3538–3545 *Nucleic Acids Research, 1996, Vol. 24, No. 18*

A mathematical model and a computerized simulation of PCR using complex templates

Eitan Rubin and Avraham A. Levy*

Department of Plant Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Received June 14, 1996; Revised and Accepted July 31, 1996

ABSTRACT

A mathematical model and a computer simulation were used to study PCR specificity. The model describes the occurrences of non-targeted PCR products formed through random primer–template interactions. The PCR simulation scans DNA sequence databases with primers pairs. According to the model prediction, PCR with complex templates should rarely yield non-targeted products under typical reaction conditions. This is surprising as such products are often amplified in real PCR under conditions optimized for stringency. The causes for this 'PCR paradox' were investigated by comparing the model predictions with simulation results. We found that deviations from randomness in sequences from real genomes could not explain the frequent occurrence of non-targeted products in real PCR. The most likely explanation to the 'PCR paradox' is a relatively high tolerance of PCR to mismatches. The model also predicts that mismatch tolerance has the strongest effect on the number of non-targeted products, followed by primer length, template size and product size limit. The model and the simulation can be utilized for PCR studies, primer design and probing DNA uniqueness and randomness.

INTRODUCTION

The polymerase chain reaction (PCR) allows the amplification of a specific region (target) from a DNA template, using two oligonucleotides (primers) that anneal to opposite strands (1,2). The reaction is based on multiple cycles of DNA synthesis; each includes denaturation of the template, annealing of the primers to complementary sites in the template and primer extension. The high sensitivity of the reaction and its low cost in time and reagents make PCR one of the most significant innovations in molecular biology during the past decade (3). Nevertheless, for any new primer–template combination, the behavior of the reaction is not completely predictable; non-targeted products are often amplified (4), particularly when complex templates, such as genomic DNA, are involved in the reaction. This problem has been addressed by empirically optimizing the components of the reaction (5), or by experimentally investigating the specificity of the priming process (6–12). Despite some progress made through these approaches, we still have a limited understanding of the factors that govern PCR specificity.

Several computer programs have been used to predict the formation of non-targeted products (12–18). Such products may occur when two opposite regions in the template, situated within a certain size limit, are similar enough to the primer to serve as annealing sites (19). The currently available primer design programs cannot handle complex templates and, therefore, have a limited prediction capability. This is unfortunate, as DNA sequence databases (DBases) such as GenBank (NCBI, 8600 Rockville Pike, Bethesda, MD 20894, USA) and EMBL (EBI, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK), contain an increasing amount of information which could be used to more accurately predict the amplification of non-targeted products. These DBases contain a very large collection of $>2.5 \times 10^8$ nucleotides (nt) from numerous species, including both a biased sample of the genome and a rapidly growing assembly of unbiased sequences in the form of complete chromosomes (20–24). Thus, sequence DBases are the best approximation of complex templates such as genomic DNA or cDNA libraries.

We combined two approaches to study the amplification of non-targeted PCR products as a result of random primer–template interactions: first, through a mathematical model, and second, through a computerized simulation of PCR (simPCR). The model was developed to assess the relative effect(s) of various parameters on the amplification of non-targeted products. It was also used to determine the expected frequency of non-targeted products when the template is a random sequence. This frequency was then compared with that of non-targeted products obtained by scanning real sequence databases with the computerized PCR simulation. We reached the following conclusions: (i) the expected probability of obtaining a non-targeted PCR product under stringent annealing conditions is extremely low; (ii) the frequent amplification of non-targeted products in real PCR is not caused by deviations from randomness in nucleotides order or composition, but rather by the tolerance of PCR to mismatches; and (iii) based on the model predictions, mismatch tolerance is the most significant factor affecting PCR specificity, followed by primer length, template size and product size limit. We discuss how the predictions based on the model and the simulation could be useful for future improvement in PCR specificity and primer design. In addition, the equations and simulation that we developed can be used to study many other biological processes

* To whom correspondence should be addressed

which, similarly to PCR, involve the recognition of sequences along the DNA in complex genomes.

MATERIALS AND METHODS

Definitions

Mismatch tolerance: the maximal number of mismatches allowed between the primer and a sequence in the template. *Find*: a sequence in a database entry identified with Findpatterns as similar to the primer within mismatch tolerance. *simPCR product*: the sequence between two opposing primers. This includes *solo simPCR products*: the sequence defined by a single primer repeated in the template in an inverted orientation; or *XY-simPCR products:* the sequence defined by two different opposing primers. *Degenerate primers:* mix of oligonucleotides, each containing alternative bases at specific sites. All possible combinations are equally represented in the mix. A degeneracy is mismatched if none of the possible bases at a location is identical to the target.

Computer algorithm for simPCR

The Findpatterns program from the GCG package is first used to search for annealing sites in the DBases (mismatches are allowed by using the appropriate option). SimPCR then creates for each entry an array of finds, each represented by the annealed primer number, position, number of mismatches and orientation. All the pairs in the array are examined, and a 'product' is reported when the two sites are opposite and satisfy all search parameters. The user defines mismatch tolerance and product size limit. The limitations of this simulation as a representation of PCR are described in the discussion. Flowcharts of the program are available through the WWW (see below).

Databases

Two databases were used in all simPCR searches. The first contained selected subdivisions from the GenBank database (release 90) omitting subdivisions which contain, on average, entries <1000 bp (EST, STS, UNA, PAT, RNA and SYN). The remaining subdivisions (BCT, INV, MAM, VRT, PHG, PLN, PRI, ROD and VRL) contain 2.38×10^8 bp arranged in 168 434 entries. The second DBase contained random sequences containing 2.5×10^6 bp organized in 250 entries. Each entry of this DBase was created by concatamering 2500 ACGT repeats and randomizing nucleotide order, using the SHUFFLE routine from GCG. Randomness was confirmed by checking the score distribution of FASTA comparisons between random sequences against the complete database, all of which showed a uniform distribution of scores (data not shown).

