

A Stochastic Approach to Count RNA Molecules Using DNA Sequencing Methods

Boris Hollas and Rainer Schuler

Abteilung Theoretische Informatik, Universität Ulm, D-89069 Ulm, Germany
{schuler,hollas}@informatik.uni-ulm.de

Abstract. Estimating the number of congeneric mRNA molecules in a preparation is an essential task in DNA technology and biochemistry. However, DNA sequencing methods can only distinguish molecules with different sequences or of different lengths. Recently, it was shown that it is possible to combine RNA molecules with short DNA molecules (tags) such that, with high probability, any two RNA molecules are combined with tags having different sequences. For this technique, we propose a statistical estimator and a confidence interval to determine the number of mRNA molecules in a preparation.

In a second approach, the mRNA molecules are lengthened by attaching a random number of linker oligonucleotides. The original number of mRNA molecules is then determined by the number of different lengths obtained from the experiment. We also give estimator and confidence interval for this method.

Both methods can be implemented using recent as well as established methods from DNA technology. The methods can also be applied to a larger number of molecules without the need to exhaust the complete preparation. The computation of estimators and confidence intervals can be accomplished by dynamic programming.

1 Introduction

Using DNA sequencing methods it is possible to determine the number of different nucleic acid molecules in a preparation. However, it remains difficult to estimate the number of molecules (or the concentration) of a single mRNA species. Methods currently used (e.g., measuring uv-absorption, gel electrophoresis with DNA size markers, northern blotting, linear PCR, or chip hybridization) are not sufficiently reliable or may require more RNA than available. If the number of molecules is increased by PCR, the estimation error roughly doubles in each amplification round. In this case, a quantitative analysis of the probe is even more difficult.

Recently, it was shown that it is possible to sequence many DNA molecules with different sequences simultaneously [3,2,10]. Short oligonucleotides (DNA molecules) are attached to the DNA molecules. If the number of different oligonucleotide sequences is large enough (100 times larger than the number of DNA molecules), almost all oligonucleotides that are combined with a DNA molecule

will contain a unique sequence. The oligonucleotide sequences which are selected by this process are a random subset of the set of all oligonucleotide sequences. Furthermore, the sequences can be chosen such that the combined molecules can be separated by the oligonucleotide sequences (by hybridization to reverse complementary sequences which are bound to the surface of micro beads). This fact is used in [3], for further DNA sequencing.

As discussed in [7], this technique can be used to determine the number of molecules of a single mRNA species. Molecules with the same sequence of base pairs are combined with oligonucleotides with many different (tag-) sequences. The number of different tag sequences coupled to the mRNA molecules depends on the number of mRNA molecules in the preparation. However, in some cases several mRNA molecules might be coupled with oligonucleotides having the same sequence.

A different technique to determine the number of molecules is proposed in [7]. In this approach molecules of different length are prepared from the RNA molecules. This can be achieved, e.g., by ligation of multimeric linkers which can be used to assemble DNA of different length and sequences. Short linker oligonucleotides are attached to a molecule until ligation is terminated by adding a stop linker. The length of a molecule is random and depends on the ratio of stop linkers. Again, in some cases several molecules will have the same length.

In this paper we consider the relationship between the observed number of different tag sequences or molecule lengths and the number of molecules in the preparation.

Uniform Model. In the first approach described above, the combination of molecules with oligonucleotides containing different tag sequences is random. Moreover, each tag sequence is chosen with a certain probability, depending on the length and the concentration of oligonucleotides with that sequence. If all oligonucleotides containing the tag sequences are of the same length and each tag sequence is present in the same quantity then each tag sequence is chosen with the same probability. The maximal number of tag sequences can be quite large, e.g., 10^6 tag sequences can be distinguished by a DNA sequencing method (MPSS) [3]. We describe the outcome of the experiment with the uniform model.

In the *uniform model*, s denotes the number of molecules (we wish to estimate) and T_i denotes the tag sequence of molecule i , $1 \leq i \leq s$. We assume that the tag sequences are independently and uniformly (equiprobably) selected. Furthermore, we assume that all random variables T_i are identically distributed, i.e., that the ratio of the tag sequences does not change.

