



# Technical Bulletin 1

Randomized DNA libraries



# Technical Bulletin 1

## Synthesis

The success of a DNA library depends on a well planned design strategy. The aim is to find a good compromise between the desired number of amino acid variations that are to be explored and a sequence that can be synthesized and handled efficiently.

Essentially there are two aspects to a DNA library design: Synthesis of the library and representation of all desired variants. As for synthesis, the randomized positions must be laid out in a way that still allows the assembly of the full-length gene. For instance, clustering too many consecutive randomized positions will hinder synthesis. We recommend not more than eight consecutive randomized nucleotides.

...ATTCGNNNNNNNNCTT... ✓  
...ATTCGNNNNNNNNNNCTT... ✗

(Fig. 1: Upper row shows recommended randomized nucleotides, lower row shows not recommended randomized nucleotides. We use the IUPAC codes for wobble nucleotides, see Appendix.)

Even then, randomized nucleotides should be separated by at least 20 nucleotides. You can organize randomized nucleotides in blocks.

...ATT **NNANNCN** TTGATTGCCTACCAGTTAGCGGAT **NNTNNCNN** ATTG...

(Fig. 2: Minimum separation of randomized nucleotides in blocks.)

A common strategy is to randomize the first two positions of a codon and leave the third one constant. However, please note that interspersing fixed nucleotides at every third position does not make synthesis much easier, therefore such a stretch of NNX codons is considered a consecutive block of randomized nucleotides (see again Fig. 2).



## Technical Bulletin 1

In most cases, the aim is to represent a set of amino acids at a given position. Please note that for most sets of amino acids, it is impossible to avoid nucleotide combinations that encode additional amino acids or even stop codons. For example, to represent all 20 amino acids, the best combination is NNK or NNS, which unfortunately codes for the amber stop codon TAG as well.

Using a subset of the four nucleotides, such as G and T (IUPAC code “K”), can reduce the complexity of the library significantly and is therefore advisable where possible. However, please keep in mind that two-nucleotide combinations are almost as difficult in synthesis as four-nucleotide combinations.

The total number of random nucleotide positions in a given sequence is limited both by synthesis and by the screening phase. For synthesis, the limit depends on the exact layout of the randomized positions, but is generally in the range of 10-20 nucleotides. For screening, the limit depends on how well the screening protocol can be automated, whether there is a selection step that works without manual intervention and how many DNA molecules can be synthesized and processed.

The overall length of a randomized DNA sequence should also be limited. Very long sequences are difficult to synthesize even without wobble positions. This difficulty affects the quality of the DNA library. In some cases, very long genes with random positions cannot be correctly assembled at all. A good rule of thumb is to limit randomized genes to 800-1000 bp in length.



## Technical Bulletin 1

### Non-uniform nucleotide distributions

By default, all nucleotides at a randomized position are represented in equal amounts. However, in some cases it may be desirable to skew that distribution. For instance, when a sequence contains many randomized positions, it could be desirable to reduce the probability of a certain variation at each position, so that the overall probability of the sequence to contain two or more 'mutations' is reasonably low. Therefore, distributions such as 80% A, 20% T or 85% A, 5% C, 5% G, 5% T are often used. The distribution of nucleotides can easily be adjusted accordingly in the oligonucleotide synthesis process.



## Technical Bulletin 1

### Library screening

When synthesizing a DNA library, Entelechon takes great care to use oligonucleotides of the highest possible quality. We have developed specific synthesis protocols which ensure a low error rate and a uniform distribution of randomized nucleotides.

However, there is a technical limit for the quality of a randomized DNA library. A library will always contain a certain level of mismatches ('mutations') at non-randomized positions, due to the limited perfection of the oligonucleotide synthesis process. Typically this affects less than 20% of all DNA molecules, but this number may be significantly higher or lower for a particular library.

Also, the exact distribution of nucleotides at a particular randomized position usually deviates to some extent from the theoretical distribution. Often the distribution is still within a reasonable margin, such as 20:35:15:30 instead of 25:25:25:25. However, this cannot always be guaranteed and a possible deviation should be compensated by a reasonably high safety factor for the number of contained variants (see below).



# Technical Bulletin 1

## Screening

The number of variants increases exponentially with the number of randomized positions, and is generally calculated as

$$n = 2^{c_2} * 3^{c_3} * 4^{c_4}$$

where  $c_2$ ,  $c_3$ , and  $c_4$  are the number of positions with two, three, and four wobble nucleotides respectively. For example, a sequence with four N positions yields 256 variants, but a sequence with ten N positions already yields over a million variants.

More variants not only make synthesis more difficult, but they also increase the risk that the library does not cover all variants and is therefore skewed. If the number of variants becomes inconveniently large, it is advisable to consider the synthesis of several libraries, each covering a smaller number of randomized positions. This does of course not allow to select variants based on the interaction between separate positions, but it could be used as a first step, in order to identify positions which are likely to have an impact on the overall 'fitness' of a gene or protein.

An important consideration is whether a library can represent all intended variations. In order to verify this, first calculate the number of variants according to the formula above. The randomized DNA follows a binomial distribution, which allows to calculate the number of individual DNA molecules  $m$  needed for a given number of variants and a desired probability  $p$  to represent each variant at least once in the library:

$$m = \log_{(1-p)/n} ((1-p)/n)$$

$p$  should be close to 1, e.g. 0.9999. Now, we have to ensure that the library contains at least this many DNA molecules. We get the required amount in mols by dividing by the Avogadro constant  $6.0221415 * 10^{23}$  mol<sup>-1</sup>. Dividing the synthesis scale (for instance, the default synthesis



## Technical Bulletin 1

scale for genes of up to 1000 bp at Entelechon is approximately 10 nmol ) by this amount yields a final factor.

If that factor is greater than one, the library should contain all variants at least once with the previously specified probability. Note that a number of steps in the synthesis and screening process can distort the library distribution. For instance, cloning into a vector, PCR, and propagation through host cells will inevitably lead to positive and negative selection of some variants.

Therefore, the calculated factor should be significantly higher than one. If, for example, we expect the underrepresentation of a particular variant due to negative selection about a factor of ten, the calculated value should be greater than ten. In practice, a factor of  $10^4$  to  $10^6$  is recommended.

### We assist with the library design

A library design has many aspects, and oftentimes it is not easy to select the perfect design. Entelechon has a long track record of creating libraries successfully, and we are happy to design a library for you or review a design made by you.



## Contact data

Entelechon GmbH  
Industriestr. 1  
93077 Bad Abbach  
Germany

Tel. +49 (9405) 96 999 10  
Fax +49 (9405) 96 999 28

[contact@entelechon.com](mailto:contact@entelechon.com)

[www.entelechon.com](http://www.entelechon.com)