Primers

The sequence of the primers used in this work and a number of PCR-related parameters are described in Table 1. The performance of the simple primers in pairs S, and the degenerate primers in pair D was tested in real PCR reactions. Primers of type R were randomly generated and were never used in real PCR. Primers in pair S were designed based on the nucleotide sequence of their target. Pair D is degenerate, as its sequence was deduced from the amino acid sequence of its target. In real PCR, pair S was successful in specifically amplifying its target (13), but pair D was not (G. Benet, personal communication).

Availability

Sources of the simPCR program in the C programming language, together with related materials, can be obtained directly from the authors by anonymous FTP to bioinformatics.weizmann.ac.il in the directory/pub/software, or through the world wide web in http://dapsas1.weizmann.ac.il/~bcrubin/simPCR/simPCR.html. It is also possible to run a demo version of the program though the WWW server.

RESULTS

A model of the PCR

A model was developed to describe the formation of unspecific PCR products as a function of several parameters (Fig. 1). The following conditions were used: (i) the template is a double-stranded DNA sequence made of 4 nt (A, C, G, T) in an equal ratio and a random order; (ii) annealing may occur at any site of the template which is similar to the primer within mismatch tolerance, with successful annealing always leading to priming; and (iii) any two opposing sites within product size limit give a PCR product. These conditions were chosen for the sake of simplicity. The limitations of this model, together with possible improvements, are discussed below (see Discussion).

The formation of a PCR product with a size limit L_s using a template of length L_t requires the annealing of two primer molecules of length l to opposing strands (Fig. 1). Notwithstanding mismatch tolerance, there are L_t annealing sites on each strand as usually $L_t >> l$ and hence the effect of template ends is negligible. Consider the total number of annealing-site pairs, N_{pairs}, which are in a correct orientation and within product size limit (see parallelogram area in Fig. 1C). $N_{pairs} = L_s L_t$ for templates where the effect of template ends is negligible, i.e. when $L_t >> L_s$ (see black triangle area in Fig. 1C).

When two different primers, X and Y, are used in a reaction, three types of products are possible: XY-products are formed when primer X anneals to one strand and primer Y to the other, and XX or YY solo-products are formed when the same primer anneals to both strands. When X≠Y, two different XY-products can be obtained depending on the type of annealing: X may anneal to the (+) and Y to the (–) strand, or Y may anneal to the (+) and X to the (–) strand. Therefore, there are $2N_{pairs}$ possible XY-products. For solo-products, the two types of annealing give identical products, i.e. there are N_{pairs} possible solo-products for either X or Y.

PCR product is amplified only if both primers anneal within the range of mismatch tolerance. The probability that a site will anneal to a specific primer with precisely m mismatches, $p(l,m)$, can be derived from the binomial distribution. Annealing can be considered as l experiments, of which $l–m$ must 'succeed', and m must 'fail'. The probability for each base to 'succeed' in finding an identical base in the template is $\frac{1}{4}$, and the probability of a 'failure' (a mismatch) is $\frac{3}{4}$. Therefore $p(l,m)$ is:

$$p(1, m) = \binom{l}{m}\left(\frac{1}{4}\right)^{l-m}\left(\frac{3}{4}\right)^{m} = \binom{l}{m}3^{m}\left(\frac{1}{4}\right)^{l} \quad \mathbf{1}$$

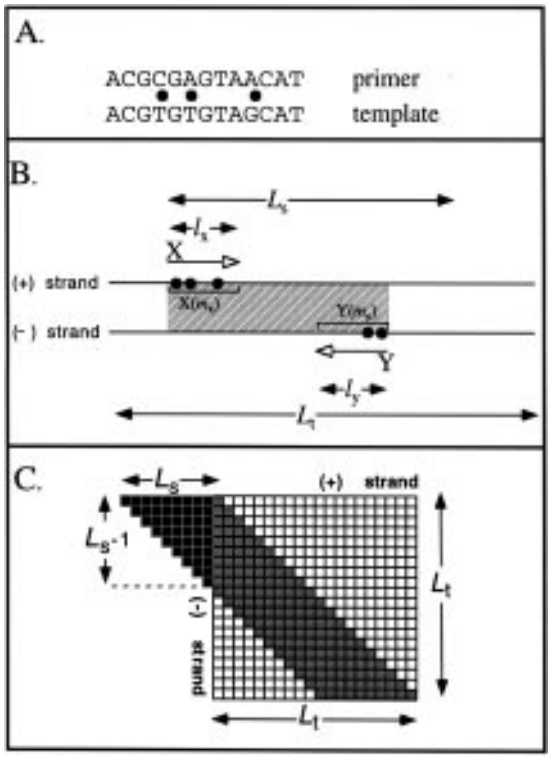

Figure 1. Parameters considered in the PCR model. (**A**) An example of primer–template annealing site with three mismatches (indicated as dots). (**B**) A PCR product (shown by the filled box) obtained with primers X and Y (indicated by empty-headed arrows). Filled arrows indicate primers length (l_x and l_y), maximal product size (L_s) and template size (L_t). The PCR product in this example is formed following annealing of X with $m_x = 3$ mismatches to the (+) strand of the template and of Y with $m_y = 2$ mismatches to the (–) strand. (**C**) Representation of all annealing sites pairs (small grey squares) which can form a PCR product. One site in each pair occurs on the (+) strand and the other on the complementary strand (–). In very long templates ($L_t >> L_s$), the number of PCR products can be calculated from the area of a parallelogram shown by the combined grey and black areas. The small black squares represent imaginary pairs with one annealing site outside of the template. Their number $[L_s(L_s - 1)/2]$ is negligible in large templates but should be subtracted from the parallelogram area when $L_t > L_s$.

