Volume 17 Number 6 1989 Nucleic Acids Research

# Polymerase chain reaction: replication errors and reliability of gene diagnosis

Michael Krawczak, Jochen Reiss*, Jörg Schmidtke and Uwe Rösler[1]

Institut für Humangenetik, Gosslerstrasse 12 d, D-3400 Göttingen and [1]Institut für Mathematische Stochastik, Lotzestrasse 13, D-3400 Göttingen, FRG

Received January 6, 1989; Revised and Accepted February 16, 1989

## ABSTRACT

The impact of replication errors on the reliability of polymerase chain reaction (PCR) data is studied theoretically. Practical applications of our results to RFLP analysis and oligonucleotide probing confirm that for practical purposes replication errors can be neglected if a large number of starting templates (e.g. 100,000) is being used. For single locus analysis in single cells, however, the probability of false diagnosis due to such errors is of the order of 1 percent.

## INTRODUCTION

The introduction of thermostable *Taq* DNA polymerase (1) into the polymerase chain reaction (PCR) technique (2) led to its widespread use in DNA analysis, including molecular cloning, nucleotide sequencing, and restriction enzyme analysis, and to their practical applications in the diagnosis of inherited diseases and forensic medicine. An estimate of the error rate introduced by false replications is therefore of interest. Until recently, only cloned sequences have been investigated for replication errors (3,4,5). Newton et al. (6) performed a direct analysis of PCR products, but their results do not allow a precise quantitative evaluation of the error rate. Recently, Tindall and Kunkel (7) have determined empirically the fidelity of DNA synthesis by *Taq* DNA polymerase using the M13 system to circumvent the cloning procedure. As compared to the former approach, this yielded a larger sample size with only few sequencing necessary. Tindall and Kunkel estimated the rate of false synthesis to be approximately $10^{-4}$ per base per cycle. However, no attempts so far have been made to study the statistical distribution of sequences correctly amplified by this technique. We present here both a mathematical method to accomplish this and numerical results for various values of the numbers of cycles, of initial templates and of the sensitivity of the detection system (e.g. recognition by a restriction enzyme or an oligonucleotide probe).

## METHODS

The amplification of DNA sequences by polymerase chain reaction can formally be regarded as the realisation of a so-called 'Galton–Watson process'. Theoretical results for this kind of branching process, which allow the study of reproducing populations, are given, for example, by Jagers (8).

*Single-strand model*

Let $S_n$ denote the total number of correct single-stranded fragments after n cycles, where 'correct' means 'correct with respect to the sensitivity of the detection system'. If the number of correct fragments resulting from the replication of the j-th fragment during the n-th

 

cycle is given by the random variable $X_{j,n}$, then

$$S_n = \Sigma_j X_{j,n}, \quad j=1...S_{n-1}.$$

Here, $S_0$ is constant and equals the initial number of single-stranded templates. All $X_{j,n}$ are independent in any generation n, and are all distributed as

$$P(X=1) = p, \quad P(X=2) = 1-p,$$

where p denotes the probability of a false replication within the primed sequence.
Let m and $\sigma^2$ denote the mean and variance of X, respectively. Then,

$$E(S_n) = S_0 m^n \text{ and}$$
$$Var(S_n) = S_0 \sigma^2 m^{n-1}(m^n-1)/(m-1) \quad (1)$$

The distribution of X has mean $m = 2-p$ and variance $\sigma^2 = p(1-p)$.
Thus, formulae 1 yield

$$E(S_n) = S_0(2-p)^n \text{ and}$$
$$Var(S_n) = S_0 p(2-p)^{n-1}[(2-p)^n-1].$$

However, we are not interested in the total number $S_n$ but in the proportion $s_n$ of correct fragments among all replicates. Thus

$$E(s_n) = E(S_n/[2^n S_0]) = (1-p/2)^n \text{ and}$$
$$Var(s_n) = Var(S_n/[2^n S_0]) \sim p(1-p/2)^{2n-1}/(2S_0) \text{ for n large.}$$