Geometric Model. In the second approach, DNA molecules are lengthened at random. Linker oligonucleotides and stop linkers are mixed with a certain ratio and added to the DNA molecules. Linker oligonucleotides are attached to the end of the molecule until the lengthening is terminated by adding a stop linker. The probability that the assembly is terminated depends on the concentration of the terminating linkers. Thus, we obtain a random number of different lengths. The maximal number of different lengths that can be distinguished, e.g., by gel electrophoresis, is approximately 1000. For the mathematical modelling, we

neglect several experimental factors, e.g., that the addition of stop linkers is length dependent and that the maximum length of the molecules is bounded.

For the *geometric model*, let again s denote the number of molecules. The random variables $L_1, \dots, L_s$ denote the number of linker oligonucleotides that have been attached to the molecules. We assume that molecules are independently lengthened with termination probability $1 - p$, thus we assume that $L_1, \dots, L_s$ are independent and geometrically distributed. Hence, $L_1, \dots, L_s$ are geometrically distributed. We denote the maximal number of different lengths that can be distinguished experimentally by $k_{\max}$.

For both models, we derive a maximum-likelihood-estimator and an approximate confidence interval for the unknown number of molecules.

2 Uniform Distribution

Let $T_1, \dots, T_s$ be random variables that are independent and uniformly distributed on the set of tags $\{1, \dots, t\}$, that is, all tags are equiprobable, and let

$$N_s = |\{T_1, \dots, T_s\}|$$

be the (random) number of different tag sequences obtained from an experiment on s sequences. Accordingly, denote by

$$\mathcal{P}_{s,n} = \{(t_1, \dots, t_s) \mid 1 \leq t_1, \dots, t_s \leq t \wedge |\{t_1, \dots, t_s\}| = n\}$$

the set of tags assuming exactly n different values. Then

$$\begin{aligned} P(N_s = n) &= \sum_{(t_1, \dots, t_s) \in \mathcal{P}_{s,n}} P(T_1 = t_1, \dots, T_s = t_s) \\ &= \left(\frac{1}{t}\right)^s S_{s,n} \binom{t}{n} n! \end{aligned}$$

whereby $S_{s,n}$ is the number of ways to partition the set $\{t_1, \dots, t_s\}$ into n nonempty subsets (*Stirling numbers of the second kind* [6]).

Note that $\{T_1 = T_2 = 5, T_3 = 1\}$ and $\{T_1 = 1, T_2 = T_3 = 5\}$ are different events, each having the same probability.

To calculate $P(N_s = n)$ we consider the two cases $T_s \in \{T_1, \dots, T_{s-1}\}$ and $T_s \notin \{T_1, \dots, T_{s-1}\}$. Since $S_{s,n} = S_{s-1,n-1} + nS_{s-1,n}$ and $\binom{t}{n} = \binom{t}{n-1} \frac{t+1-n}{n}$, we get

$$\begin{aligned} P(N_s = n) &= \left(\frac{1}{t}\right)^{s-1} S_{s-1,n-1} \binom{t}{n-1} (n-1)! \frac{t+1-n}{t} \\ &\quad + \left(\frac{1}{t}\right)^{s-1} S_{s-1,n} \binom{t}{n} n! \frac{n}{t} \\ &= P(N_{s-1} = n-1) \frac{t+1-n}{t} + P(N_{s-1} = n) \frac{n}{t} \end{aligned}$$

For the edges ($s = 1$ or $n = 1$) holds

$$P(N_1 = n) = \begin{cases} 1 & n = 1 \\ 0 & n > 1 \end{cases} \quad \text{and} \quad P(N_s = 1) = \left(\frac{1}{t}\right)^{s-1}$$

Though this recursion can readily be computed, it is much more efficient with regards to time and space to build up an array containing the values $P(N_s = n)$ (dynamic programming [4]). Note that, since necessarily $s \geq n$, only half of the array elements need to be computed. Then, a maximum-likelihood-estimator

$$\hat{s}(n) = \arg \max_{1 \leq s \leq s_{max}} P(N_s = n)$$

for the unknown number of sequences s , given n different tags obtained, can be computed. The upper limit s_{max} must be sufficiently large to ensure that the maximum is really attained.