The probability of both primers, X and Y, annealing to any pair with precisely m_x and m_y mismatches, respectively, is $p(l_x,m_x)p(l_y,m_y)$. Therefore, the number of XY-products $u_{xy}(m_x,m_y)$, obtained with precisely m_x and m_y mismatches, is:

$$u_{xy}(m_x,m_y) = 2N_{pairs}p(l_x,m_x)p(l_y,m_y) \qquad \mathbf{2}$$

The number of products obtained within mismatch tolerance, U_{xy}, is the sum of $u_{xy}(m_x,m_y)$ for all combinations of m_x and m_y which satisfy $0 \le m_x \le m_{max}$, $0 \le m_y \le m_{max}$:

$$U_{xy} = \sum_{m_x=0}^{m_{max}} \sum_{m_y=0}^{m_{max}} 2N_{pairs}p(l_x, m_x)p(l_y, m_y) =$$

$$= 2L_tL_s\left(\frac{1}{4}\right)^{(l_x+l_y)} \sum_{m_x=0}^{m_{max}} \sum_{m_y=0}^{m_{max}} \binom{l_x}{m_x}\binom{l_y}{m_y} 3^{(m_x+m_y)} \qquad \mathbf{3}$$

Solo-PCR products can be obtained from the primer annealing with m_1 mismatches to one strand and with m_2 mismatches to the other, where $0 \le m_1 \le m_{max}$ and $0 \le m_2 \le m_{max}$. There are only N_{pairs} pairs which may give a solo product (see above explanation for X ≠ Y). Therefore, the number of XX-products, U_{xx}, is:

$$U_{xx} = L_tL_s\left(\frac{1}{4}\right)^{2l_x} \sum_{m_1=0}^{m_{max}} \sum_{m_2=0}^{m_{max}} \binom{l_x}{m_1}\binom{l_x}{m_2} 3^{m_1+m_2} \qquad \mathbf{4}$$

Similarly, the number of YY-products, U_{yy}, is:

$$U_{yy} = L_tL_s\left(\frac{1}{4}\right)^{2l_y} \sum_{m_1=0}^{m_{max}} \sum_{m_2=0}^{m_{max}} \binom{l_y}{m_1}\binom{l_y}{m_2} 3^{m_1+m_2} \qquad \mathbf{5}$$

In total, the number of PCR products of all three types, U, is

$$U = U_{xy} + U_{xx} + U_{yy} \qquad \mathbf{6}$$

For further comparisons, the number of annealing sites was calculated. In a double-stranded template, there are $2L_t$ annealing sites which will anneal with the primer, notwithstanding mismatch tolerance, and considering that the effect of the ends is negligible as $L_t >> l$. The probability of primer X with length l_x annealing with precisely m_x mismatches is $p(l_x,m_x)$, therefore the number of sites which anneal to primer X with no more than m_{max} mismatches, A_x, is:

$$A_x = \sum_{m=0}^{m_{max}} 2L_tp(l_x, m_x) = 2L_t \sum_{m=0}^{m_{max}} \binom{l_x}{m} 3^m \left(\frac{1}{4}\right)^{l_x} \qquad \mathbf{7}$$

Similarly, for primer Y

$$A_y = 2L_t \sum_{m=0}^{m_{max}} \binom{l_y}{m} 3^m \left(\frac{1}{4}\right)^{l_y} \qquad \mathbf{8}$$

In total, the number of annealing sites for both primers, A, is:

$$A = A_x + A_y \qquad \mathbf{9}$$

Modification of the model for degenerate primers. Primers containing degenerate bases require special treatment, since a mismatch in a degenerate base has a different effect than a mismatch in a non-degenerate one. Nevertheless, we used the equations described above by calculating effective primer length. Fully degenerate bases ('N') are ignored, and 2-fold degenerate bases (e.g. A or T) are considered as having a length of 0.5 bases. For example, the effective length of primer D.X shown in Table 1 is 15.5 bases, and of D.Y is 13.5 bases. $\binom{l}{m}$ of non-integer primer length is calculated by replacing factorials with the λ_x function. For example, for any integer m

$$\left(m - \frac{1}{2}\right)! = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2m-1)}{2^m}\sqrt{\pi} \qquad \mathbf{10}$$

Modification of the model for fragmented templates. The number of pairs which can give a PCR product, $N_{pairs} = L_sL_t$, is a good approximation when $L_t >> L_s$ (see parallelogram area in Fig. 1C). When the model is used to predict unspecific amplification from fragmented templates, such as cDNA library or sequence DBases, the effect of the multiple ends should be taken into account. Consider a single fragment of length L_i. If $L_i > L_s$, the total number of possible PCR products can be calculated by subtracting $1/2L_s(L_s - 1)$, the number of pairs which contain one annealing site outside the template (Fig. 1C), from L_iL_s. For sequences where $L_i < L_s$, the total number of possible PCR products is $1/2L_i(L_i + 1)$.

Table 1. Data on primers used for simPCR reactions

Primer name[a]	Primer sequence[b]	Effective length[c]	T_m[d] (min,max)	Target gene
S.X	GTTGTGGGTCACATAACGC	19	58	*Endothelin*
S.Y	AGAGTGTGTCTACTTCTGCC	20	60	
D.X	GCNATGGGNATGAA(C,T)ATG	15.5	(50,56)	*HMG-CoA reductase*
D.Y	GC(A,G)TGNGC(A,G)TT(A,G)AANCC	13.5	(48,58)	
R	$A_5C_5T_5G_5$	20	ND	None

[a]Each primer pair is denoted by a letter representing its type: S, simple (13); D, degenerate; R, random. Each member of a pair is denoted by X or Y.
[b]Degeneracies are indicated with brackets or as N for complete degeneracy. Five primers of type R were generated. The sequence of each primer was obtained by randomizing the order of its 20 nt. It is available upon request.
[c]Effective length was calculated by considering bases degenerate 2-fold (e.g. the 15th base, C/T, in primer D.X) as half a base and ignoring completely degenerate bases (e.g. the 3rd base, N, in primer D.X).
[d]T_m was calculated using the Data Minder Shareware (Karen Usdin, National Institutes of Health, Bethesda, MD 20892, USA) based on a nearest-neighbour analysis model (35) at 50 mM Na^+. For degenerate primers T_m was calculated for the molecule with the maximal and minimal GC content.

