Nucleic Acids Research, Vol. 18, No. 4

The effect of replication errors on the mismatch analysis of PCR-amplified DNA

Jochen Reiss*, Michael Krawczak, Manfred Schloesser, Michael Wagner and David N.Cooper[1]
Institute of Human Genetics, University of Goettingen, Gosslerstrasse 12 d, D-3400 Goettingen, FRG and [1]Molecular Genetics Section, Thrombosis Research Unit, King's College Hospital School of Medicine, Denmark Hill, London SE5 8RS, UK

Received September 25, 1989; Revised and Accepted January 22, 1990

ABSTRACT

The mismatch analysis of PCR-amplified DNA has generally assumed the absence of artificially introduced base substitutions in a significant proportion of the amplification product. This technique, however, differs from the direct sequencing of amplified DNA in that non-specific substitutions will render a molecule useless in analysis. The expected signal-to-noise ratio is heavily influenced by several parameters viz. initial template copy number, number of replication cycles, eventual product yield and the type of experimental system adopted. Mathematical modelling can be used to optimize fragment length with respect to the method applied and suggests as yet undescribed improvements such as partial modification or cleavage to optimize signal detection.

INTRODUCTION

In vitro DNA amplification using the Polymerase Chain Reaction (PCR) (1) has greatly extended the scope of existing methods for the characterization of eukaryotic DNA sequences and has made possible a detailed analysis of nucleic acids starting from a single molecule. PCR may be used to amplify either genomic DNA or mRNA via a reverse-transcribed cDNA intermediate. Since in principle, PCR may be employed to manufacture virtually unlimited amounts of a given DNA sequence, analytical accuracy is dependent solely upon the quality of the amplified product. The use of the thermostable Taq DNA polymerase (2) has made possible the automation of the PCR reaction, but since this enzyme lacks a proof-reading activity, it exhibits a relatively high single base misincorporation rate of 10^{-4}, much higher than that associated with the Klenow fragment of DNA polymerase (10^{-7}) (3). We have previously examined the likely impact of such replication errors upon the reliability of PCR in a diagnostic context (4).

PCR, followed by the direct sequencing of the amplified material, has greatly facilitated the characterization of disease alleles by making cloning unnecessary (5, 6). Artefactual errors introduced during polymerization are irrelevant to the sequence determination since each individual misincorporated base will be represented only very infrequently in the population of DNA molecules to be sequenced. The sequence analysis of complex human genes (e.g. those encoding dystrophin, factor VIII) is nevertheless still a laborious procedure if exons are to be amplified and sequenced individually. PCR amplification of cDNA molecules promises to circumvent this problem by facilitating the detection of single base-pair substitutions even in genes the size of that encoding human dystrophin (approximate length 2.3 Mbp). Indeed, reports of 'illegitimate transcription' (very low basal level of transcription in tissues not normally associated with expression of the gene product; 7,8) now hold out the promise of increased ease of access to hitherto inaccessible mRNA species.

The approximate localization of the lesion within the disease gene may however be achieved by means of various screening procedures, thereby obviating the need to sequence large tracts of the gene region. Denaturing gradient gel electrophoresis of mismatched heteroduplexes formed between wild-type and mutant alleles can in principle detect all point mutations by virtue of the altered melting behaviour of these fragments. Without heteroduplex formation, the proportion of mutations detected is reduced to 50% by base exchanges which are neutral with respect to the melting temperature. RNase A cleavage of mismatched RNA/DNA duplexes also proffers a similar detection frequency. Although these approaches (reviewed in 9) are sensitive enough to be applied to genomic DNA, the exact location of the lesion within a specific fragment cannot be determined. However, a novel and elegant technique, exploiting the ability of piperidine to cleave mismatched cytosines and thymines modified by hydroxylamine and osmium tetroxide respectively, has recently been reported (10, 11). This technique utilizes PCR-amplified wild-type and mutant template DNAs and is applicable to the detection of all point mutations.

