Nucleic Acids Research, 2005, Vol. 33, No. 6 ***e58***
doi:10.1093/nar/gni058

Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage

Markus Fuhrmann*, Wolfgang Oertel[1], Peter Berthold[2] and Peter Hegemann[2]

Universität Regensburg, Kompetenzzentrum für Fluoreszente Bioanalytik, Josef-Engert-Strasse 9, 93053 Regensburg, Germany, [1]Universität Regensburg, Lehrstuhl für Genetik, Universitätsstrasse 31, 93053 Regensburg, Germany and [2]Universität Regensburg, Lehrstuhl für Biochemie I, Universitätsstrasse 31, 93053 Regensburg, Germany

Received as resubmission February 18, 2005; Accepted March 8, 2005

ABSTRACT

The success of long polynucleotide *de novo* synthesis is largely dependent on the quality and purity of the oligonucleotides used. Generally, the primary product of any synthesis reaction is directly cloned, and clones with correct products have to be identified. In this study, a novel strategy has been established for removing undesired sequence variants from primary gene synthesis products. Single base-pair mismatches, insertions and deletions were cleaved with specific endonucleases. Three different enzymes—T7 endonuclease I, T4 endonuclease VII and *Escherichia coli* endonuclease V—have been tested. As a model, a synthetic polynucleotide encoding the bacterial chloramphenicol-acetyltransferase (*cat*) was synthesized using different methods for one step polynucleotide synthesis based on ligation of oligonucleotides. The influence of enzymatic mismatch cleavage (EMC) as an error correction step on the frequency of correct products was analyzed by functional cloning of the synthetic *cat* and comparing the error rate with that of untreated products. Significant reduction of all mutation types was observed. Statistical analysis revealed that the T4 and *E.coli* endonucleases reduced the occurrence of mutations in cloned synthetic gene products. The EMC treatment was successful especially in the removal of deletions and insertions from the primary ligation products.

INTRODUCTION

In recent years, the complete *de novo* synthesis of structural genes has become an important tool in molecular engineering (1,2), vaccine development (3), gene therapy (4) and many other fields of recombinant DNA technology. In many of these cases, the synthesis of the chosen nucleic acid sequences is more economical than classical cloning and mutagenesis procedures. Despite many improvements in the basic technology developed in 1972 (5), oligonucleotide quality is still the central problem for *de novo* gene synthesis. Current methods for the *in vitro* synthesis of double-stranded polynucleotides use chemically synthesized oligonucleotides that are ligated or made double-stranded by suitable enzymes. The result depends strongly on the quality of the oligonucleotides used. If, at an average length of 40 nt in each oligonucleotide, 99% of the molecules have the correct sequence and 1% have a wrong one (e.g. due to incomplete removal of protective groups), then a typical end product of 1000 nt in length has a probability of $0.99^{25} = 77.8\%$ of having the correct sequence. If the quality is lower (as is generally the case, e.g. 95% correct oligonucleotides), then the resulting products are correct only in $0.95^{25} = 27.7\%$ of all cases. For all presently known methods of producing double-stranded polynucleotides, the quality of the product is directly and exponentially dependent on the correctness of the employed oligonucleotides. This depends on painstaking efforts during their synthesis, purification and quality control. They must be prepared individually and the by-products, primarily shorter '$n - 1$' products, are generally removed by a costly and laborious method such as HPLC or PAGE.

There are a number of methods available for detecting mismatches in heteroduplex polynucleotides. 'Resolvases' have been used to identify or evaluate mutations in nucleic acid sequences, in particular to identify heritable changes in genetic material (6). Enzymes that cleave mismatches are applied in the TILLING technique (7), allowing the identification of rare single base-pair mismatches in a large set of DNA pools. This ability of mismatch-cleaving enzymes to specifically recognize and cleave DNA near the sites of mispaired bases leaving single-stranded overhangs may be used for introducing a 'repair step' into the gene synthesis procedure. Endonuclease

*To whom correspondence should be addressed. Tel: +49 941 943 5013; Fax: +49 941 943 5018; Email: markus.fuhrmann@vkl.uni-regensburg.de

VII of bacteriophage T4 (8,9) detects all possible mismatches including C/C pairing, heteroduplex loops, single nucleotide bulges, single-stranded overhangs, branched DNAs, bulky adducts, psoralen cross-links and apurinic sites. This broad substrate specificity makes the enzyme an extremely versatile tool for mismatch detection (10). The nucleolytic activity of T4 endonuclease VII has been used successfully to detect mutations in heteroduplex DNA in cleavage assays (11). Bacteriophage T7 endonuclease I (12) and *E.coli* endonuclease V (13), among others, show similar enzymatic activity.

For the application as a repair step in primary gene synthesis, a combined strategy of specific cleavage of mispaired double strands and removal of single-stranded overhangs has been developed. A double-stranded polynucleotide with a mismatching base pair resulting from a ligation of imperfectly matching oligonucleotides in a gene synthesis reaction is cleaved by the EMC enzyme in the first step. The short overhangs thus generated dissociate at the reaction temperature of 37°C, leaving them as single-stranded overhangs. Subsequently, these single-stranded extensions are degraded by a single-strand-specific 3′–5′-exonuclease, e.g. by *E.coli* exonuclease I or the corresponding activity of proofreading polymerases (Figure 1).