Consider a template composed of N fragments, of which Q are longer than L_s and R are shorter (N = Q + R). The total number of possible PCR products is calculated by substituting L_tL_s in equations **2–5** with the N_{pairs} value which takes ends into consideration:

$$N_{pairs} = \sum_{Q}\left[L_iL_s - \frac{1}{2}L_s(L_s-1)\right] + \sum_{R}\left(\frac{L_i^2}{2} + \frac{L_i}{2}\right) \qquad \mathbf{11}$$

Variables affecting the occurrence of unspecific PCR products according to the mathematical model

Using equations **6** and **9** from the model, the number of non-targeted PCR products or of annealing sites was calculated as a function of several parameters (Fig. 2): mismatch tolerance (Fig. 2A); primer length (Fig. 2B); maximal product length (Fig. 2C); and template length (Fig. 2D). Unless otherwise specified, reaction parameters were set to be characteristic of real PCR: maximal product length, L_s, of 3000 bp and template size similar to the Human genome ($L_t = 3\times10^9$ bp). These calculations indicate that mismatch tolerance is the variable with the strongest effect on PCR specificity. At low mismatch tolerance values, the proliferation of PCR products is nearly exponential. It becomes more moderate with increasing mismatch tolerance, until a maximal value of $4L_tL_s$ is reached (see equation **6**). In the example shown in Figure 2A, the number of products reaches $4L_tL_s = 3.6\times10^{13}$. Primer pairs of different length give shifted curves: shorter primer pairs give more products at 0 mismatches, and reach their maximal value at lower mismatch tolerance (Fig. 2A). Increasing primer length causes a nearly exponential reduction in the number of PCR products (Fig. 2B): 20 base primer pairs give five products with four mismatches, whereas 11 base primer pairs give 2×10^9 products with the same number of mismatches. Nevertheless, reducing primer length has a smaller effect than increasing mismatch tolerance. For example, reducing the length by two bases has a very similar effect to increasing the mismatch tolerance by one base (Fig. 2A, B and D). Other variables, such as maximal product length (Fig. 2C) and template length (Fig. 2D) have a linear effect on the number of products.

We compared the effect of increasing mismatch tolerance on the number of PCR products and annealing sites for 20 base primer pairs (Fig. 2A). The overall shape of the two curves is similar: at low mismatch tolerance, both increase nearly exponentially though at different rates, reaching a maximum at m_{max} = 20. There are 10^9 times less PCR products than annealing sites with zero mismatches. However, with increasing mismatches, the number of products and of annealing sites reaches similar values (between seven and eight mismatches), and finally, there are L_s

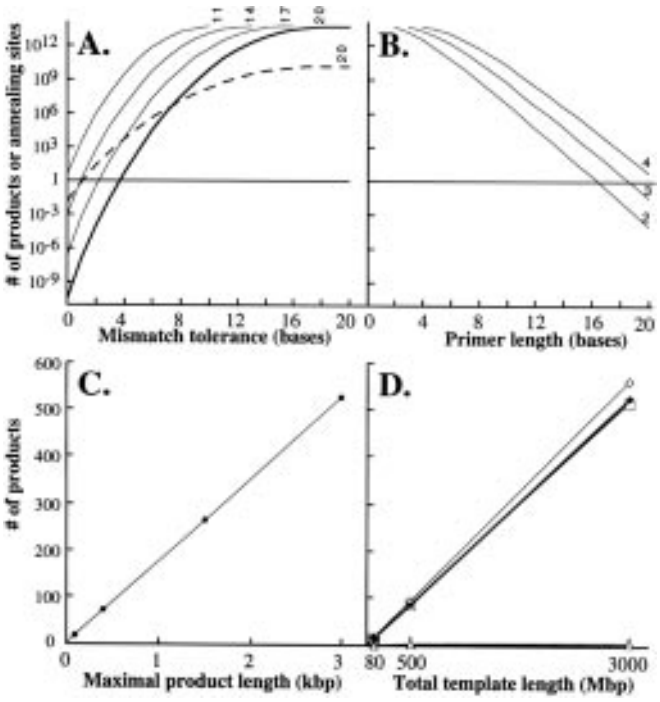

Figure 2. Effect of various factors on the number of unspecific PCR products. The total number of unspecific PCR products (*U*) predicted from the model is shown for examples in which members of a primer pair, X and Y, have the same length ($l_x = l_y = l$) and a mismatch tolerance of m_{max}. *U* is expressed as a function of mismatch tolerance (**A**), primer length (**B**), maximal product length (**C**) and template length (**D**). The number PCR products was calculated using equation **9** in the model. In (A), $L_t = 3\times10^9$ bp and $L_s = 3000$ bp. *U* is shown for primers pairs of different length as indicated above each curve. In addition, the number of annealing sites, as calculated from equation **12** is shown by a dashed line for $L_t = 3\times10^9$ bp and $l = 20$ bases. In (B), $L_t = 3\times10^9$ bp, $L_s = 3000$ bp and m_{max} is indicated at the right of each curve. In (C), $L_t = 3\times10^9$ bp, $l = 20$ bases and $m_{max} = 5$. In (D), $L_s = 3000$ bp, and different combinations of l and m_{max}. are used: $l = 20$ and $m_{max} = 5$ (full circles); $l = 18$ and $m_{max} = 4$ (diamonds); $l = 16$ and $m_{max} = 3$ (rectangles); and $l = 20$ and $m_{max} = 4$ (triangles).