Apart from the most likely number of sequences $\hat{s}$, we are also interested in the reliability of this result. A $(1 - \alpha)$ confidence interval for s must satisfy

$$\begin{aligned} 1 - \alpha &\leq P(\hat{s} + l \leq s \leq \hat{s} + r) \\ &= P(s - r \leq \hat{s} \leq s - l) \end{aligned}$$

Remember that $\hat{s} = \hat{s}(N_s)$. Thus, we get

$$P(\hat{s}^{-1}(s - r) \leq N_s \leq \hat{s}^{-1}(s - l)) \quad (1)$$

However, N_s depends on the unknown parameter s . Therefore, we construct an approximate confidence interval by replacing s with $\hat{s} = \hat{s}(n)$ in (1) and compute a, b such that

$$P(a \leq N_{\hat{s}} \leq b) \geq 1 - \alpha$$

Then, $\hat{s}(a)$ and $\hat{s}(b)$ are the bounds for this confidence interval.

Figure 1 shows the most probable number of sequences $\hat{s}$ (thick line) together with the bounds for the approximate confidence interval (thin lines). The confidence interval was constructed according to the density of $N_{\hat{s}}$. Thus the confidence interval has minimum length. The chart on the top shows the results for 500 tag sequences, the chart on the bottom for 1000 tag sequences. n denotes the number of different tags that result from the experiment.

3 Geometric Distribution

Let $L_1, \dots, L_s$ be independent and geometrically distributed random variables, that is $P(L = l) = p^l q$, $l \geq 0$, for $p \in (0, 1)$ and termination probability $q = 1 - p$. Let $N_s = \{L_1, \dots, L_s\}$ denote the number of different values (numbers) obtained from an experiment on s sequences. Let $p_{s,n,k} := P(N_s = n \wedge \forall i \in \{1, 2, \dots, s\} L_i \leq k)$. For k holds $1 \leq k \leq k_{\max}$, whereby $k_{\max}$ is the maximal number of different linker length detectable. With current technology $k_{\max}$ is approximately 1000.

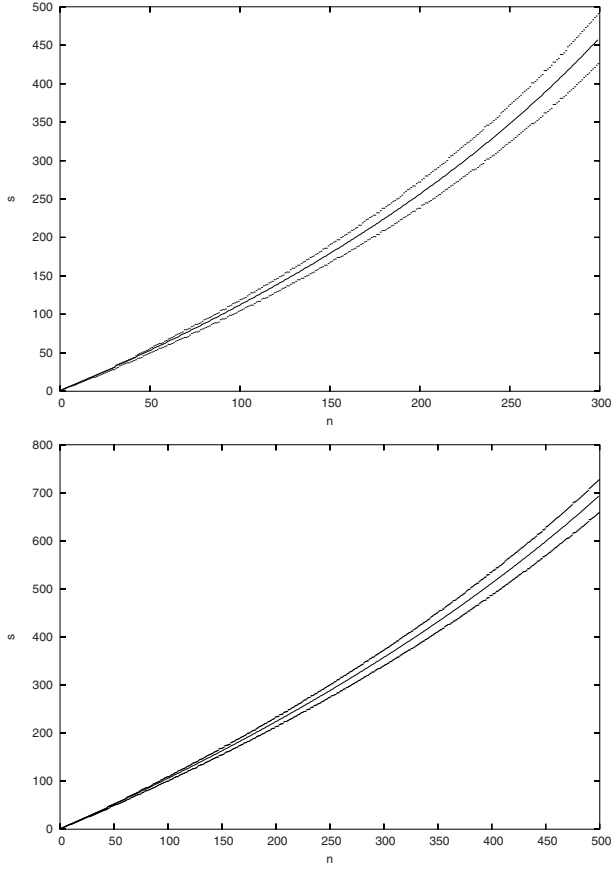


Fig. 1. Uniformly distributed lengths.