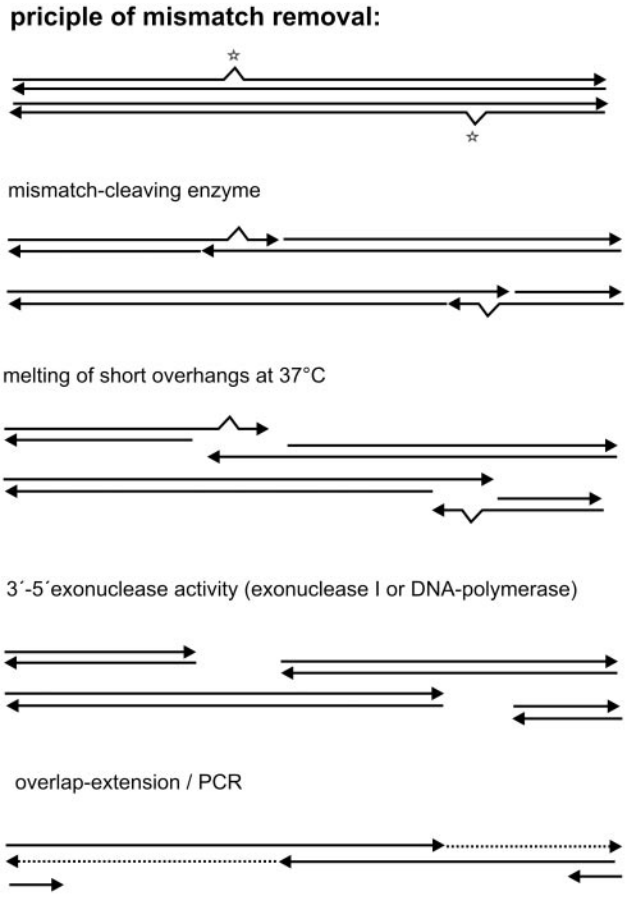


Figure 1. Principle of mismatch removal. In the first step, a double-stranded polynucleotide with a mismatching base pair, e.g. resulting from ligation of oligonucleotides in a gene synthesis reaction, is cleaved by the EMC enzyme in both strands, 2–5 bp downstream of the mismatch. The short overhangs thus generated dissociate immediately at the reaction temperature of 37°C. The resulting single-stranded overhangs can be removed by different strategies: addition of a single-strand-specific 3′–5′-exonuclease, e.g. *E.coli* exonuclease I in the EMC reaction or in a subsequent exonuclease step. Alternatively, the 3′–5′-exonuclease activity of proofreading polymerases can be used in the subsequent PCR. The proposed mechanism in this case is (i) removal of single-stranded overhangs during the initial heating step from 20 to 95°C or (ii) removal of mispaired bases by 'proofreading' in the first elongation cycle of the overlap extension PCR.

MATERIALS AND METHODS

Activity assay of mismatch cleaving enzymes

Two sets of four 69mer oligonucleotides purified by HPLC were purchased from Metabion (Planegg, Germany) and stored at −20°C until use. The individual sequences of all oligonucleotides are given in the Supplementary Material. To create all possible single-base mismatches in double-stranded DNA, suitable forward and reverse oligonucleotides were annealed by heating 625 pmol forward oligonucleotide and 625 pmol reverse oligonucleotide in 10 mM Tris, pH 7.5–8.0, 50 mM NaCl and 1 mM EDTA in a standard heating block at 90–95°C. The reaction was then allowed to cool slowly to room temperature within 45–60 min.

Enzymatic activity was tested by incubation of 25 pmol of double-stranded, annealed 69mer oligonucleotides at 37°C in the reaction buffers recommended by the supplier. The amount of enzyme was chosen to be sufficient for nearly 50% digestion within 4 h: 5 U endonuclease V (Trevigen, Gaithersburg, MD), 1000 U T4 endonuclease VII (USB Corporation, Cleveland, OH), 10 U T7 endonuclease I, (New England Biolabs, Beverly, MA). After completion, reactions were heated for 20 min at 65°C. Protein was extracted twice with equal volumes of saturated phenol (Tris–HCl, pH = 7.5) and chloroform, ethanol-precipitated and dissolved in 50 µl of 5 mM Tris–HCl pH = 7.5. A 20 µl of the reaction mixture was mixed with 2 µl of 10× loading dye (30% sucrose, 0.1% bromophenol blue). DNA fragments were separated on non-denaturing 18% polyacrylamide gels (76 mM Tris base, 100 mM glycine, 100 mM DL-serine and 100 mM L-aspartate, resulting pH = 7.4) with 25 mM Tris base and 192 mM glycine (resulting pH = 8.5), at a constant current of 26 mA (14).

Design of oligonucleotides for gene synthesis

Oligonucleotides for synthesis of an artificial gene encoding chloramphenicol-acetyltransferase from *E.coli* transposon Tn9 (GenBank accession no. V00622) were designed according to the following parameters:

(i) Back-translation of the amino acid sequence into coding sequence using a codon usage table derived from the sequence of the *Chlamydomonas reinhardtii* nuclear genome (http://www.kazusa.or.jp/codon/).
(ii) Modification of the 5′ and 3′ ends of the gene to introduce new unique EcoRI and BamHI restriction sites for later cloning.
(iii) Partitioning of the full sequence into 17 sense and 17 antisense oligonucleotides of lengths ranging from 40 to 50 bases. Exceptions are the first 5′ sense oligonucleotide and the final 3′ antisense oligonucleotide with sizes of 18 and 30 nt, respectively. Detailed information is given in the Supplementary Material. The final partitioning covers the whole sequence of the artificial gene by oligonucleotides overlapping by an average of 20–25 nt.