(L_s= 3000 in Fig. 2A) times more PCR products than annealing sites when annealing is totally unspecific (m_{max} = 20).

The number of non-targeted PCR products expected under stringent conditions (zero mismatches) is extremely low; only 10^{-11} products are expected for primer–template combinations frequently used in real PCR, i.e. $L_t = 3 \times 10^9$ bp, l = 20 bases and L_s = 3000 bp (Fig. 2A). This result is surprising in view of the frequent occurrence of unspecific products in real reactions with similar primer–template combinations. We refer to this discrepancy as the 'PCR paradox'.

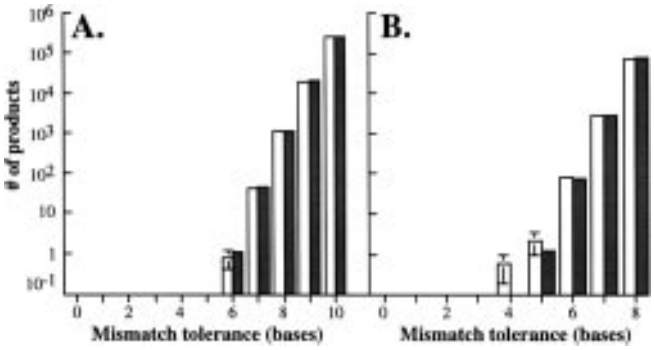

Figure 3. Model predictions versus simPCR results with random primers and random or natural sequences as templates. simPCR was run with a set of five random primers (R in Table 1). The average number of solo simPCR products are presented at various mismatch tolerance levels (empty columns), together with the number of products predicted from the model (filled boxes); the standard error of the mean (bars on top of columns). With increasing number of products the error bars are too small to be discerned. Simulations were performed with maximal product length set to 500 bp, using the DBase of random sequences totalling 2.5×10^6 bp (**A**), and a subset of GenBank as described in Materials and Methods (**B**). Model predictions were calculated using equation **4**, with L_s = 500 and l = 20.

Simulation of PCR with a complex template

The PCR paradox can be accounted for by the model provided a large number of mismatches is tolerated in real PCR. For example, a tolerance of four to five mismatches with a 20 nt primer could give one or a few non-targeted products (Fig. 2A), a value often observed in real PCR. Alternatively, it is possible that the frequent occurrence of non-targeted products stems from deviations from the model's conditions in real PCR. Two important assumptions of the model, namely, the randomness in nucleotide order and the equal representation of the four nucleotides, do not reflect real genomes. Possibly, annealing sites occur more frequently in real sequences than expected on a chance basis, reflecting biases in nucleotide composition and order. One way to test this possibility empirically is to simulate PCR with natural genomic sequences and examine whether the frequency of the obtained non-targeted products is higher than expected with a random genome. For that purpose, a program was written, simPCR, which can handle large templates such as GenBank or EMBL DBases (see Materials and Methods).

First, as a control, simPCR was run with random primers (Table 1) and a random database (Fig. 3A). The average number of solo simPCR products obtained with five different primers is shown in comparison with model predictions calculated from equation **4**. An excellent fit between the model and simPCR results is observed (Fig. 3A). The same random primer set was used in a simPCR run with natural template sequences from GenBank (see Materials and Methods). In this case, the template is fragmented, therefore the total number of annealing sites, N_{pairs}, as calculated from equation **11** was used in equations **2**–**5**. The good fit between the model and the simulation (Fig. 3B) suggests that for random primers, natural and random templates are similar with respect to non-targeted product formation.

Finally, the fit between the model and simPCR was tested for real primers with natural template sequences from GenBank (see Materials and Methods). simPCR was run with primers pairs S and D (see Table 1 and Fig. 4). For each primers pair the number of simPCR products, solo and non-solo, is shown in comparison with the model predictions calculated from equations **3**–**5** using N_{pairs} from equation **11** (Fig. 4A and C). In addition, the number of annealing sites determined from running Findpatterns with the same DBase is shown for each primer and compared with the expectations from equations **7** and **8** (Fig. 4B and D). For both primer pairs S and D, the similar mismatch response curves were observed. First, there is a specific phase during which one or a few annealing sites or simPCR products are detected. These products originate from sequences homologous to the target gene, as can be seen in the entries annotations. For example, the simPCR products detected with m_{max} = 4 for primer pair S are shown in Table 2. Only targeted genes (*endothelin*) gave products with $m_{max} \leq 3$; no unspecific products were detected, as expected from the model. Primers pair D gave no unspecific product with $m_{max} \leq 2$ (data available through WWW; see Materials and Methods). With increasing mismatch tolerance, there is a transition to a phase during which the number of simPCR products or of annealing sites becomes similar to model predictions. In the case of pair S, the *cytochrome p450* gene is detected with four mismatches (Table 2). This gene which is apparently not related to *endothelin*, should be considered the first non-targeted product. This phase is therefore referred to as the unspecific phase. Note that for both pairs, transition to the unspecific phase starts earlier for the number of annealing sites (Fig. 4B and D) than for the respective simPCR products (Fig. 4A and C).

In summary, non-targeted simPCR products, recognized through their annotations, start to be detected only when there is a reasonable chance to find them according to the model. In this respect, there is a good agreement between the model and simPCR even when non-random primers pairs and non-random templates are used. This result was further supported by the analysis of 20 additional primers which showed the same mismatch response curve (data available through WWW; see Materials and Methods).

