A Computer Support for Genotyping by Multiplex PCR

Pierre Nicodème
INRIA-Rocquencourt

January 16, 1995

[summary by Pierre Nicodème]

Abstract

The Polymerase Chain Reaction, PCR for short, is able to produce million copies of a specified DNA segment. Grouping (multiplexing) numerous PCR in a few experiments would decrease the PCR costs and save time. Starting from a biological model for the multiplexing conditions, we transform the problem to a combinatorial one, show that the problem is $NP-$complete, give an approximation algorithm, and show its quasi-optimality.

1. Introduction

Devised in the mid-1980s, the Polymerase Chain Reaction, PCR for short[1], is able to produce enormous numbers of copies of a specified DNA sequence. The method is sensitive to very small amounts of DNA, and has numerous applications (diagnostics, etc); however, in most of the PCR experiments performed by biologists, the amplification of each target fragment of DNA requires a separate and costly PCR experiment, with the corresponding manipulations, and the immobilization of an automat [4].

PCR exploits certain features of DNA replication. Single-stranded DNA is used as a template for the synthesis of a complementary new strand. These single stranded DNA templates can be produced by simply heating double-stranded DNA to temperature near boiling. Then we require a small section of double stranded DNA to initiate ("prime") synthesis.

The starting point for DNA synthesis can be specified by supplying an oligonucleotide primer that anneals to the template at this point. Both DNA strands can serve as templates for synthesis provided an oligonucleotide primer is supplied for each strand. Each cycle of PCR duplicates the segments under amplification; so, starting from one segment, n cycles of PCR produce 2^n segments. Figure 1 shows the synthesis initiated by the forward primer 5'-ACACA...AGCAA-3' on the 3'-5' strand of a segment of DNA[2].

Primers cannot be chosen at will inside a locus (a portion) of a gene: they must respect conditions permitting a correct amplification by PCR; the temperature of hybridization at which the polymerase synthesises the new DNA strands is one of these conditions; this temperature depends on the composition of the primer, and more specifically on the respective percentage of the bases A and T, versus the bases G and C; a more accurate method relates the hybridizing temperature

[1] We refer to [5] for a detailed introduction to the subject of PCR.

[2] The 3' extremity of a chain is N-terminal; the 5' extremity is C-Terminal; the numbers 3 and 5 refer to the position of the carbon connected to the N-termination and to the C-termination inside the 5-carbon sugar constitutive of the bases of DNA (other components of a base of DNA are a phosphate group and one out of four organic bases).

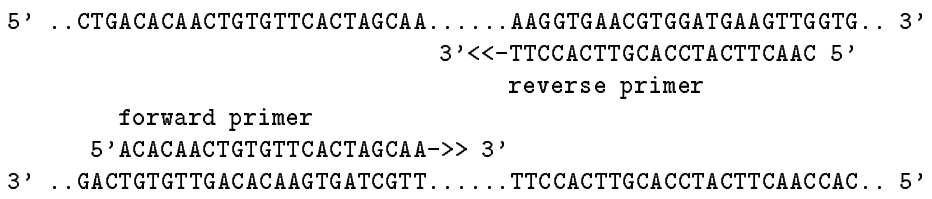


FIGURE 1. Primers for DNA polymerase

to experimental measurements of base-stacking energy. Anyhow, when choosing a pair of primers, the hybridizing temperatures of the two primers should be about the same. Another condition relates to homology between the two primers and to self-homology; such homology would very often prevent a correct amplification, the primers hybridizing to each other, or identical copies of a self-homologous primer hybridizing together.

Several software programs are available to predict which pair of primers to choose inside a given locus. The conditions which hold for a one-locus PCR amplification still have to hold for multi-loci amplification.

Starting from a set S of n loci, we want to find the subset C_{max} of maximum size of S, such that in each locus of C_{max} we can select a pair of compatible primers, and such that the $2n$ selected primers are each other compatible.

We made an extension the program PRIMER, of S. E. Lincoln, M. J. Daly, and E. S. Lander [1] in a MULTIPCR program. PRIMER is a two-step program; step-1 selects forward and reverse candidates primers; step-2 chooses a best pair of one forward and one reverse primer among all the possible pairs of candidates. MULTIPCR takes as input the output of PRIMER step-1, and chooses for each locus a forward and a reverse primer compatible with the primers chosen for the other loci, whenever this is possible.

2. Multiplexing the Polymerase Chain Reaction

2.1. Requirements. We detail in this section a model of compatibility between primers that Gilles Thomas[3] proposed to us and the corresponding requirements.