*Double-strand model*
Let $D_n$ denote the total number of correct double-stranded fragments after n cycles. If $S_0=2D_0$ then

$$D_n = \Sigma_j Y_{j,n}, \quad j=1...S_{n-1}, \quad (2)$$

where the random variables $Y_{j,n}$ are again independent for fixed n and identically distributed as

$$P(Y=0) = p, \quad P(Y=1) = 1-p.$$

Replacing $X_{j,n}$ by $Y_{j,n}$ reflects the suppression of single-strand formation after the n-th cycle.
Applying Wald's identities (9) to formula 2 yields

$$E(D_n) = E(S_{n-1})E(Y) \text{ and}$$
$$Var(D_n) = Var(S_{n-1})E(Y)^2 + E(S_{n-1})Var(Y). \quad (3)$$

We define $d_n=D_n/(2^n D_0)$ as the proportion of correct copies among all double-stranded fragments. From formulae 3 we obtain

$$E(d_n) = (1-p)(1-p/2)^{n-1} \text{ and}$$
$$Var(d_n) \sim p(1-p)^2(1-p/2)^{2n-3}/(4D_0) \text{ for n large.}$$

*Error probability*
If we assume that for correct diagnosis, i.e. unambigious allele assignment, either $s_n$ or $d_n$ has to be larger than a fixed proportion r, then we are able to give an upper limit for the probability of false diagnosis due to replication errors. This is done by the application of Tchebychev's inequality (9).
Let $e=E(s_n)-r$ or $e=E(d_n)-r$, repectively. Then

$$P(s_n \leq r) \leq P(|s_n-E(s_n)| \geq e) \leq Var(s_n)/e^2 \text{ and}$$
$$P(d_n \leq r) \leq P(|d_n-E(d_n)| \geq e) \leq Var(d_n)/e^2. \quad (4)$$

## RESULTS

The number of PCR cycles performed usually varies between 20 and 50. To derive practical figures from the formulae given above, we base further calculations on the protocols used in our laboratory. Starting with 500 ng of total genomic DNA we obtain 5 $\mu$g of the target sequence. Given a typical fragment size of 200 base pairs this corresponds to approximately

**Table I** Expectations of the proportions $s_n$ and $d_n$ of correct fragments.

| | p | n 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| $d_n$ | $6\times10^{-4}$ | 0.9937 | 0.9907 | 0.9878 | 0.9848 |
| $s_n$ | $2\times10^{-3}$ | 0.9802 | 0.9704 | 0.9608 | 0.9512 |

n: number of cycles, p: probability of false replication per sequence per cycle.

30 cycles of perfect replication. The initial number of templates is usually large (e.g. 100,000 double-stranded fragments) but may in some cases be as small as two single-strands (single haploid cell).

For RFLP analysis the double-strand model is appropriate with a maximum value of $p=6\times10^{-4}$, corresponding to the loss of a recognition sequence by each false replication within the site. This figure applies to the majority of restriction enzymes used in practice, for which the length of the recognition sequence is 6 base pairs or less.

The single-strand model corresponds to oligonucleotide probing. Given an oligonucleotide length of 20 bases, the maximum value of p is $2\times10^{-3}$, assuming hybridization failure caused by each misincorporation within the target sequence. The expectations of $s_n$ and $d_n$ resulting from the above p-values and for several numbers of cycles are given in Table I. These figures represent the average proportion of correct fragments after a given number of cycles. For reliability estimation, however, the variation and not the average is the key parameter.

For correct allele assignment we regard a minimum of 80% correct copies of the polymorphic sequence as sufficient. This proportion allows an unambiguous discrimination between different genotypes because any correct signal is guaranteed to appear twice as strong as any false signal.

Upper limits of $P(d_n \leq 0.80)$ for $p=6\times10^{-4}$ and $D_0=100{,}000$ are given in Table II. It can be inferred from Table II, that the reliability of PCR is not substantially influenced by replication errors if the number of starting templates is large. The probability of false diagnosis due to such errors is less than one in 25 million. At this or smaller orders of magnitude the precise value is irrelevant for practical purposes.