We demonstrate here, however, that the utility of this technique is critically dependent upon the final frequency of the original mismatches (point mutations plus possible DNA polymorphisms) as compared with the frequency of mismatches introduced as a result of misincorporation events during PCR amplification. The latter frequency is influenced both by the length of the sequence under study and by the number of amplification cycles. If the

* To whom correspondence should be addressed

amount of starting material is limited (eg. rare mRNA species) and thus, a large number of cycles is required in PCR, or if the sequence of interest is rather long, important implications will be shown to arise for the statistical expectations of the relative frequencies of correct and incorrect copies. We assess the implications of our findings in the context of optimizing the signal/noise ratio in mismatch analyses following PCR. Finally, we demonstrate that incomplete degradation reactions may in principle be employed to extend the upper limit in fragment length restricting successful analysis.

METHODS

Calculations of the distribution of 'correct' and 'incorrect' copies were originally confined to the consideration of a specified region of 4 to 20 bases in a population of amplified DNA molecules (4). Mismatch detection, either by altered melting behaviour (12) or by chemical means (10, 11) does not discriminate between alleles at a particular site. Rather, it detects sequence deviations present anywhere within the entire amplified fragment without distingiushing between an original mutation and a substitution introduced during in vitro synthesis.

For the evaluation of the proportion of such sequence alterations in the amplification end-product, a number of parameters will be considered that are denoted by the following abbreviations.

b = number of bases per single strand in amplified DNA
n = number of perfect PCR cycles
p = error rate per base per cycle
S = number of single-stranded copies before amplification
c = proportion of mismatches detected by a given method

An additional parameter was introduced for the chemical mismatch analysis, which allows for variation in the extent of chemical degradation, thus simulating controlled partial degradation.

r = fraction of cleavage (degraded proportion of the total detectable mismatches).

In what follows, the term 'signal' is used to refer to signals other than those created by PCR misincorporation. This is distinct from 'noise', which refers to background smear or visible artefacts due to early replication errors.

Distribution of replication errors causing detectable mismatches

The following calculations neglect the possibility that any mismatch (i.e. an original mutation or a mismatch caused by a replication error) is 'repaired' by a replication error. Such an event has been shown to be rather unlikely (4).

The probability, p', that a strand of DNA is replicated with at least one error causing a detectable mismatch, is given by

$$p' = 1-(1-c*p)^b. \quad (1)$$

If π denotes the proportion of copies without replication errors causing detectable mismatches, then the mean and variance of π can be calculated using previously published formulae (4):

$$\text{mean}(\pi) = (1-p'/2)^n \text{ and } \text{variance}(\pi)=p'(1-p'/2)^{2n-1}/2S. \quad (2)$$

The probability, q, that a specific base on a strand chosen at random from the amplified population, was falsely replicated such that the resulting mismatch is detectable, can be calculated using a similar formula:

$$q = c\,[1-(1-p/2)^n]$$

From this, the probability of k errors causing a degradable mismatch on a single, 'random' strand is determined by

$$p_k = \beta(b,k)*q^k(1-q)^{b-k}, \quad (3)$$

where $\beta(b,k)$ denotes the binomial coefficient, i.e. the number of ways in which k elements can be selected from a total of b elements. The above equation assumes stochastic independence of replication errors at different sites, which holds approximately true if n is large enough (e.g. $n>20$) and if k is small compared with b.

Results expected for a mismatch analysis

If one original deviation (mutation), causing a detectable mismatch in heteroduplices with wild-type DNA is present in the template DNA, then the expected proportion of non-degraded DNA molecules is

$$nd(n,1) = \Sigma_k(1-r)^{k+1}p_k$$

with k theoretically ranging from zero to b. Terms with k larger than 10, however, are so small that they are irrelevant for summation (see above).

