PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences

L A Clarke, C S Rebelo, J Gonçalves, M G Boavida, P Jordan

Abstract

The polymerase chain reaction (PCR) is used universally for accurate exponential amplification of DNA. We describe a high error rate at mononucleotide and dinucleotide repeat sequence motifs. Subcloning of PCR products allowed sequence analysis of individual DNA molecules from the product pool and revealed that: (1) mononucleotide repeats longer than 11 bp are amplified with decreasing accuracy, (2) repeats generally contract during PCR because of the loss of repeat units, (3) Taq and proofreading polymerase Pfu generate similar errors at mononucleotide and dinucleotide repeats, and (4) unlike the parent PCR product pool, individual clones containing a single repeat length produce no “shadow bands”. These data demonstrate that routine PCR amplification alters mononucleotide and dinucleotide repeat lengths. Such sequences are common components of genetic markers, disease genes, and intronic splicing motifs, and the amplification errors described here can be mistaken for polymorphisms or mutations.

Keywords: polymerase chain reaction; nucleotide repeat sequences; shadow bands

The polymerase chain reaction (PCR) is one of the most widely used techniques in molecular biology and has made possible a great variety of both diagnostic and research applications. Examples are the detection of gene mutations, the analysis of polymorphic markers and microsatellite loci, analysis of gene expression, DNA cloning, and site directed mutagenesis. Because PCR involves the exponential amplification of target sequences, a high degree of polymerase fidelity is essential if the introduction of a large number of replication errors during the PCR reaction is to be avoided. Most commercially available Taq polymerases introduce errors at the rate of approximately $10^{-3}$ to $10^{-4}$ point mutations/bp/duplication, with higher fidelity polymerases such as Pfu and Deep Vent generating up to eight times fewer errors. In contrast, we found that mononucleotide and dinucleotide repeats were not faithfully reproduced during PCR.

Methods and results

A stretch of 26 A nucleotides in intron 5 of the hMSH2 gene, termed Bat-26, was amplified from the genomic DNA of several healthy individuals under routine PCR conditions (200 ng genomic DNA and 1.25 U AmpliTaq (Perkin Elmer, Branchburg, New Jersey, USA) in 50 µl containing 1.5 mM MgCl₂, 300 ng each primer, and 200 µM dNTPs). Direct sequencing of the PCR product produced an illegible sequence after the mononucleotide repeat, indicating a possible difference between two alleles. Such allele differences, along with any polymerase slippage incurred during the PCR reaction, can both be visualised using gene scanning technology. However, for our study a strategy of subcloning and sequencing of individual clones of the PCR product was used for the qualitative determination of the composition of the PCR product pool, by identifying the sequences of individual DNA molecules. This revealed differences in the length of the Bat-26 poly-A stretch, which varied from 19 to 28 bp, which is incompatible with the concept of simple polymorphism or mutation (table 1). The sequencing of more than 30 individual clones revealed that the repeat was predominantly shortened, with only 35% of the clones containing the predicted sequence of (A)_{26} (table 1). These data suggested that a systematic polymerase error had taken place during the PCR reaction, which was specific to the mononucleotide repeat.

To determine in more detail the performance of PCR amplification at similar sequences, repeats of 21, 13, 11, and nine mononucleotides were amplified and subcloned. The polyprymidine tract of the hMLH1 intron 11 splice acceptor site contains a (T)_{21} mononucleotide repeat, termed Bat-21 (AcNb U40971), with an adjacent (TA)_{11} dinucleotide repeat. The polyprymidine tract preceding exon 2 in the gene hMSH2 contains a (T)_{21} stretch, termed Bat-21 (AcNb U41207), and intron 4 of the human RAC1 gene contains both (T)_{11} and (T)_{21} runs (AcNb AJ132695). As shown in table 1, sequencing of individual cloned PCR products revealed incorrect amplification of the (T)_{21} and (T)_{11} repeats, whereas (T)_{11} was replicated faithfully. The limit for correct amplification was reached with (T)_{11}, where 90% of the cloned products contained the predicted number of thymidines. The predominant observation was of repeat contraction; the (T)_{21} repeat appeared to expand, but this tendency was accounted for upon inspection of individual Bat-21 clones by A-T transversions in the adjacent TA repeat. Two different amplification errors at this combined repeat therefore led to an overall expansion of the poly-T stretch.