Synthesis of single-stranded and double-stranded DNA

Synthesis of single-stranded DNA was performed as described previously (15). Synthesis of double-stranded DNA was done according to a modified published procedure (16). Reactions included 85 pmol sense oligonucleotides (in equal molar ratio, i.e. 5 pmol of each sense oligonucleotide), 10 µl polynucleotide kinase (PNK) buffer (NEB), 10 µl ATP (25 mM) and 20 U polynucleotide kinase (NEB) in a total volume of 100 µl, and were incubated for 3 h at 37°C, then 5 min at 95°C. After heat-inactivation, the reaction was immediately chilled on ice. An equivalent phosphorylation reaction was performed for 85 pmol antisense oligonucleotides. Synthesis of double-stranded DNA was performed by thermal cycling of 10 µl phosphorylated sense oligonucleotides, 10 µl phosphorylated antisense oligonucleotides, 3 µl 10× Taq DNA ligase buffer (NEB) in a total volume of 30 µl. The reaction was denatured for 1 min at 95°C without enzyme, then cooled to 80°C for 1 min, after which 60 U of Taq DNA ligase (NEB) were added. The reaction mixture was incubated for 5 cycles of 1 min at 95°C, 6 min at 70°C, stepwise cooling to 56°C within 1 h and then 6 min at 56°C. After the final step, the reaction was cooled to 4°C and used for further experiments directly.

Error removal by enzymatic cleavage of mismatches

A 5 µl aliquot of the synthesis reaction was incubated with 5 U mismatch-cleaving enzyme *E.coli* endonuclease V (Biozol, Germany; manufactured by Trevigen), or 1000 U T4 endonuclease VII (USB Corporation), in a total volume of 20 µl. Cleavage was performed at 37°C in the buffer recommended by the manufacturer, for up to 24 h. For removal of single-stranded DNA, *E.coli* exonuclease I (Fermentas, St Leon-Rot, Germany) was included in some reactions. Post-EMC amplification by PCR was done with *Taq* DNA polymerase (Fermentas) or Vent DNA polymerase (NEB) as indicated.

Functional analysis of the synthetic *cat* gene

Two consecutive survival assays served our primary test system for error correction. The synthesized genes were cloned between the unique EcoRI and BamHI restriction sites in the cloning vector pUC18. The synthetic gene was designed so that standard isopropyl-thiogalactoside (IPTG)-induced expression of the *lacZ* fragment of pUC18 leads to in-frame translation of the synthetic gene.

The first test was done by transferring transformed (i.e. ampicillin resistant) *E.coli* to Luria–Bertani (LB)-agar plates containing 100 µg/ml ampicillin, 40 µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal) and 0.1 mM IPTG to induce *lacZ* expression. The number of surviving white clones from step one represents the 'number of analyzed clones'.

In the second test, white ampicillin-resistant clones from step 1 were transferred to LB-agar plates containing 100 µg/ml ampicillin, 0.1 mM IPTG and 34 µg/ml chloramphenicol. The number of surviving clones from test two represents the 'number of active clones'.

Statistical analysis of DNA sequences

The relative error frequency (f) per 1 kb was calculated with n (number of sequenced clones), x_i (number of errors in clone i), and l_i (length of sequence data from clone i):

$$f = \frac{\sum_{i=1}^{n} x_i \frac{1000}{i}}{n}. \quad \mathbf{1}$$

For a statistical evaluation of each individual method, the experimental data were approximated to a normal distribution with the determined relative error frequency as the expectation value $\mu := f$. Generally, the SD σ of a random variable y can be estimated from an average sample of n according to Equation 2:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \quad \text{with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \quad \mathbf{2}$$

This can be converted according to the statistical displacement law into the more easily calculated variant Equation 2a:

$$\sigma^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} \quad \text{with } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \quad \mathbf{2a}$$

Replacing y_i with the normalized error number $x_i \times (1000/l_i)$ introduced in Equation 1, σ is immediately defined by Equation 3:

$$\sigma = \sqrt{\frac{n \sum_{i=1}^{n} \left(x_i \frac{1000}{l_i}\right)^2 - \left(\sum_{i=1}^{n} x_i \frac{1000}{l_i}\right)^2}{n(n-1)}}. \quad \mathbf{3}$$

The resulting calculated normal distributions for all three methods are shown in Figure 4. The density functions are defined by Equation 4:

$$\rho(X; \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)/\sigma]}. \quad \mathbf{4}$$

From these distributions, the probability (p) for obtaining clones with a maximum number of k errors can be calculated by Equation 5 for all three methods by integration of the corresponding density functions. Integration is performed from $-\infty$ up $k + 0.5$ to round for whole-numbered errors k.

$$p(f \leqslant k) = \int_{-\infty}^{k+0.5} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)/\sigma]} dx, \quad k \in N_0. \quad \mathbf{5}$$

RESULTS

Mismatch-cleaving activities of endonucleases

Model substrates were created by annealing complementary 69mer oligonucleotides with a single mismatched base pair at the central position. We then tested the abilities of T7 endonuclease I, T4 endonuclease VII and *E.coli* endonuclease V to perform a specific and quantitative digestion of the model substrates. The relative efficiencies with which these enzymes recognize mismatches were reported previously (6,13,17–19). We chose the C/C pairing at first. This mismatch has been reported to be the least efficiently detected and therefore represents a suitable model for a quantitative assay of cleavage. After 4 h, nearly quantitative cleavage was detected only with *E.coli* endonuclease V (Figure 2A). With T4 endonuclease VII, ~30% of the substrate remained uncleaved, whereas with T7