We will speak of *locus* amplification when considering the amplification of a single segment; only one amplification is allowed inside a given locus; to each locus amplification correspond a *forward* and a *reverse* primer. We define a *subprimer* as a subsequence of a primer and we consider in the following that all subprimers of a multiplexing experiment have the same length σ. In practical experimentations, σ will have values 4 or 5. We define a *3'-subprimer* as the subprimer ending a primer at its 3' extremity (primers being always read in the direction 5' $\Rightarrow$ 3').

The requirements are the following:

(1) *Locus* amplification requirements:
   (a) The distance between the forward primer and the reverse primer is between 150 and 450 bases (these minimum and maximum values are given as parameters and correspond to the "product range size" taken as input by the program PRIMER).
   (b) The primers satisfy the conditions of non-palindromicity; such a palindromicity would cause self-homology.
   (c) The 3'-subprimers are not reverse complementary with any of the subprimers (subprimers as 3'-subprimers are assumed to be of length σ bases).

(2) *Multi-locus* amplification or *experiment* requirements:

[3]Laboratoire de Génétique des Tumeurs, Institut Curie, 26, rue d'Ulm, 75005 Paris.

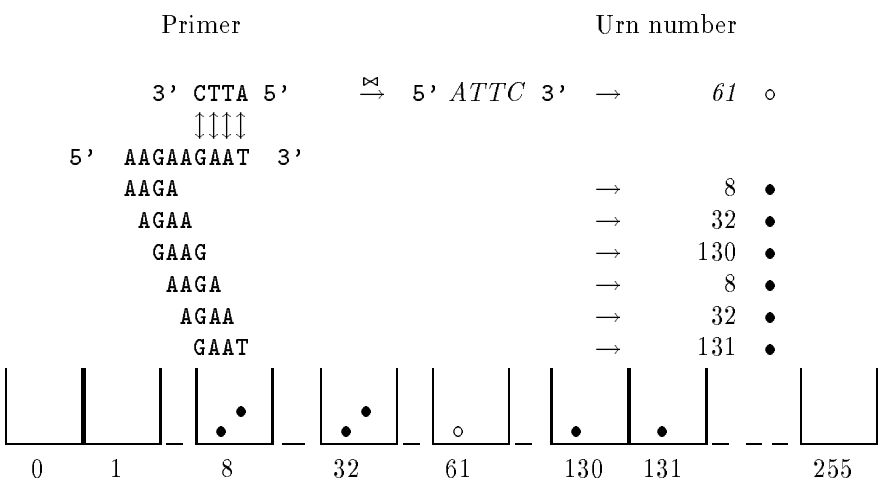


FIGURE 2.

(a) Any 3'-subprimer of an experiment is not reverse complementary with any subprimer, including itself, of any primer of the experiment; this would initiate hybridization of the primers themselves. An example for this condition is given in subsection 2.2. Note that subprimers (including the 3'-subprimers) may be identical between different loci, or inside a locus.

(b) The temperatures of denaturation, or the GC/AT percentage in the primers of a multi-locus PCR amplification have to belong to a limited range of values (by instance $48\% - 52\%$).

(c) Electrophoresis[4] distance: the difference of lengths between any two segments amplified in the same multi-locus PCR amplification is greater than δ bases; this is necessary to allow a correct differentiation of the amplified segments after electrophoresis. This distance supposes that the loci are not polymorphic, in which case the problem of differentiating the amplified segments has to be handled in a different way.

2.2. An urn model to solve the problem of compatibility between primers. We give here a constructive example of our algorithm.

- Using the mapping (A $\Rightarrow$ 0, C $\Rightarrow$ 1, G $\Rightarrow$ 2, T $\Rightarrow$ 3), We transform each subprimer of length $\sigma = 4$ in a number in base 4 between 0 and $4^4 = 256$, and each subprimer of length $\sigma = 5$ in a number between 0 and $4^5 = 1024$; the resulting numbers are converted in base 10 (TTA $\Rightarrow 330_4 \Rightarrow 60_{10}$).
- We then consider a model of 256 urns, when $\sigma = 4$, or a model of 1024 urns, when $\sigma = 5$. For each subprimer, we compute the associated number as described above, and we throw a ball in the corresponding urn.