For the number $|\{L_1, L_2, \dots, L_s\} \cap \{1, 2, \dots, k_{\max}\}|$ of different values obtained from the experiments we get

$$\begin{aligned}
 & P(|\{L_1, L_2, \dots, L_s\} \cap \{1, 2, \dots, k_{\max}\}| = n) \\
 &= \sum_{d=0}^s P(\exists I \subseteq \{1, \dots, s\}, |I| = d \wedge \\
 &\quad \forall i \notin I \ L_i > k_{\max} \wedge \forall i \in I \ L_i \leq k_{\max} \wedge |\bigcup_{i \in I} \{L_i\}| = n) \\
 &= \sum_{d=0}^s \binom{s}{d} P(L_1 > k_{\max})^{s-d} \cdot p_{d,n,k_{\max}} \\
 &\text{since } P(L_1 \leq k_{\max}) = \sum_{i=0}^{k_{\max}} p^i q = q \frac{1-p^{k_{\max}+1}}{1-p} = 1 - p^{k_{\max}+1}, \text{ this equals} \\
 &= \sum_{d=0}^s \binom{s}{d} p^{(k_{\max}+1)(s-d)} \cdot p_{d,n,k_{\max}}
 \end{aligned}$$

To compute $p_{s,n,k}$, observe that either some lengths are equal to k and the remaining lengths are shorter than k , or all lengths are shorter than k .

$$\begin{aligned}
p_{s,n,k} &= P(N_s = n \wedge \exists I \subseteq \{1, \dots, s\} \forall i \in I L_i \leq k-1 \wedge \forall i \notin I L_i = k) \\
&= \sum_{d=n-1}^{s-1} \sum_{I, |I|=d} P(|\bigcup_{i \in I} \{L_i\}| = n-1 \wedge \forall i \in I L_i \leq k-1 \\
&\quad \wedge \forall i \notin I L_i = k) + p_{s,n,k-1} \\
&= \sum_{d=n-1}^{s-1} \binom{s}{d} p_{d,n-1,k-1} \cdot (p^k q)^{s-d} + p_{s,n,k-1}
\end{aligned}$$

For the edges holds

$$\begin{aligned}
p_{s,0,k} &= \begin{cases} 1, & \text{if } s = 0, k \geq 0 \\ 0, & \text{otherwise.} \end{cases} \\
p_{s,n,0} &= \begin{cases} q^s, & \text{if } n = 1, s \geq 1 \\ 0, & \text{if } n > 1. \end{cases}
\end{aligned}$$

The recursion to calculate the values of $p_{s,n,k}$ can be computed using a three-dimensional array. However for larger values of n , s , and k space can be saved. The array is computed for increasing values of s' , k' and n' . For the calculation of $p_{s',n',k'}$ only the values $p_{d,n,k-1}$ and $p_{d,n-1,k-1}$ for any $d < s'$ are needed.

A maximum-likelihood-estimator for the unknown number of sequences s can be defined by taking the number $\hat{s}$ with maximum probability.

$$\hat{s}(n) = \arg \max_{1 \leq s \leq s_{\max}} P(|\{L_1, L_2, \dots, L_s\} \cap \{1, 2, \dots, k\}| = n)$$

The upper limit $s_{\max}$ must be sufficiently large to ensure that the maximum is really attained.

Using the approach from section 2 we construct an approximate confidence interval by replacing s with $\hat{s} = \hat{s}(n)$ and compute an interval a, b such that

$$P(a \leq |\{L_1, L_2, \dots, L_{\hat{s}}\} \cap \{1, 2, \dots, k\}| \leq b) \geq 1 - \alpha$$

Then, $\hat{s}(a)$ and $\hat{s}(b)$ are the bounds for this confidence interval.

Figure 2 shows the most probable number of sequences $\hat{s}$ (thick line) together with the bounds for the approximate confidence interval (thin lines). The chart on the top shows the results for termination probability $1 - p = 1/50$ and maximum length $k_{\max} = 800$, the chart on the bottom for $1 - p = 1/80$ and $k_{\max} = 1100$. n denotes the number of different tags that result from the experiment. For these numbers the probability that all lengths are less than $k_{\max}$ is at least 0.99998 and 0.9998, respectively.