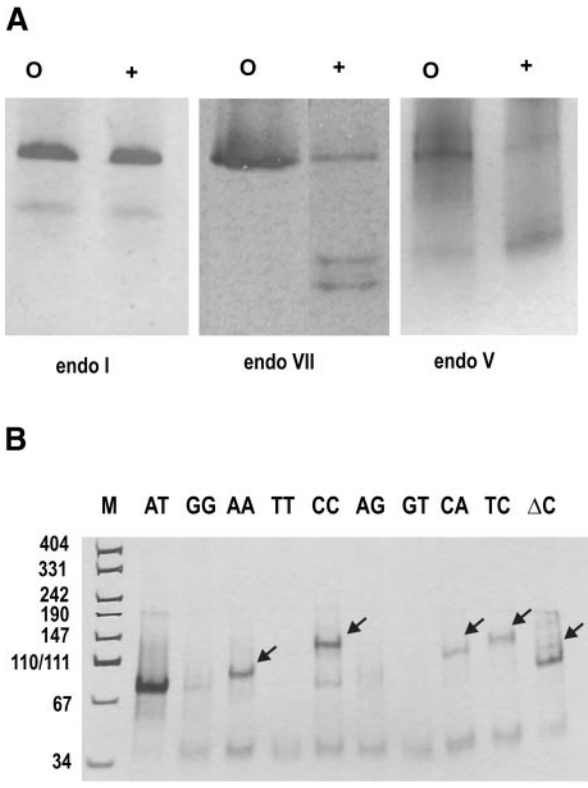
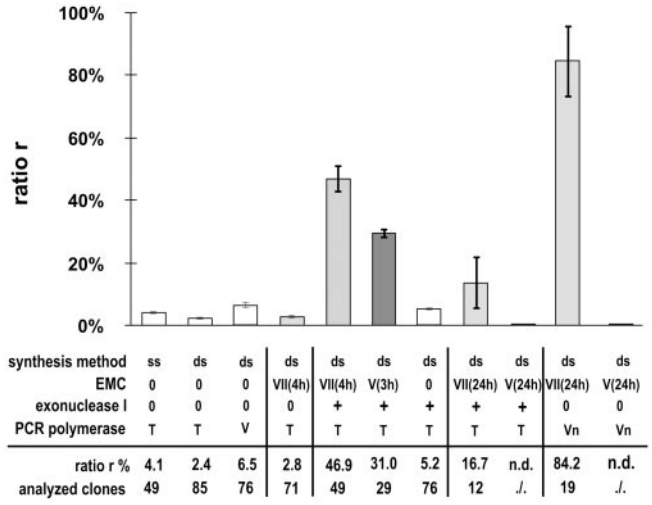


Figure 2. Cleavage assay of double-stranded oligonucleotides containing single base mismatches. (**A**) The 69mers with a CC pairing at the central position were cleaved for 24 h with 1000 U endonuclease VII, 5 U endonuclease V and 10 U endonuclease I, respectively (+). Controls without enzyme were performed in parallel (O). (**B**) The 69mers with single base mismatches at the central position were incubated with endonuclease V for 4 h. Labeling of the lanes corresponds to the respective mismatched base pair (AA = both strands contain an adenine at the central position; the fully paired control has an AT base pair at this position; M = molecular size standard). Arrows indicate unexpected size shifts in cleavage products.



synthesis method	ss	ds	ds	ds	ds	ds	ds	ds	ds	ds	ds
EMC	0	0	0	VII(4h)	VII(4h)	V(3h)	0	VII(24h)	V(24h)	VII(24h)	V(24h)
exonuclease I	0	0	0	0	+	+	+	+	+	0	0
PCR polymerase	T	T	V	T	T	T	T	T	T	Vn	Vn
ratio r %	4.1	2.4	6.5	2.8	46.9	31.0	5.2	16.7	n.d.	84.2	n.d.
analyzed clones	49	85	76	71	49	29	76	12	./.	19	./.

Figure 3. Synthesis of a functional chloramphenicol acetyltransferase gene with changed codon composition. The ratio *r* of 'active clones' to 'analyzed clones' as described in the text is shown for different gene synthesis methods with or without an EMC step. A significant increase of *r* can be observed only in the cases where EMC is combined with an exonuclease activity present in the reaction or in the later amplification reaction. Prolonged incubation with *E.coli* endonuclease V results in no detectable product after the amplification steps (ss, single-stranded synthesis, ds, double-stranded synthesis; VII, T4 endonuclease VII; V, *E.coli* endonuclease V; T, Taq DNA polymerase; and Vn, Vent DNA polymerase).

endonuclease I apparently 100% of the substrate was still present. In a second step, the specificity and cleavage efficiency of *E.coli* endonuclease V were analyzed using nine oligonucleotides with all possible mispairings or a deletion, and one fully matching oligonucleotide as a control substrate. Again, after 4 h of cleavage the reactions were separated by PAGE (Figure 2B). Quantitative cleavage was detected only for the T/T and G/T pairings, whereas with A/A, A/G, C/A and T/C a size shift was found in addition to the cleavage products. The C/C mismatch and the deletion were partially cleaved, whereas the control oligonucleotide remained virtually unchanged. Extension of the incubation for a longer period (i.e. 24 h) did not improve the overall results, because nonspecific destruction of the substrates increased significantly over time.

Error frequency in the products of primary gene synthesis

The bacterial chloramphenicol-acetyltransferase (*cat*) of transposon Tn9 was chosen as a test gene, since a gene conferring antibiotic resistance to a host cell is easily screened for its general function by a survival assay. To determine the error frequency in the genes synthesized without an EMC repair step, polynucleotide synthesis of *cat* was performed by standard single-stranded and double-stranded synthesis using HPLC-purified oligonucleotides. The resulting polynucleotides were amplified by PCR and ligated via unique EcoRI and BamHI restriction sites at their termini into the respective sites in pUC18. This created an in-frame translational fusion of the *lacZ* fragment in pUC18 with the inserted *cat* gene. Upon induction with IPTG, the bacteria produced an active chloramphenicol-acetyltransferase fusion protein, raising the bacterial resistance to chloramphenicol.