The compatibility constraint (requirement 2(a) of §2.1) is then transformed as shown in Fig. 2 (when $\sigma = 4$). The complementary of the 3'-subprimer is taken (CTTA in Fig. 2) and reversed

[4]Electrophoresis is a migration method which allows short segments to move faster than the long ones; this method allows the differentiation of segments of different lengths, from a mixture of them, but it has a limited precision corresponding to our parameter δ.

(ATTC in Fig. 2), with $\bowtie$ being the palindromic operation on a chain). Ordinary subprimers generate black balls, while reversed complementary 3'-subprimers generate white balls.

The compatibility rule implies that an urn can never contain simultaneously black and white balls.

2.3. An algorithm deriving from the urn model. We propose in this section an approximate algorithm with high efficiency in practical computations; this algorithm is likely to be almost optimal.

Our algorithm is as follows: we sort our set of loci in increasing order along the number of candidates pairs of primers; we process our set of ordered loci, locus after locus; for each locus, we try each possible pair of primers with respect to the conditions, including the distance condition (requirement 1(a)).

For each pair, we "throw white and black balls in urns", along the model described above; we eliminate the pairs which cause "black and white" collisions; among the acceptable pairs of primers, we select the pair of primers which minimizes, in the following order:

(1) the number of urns containing white balls;
(2) the number of urns containing black balls, whenever the number of "white urns" is identical for two pairs.

The "white and black balls" corresponding to pairs of primers already selected remain in the urns when processing a new locus.

The loci providing no compatible pair with the pairs of the loci already chosen for the current experiment are left apart and processed in a next experiment.

Experimental result shows that, when processing 248 loci of Genbank, it would be theoretically possible to amplify simultaneously 245 loci, with $\sigma = 5$; the average size of the loci is 4000 bp., with an average number of 20,000 admissible pairs of primers. However, it is biologically unrealistic to think to amplify simultaneously much more than ten loci.

3. Determining the pairs of primers which maximize the number of loci in a single experiment is a NP-complete problem

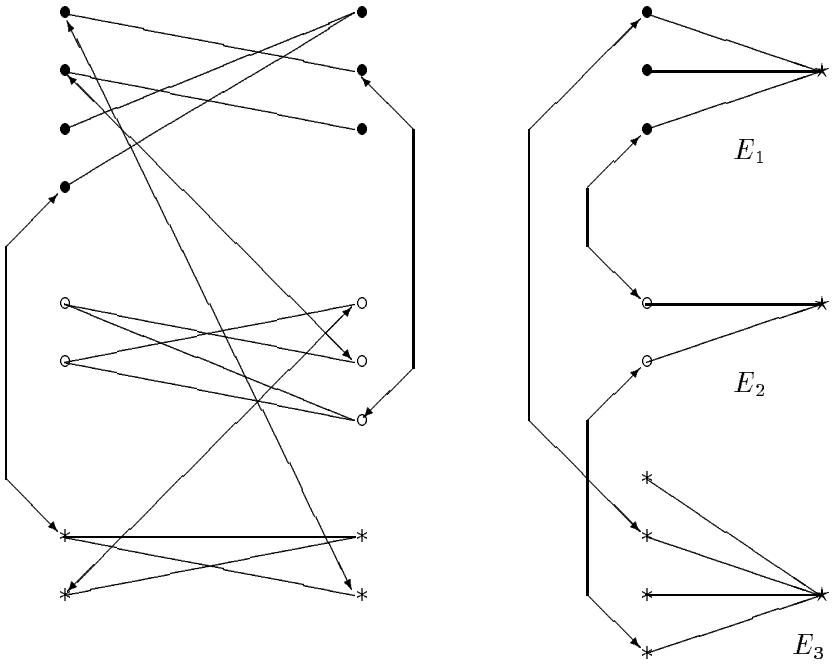
We model our problem as a set of bipartite subgraphs with additional edges (Figure 3 (a)); in this graph, each primer is represented by a vertex; the set of vertices is partitioned by locus, each locus corresponding to a bipartite subgraph; in our example, vertices belonging to the same locus are represented by the same character ($\bullet$ for locus 1, $\circ$ for locus 2, $*$ for locus 3), the forward primers being represented on the left part of the figure (Figure 3 (a)), while the reverse primers are represented on the right part. There are two kind of edges:

- acceptance edges, inside the bipartite subgraph restricted to a single locus; such a non-arrowed edge indicates that the forward and the reverse primers joined by the edge are compatible;
- incompatibility edges, joining a vertex of a locus to a vertex of a different locus; these edges with arrowed extremities indicate that the primers they join are not compatible.