DISCUSSION

We have developed a mathematical model and a simulation to describe the formation of non-targeted PCR products which occur as a result of random primer–template interactions. This approach, which allows a number of predictions, has not been used previously to study PCR specificity. Before describing those predictions, the limitations of the model and suggestions for its improvement are discussed.

Table 2. Data available in the simPCR output—example of a run with primer pair S[a]

GenBank entry[b]				simPCR product[c]		
Accession	Length	Species	Putative function	X(m_x)	Y(m_y)	Length
X59931	1639	*O.cuniculus*	**endothelin 1**	692(4)–	296(1)+	396
Y00749	1167	*H.sapiens*	**endothelin 1**	858(0)–	435(0)+	423
J05008	12461	*H.sapiens*	**endothelin 1**	9284(0)–	5703(0)+	3581
S56805	1251	*H.sapiens*	**preproendothelin 1**	942(0)–	519(0)+	423
X16699	2130	*H.sapiens*	cytochrome P-450	547(3)+	918(4)–	371
J02871	2084	*H.sapiens*	lung cytochrome P-450	559(3)+	930(4)–	371

[a]Primer pair S: X = GTTGTGGGTCACATAACGC and Y = AGAGTGTGTCTACTTCTGCC (Table 1).
[b]Entries with a putative function related to the target are shown in bold.
[c]SimPCR products are shown for a mismatch tolerance of $m_{max} = 4$. For simPCR products, simPCR describes the position of primers X and/or Y annealing sites within the GenBank entry (X pos and Y pos, with + or – indicating plus or minus strand), the number of mismatches between the entry and the primer sequence (m_x and m_y), and the length of the simPCR product.

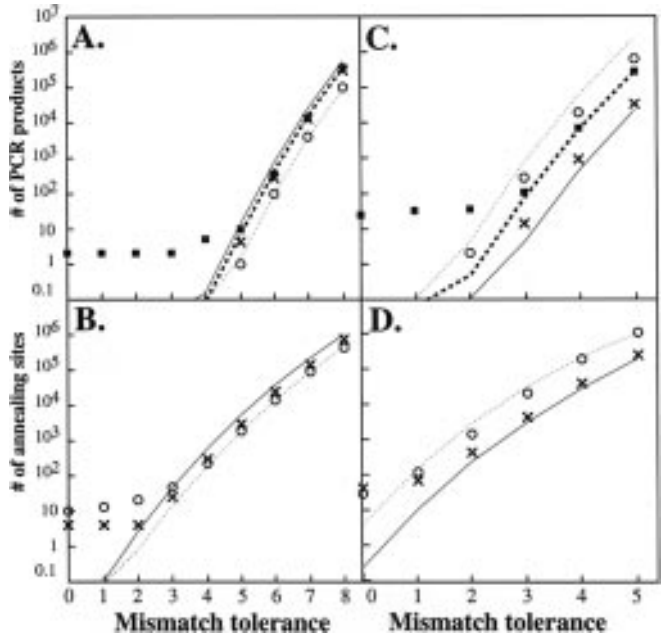

Figure 4. Number of simPCR products and annealing sites obtained with real primers, using natural sequences as templates, versus expected values from the model. The number of simPCR products (**A** and **C**) and annealing sites (**B** and **D**) is shown, using GenBank as a template. In (A) and (B), non-degenerate primers (pair S in Table 1) were used. In (C) and (D) data are shown for degenerate primers (pair D in Table 1). The number of PCR products was determined for different mismatch tolerance levels (m_{max}). In (A) and (C) the number of PCR products predicted from the model is presented as lines. It was calculated using equations **3**–**5**, with $L_s = 500$, $l_x = 19$ and $l_y = 20$. In (B) and (D), the number of annealing sites predicted from the model is presented as lines. It was calculated using equations **7** and **8**, with $L_t = 2.2 \times 10^8$. The number of observed solo-simPCR products or of annealing sites is indicated for primer X (xs), and for primer Y (circles). The number of XY simPCR products is also indicated (full rectangles). Expected values are shown for solo products or annealing sites with primer X (full lines) and primer Y (fine dashed lines). Expected number of XY-products are also shown (thick dashed lines).

Limitations of the model and possible improvements

To describe the process of amplification by PCR, a number of simplified conditions were used (see results), which do not fully describe real PCR conditions. (i) The DNA template is not a random sequence composed of equal ratios of all four nucleotides. This limitation, however, applies mainly to the model and is less critical when real sequence databases are used in the simulation. Further improvement of the model should consider biases in nucleotide composition of real genomic sequences, including mononucleotides and K-tuples composition (25). (ii) Not all mismatches equally effect annealing to the template and priming of DNA synthesis. For example, mismatches at the 5′-end of the annealing site are less restrictive to priming than mismatches at the 3′-end. (8,11,12). It is, however, difficult to formulate a general and quantitative model based on these works, because they supply qualitative data (11), are based on a small number of examples (12) or arerestricted to the effect of single mismatches on relatively short primers (8). Nevertheless, an attempt was made to incorporate into a primer design package, a simple weight function for mismatches based on a 3′→5′ gradient (14). This function however does not accurately represent real priming as suggested by the complex patterns of mismatch position described by Dyachenko *et al*. (8). In addition, there is also evidence that non-Watson–Crick interactions can make certain mismatches less restrictive to priming than others (9). An improved model and simulation, more accurately representing the effect of mismatches, can only be developed when experimental data will provide a more quantitative and general description of the priming process. (iii) Distance limitation is not the only factor which determines the production of a PCR product. Some DNA regions are less efficiently amplified in PCR (26). Although it is generally believed that DNA secondary structures may explain these results, it is not yet possible to predict the efficiency of amplification from a given template. Despite the limited description of PCR by the model and the simulation presented in this work, various predictions can be made as discussed below.

