*. (2005) Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: identification of novel zinc finger proteins. *Nucleic Acids Res.*, **33**, e32.

Lutz,S. and Patrick,W.M. (2004) Novel methods for directed evolution of enzymes: quality, not quantity. *Curr. Opin. Biotechnol.*, **15**, 291–297.

Moore,G.L. and Maranas,C.D. (2004) Computational challenges in combinatorial library design for protein engineering. *AIChE J.*, **50**, 262–272.

Patrick,W.M. *et al*. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein–encoding libraries. *Protein Eng.*, **16**, 451–457.