Our "*Compatible Primers Problem*", in short CPP, has the following description:

Instance of the problem: a graph composed of a set of bipartite graphs $B_1, B_2, \ldots, B_J$; the edges of these graphs constitute a set of acceptance edges A; a set of incompatibility edges, these edges joining pairs of vertices which do not belong to the same bipartite subgraphs; an integer K.

Question: is it possible to choose a subset of acceptance edges $A' \subseteq A$ with $|A'| \geq K$ such that A' contains at most one edge from each B_i, $1 \leq i \leq J$, and such that no two vertices belonging to these edges are extremities of an incompatibility edge.



(a) A representation of the compatible primer problem

(b) A transform of the multiple choice matching problem

FIGURE 3. Transformation of any graph to a set of bipartite subgraphs modelling the compatible primers problem

A typical graph of CPP is shown in Figure 3(a); any edge of a *"Multiple Choice Matching Problem"* (in short MCMP) graph is transformed in a vertex of a CPP graph; dummy nodes (figured by $\star$) are added, one for each subset of vertices of MCMP. Figure 3(b) represents such a transform of a MCMP graph (not represented in the figure) to a CPP one. This transformation is detailed in [2, 3]. Therefore, solving CPP in polynomial time would also solve MCMP in polynomial time, which contradicts the NP-completeness of MCMP. We hence proved $NP-$completeness of CPP.

4. Evaluating the limit probability of rejection of a locus

The experimental results obtained with 248 loci show that about 50 loci are enough to fill almost completely the system of urns. We want to evaluate the probability of rejection of a locus in such a saturated system of urns.

if s is the number of subprimers of a primer (practically, if $\sigma = 4$, $s = 17$ for primers of length 20), with $\pi_{1,b,n}$ the probability of acceptation of a primer by a system of U urns containing either white, or black balls, we have

$$\pi_{1,b,n} = \frac{b}{U}\left(1 - \frac{b}{U}\right)^s \quad \text{and} \quad \pi_{1,30,226} = 0.014. \tag{1}$$

The probability π_{11} of compatibility of two primers between themselves, when considering an empty system of urns, is

$$\pi_{11} = \frac{1}{U}\left(1 - \frac{1}{U}\right)^{2s} + \left(1 - \frac{1}{U}\right)\left(1 - \frac{2}{U}\right)^{2s} = 0.766. \tag{2}$$

The MULTIPCR algorithm considers a number V of forward primers for a locus, and, for each forward primer, a number R of reverse primers at an acceptable distance of this primer (between 150 and 450 bp); depending of the locus length, V is between 100 and 500, while R remains close to 50.

The small value of $\pi_{1,30,226}$ allows us to apply the Poisson approximation to the binomial distribution of the number of accepted forward and reverse primers, with respective parameters $\nu = V\pi_{1,b,n}$ and $\rho = R\pi_{11}\pi_{1,b,n}$.

The probability Π of rejection of a locus is then

$$\Pi(\nu(V), \rho(R)) = \sum_{i=0}^{\infty} (\Pr\{v = i\} \times (\Pr\{r = 0\})^i) = e^{-\nu + \nu e^{-\rho}}, \tag{3}$$

probability whose some values for $R = 50$ are

V	250	300	350	400	450
$\Pi(\nu(V), \rho(50))$	0.230	0.171	0.128	0.095	0.071

Considering our experimental results on 248 loci, this shows that our algorithm is quasi-optimal.

Bibliography

[1] Lincoln (S. E.), Daly (M. J.), and Lander (E. S.). – *PRIMER: A Computer Program for Automatically Selecting PCR Primers.* – MIT Center for Genome Research and Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, 1991.

[2] Nicodème (P.). – *A computer support for genotyping by multiplex PCR.* – Technical Report n° LIX/RR/93/09, LIX, École polytechnique, France, 1993.

[3] Nicodème (P.). – Un support informatique pour le multiplexage de la PCR. *Technique et Science Informatique*, 1995. – Numéro special bioinformatique, à paraître.

[4] Olschwang (S.), Delaitre (O.), Melot (T.), Peter (M.), Schmitt (A.), Frelat (G.), and Thomas (G.). – Description and use of a simple laboratory-made automat for *in vitro* DNA amplification. *Methods in Molecular and Cellular Biology*, vol. 1, n° 3, May/June 1989, pp. 121–127.

[5] Watson (J. D.), Witkowski (J.), Gilman (M.), and Zoller (M.). – *Recombinant DNA.* – Scientific American Books, 1992, 2nd edition, 79–98p.