To determine the performance of the high fidelity Pfu polymerase at such sequences, the
strategy was repeated for the above markers (with 200 ng genomic DNA and 2.5 U Pfu DNA polymerase (Stratagene, Amsterdam, the Netherlands) in 100 µl containing 2mM MgCl₂, 600 ng of each primer, and 250µM dNTPs for 30 cycles (95°C for 45 seconds; annealing at 60°C for one minute; 72°C for two minutes) and showed that the limit for correct amplification by Pfu was raised from (T)₁₁ to (T)₁₃, but that longer repeats were also incorrectly amplified (table 1).

Dinucleotide repeats are another group of frequently amplified genetic markers. We amplified the dinucleotide repeat D15S128 (AcNb Z17197) in an individual previously determined to be homozygous for (CA)₁₈, the allele with the highest frequency. PCR was carried out with both Taq and Pfu polymerases, followed by subcloning and sequencing of individual inserts. Both enzymes were found to amplify this repeat with a high error rate (table 1), as seen for the longer mononucleotide repeats.

Trinucleotide (CAG)n repeats occur in the coding regions of disease causing genes, such as the androgen receptor or Huntington's and Machado Joseph disease genes. Expansion of the triplets is associated with severe neurodegeneration, and genetic analysis of patients relies on PCR amplification of the repeats from genomic DNA. To test for possible errors introduced by in vitro amplification, the (CAG)n and (GGC)n trinucleotide repeats in the first coding exon of the androgen receptor gene (AcNb NM_000044) were amplified by both Taq and Pfu, revealing that over 80% of clones contained the correct sequences (data not shown), a level of error compatible with direct sequencing. These data are also consistent with previous observations that longer nucleotide repeat units are associated with less polymerase error.

An accepted feature of nucleotide repeat marker analysis, both in mutation detection or analysis of genetic polymorphisms, is the occurrence of staining patterns called “shadow bands”. When PCR products of such markers are separated by non-denaturing polyacrylamide gel electrophoresis they appear as broad bands or separate into a series of individual shadows. To determine the contribution of PCR polymerase errors to this phenomenon, a Bat-26 PCR product amplified with Taq was run on a polyacrylamide gel alongside several individual clones of the same product with a known number of A nucleotides. The latter were excised from the cloning vector by restriction digestion. As shown in fig 1, the PCR product appeared as a broad, smeared area of adjacent shadow bands, whereas the cloned fragments of defined length migrated as discrete single bands.

Table 1  Repeat length of individual DNA sequences subcloned from a PCR product pool after amplification of mononucleotide or dinucleotide repeat markers

<table>
<thead>
<tr>
<th>Locus</th>
<th>Repeat</th>
<th>Repeat length and corresponding number of clones</th>
<th>N</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAC1</td>
<td>(T)₉</td>
<td>9</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>(T)₁₁</td>
<td>10</td>
<td>11</td>
<td>90</td>
</tr>
<tr>
<td>Bat-13</td>
<td>(T)₁₁</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>(T)₁₃</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Bat-21</td>
<td>(T)₂₃</td>
<td>23</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>(T)₁₃</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Bat-21</td>
<td>(T)₂₃</td>
<td>23</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>(T)₁₃</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Bat-21</td>
<td>(T)₁₃</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>D15S128</td>
<td>(CA)₂₆</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>(CA)₂₆</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
</tbody>
</table>

Each locus was amplified by PCR with either Taq or Pfu, as indicated. The number of clones found containing varying numbers of repeat units is shown, with the correct number of repeats for each locus shown in bold. N, total number of cloned inserts sequenced for each locus; % Correct, the percentage containing the exact repeat length.