The expected proportion of DNA molecules degraded only at the original mismatch is given by

$$cd(n,1) = \Sigma_k r(1-r)^k p_k \qquad k=0, \ldots, b.$$

To distinguish this 'one mismatch pattern' from other cases, i.e. zero or two original mismatches (mutations, polymorphisms etc.) in template DNA, the following equations were employed:

— two original mismatches in template DNA

Expected proportion of non-degraded DNA:

$$nd(n,2) = \Sigma_k(1-r)^{k+2}p_k \qquad k=0, \ldots, b.$$

Expected proportion of DNA degraded only at one specific original mismatch:

$$cd(n,2) = \Sigma_k r(1-r)^{k+1}p_k \qquad k=0, \ldots, b.$$

Expected proportion of DNA degraded at both original mismatches only:

$$cd'(n,2) = \Sigma_k r^2(1-r)^k p_k \qquad k=0, \ldots, b.$$

— No mismatch in template DNA

Expected proportion of non-degraded DNA after n cycles :

$$nd(n,0) = \Sigma_k(1-r)^k p_k \qquad k=0, \ldots, b.$$

Replication errors in the first cycle have the greatest effect on the composition of the sequences comprising the single strand population. These occur, using m to define the number of errors in the first cycle causing a detectable mismatch, with probability

$$P(m) = \beta(S*b,m)*(c*p)^m(1-c*p)^{S*b-m}.$$

To assume exactly one replication error causing a degradable mismatch in the first cycle means that a proportion 1/(2S) of DNA molecules carry this mismatch. Thus the total proportion of non-degraded DNA is,

$$nd(n-1,1)/2S + nd(n-1,0)*[1-1/2S]$$

and the expected proportion of DNA degraded at the mismatch caused by the early replication error is,

$$cd(n-1,1)/2S.$$

If the complete strand is labelled (e.g. by adding labelled nucleotides for the last few PCR cycles) the expected proportions,

i.e. relative signal intensities, calculated above are applicable without further comment. For end-labelling procedures, however, the presented data are worst-case figures since only those detectable mismatches, introduced between label and original deviation, diminish the signal and contribute to noise. If the original deviation is very close to the label, most of the additional mismatches are negligable and the relative signal intensities can be expected to be much higher.

RESULTS

A constant error rate of $p = 10^{-4}$ (3) was used throughout in all calculations. If this parameter is kept constant, then the proportion of copies in amplified DNA, which are correct replicates of the original template sequence, is largely a function of two key parameters. The first of these is the number of PCR cycles, which depends on the initial number of template copies and the desired yield of amplified material. The second critical parameter is fragment length, which assumes greater importance if large fragments are to be analyzed in a search for mutations whose locations are unknown. The decrease in the proportion of completely correct replicates with increasing cycle number is demonstrated in Figure 1 for various fragment lengths. It can be seen, that the amplification of fragments larger than 1000bp will result in a final strand population with only a minor fraction of completely correct copies.

Since the denaturing methods of mismatch detection are not base-specific, the maximum signal intensities can be directly taken from Figure 1. In practice, however, the capability of mismatch detection in these systems might be reduced by the inability to detect replication errors, which do not effect melting behaviour. The impact of these limitations on mismatch analysis has to be determined empirically.

Chemical mismatch analysis of a putative mutational DNA sequence involves hybridization with the corresponding wild-type DNA fragment, either amplified or not (e.g. cDNA). One of these DNAs is labelled and the other is used in excess to suppress self-annealing of the labelled molecules. Our model predicts that 30 cycles of replication of a 1000 bp fragment will generate an amplification product with approximately 23% of perfect replicates (see Figure 1). This proportion provides a direct measure of the signal intensity in a system capable of detecting 100% of all mismatches. The base substitution frequency associated with the production of cDNA (less than 10^{-7}) may be neglected. If, however, the control DNA is also amplified (10, 11), the reannealing of two distinct PCR products each with a content of 23.2% correct replicates prior to analysis, leaves 0.232*0.232 = 5.4% of labelled heteroduplices with both strands representing their original sequences. These double-stranded DNAs alone exhibit solely the original deviation from the analyzed sequence as compared with the wild-type. Use of base-specific analysis, however, serves to improve the predicted signal intensity corresponding to non-artefactual mismatches. Selective modification by osmium tetroxide and hydroxylamine followed by piperidine cleavage can in our model be expressed by c=0.5 reflecting the fact that only one half (2*(1/4*1/3 + 1/4*1/3 + 1/4*1/3 + 1/4*0/3)) of the PCR-introduced mismatches will be detected. Equations (1) and (2) applied to a 1000 bp fragment and 30 cycles of amplification yield 47.4% of replicates without replication errors causing detectable mismatches; heteroduplex formation with an amplified control will produce 0.474 * 0.474 = 22.8% of heteroduplices reflecting one original mismatch.