The ratio *r* of 'active clones' to 'analyzed clones' was determined as 4.1% for single-stranded (ss-) synthesis (12) and 2.4% for double-stranded (ds-) synthesis, respectively (Figure 3, columns 1 and 2). The number and nature of the occurring errors was determined by individual sequencing of the *cat* inserts from four clones of single-stranded synthesis and five clones of double-stranded synthesis, which all were enzymatically inactive (Table 1). On average, >4 errors were found among the 616 sequenced positions of the synthetic *cat* gene. Most frequently, the errors resulted from deletions, insertions, and C to T and T to C transitions.

Enzymatic mismatch-cleavage of primary synthesis products

Instead of directly amplifying the primary ligation product by PCR, an additional repair step was used in a set of four consecutive experiments: the product of the gene synthesis reaction was incubated with *E.coli* endonuclease V or T4 endonuclease VII to cleave any mispaired double strands resulting from imperfect oligonucleotide synthesis. We first tested whether active removal of the resulting single-stranded overhangs by exonucleases was necessary. The first control

Table 1. Sequencing results of synthetic *cat* genes with or without EMC treatment of the primary ligation product

Clone #: number of errors (x_i)	Sequence length (l_i)	Endonuclease treatment	Nature of the errors
ss#1: 4	616	0	A>G, C>T, C>T, C>T
ss#2: 5	616	0	Del C, T>C, C>T, C>T, G>A
ss#3: 7	616	0	Del TA, Del C, T>C, Del A, T>C, C>T, Del T
ss#4: 6	616	0	Del CC, C>T, T>G, C>T, C>T, T>C
ss#5: 3	616	0	Del T, G>T, A>G
ss#6: 4	616	0	G>T, C>T, G>A, Ins C
ss#7: 4	531	0	C>T, T>C, G>A, C>T
ss#8: 3	616	0	Del A, Del C, C>T
ss#9: 6	616	0	G>A, G>A, Del T, C>T, A>T, G>T
ss#10: 2	616	0	Ins C, Del G
ds#1: 3	616	0	Del A, Del G, A>T
ds#2: 5	616	0	Ins G, C>G, Del T, C>T, C>G
ds#3: 2	616	0	Ins C, G>A
ds#4: 3	616	0	Ins C, A>T, C>T
ds#5: 3	616	0	Ins C, G>T, C>G
ds#6: 4	616	0	G>T, C>T, T>A, C>T
ds#7: 3	398	0	C>T, Del CC, A>G
ds#8: 3	616	0	G>A, C>T, C>T
ds#9: 5	616	0	Del G, Del C, C>T, C>T, T>C
ds#10: 3	571	0	C>T, T>A, Ins C
e7#1: 0	616	24 h endo VII	
e7#2: 0	616	24 h endo VII	
e7#3: 3	616	24 h endo VII	G>T, C>T, G>A
e7#4: 2	616	24 h endo VII	C>T, A>T
e7#5: 0	616	24 h endo VII	
e7#6: 1	616	24 h endo VII	T>A
e7#7: 1	616	24 h endo VII	C>T
e7#8: 2	616	24 h endo VII	C>T,G>A
e7#9: 0	159	24 h endo VII	
e7#10: 1	616	24 h endo VII	C>T
e5#1: 0	616	4 h endo V	
e5#2: 3	616	4 h endo V	Del C, G>T, C>T
e5#3: 0	616	4 h endo V	
e5#4: 1	616	4 h endo V	T>A
e5#5: 1	613	4 h endo V	A>G
e5#6: 3	616	4 h endo V	G>A, C>G, C>T
e5#7: 1	513	4 h endo V	G>T
e5#8: 0	526	4 h endo V	
e5#9: 2	616	4 h endo V	C>T, A>T
e5#10: 1	613	4 h endo V	C>T

Error notation: single nucleotide changes are shown as 'correct base' > (was changed to) 'detected base', e.g. A>G means instead of an expected A, a wrong G was found. Insertions of a nucleotide are indicated as 'Ins' followed by the nucleotide that are found additionally, e.g. Ins C means the detection of an additional C. Deletions of nucleotides are indicated as 'Del' followed by the missing nucleotide(s), e.g. Del CC means that two consecutive Cs are missing. All endonuclease reactions contained no additional exonuclease. Post-EMC amplifications were done with Vent DNA polymerase.

reaction was performed with T4 endonuclease VII, but without any exonuclease treatment (Figure 3, column 3). This reaction did not result in any increase of correct clones (2.8%), as compared with the standard double-stranded synthesis. Since no difference in the untreated sample was observed, we concluded that active removal by single-strand-specific exonucleases was essential.

In a second set of experiments, the efficiencies of T4 endonuclease VII and *E.coli* endonuclease V in error removal were compared. For active removal of the single-stranded overhangs, *E.coli* exonuclease I (10 U) was included during the EMC treatment. After 4 h, the EMC treatment was stopped by heat inactivation. The ligation product was amplified by PCR using Taq DNA polymerase, cloned into pUC18, and analyzed in the survival assay. Both enzymes led to a significant increase in the ratio of 'active clones' to 'analyzed clones' as described above. The ratio increased >10-fold (from 4.1 to 46.9%) for T4 endonuclease VII and 8-fold (from 4.1 to 31.0%) for *E.coli* endonuclease V (Figure 3, columns 5 and 6).

Subsequently, we increased the time of the EMC treatment to 24 h. However, this did not further improve the result for either enzyme. Instead, the ratio of 'active clones' to 'analyzed clones' declined to 16.7% (Figure 3, T4 endonuclease VII, column 8), still being ~4-fold higher than without the EMC step. Using *E.coli* endonuclease V in a 24 h EMC step, no product was obtained after PCR amplification, suggesting that nonspecific degradation in a side reaction was excessive (Figure 3, column 9).