Non-randomness of the DNA cannot explain the PCR paradox

According to the PCR model presented here, unspecific amplification of PCR products should virtually never occur in reactions with no mismatches and with typical primers (Fig. 2A). This result is surprising as under such conditions, real PCR often gives unspecific products, even when reactions are optimized for stringency. The great discrepancy between real PCR behavior and

the model predictions, the so called PCR paradox, can be explained in two ways which were tested in this work. First, real PCR primers and templates might share non-random features which cause the occurrence of more annealing sites than expected (27). Second, PCR can tolerate several mismatches, even under presumably stringent conditions. simPCR output, using template sequences from natural genes and real primers, allowed us to assess the effect of the non-randomness of genomic sequences. A good fit was found between simulation results and model predictions for the two primers pairs presented here (Fig. 4 and Table 2), although the fit is better for the non-degenerated pair. This may result from the use of effective length approximation, which do not fully represent the effect of degenerated bases on the probability of chance annealing. Analysis of 20 additional real primers also showed good agreement between model prediction and simulations results (data available through WWW; see Materials and Methods). As the model predictions are based on the assumption that DNA is a random sequence, the good fit between the model and the simulation rules out the possibility that non-randomness of the genome accounts for the PCR paradox. Interestingly, the fact that real DNA behaves almost as a random template suggests that the model, despite its over-simplified assumptions, is adequate for the prediction of non-targeted primer–template interactions. The non-randomness of the genome probably has some effect on the amplification of non-targeted products (25). This effect, however, must be minor compared with the deviations mentioned as the 'PCR paradox'. In summary, the most likely explanation for the PCR paradox is high tolerance of the reaction to mismatches. These conclusions are supported by experimental data indicating that mismatches occur frequently in PCR (see below).

Relative weight of factors affecting PCR specificity—importance of mismatches

According to the model, the effect of template length on specificity is linear (Fig. 2D). This is in agreement with data from real PCR, as the problem of non-targeted product amplification is less frequent with short templates than with larger ones. Maximal product length is also expected to have a linear effect on specificity (Fig. 2C). Currently, there is no good experimental data on the relationship between reaction conditions and product length that enable to confirm model predictions. The length of the primers affects exponentially the number of PCR products expected from the model (Fig. 2B). From these predictions, it could be assumed that any increase in primer length improves specificity. In real PCR, short primers, such as 10 or 11mers, used in RAPDs, are known to give several products (28,29); using longer primers indeed increases specificity. However, real PCR data suggest that specificity cannot be increased indefinitely by using longer primers: a 30 base primer was shown to amplify its target with eight mismatches at annealing temperature of 10°C above calculated T_m (9). This unexpected low specificity requires further experimental research, but might be explained if increased primer length is accompanied by increased mismatch tolerance even under stringent conditions.

Of all the factors considered, the number of mismatches tolerated in the reaction had the strongest effect on amplification of non-targeted products (Fig. 2A). In real PCR, factors that reduce mismatch tolerance, like increasing annealing temperature, were found to improve specificity (6), in agreement with the model. This raises the question of the extent of mismatch tolerance in real PCR. Experimental works have shown that mismatches were tolerated under supposedly stringent conditions with 30 base primers, as described above (9). Similarly, using a 20 base primer, a PCR product was amplified with only 13 bp shared between the primer and the template (7). Under less stringent conditions, at 37°C, a 17 base primer was found to amplify a product with nine mismatches distributed throughout the primer (11). These experimental data suggest that in many reactions mismatches cannot be prevented, further supporting the above proposal that mismatch tolerance can resolve the PCR paradox. Reducing mismatch tolerance might therefore be the most significant means to improve PCR specificity. This might become possible by stabilizing perfect matches with chemical components added to the reaction, or with heat-stable enzymes (30).

Utilization of simPCR for primer design

An important aspect of primer design is to identify unwanted annealing sites that might give rise to a non-targeted PCR product, prior to primer synthesis. Several programs can handle this task (13–15,18,31–34), but not with complex templates. Therefore, DBases screening for 'suspicious' homologies to individual primers is sometimes performed using Findpatterns, or more sophisticated programs which monitor single annealing sites under various T_m conditions (17). Dbase screening for single annealing sites becomes unpractical with mismatch levels tolerated in PCR as hundreds or thousands of entries are detected (see example in Fig. 4). Compared with Findpatterns, the simPCR output has the advantage of reporting only putative PCR products, and thus being more compact and informative. The utility of simPCR will be enhanced when the sequence of complete genomes becomes available and a better understanding of the reaction is gained.

Probing uniqueness and randomness of DNA sequences using mismatch response curves

PCR can be considered a private case of reactions involving recognition of specific sites along the DNA. A wide range of biological reactions that involve such recognition sites can be studied using the approach presented in this work. Initiation of transcription, processing of introns, and several other processes require at least two different motifs positioned within a certain distance and orientation in a specific manner. Like PCR, the recognition of each motif tolerates mismatches, and the distance between the motifs may vary. Each motif is analogous to an annealing site, whose chance occurrence can be described by equation **8**. A composite structure is analogous to the formation of non-targeted PCR products, and can be mathematically described, with minor modifications, by equation **3**. Thus, equations **8** and **3** allow the uniqueness of a recognition site or of a composite structure in the genome to be predicted. Consider a transcription unit composed of a 19 base promoter and a 20 base enhancer, both occurring on the same strand and within 500 bp distance. This structure has the same mismatch response curve as shown for XY-products of pair S in a 2.2×10^8 bp genome (Fig. 4A, bold dashed line). From this theoretical curve, it can be predicted that such a structure will be unique only if each motif tolerates no more than four mismatches. Furthermore, when comparing the theoretical curve with simulation results (Fig. 4A), deviations from the model predictions indicate the extent of

non-randomness of the DNA studied. In the future, when DBases contain complete genomic sequences, these comparisons will enable to probe DNA non-randomness more accurately. In summary, the combined use of response curves for mismatches, distance limitations and motif length obtained from mathematical modeling and DBase scanning, is a new approach to probe uniqueness and randomness of DNA sequences in complex genomes.