Figure 1  Non-denaturing polyacrylamide gel electrophoresis comparing the Taq polymerase amplified Bat-26 marker and individual clones obtained after subcloning this PCR product. The original PCR product (indicated by the arrow) was a pool of molecules with different repeat lengths and it migrates as a blurred area composed of “shadow bands”. In contrast, cloned individual molecules containing a known number of repeated adenines (indicated by numbers) are seen as discrete bands. The individual clones contain an extra 11 bp of vector sequence after restriction digestion of the cloned insert with EcoRI and thus migrate more slowly than the original PCR product.
forming a size ladder according to the length of the poly-A stretch. This demonstrates that the phenomenon of shadow bands commonly observed in PCR-based genetic analysis of microsatellite loci can be attributed at least in part to artefactual PCR amplification errors. In addition, others have shown that improper annealing of PCR products during non-denaturing gel electrophoresis can further compound the occurrence of multiple band patterns.1,13

Discussion
In our study, we describe a high error rate in PCR amplification of mononucleotide and dinucleotide repeats from genomic DNA. Our findings relate mainly to monothymidine/methylenosine repeats, and suggest that the longer the repeat, the greater the errors made during amplification. Although our additional data include only TA and CA dinucleotide repeats, we believe it is possible to extrapolate the general rule that PCR amplification of mononucleotide or dinucleotide repeats results in error. Use of the high fidelity proofreading polymerase Pfu in place of Taq restricted the occurrence of such errors only in the case of short monothymidine repeats. We think that contraction of the repeat itself is the most common type of error occurring during PCR amplification. This type of error is commonly referred to as polymerase “slippage” and is probably caused by slipped strand mispairing.14,15 Alternatively, the loss or gain of nucleotide repeat units without affecting the surrounding sequence16,17 can be explained by mega-priming. In this case, fragments that were either incompletely synthesised or broken in their repetitive element during PCR can anneal under formation of mismatches, then extend, resulting in repeat length variation, independent of polymerase type or PCR conditions. This conclusion was drawn from work using synthetic oligonucleotides in the absence of genomic DNA.18,19 Previous work has suggested that such polymerase error can be reduced by optimisation of PCR reaction conditions and buffer composition.13,18–25 In our present study, a fixed set of routine PCR conditions were used to determine the effect that repeat size and length alone would have on polymerase error rate. Our data lead to the conclusion that polymerase performance itself imposes considerable constraints upon PCR amplification fidelity independent of PCR conditions, primer composition, or PCR product length.

Although the molecular mechanism of the described errors during amplification of repetitive DNA sequence motifs remains a matter of discussion, such errors are frequently encountered during cloning or genetic analysis of introns, polyprymidine tracts, or microsatellite markers. In addition, they are of great interest during analysis and diagnosis of various diseases. For example, mononucleotide and dinucleotide microsatellite sequences are hotspots for mammalian polymerase errors during in vivo DNA replication,26 and are highly unstable in tumours resulting from mismatch repair deficiencies.27 On this basis, they are widely used as markers for certain cancer syndromes,28 including hereditary non-polyposis colorectal cancer.29 Furthermore, dinucleotide repeats are also used as polymorphic markers to determine heterozygosity. With the increased use of automated DNA sequencing, the described in vitro amplification errors at repeat motifs can easily be mistaken, in research and diagnostics, for polymorphism or mutation.

We thank C Caldas for comments on the manuscript, L Vieira for donating D15S128 primers, S Pedro for running the ABI sequencing apparatus, and S Beck and S Vieira for assistance with high performance liquid chromatography analysis. This study was supported by PRAXIS XXI grant 2/2.1/SAU/1997/05 from the Fundação para a Ciência e a Tecnologia, and PRAXIS XXI postdoctoral fellowship BPD/1140/96 to LAC.