Nucleic Acids Research, 1996, Vol. 24, No. 17 **3437–3438**

Representation of unique sequences in libraries of randomized nucleic acids

Martin Tabler[1,*], Panayiotis Benos[1,3] and Martin Dörr[2]

[1]Institute of Molecular Biology and Biotechnology, [2]Institute of Computer Science, Foundation for Research and Technology, Hellas and [3]Department of Biology, University of Crete, PO Box 1527, GR-71110 Heraklion, Crete, Greece

Received April 16, 1996; Revised and Accepted July 12, 1996

ABSTRACT

From a library of nucleic acid molecules, which are randomized in parts of their sequence, unique sequence variants can be selected for specific properties. The planning of such an *in vitro* selection experiment requires some consideration regarding how much DNA template or RNA transcript should be used initially. The amount applied depends on the number of randomized nucleotides and on the expectations of how often each conceivable and unique sequence combination should be represented in the experimental pool. We display graphs describing the probability for the representation of unique nucleic acid molecules in a randomized pool as a function of the mean representation *k*, defined by the ratio of sampled nucleic acid molecules to conceivable sequence combinations and we summarize the amounts required to represent unique sequences with 99% likelihood. The probability of representation, $P = 1 - e^{-k}$, can be applied also to 'sub- saturated' pools ($k < 1$) of nucleic acids with long randomized domains, where it is impossible to provide sufficient material for full sequence representation.

An RNA molecule provides sequence information but it may also have functional properties, for example the capability of self-cleavage or being a ligand. These RNAs can be identified by *in vitro* selection procedures (reviewed in 1–3). The RNA of interest is selected from a pool (library) of RNA molecules that differ in sequence. For selection, a DNA molecule is synthesized containing, at defined positions, a completely random mixture of all four bases, A, C, G, T. Fixed sequences at the termini allow *in vitro* transcription, reverse transcription and PCR amplification. The pool of RNA transcripts is subjected to selection: for example binding to an immobilized ligand (SELEX) (4) or for different catalytic properties (5–7). Selected RNAs are reverse transcribed, the resulting cDNAs are PCR-amplified and used as template for RNA synthesis in a new round of selection.

The length of the randomized sequence determines the conceivable number of unique sequence variants in a pool of nucleic acids. However, for practical reasons, the amount of RNA (or DNA) that can be actually synthesized and subjected to a selection experiment is limited. In view of the enrichment during the selection procedure, it does not matter—at least in theory—whether the pool contains this unique sequence several-fold or just once. It is therefore of interest to determine the likelihood that a particular sequence is not represented in a library of nucleic acids. This probability, $P_{0,n}$, approximates e^{-k}, where the representation factor k, given as $k = n/4^L$ is the ratio of molecules, n, in a pool to the conceivable sequence combinations depending on the number of randomized nucleotides, L.

In a pool with $k = 1$, which contains as many molecules as there are unique sequence combinations possible, a unique sequence is not represented with a 36.8% chance, but 63.2% of all sequence combinations are represented at least once. Each further increase of the pool size by a factor of 2.3 (~ln 10) will reduce the number of unrepresented sequences by a factor of 10. Figure 1A demonstrates the relationship between the representation factor and the probability that an RNA sequence is not included in an experimental pool.

The probability of representation is only a function of k and is independent of the number of randomized nucleotides, L, as long as the number of molecules sampled in a pool increases with the number of conceivable sequence combinations. For example, each extension of the randomized sequence by one nucleotide (increase of L by 1), requires the 4-fold increase of nucleic acid molecules in the library to maintain the same likelihood of sequence representation. These pools are characterized by the same k factor. The amounts of nucleic acids required to achieve 99% sequence representation depending on the number of randomized nucleotides is summarized in Table 1.

$P_{0,n}$ can be calculated also for 'sub-saturated' pools ($k < 1$; Fig. 1B), for example if the randomized sequence is large ($L > 25$), where it is impossible to provide sufficient material for full sequence representation. Here, k also represents the upper limit of the possible conceivable sequence combinations.

Some researchers prefer pool sizes with simultaneous representation of almost all conceivable sequences to ensure that all sequence variants are subjected to the selection process (8). A formula to calculate the required pool sizes is provided in http://www.imbb.forth.gr/jol/sel.html . However, the probability of identifying the best performing sequence is solely described by $P = 1 - e^{-k}$ and is independent of whether the residual sequences are present or not. Therefore, simultaneous representation of all sequences is not relevant.