Since the optimal reaction conditions for the endonucleases differ from the reaction conditions of *E.coli* exonuclease I, with possibly undesirable side effects, we decided to omit the exonuclease during the EMC step. Instead, the final PCR amplification was performed using Vent DNA polymerase, a thermostable polymerase with intrinsic 3′–5′-exonuclease activity. In this case, 24 h incubation of the primary ligation product with T4 endonuclease VII led to an increase in the ratio of 'active clones' to 'analyzed clones' by a factor of 20–84.2% (Figure 3, column 10). When *E.coli* endonuclease V was used, no product was obtained after PCR amplification with Vent DNA polymerase (Figure 3, column 11), as was the case with Taq.

Sequence analysis of EMC-treated clones

For a more detailed analysis of the EMC repair efficiency, sequences of individual *cat* products were determined by DNA sequencing. The number and nature of any differences from the correct sequence were determined (Table 1). The relative error frequency (f) per 1 kb with a given oligonucleotide quality was determined as $f_{ds} = 5.8$ for the double-stranded synthesis protocol, and $f_{ss} = 7.2$ for single-stranded synthesis after analysis of 10 (ds) and 10 (ss) clones (according to Table 1). Statistically, there is no difference expected between these two methods regarding the relative error frequency, provided that the quality of sense and antisense oligonucleotides is equal. Since all EMC experiments were performed with double-stranded ligation products, all consecutive calculations were done only for this method.

It was assumed from the simple plating assay that the average number of errors should be decreased in clones derived by EMC methods. This assumption was proven by the relative-error frequencies determined for the two analyzed EMC protocols. Using T4 endonuclease VII (24 h), the value of f decreased to $f_{ds\text{-}endoVII} = 1.62$ and with *E.coli* endonuclease V (4 h) to $f_{ds\text{-}endoV} = 1.98$ (calculated from Table 1).

The SD for the relative error frequency with the double-stranded synthesis protocol was calculated as $\sigma_{ds} = 1.65$. For both EMC protocols, σ was determined accordingly, for the endonuclease VII protocol $\sigma_{ds\text{-}endoVII} = 1.71$ and for the endonuclease V protocol $\sigma_{ds\text{-}endoV} = 1.84$. The resulting calculated normal distributions for three methods are shown in Figure 4.

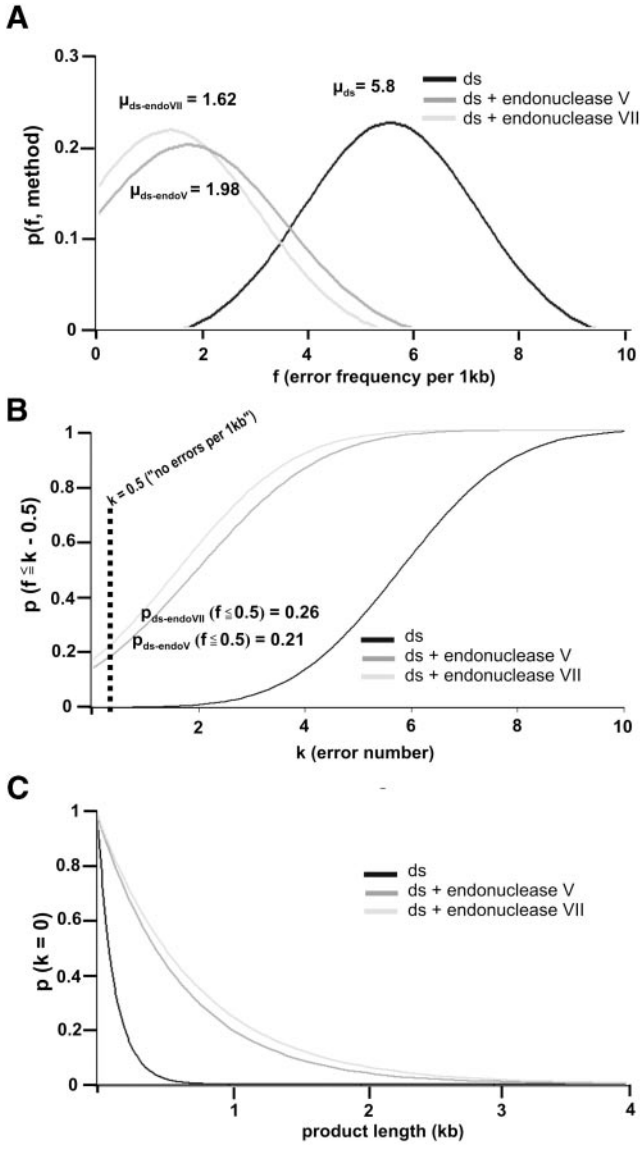


Figure 4. Statistical evaluation of error numbers in synthetic genes as determined by DNA sequencing. The different methods of synthesis and EMC treatments are compared. (**A**) Experimental data are fitted to a normal distribution using the relative error frequency f of the respective method as expectation value μ, and the calculated σ values determined. (**B**) Integration of the density function of the normal distribution from (A) leads to the probability function $p(k)$ that the relative error frequency per 1 kb of a synthesis product is smaller than k. (**C**) Dependence of the probability for an error-free clone ($k = 0$) on the length of the synthetic product and the method used.

From these distributions, the probability (p) of obtaining clones with a maximum number of k errors can be calculated by Equation 5 for all three methods by integration of the corresponding density functions.