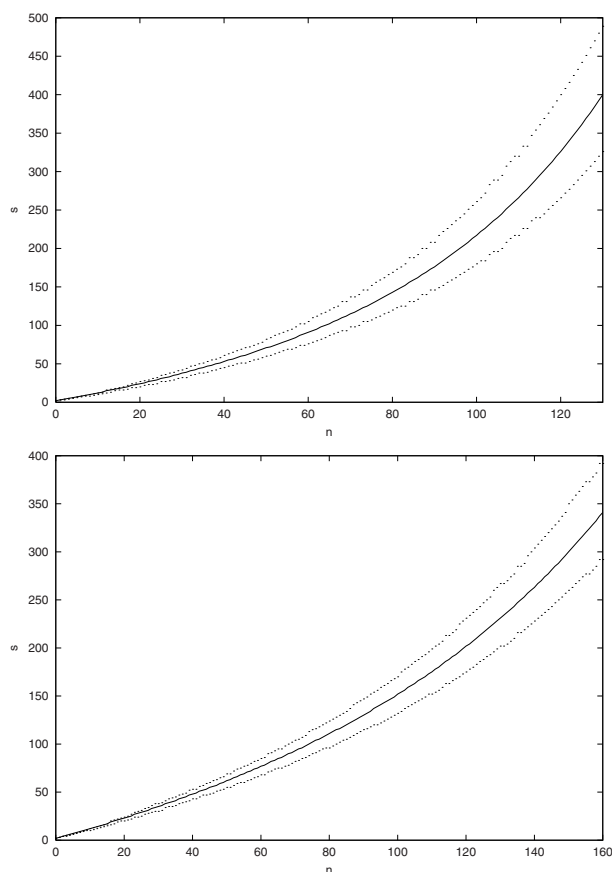


Fig. 2. Geometrically distributed lengths.

4 Discussion

The technology to obtain the data is already established. Oligonucleotides which contain the tag sequences can be synthesized (by repeatedly adding 12 different sequences of length 4) [3]. Detecting the number of different tag sequences using the MPSS method from [3] is only feasible in laboratories with sophisticated equipment. However, smaller number of tag sequences can be separated using DNA chips.

Similarly, the technique to assemble DNA sequences of different length is well established in DNA sequence analysis [9]. We note however, that the exact mixing ratio of terminating linkers has to be obtained experimentally. Furthermore, the assembly of multimeric linkers, i.e., the addition of linker oligonucleotides and stop linkers is dependent on the length of the molecules. That is, the lengths of the molecules are not truly geometrically distributed. This can

be taken into account if we can rearrange the distorted distribution P_L to a geometric distribution: e.g., if there is a permutation $\pi : \mathbb{N} \rightarrow \mathbb{N}$ such that $\pi(L)$ is geometrically distributed, the results of section 3 are valid even for P_L since $N_s = |\{L_1, \dots, L_s\}| = |\{\pi(L_1), \dots, \pi(L_s)\}|$.

The proposed methods are particularly suitable for small numbers of molecules. This is important since often only a small amount of RNA is available. By drawing a sample from a large preparation, our method can also be applied without the need to exhaust the complete preparation.

Acknowledgment

We thank the anonymous referee for helpful comments.

References

1. R.S. Braich, N. Chelyapof, C. Johnson, P.W.K. Rothmund, and L.M. Adleman. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, 96:478–479, 2002.
2. S. Brenner, M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18:630–634, 2000.
3. S. Brenner, S.R. Williams, E.H. Vermaas, T. Storck, K. Moon, C. McCollum, J. Mao, S. Luo, J. Kirchner, S. Eletr, R.B. DuBridge, T. Burcham, and G. Albrecht. In vitro cloning of complex mixtures of dna on microbeads: physical separation of differentially expressed cDNA's. *Proc. Natl. Acad. Sci. USA*, 97:1665–1670, 2000.
4. T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. McGraw Hill, 1990.
5. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. Wiley, 1971.
6. R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1998.
7. H. Hug and R. Schuler. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *Journal of Theoretical Biology*, 221(4): 615 - 624, 2003.
8. William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1999.
9. F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74:5463–5467, 1977.
10. S. Tyagi. Taking consensus of mRNA populations with microbeads. *S. Nat. Biotechnol*, 18:597–598, 2000.
11. S. Wilks. *Mathematical Statistics*. Wiley, 2nd edition, 1963.