With increasing length of amplified DNA, eventually no detectable trace of correct signal will remain since the PCR product after chemical degradation only produces a 'smear' representing the spectrum of replication errors. Incomplete degradation is, however, still potentially capable of producing distinct signals, e.g. electrophoretic bands. An example of this is given in Table 1. If the critical fragment length (L) denotes the minimum length for which the relative signal intensity attains its maximum at a fraction of cleavage (r) less than one, then a partial reaction should be considered for the analysis of fragments larger than L in order to improve the relative signal intensity. Figure 2a) illustrates that for an all-detecting system (c=1.0) L is approximately 1000bp. For the chemical mismatch analysis with a base-specific reaction covering 50% of all possible mismatches (c=0.5), and where partial reactions seem feasible, L increases to 1500−2000bp (Figure 2b). In a base- and strand-specific procedure (c=0.25), however, improvements are to be expected beyond 5000bp (Figure 2c).

If several fragments amplified from a certain gene are candidates for bearing a mutation, the question of the exclusion potential of mismatch analysis arises. Early replication errors may

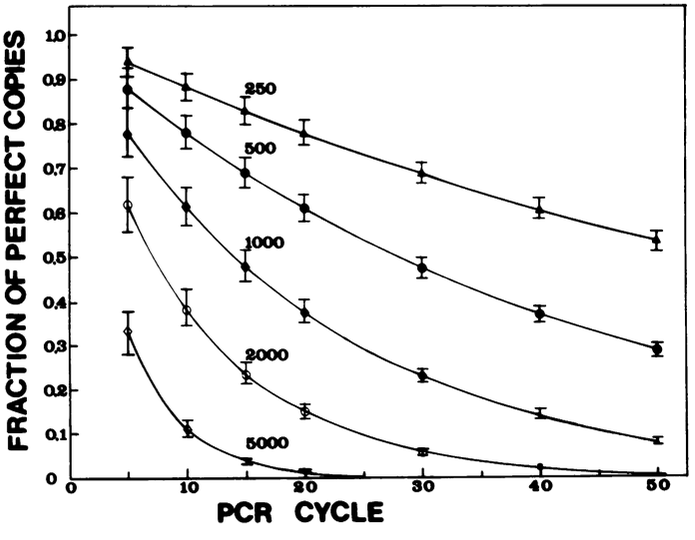

Fig. 1. Mean proportion (%) of correct replicates as a function of cycle number and fragment length (calculated using formulae (1) and (2) from text). Vertical bars represent twice the standard deviation.

Table 1. Results expected for a mismatch analysis following PCR with variable fraction of cleavage (r). One original deviation, 2000 bp per strand, 30 cycles, 10 starting copies, detection proportion c=1.0 (strands degraded only at the original mismatch as a proportion (%) of the total degraded strands is given in brackets).

r	Proportion (%) of strands degraded	Proportion (%) of strands degraded only at the original mismatch (i.e. yield of signal)
0.1	33.32	7.41 (22.24)
0.2	56.08	10.98 (19.58)
0.3	71.53	12.20 (17.06)
0.4	81.92	12.05 (14.71)
0.5	88.84	11.16 (12.56)
0.6	93.38	9.92 (10.63)
0.7	96.32	8.58 (8.90)
0.8	98.19	7.26 (7.39)
0.9	99.33	6.05 (6.09)
1.0	100.00	4.98 (4.98)