* To whom correspondence should be addressed

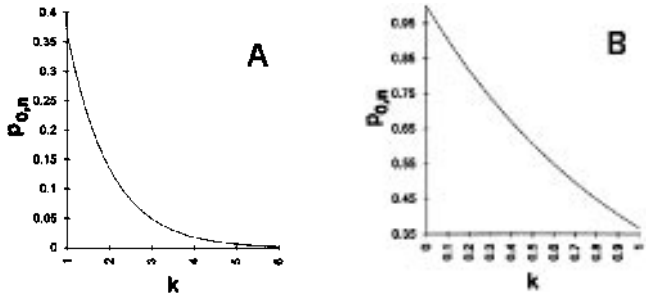

Figure 1. Relationship between the representation factor k and probability $P_{0,n}$, which indicates the likelihood that a nucleic acid is not represented in a sequence-randomized library. (**A**) The relationship for 'saturated' libraries, in which the number of sampled molecules is greater than the number of conceivable sequence combinations ($k \geq 1$). (**B**) The relationship for 'sub-saturated' libraries, in which the number of sampled molecules is smaller than the number of conceivable sequence combinations ($k \leq 1$).

Table 1. Nucleic acids required for 99% likelihood of sequence representation ($P_{0,n} = 0.01$)

Randomized nucleotides (L)	Sequence combinations (4^L)	Size of library[a] (M)
6	4.10×10^{3}	3.13×10^{-20}
7	1.64×10^{4}	1.25×10^{-19}
8	6.55×10^{4}	5.01×10^{-19}
9	2.62×10^{5}	2.00×10^{-18}
10	1.05×10^{6}	8.02×10^{-18}
11	4.19×10^{6}	3.21×10^{-17}
12	1.68×10^{7}	1.28×10^{-16}
13	6.71×10^{7}	5.13×10^{-16}
14	2.68×10^{8}	2.05×10^{-15}
15	1.07×10^{9}	8.21×10^{-15}
16	4.29×10^{9}	3.28×10^{-14}
17	1.72×10^{10}	1.31×10^{-13}
18	6.87×10^{10}	5.25×10^{-13}
19	2.75×10^{11}	2.10×10^{-12}
20	1.10×10^{12}	8.41×10^{-12}
21	4.40×10^{12}	3.36×10^{-11}
22	1.76×10^{13}	1.35×10^{-10}
23	7.04×10^{13}	5.38×10^{-10}
24	2.81×10^{14}	2.15×10^{-9}
25	1.13×10^{15}	8.61×10^{-9}
26	4.50×10^{15}	3.44×10^{-8}

[a]For a pool size of $\ln 100 \times k$; for each 2-fold increase, the sequences that are not represented are reduced by a factor of 100.

REFERENCES

1 Gold,L., Polisky,B., Uhlenbeck,O. and Yarus,M. (1995) *Annu. Rev. Biochem.*, **64**, 763–697.
2 Kumar,P.K. and Ellington,A.D. (1995) *FASEB J.*, **9**, 1183–1195.
3 Burgstaller,P. and Famulok,M. (1995) *Angew. Chem. Int. Ed. Engl.*, **34**, 1189–1192.
4 Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–510.
5 Robertson,D.L. and Joyce,G.F. (1990) *Nature*, **344**, 467–468.
6 Beaudry,A.A. and Joyce,G.F. (1992) *Science*, **257**, 635–641.
7 Berzal-Herranz,A., Joseph,S. and Burke,J.M. (1992) *Genes Dev.*, **6**, 129–134.
8 Irvine,D., Tuerk,C. and Gold,L. (1991) *J. Mol. Biol.*, **222**, 739–761.
9 Bronstein,I.N. and Semendjajew,K.A. (1979) *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun und Frankfurt/Main, Germany.

APPENDIX

Regardless of the theoretical mean representation in the experimental pool given by k, each defined unique sequence variant is represented with a distinct appearance, m, which is either 0, 1, 2 or higher. In an experimental pool consisting of n molecules, in which each conceivable sequence combination has the probability $P = 1/4^L$ of occurrence for each of the n molecules, the probability for m-fold occurrence of the particular sequence combination within the entire pool of n molecules can be described by the general formula of the binominal distribution (9):

$$P_{n,m} = \binom{n}{m} \cdot P^m \cdot (1-P)^{n-m}$$

Considering the definitions above, $P_{0,n}$ is described as:

$$P_{0,n} = \binom{k \cdot 4^L}{0} \cdot (\frac{1}{4^L})^0 \cdot (1-\frac{1}{4^L})^{k \cdot 4^L - 0} \quad \mathbf{1}$$

Since $\binom{k \cdot 4^L}{0} = 1$ and $(\frac{1}{4^L})^0 = 1$

equation **1** can be simplified to equation **2**:

$$P_{0,n} = (1-\frac{1}{4^L})^{k \cdot 4^L} \quad \mathbf{2}$$

This can be transformed:

$$P_{0,n} = \left((1-\frac{1}{4^L})^{4^L}\right)^k \quad \mathbf{3}$$

For equation **3**, one can use the relationship:

$(1-\frac{1}{x})^x = \frac{1}{e}$ for large values of x (9)

to convert equation **3** into equation **4**:

$$P_{0,n} = (\frac{1}{e})^k \quad \mathbf{4}$$

which is equivalent to the general formula:

$$P_{0,n} = e^{-k}$$