ACKNOWLEDGEMENTS

We are thankful to D. Lancet, S. Pietrokovski, Y. Elkind, L. Segal, M. Edelman and O. Yarden for fruitful discussions and critical reading; to G. Benet for providing unpublished results; to A. Rubin and O. Asor for help in developing the model; to the bioinformatics unit for technical support in programming and database handling; and to Y. Avivi and V. Levy for carefully editing the manuscript. This work was supported by a doctoral fellowship from the Feinberg graduate school to E.R. and an Yigal Alon Fellowship to A.A.L.

REFERENCES

1 Saiki, R.K., Scharf, S., Faloon, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) *Science*, **230,** 1350–1354.
2 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.
3 Carr, K. (1993) *Nature*, **365,** 685.
4 Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) *Cold Spring Harbor Symp. Quant. Biol*., **51,** 263–273.
5 Innis, M.A. and Davis, H.G. (1990) In Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (eds), *PCR protocols.* Academic Press Inc., San Diego, pp 3–12.
6 Rychlik, W., Spencer, W.J. and Rhoads, R.E. (1990) *Nucleic Acids Res.*, **18,** 6409–6412.
7 Rychlik, W. (1995) *Biotechniques*, **18,** 84–90.
8 Dyachenko, L.B., Chenchik, A.A., Khaspekov, G.L., Tatarenko, A.O. and Bibilashvili, R.S. (1994) *Mol. Biol. Eng. Tr.*, **28,** 654–660.
9 Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J.J. (1990) *Nucleic Acids Res.*, **18,** 999–1005.
10 Wu, D.Y., Ugozzoli, L., Pal, B.K., Qian, J. and Wallace, R.B. (1991) *DNA Cell Biol*, **10,** 233–238.
11 Sommer, R. and Tautz, D. (1989) *Nucleic Acids Res.*, **17,** 6749.
12 Sakuma, Y. and Nishigaki, K. (1994) *J. Biochem. (Tokyo)*, **116,** 736–741.
13 Lowe, T., Shrefkin, J., Yang, S.Q. and Dieffenbach, C.W. (1990) *Nucleic Acids Res.*, **18,** 1757–1761.
14 Engels, W.R. (1993) *Trends Biochem. Sci.*, **18**, 448–450.
15 Hillier, L. and Green, P. (1991) *PCR Methods Appl*., **1,** 124–128.
16 Lincoln, S.E., Daly, M.J. and Lander, E.S., (1991) *PRIMER: A Computer Program for Automatically Selecting PCR Primers (0.5)*—program and manuals, MIT Center for Genome Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
17 Mitsuhashi, M., Cooper, A., Ogura, M., Shinagawa, T., Yano, K. and Hosokawa, T. (1994) *Nature*, **367,** 759–761.
18 Montpetit, M.L., Cassol, S., Salas, T. and O'Shaughnessy, M.V. (1992) *J. Virol. Methods*, **36,** 119–128.
19 Cooper, D.L. and Baptist, E.W. (1991) *PCR Methods Appl*., **1,** 57–62.
20 Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J., Banrevi, A., Bolle, P.A., Bolotinfukuhara, M., Bossier, P. *et al*. (1994) *Nature*, **369,** 371–378.
21 Feldmann, H., Aigle, M., Aljinovic, G., Andre, B., Baclet, M.C., Barthe, C., Baur, A., Becam, A.M., Biteau, N., Boles, E. *et al.* (1994) *Embo J.*, **13,** 5795–5809.
22 Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Kucaba, T., Hillier, L., Jier, M., Johnston, L. *et al.* (1994) *Science*, **265,** 2077–2082.
23 Oliver, S.G., Vanderaart, Q., Agostonicarbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J., Benit, P. *et al*. (1992) *Nature*, **357,** 38–46.
24 Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., C, o.M., Copsey, T., Cooper, J. *et al.* (1994) *Nature*, **368,** 32–38.
25 Griffais, R., Andre, P.M. and Thibon, M. (1991) *Nucleic Acids Res.*, **19,** 3887–3891.
26 He, Q., Marjamaki, M., Soini, H., Mertsola, J. and Viljanen, M.K. (1994) *Biotechniques*, **17,** 82.
27 Karlin, S. and Cardon, L.R. (1994) *Annu. Rev. Micro.*, **48,** 619–654.
28 Welsh, J., Chada, K., Dalal, S.S., Cheng, R., Ralph, D. and McClelland, M. (1992) *Nucleic Acids Res.*, **20,** 4965–4970.
29 Birkenmeier, E.H., Schneider, U. and Thurston, S.J. (1992) *Mamm. Genome*, **3,** 537–545.
30 Angov, E. and Cameriniotero, R.D. (1994) *J. Bacteriol*., **176,** 1405–1412.
31 Lucas, K., Busch, M., Mossinger, S. and Thompson, J.A. (1991) *Comput. Appl. Biosci*., **7,** 525–529.
32 Dopazo, J., Rodriguez, A., Saiz, J.C. and Sobrino, F. (1993) *Comput. Appl. Biosci*., **9,** 123–125.
33 Eberhardt, N.L. (1992) *Biotechniques*, **13,** 914–917.
34 Grob, U. and Gartmann, C.J. (1991) *Comput. Appl. Biosci*., **7,** 379–381.
35 Rychlik, W. and Rhoads, R.E. (1989) *Nucleic Acids Res.*, **17,** 8543–8551.