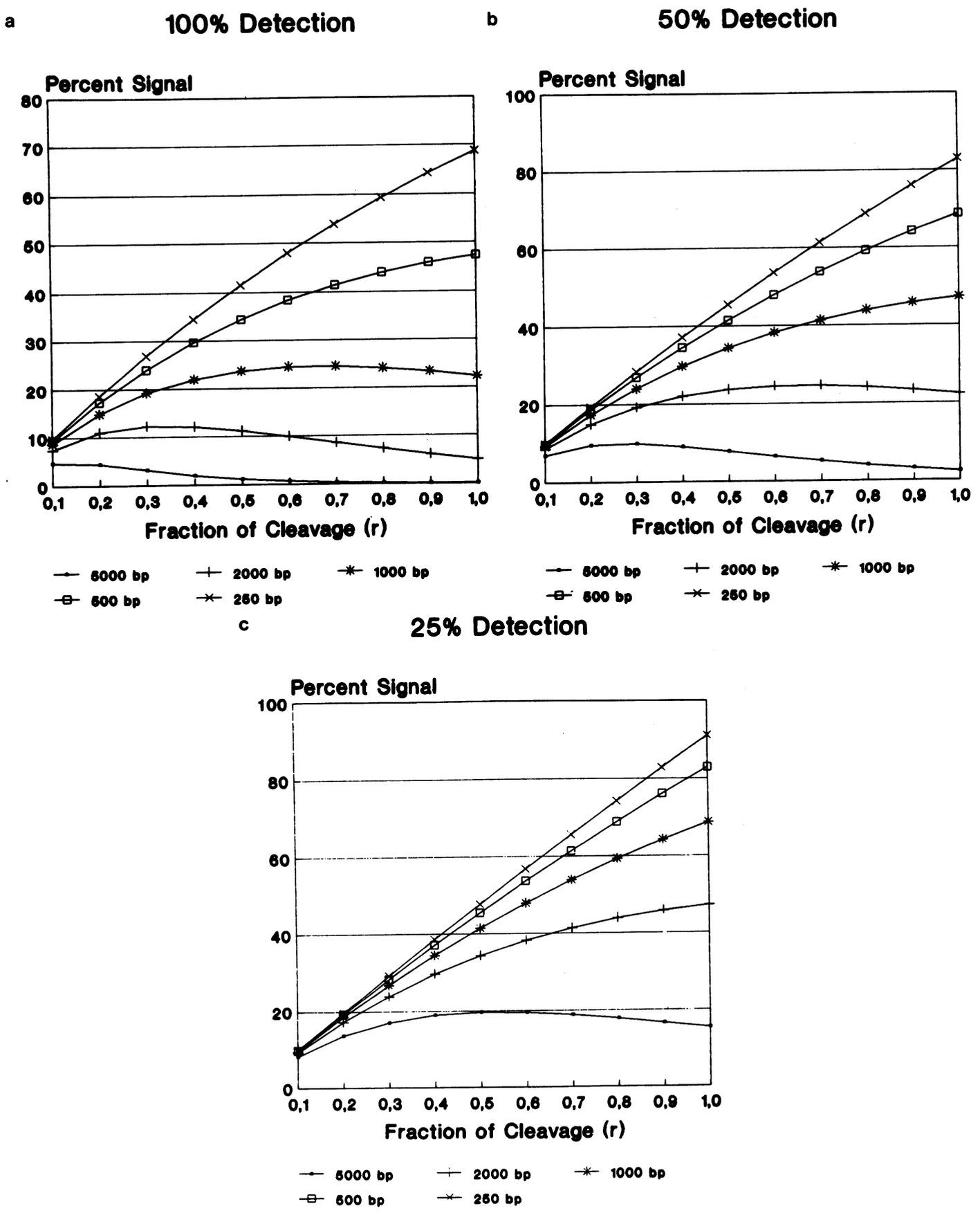

Fig. 2. One original deviation: ratio (%) of fragments degraded only at the original mismatch vs. fraction of cleavage (r). Calculations are based on 30 cycles of perfect (not experimental) cycles.