Using this approach, the probability of synthesizing an error-free product of 1 kb with double-stranded synthesis was calculated as $p(f \leqslant 0)_{ds} = 0.0006$ (ds-synthesis), $p(f \leqslant 0)_{ds\text{-}endoVII} = 0.26$ (endonuclease VII) and $p(f \leqslant 0)_{ds\text{-}endoV} = 0.21$ (endonuclease V). In a more practical sense, these data mean that dramatically fewer clones need to be analyzed using any EMC repair protocol. For example, the number j of individual clones of 1 kb that one has to analyze to get at least one without errors ($k = 0$), with a probability of >90%, directly from gene synthesis, can be calculated using inequality Equation 6.

$$q^j \leqslant 0.1, \quad q = 1 - p\left(f \leqslant k\right). \qquad \mathbf{6}$$

Consequently, one will have to screen 8 or 10 clones of 1 kb treated with EMC. However, to achieve the same result following double-stranded synthesis without an EMC step, the corresponding number is >3837 clones.

DISCUSSION

We have developed a strategy to improve the quality of long ligation-based primary polynucleotide synthesis products. The main feature of this novel gene synthesis protocol is based on the enzymatic mismatch-cleavage activities of certain bacterial and phage resolvases. Two phage enzymes (T4 endonuclease VII, T7 endonuclease I) and one bacterial enzyme (*E.coli* endonuclease V) have been tested for their ability to cleave double-stranded DNA containing single mismatched base pairs. Since T7 endonuclease I showed no detectable activity in an initial activity test with model substrates, its use was not explored further. The remaining two enzymes were active in the oligonucleotide cleavage test, but the observed size shifts of A/A, A/G, C/A and T/C were previously reported only for the formation of different enzyme–DNA complexes under non-denaturing conditions (13). The substrate-dependent efficiency of second-strand cleavage observed for endonuclease V (Trevigen, product data sheet) may result in single-strand cuts that are not resolved by non-denaturing gel electrophoresis. Also, cleavage of A/A was not perfect, though it should be the best-recognized and best-cleaved mismatch. We used oligonucleotide substrates with a G/C pair 5′ to the mismatch. This type of substrate was previously reported to substantially reduce the cleavage activity of endonuclease V (18). However, T4 endonuclease VII and *E.coli* endonuclease V performed similarly, though not identically, when used in combination with a polynucleotide synthesis protocol. The bacterial enzyme was efficient only when relatively short (up to 4 h) treatments were used. Alternative assays affected product quality differently. In the functional plating screen, the results of a 4 h treatment with endonuclease V were comparable to a 4 h treatment with T4 endonuclease VII. By sequence analysis, 4 h EMC with endonuclease V was nearly as efficient as a 24 h treatment with T4 endonuclease VII. However, endonuclease V treatment could never be successfully extended to 24 h. No PCR products could be generated after 24 h incubation with endonuclease V. The enzyme may exhibit reduced specificity during long incubation times, or there may be other activities in the commercial product. Such an effect would be similar to the frequently observed relaxation of specificity of certain restriction endonucleases, also referred to as star activity (20,21). Alternatively, the reported inherent nonspecific 5′exonuclease activity of endonuclease V (~10% of the endonuclease activity measured as cleavage of terminal nucleotides) might be the reason for the difficulties encountered with longer incubation times. It should be noted that for T4 endonuclease VII as well, the enzyme's stability was critical. In our hands, even the enzyme stock, when stored at −20°C, was not stable over a

period of more than 6 months, as measured by the oligonucleotide cleavage test (data not shown). For optimal results, we therefore suggest to use endonuclease V in a time course from 1 to 24 h of incubation. This protocol has to be adjusted for every set of oligonucleotides in a gene synthesis reaction. The latest time point yielding sufficient amounts of PCR products for cloning should be considered optimal for error removal.

Gene synthesis with an integrated EMC step offers several advantages over older methods. Chemical oligonucleotide synthesis is prone to error, and strongly dependent on the instrumentation used and the exact conditions of the individual synthesis steps. Purification procedures including HPLC or PAGE are laborious, expensive and cannot fully exclude by-products of similar size, e.g. mutations arising from base exchanges or uncleaved protection groups. This novel gene synthesis strategy reduces the dependence of polynucleotide synthesis on the quality of the oligonucleotides used as a starting material, leading to more cost-effective polynucleotide synthesis for various applications in the future. In particular, the need for tedious, time-consuming and expensive purification procedures for the oligonucleotides may be eliminated by EMC treatment.

Furthermore, the ligation temperature is usually chosen as close to the melting temperature of the oligonucleotides as possible. This is supposed to ensure a maximum stringency during hybridization, possibly reducing the incorporation of mismatched oligonucleotides. However, it is usually difficult to design a set of 25 or more oligonucleotides in a way that all overlaps are within ±1°C, even if special oligonucleotide design software tools, like 'Gene2Oligo' (22), are used. Additionally, the calculated difference in hybridization temperatures for a 20–25 bp overlap with versus without a single mismatching base is quite low. Although the general effect of single base-pair mismatches is difficult to quantify, a general working rule sets the average difference to ~1°C per % mismatch. For a 25mer, this would mean a decreased melting temperature of 4°C, being well within the recommended variation of the parameters for 'Gene2Oligo' of ±3–4°C. Therefore, a significant reduction of single base-pair mismatches will be difficult to achieve by optimizing the annealing and ligation temperatures of the oligonucleotides.

The EMC repair process will be particularly useful for the production of long polynucleotides in a single step. The relatively high error content of synthesized oligonucleotides and the limited range of sequencing reactions are frequently circumvented by stepwise gene synthesis protocols. Short elements of <500 bp are synthesized as building blocks. Their sequences are verified individually, and correct blocks are assembled into a large polynucleotide. Using EMC to improve the quality of primary synthesis products allows one to increase the size of single-step assemblies up to >1 kb, since the error frequency can be reduced very considerably.

The use of EMC is not limited to polynucleotide synthesis protocols based on ligation. PCR-based methods could also benefit from the incorporation of an enzymatic repair step. The implementation of such a strategy would require the initial production of a small pool of full-length polynucleotides by PCR. These polymerase products could be separated into single strands by high temperature. The single strands could then be re-hybridized to create double strands containing mismatches, which could be repaired using EMC.