With a much smaller number of templates to start with, the application of Tchebyshev's inequality is inefficient. If, for example, $p=2\times10^{-3}$, $S_0=2$, and $n=40$ we have

$$P(s_n \leq 0.80) \leq 1.79\times10^{-2}.$$

This upper limit of one in 56 might be unacceptably high for certain applications, and a better characterization of the true value (for example by a lower limit) is needed. If a replication error occurs during the first cycle then the proportion of correct single-stranded fragments is at most 3/4 in all following cycles. The probability of such an event is

**Table II** Upper limits for the probability of yielding less than 80% correct fragments (double-strand model, $p=6\times10^{-4}$, $D_0=100{,}000$)

| n | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| ul($\times10^{-8}$) | 3.95 | 4.05 | 4.15 | 4.26 |

ul: upper limit for $P(d_n \leq 0.80)$, n: number of cycles.

approximately 2p. Thus,

$$4\times10^{-3} \leq P(s_n \leq 0.80) \leq 1.79\times10^{-2}.$$

In our example the probability of false diagnosis due to replication errors is therefore at least one in 250 and at most one in 56. For the current state of technology, these are the limits for the accuracy of single locus analysis in single cells.

## DISCUSSION

In this study, we examined the replication reliability of *Taq* DNA polymerase in PCR. No proof-reading activity of this enzyme has been detected, although it is not clear whether an undetected exonuclease perhaps dissociated during purification (7). Nevertheless, the *Taq* DNA polymerase is the enzyme of choice for *in vitro* amplification of DNA. Its thermostability precludes the necessity of adding fresh enzyme after each denaturation step. Furthermore, compared to protocols using Klenow enzyme the high elongation temperature results in a more specific interaction with the target DNA (5). Considering the observation that Klenow fragment produces artefacts of a different type, we limited our examination to the misincorporation rate of *Taq* DNA polymerase, assuming a dominant use of this enzyme and neglecting the minor fraction of frameshift mutations produced (5).

Other studies dealing with the error rate of *Taq* DNA polymerase replication analysed the PCR products after cloning individual fragments (3,4,5). This approach results in a loss of distribution patterns, which can be overcome only by laborious repetitions. The importance of PCR for all applications is based on the circumvention of cloning procedures. Hence, replication errors also have to be analysed directly, but to date only little such data have been available (6).

No base-specific bias of the errors has been considered in our assumptions. This approach must be modified for the study of a specific DNA region. Data available so far indicate a high percentage of T−C transitions, but differ in the percentage of A−G substitutions observed (4,7). The rapid accumulation of PCR derived data will soon provide more detailed knowledge. The described model does not consider reverse mutations restoring the original DNA sequence, because in the majority of cases such events are at least one order of magnitude less likely than the loss of the site: for a unique recognition sequence of 4 base pairs the probability ratio (restoration/loss) is 1/4 times 1/3 = 1/12.

Our estimates, although possibly not specific enough for certain sites or applications, have more clearly defined the limits of PCR reliability. These estimates are consistent with the common expectation regarding the correlation between the number of starting copies, replication rounds necessary and diagnostic accuracy. Although we have used worst-case figures in the assumptions made above, it must be emphasized that our considerations were limited only to the effect of replication errors. For an overall evaluation of the reliability of the PCR technique, it would be necessary to examine several other possible causes of wrong allele assignment, (e.g. contamination with foreign DNA or partial enzyme digestion).

## ACKNOWLEDGEMENT

We thank Drs David N. Cooper, London, and Ryszard Slomski, Poznan, for their helpful discussions.

*To whom correspondence should be addressed

## REFERENCES


1. Chien, A., Edgar, D.B. and Trela, J.M. (1976) J. Bacteriol. *127*, 1550–1557
2. Saiki, R.K., Scharf, S.J., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) Science *230*, 1350–1354
3. Scharf, S.J., Horn, G.T. and Erlich, H.A. (1986) Science *233*, 1076–1078
4. Dunning, A.M., Talmud, P. and Humphries, S.E. (1988) Nucl. Acids Res. *16*, 10393
5. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Science *239*, 487–491
6. Newton, C.R., Kalsheker, N., Graham, A., Powell, S., Gammack, A., Riley, J. and Markham A.F. (1988). Nucleic. Acids Res. *16*, 8233–8243
7. Tindall, K.R. and Kunkel, T.A. (1988) Biochemistry *27*, 6008–6013
8. Jagers, P. (1975) Branching processes with biological applications, Wiley, London.
9. Chow, Y.S. and Teicher, H. (1978) Probability theory, Springer-Verlag, New York.


**This article, submitted on disc, has been automatically converted into this typeset format by the publisher.**