a) 100% detection (c=1.0, denaturing methods or combination of two base-specific mismatch reactions analysing both strands)

b) 50% detection (c=0.5, applicable to base-specific detection as described (10, 11)

c) 25% detection (c=0.25, for strand- and base-specific procedures, e.g. single base-specific mismatch analysis with single strand)

produce visible artefacts, but the latter will decrease proportional to the true signal because further misincorporation events will serve to diminish these subpopulations, too. Table 2 shows the artefactual signal intensities expected from a single error during the first cycle of amplification of a 1 kb fragment, not bearing an original deviation. Although detectable errors in the first PCR cycle are very likely (Table 2a), a false-positive signal is less than 3.5% of the total label with all applied detection systems (Table 2b).

Two original sequence deviations within an amplified fragment might be a rare event, but nevertheless theoretically possible. Table 3 shows, that both deviations are detectable with a certain intensity, if the complete strand is labelled. Obviously, only one deviation will be detected by a combination of end-labeling and complete degradation (r=1). Partial modification or cleavage, however, will reflect both deviations and the fraction of cleavage (r) can again be used to maximize the corresponding signals.

DISCUSSION

We have demonstrated that the relatively high error rate of Taq DNA Polymerase can hinder mismatch analysis after amplification of a DNA sequence potentially harbouring a mutation. Our theoretically derived results suggest that false-positive signals will not reach significant intensities. The utility of the analysis may be optimized by paying attention to two main factors. Firstly, it is important, that in most cases where the wild-type sequence is available, only the target DNA (e.g. patient DNA) should be amplified. Use of two amplified control DNA species (patient and wild-type DNA) for heteroduplex formation would decrease the correct signal by the square.

Secondly, we suggest, depending upon the method, optimal fragment lengths and/or partial degradation strategies in case of chemical mismatch analysis. Our study of the influence of partial reactions might not only be relevant to the deliberate performance

Table 2. Probability of early replication errors (a) and expected false-positive signals in mismatch analysis of a correct fragment (i.e. no original deviation) (b). 1000 bp per strand, 30 cycles, 10 starting copies.

a) Probability of m detectable replication errors in first cycle.

	detection proportion (c)		
m	0.25	0.5	1.0
0	0.778	0.606	0.368
1	0.194	0.303	0.368
2	0.024	0.076	0.184
3	0.002	0.013	0.061

b) Expected proportion of strands degraded assuming exactly one replication error causing a detectable mismatch in first cycle.

	Proportion (%) of strands degraded			Proportion (%) of strands degraded only at the false mismatch		
		c			c	
r	0.25	0.5	1.0	0.25	0.5	1.0
0.2	7.9	14.4	25.9	0.93	0.87	0.75
0.4	15.2	26.7	45.1	1.73	1.50	1.12
0.6	21.9	37.2	59.3	2.41	1.94	1.26
0.8	28.1	46.2	69.9	2.99	2.24	1.25
1.0	33.8	54.0	77.7	3.48	2.42	1.17

Table 3. Results expected for a mismatch analysis of a fragment with two original deviations. 1000bp per strand, 30 cycles, 10 starting copies. The upper value in the right column refers to degradation at one original mismatch only, the lower value to degradation at both original mismatches only.

	Proportion (%) of strands degraded			Proportion (%) of strands degraded only at the original mismatch(es)		
		c			c	
r	0.25	0.5	1.0	0.25	0.5	1.0
0.2	40.6	44.9	52.6	14.8	13.8	11.9
				3.7	3.4	3.0
0.4	69.0	73.3	80.0	20.6	17.8	13.2
				13.7	11.9	8.8
0.6	87.2	89.8	93.5	19.1	15.3	9.8
				28.7	23.0	14.6
0.8	97.0	97.8	98.8	11.8	8.8	4.8
				47.4	35.1	19.3
1.0	100.0	100.0	100.0	0.0	0.0	0.0
				68.7	47.2	22.3