The improvement of gene synthesis for use in high-throughput applications requires reductions in the costs of preparing oligonucleotides as well as additional streamlining of the process itself. The quality of the primary synthesis products (as judged by sequencing cloned molecules) should be sufficient to avoid the need for laborious post-synthesis editing and re-cloning steps. The implementation of an EMC repair step during automated gene synthesis will probably improve both parameters significantly.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank USB for T4 endonuclease VII for beta testing. We would like to thank Dr Paul Kretchmer (kretchmer@sfedit.net) at San Francisco Edit for his assistance in editing this manuscript. This work was funded in parts by grant KFB2.1 of the Bavarian Ministry for Trade and Commerce. Funding to pay the Open Access publication charges for this article was provided by the Bavarian Ministry for Trade and Commerce.

Conflict of interest statement. None declared.

REFERENCES

1. Scheller,J., Guhrs,K.H., Grosse,F. and Conrad,U. (2001) Production of spider silk proteins in tobacco and potato. *Nat. Biotechnol.*, **19**, 573–577.
2. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.
3. Casimiro,D.R., Tang,A., Perry,H.C., Long,R.S., Chen,M., Heidecker,G.J., Davies,M.E., Freed,D.C., Persaud,N.V., Dubey,S. *et al.* (2002) Vaccine-induced immune responses in rodents and nonhuman primates by use of a humanized human immunodeficiency virus type 1 pol gene. *J. Virol.*, **76**, 185–194.
4. Yew,N.S., Zhao,H., Przybylska,M., Wu,I.H., Tousignant,J.D., Scheule,R.K. and Cheng,S.H. (2002) CpG-depleted plasmid DNA vectors with enhanced safety and long-term gene expression *in vivo*. *Mol. Ther.*, **5**, 731–738.
5. Agarwal,K.L., Yamazaki,A., Cashion,P.J. and Khorana,H.G. (1972) Chemical synthesis of polynucleotides. *Angew Chem. Int. Ed. Engl.*, **11**, 451–459.
6. Youil,R., Kemper,B.W. and Cotton,R.G. (1995) Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII. *Proc. Natl Acad. Sci. USA*, **92**, 87–91.
7. McCallum,C.M., Comai,L., Greene,E.A. and Henikoff,S. (2000) Target induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol.*, **123**, 439–442.
8. Golz,S., Birkenbihl,R.P. and Kemper,B. (1995) Improved large-scale preparation of phage T4 endonuclease VII overexpressed in *E. coli*. *DNA Res.*, **2**, 277–284.
9. Kemper,B. and Garabett,M. (1981) Studies on T4-head maturation. 1. Purification and characterization of gene-49-controlled endonuclease. *Eur. J. Biochem.*, **115**, 123–131.
10. Cotton,R.G.H. (1997) *Mutation Detection.* Oxford University Press, Oxford, UK.
11. Youil,R., Kemper,B. and Cotton,R.G. (1996) Detection of 81 of 81 known mouse beta-globin promoter mutations with T4 endonuclease VII—the EMC method. *Genomics*, **32**, 431–435.
12. West,S.C. (1992) Enzymes and molecular mechansims of genetic recombination. *Annu. Rev. Biochem.*, **61**, 603–640.

13. Yao,M. and Kow,Y.W. (1997) Further characterization of *Escherichia coli* endonuclease V. Mechanism of recognition for deoxyinosine, deoxyuridine, and base mismatches in DNA. *J. Biol. Chem.*, **272**, 30774–30779.
14. Ahn,T., Yim,S.K., Choi,H.I. and Yun,C.H. (2001) Polyacrylamide gel electrophoresis without a stacking gel: use of amino acids as electrolytes. *Anal. Biochem.*, **291**, 300–303.
15. Fuhrmann,M., Oertel,W. and Hegemann,P. (1999) A synthetic gene coding for the green fluorescent protein (GFP) is a versatile reporter in *Chlamydomonas reinhardtii*. *Plant J.*, **19**, 353–361.
16. Sutton,D.W., Havstad,P.K. and Kemp,J.D. (1995) Rapid gene synthesis using ampligase thermostable DNA ligase. *EPICENTRE Forum*, **2**, 1.
17. Mashal,R.D., Koontz,J. and Sklar,J. (1995) Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. *Nature Genet.*, **9**, 177–183.
18. Yao,M. and Kow,Y.W. (1994) Strand-specific cleavage of mismatch-containing DNA by deoxyinosine 3′-endonuclease from *Escherichia coli*. *J. Biol. Chem.*, **269**, 31390–31396.
19. Yao,M. and Kow,Y.W. (1996) Cleavage of insertion/deletion mismatches, flap and pseudo-Y DNA structures by deoxyinosine 3′-endonuclease from *Escherichia coli*. *J. Biol. Chem.*, **271**, 30672–30676.
20. Woodbury,C.P.,Jr, Hagenbuchle,O. and von Hippel,P.H. (1980) DNA site recognition and reduced specificity of the EcoRI endonuclease. *J. Biol. Chem.*, **255**, 11534–11548.
21. Polisky,B., Greene,P., Garfin,D.E., McCarthy,B.J., Goodman,H.M. and Boyer,H.W. (1975) Specificity of substrate recognition by the EcoRI restriction endonuclease. *Proc. Natl Acad. Sci. USA*, **72**, 3310–3314.
22. Rouillard,J.M., Lee,W., Truan,G., Gao,X., Zhou,X. and Gulari,E. (2004) Gene2Oligo: oligonucleotide design for *in vitro* gene synthesis. *Nucleic Acids Res.*, **32**, W176–W180.