of these kinds of experiments, but will perhaps also prove helpful in the interpretation of reactions planned but not performed with 100% efficiency. Moreover, it is relevant for special events, e.g. the detection of more than one deviation in a fragment analyzed by end-labelling procedures. The influence of the inevitable misincorporation during signal intensification by PCR varies with the base-specifity of the detection system. To date, it seems likely that future gene analysis in research and diagnostic medicine will involve the amplification of distinct segments such as widely spaced exons . For rapid identification of sequence deviations (e.g. mutations) the first step would be the identification of the mismatch-carrying fragment from several possible candidates. A rough localisation of the sequence deviation within this fragment without additional effort is highly desirable. To this end, chemical mismatch analysis appears to be superior to the denaturing or melting approaches.

We distinguished between 100%, 50% and 25% models describing various experimental approaches. For large screening projects, an 'all-in-one' strategy is to be preferred. Such a 100% model is provided by the denaturing approach, although the latter has a limited exclusion potential. The two base-specific reactions of chemical mismatch analysis can be combined in a powerful, informative and convenient first step analysis. If used in two separate reactions, the 50% model is applicable, thereby doubling the number of samples to be processed, but yielding more information. No experimental system to date fits the 25% model. However, a combination of generating single-stranded DNA by PCR with unequal primer concentrations (13) followed by chemical mismatch analysis, with separate reactions, could be considered. The 25% model provides unambiguous information about the nature of a base substitution. Provision of single base resolution in the size determination of degraded fragments (as in sequencing) makes a one-step mutation identification strategy feasible. This could in principle substitute for RFLP analysis with all its drawbacks and/or laborious DNA sequencing, thereby revolutionizing mutation detection and diagnostic medicine.

ACKNOWLEDGEMENTS

We thank David Millar for helpful discussions and Doris Immke for secretarial assistance.

REFERENCES

1. Saiki R, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354
2. Chien A, Edgar, DB, Trela JM (1976) Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. J. Bacteriol. 127:1550–1557
3. Tindall KR, Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. Biochem. 27:6008–6013
4. Krawczak M, Reiss J, Schmidtke J, Rösler U (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. Nucleic Acids Res. 17:2197–2201
5. Gibbs RA, Nguyen PN, McBride LJ, Koepf SM, Caskey CT (1989) Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of in vitro amplified cDNA. Proc. Natl. Acad. Sci. USA 86:1919–1923
6. Green PM, Bentley DR, Mibashan RS, Nilsson IM, Giannelli F (1989) Molecular pathology of haemophilia B. EMBO J. 8:1067–1072
7. Chelly J, Concordet JP, Kaplan JC, Kahn A (1989) Illegitimate transcription: Transcription of any gene in any cell type. Proc. Natl. Acad. Sci. USA 86:2617–2621
8. Sarkar G, Sommer SS (1989) Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. Science 244:331–334
9. Myers RM, Sheffield VC, Cox DR (1988) Detection of single base changes in DNA:ribonuclease cleavage and denaturing gradient gel electrophoresis. Chap.5 in Genome analysis—a practical approach. Ed. Davies KE. IRL Press, Oxford.
10. Montandon AJ, Green PM, Giannelli F, Bentley DR (1989) Direct detection of point mutations by mismatch analysis: application to haemophilia B. Nucleic Acids Res. 17:3347–3358
11. Grompe M, Muzny DM, Caskey, CT (1989) Scanning detection of mutations in human ornithine transcarbamoylase by chemical mismatch cleavage. Proc. Natl. Acad. Sci. USA 86:5888–5892
12. Attree O, Vidaud D, Vidaud M, Amselem S, Lavergne JM, Goossens M (1989) Mutations in the catalytic domain of human coagulation factor IX: Rapid characterization by direct genomic sequencing of DNA fragments displaying an altered melting behavior. Genomics 4:266–272
13. Gyllensten UB, Erlich HA (1988) Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus Proc. Natl. Acad. Sci. USA 85:7652–7656