Selecting Optimal Oligonucleotide Primers for Multiplex PCR

Pierre Nicodème[*][†]
INRIA and LIX École Polytechnique (Email: Pierre.Nicodeme@inria.fr)
and
Jean-Marc Steyaert
LIX École Polytechnique, 91128 Palaiseau, France (Email: Steyaert@lix.polytechnique.fr)

Abstract

We investigate the problem of designing efficient multiplex PCR for medical applications. We show that the problem is NP-complete by transformation to the Multiple Choice Matching problem and give an efficient approximation algorithm. We developed this algorithm in a computer program that predicts which genomic regions may be simultaneously amplified by PCR. Practical use of the software shows that the method can treat 250 non-polymorphic *loci* with less than 5 simultaneous experiments.

Keywords. Multiplex PCR, diagnostic, heuristic algorithms, NP-completeness.

Introduction

Introduced in the mid-1980s, the Polymerase Chain Reaction, PCR for short, is able to amplify segments of DNA a million times. The method is suitable for very small amounts of DNA, and is considerably faster than other methods. However, amplification of the target fragments of DNA requires separate and costly experiments. PCR has numerous applications as for instance genotyping in order to characterize pathological genes which suffer from important deletions. Genotyping requires PCR amplifications of many different *loci*; whenever it is possible to group these PCR amplifications in a same experiment, called a multiplex PCR, time and money can be saved. We study in this paper the conditions required for multiplex PCR and propose an algorithm that performs an almost optimal choice of PCR primers, with the aim of minimizing the number of PCR multiplex operations. Besides genotyping applications, multiplex PCR has been used to detect multigene families (Harrison, Pearson, & Lynch 1991). More recently, multiplex PCR has also been used to speed up the ordering of contigs in DNA physical mapping (Sorokin *et al.* 1996). The algorithmic problems corresponding to these two applications are described in (Grebinski & Kucherov 1996; Pearson *et al.* 1995).

This paper is organised as follows. In section 2, we present the constraints of multiplex PCR, and describe an efficient heuristic algorithm. In section 3, we give experimental results. In section 4, we show that the problem is NP-complete and give a probabilistic analysis of the model.

Multiplexing the Polymerase Chain Reaction

We refer to J. D. Watson and al., *Recombinant DNA* (Watson *et al.* 1992), for a survey of the PCR subject. Let us just mention that, generally speaking, primers cannot be chosen at will inside a *locus* : they must respect constraints permitting a correct amplification by PCR, fulfilling hybridization temperature conditions and auto or hetero-hybridization prevention. We put the emphasis on problems encountered when selecting n pairs of primers for n loci, which implies the verification of $\binom{2n}{2}$ compatibility conditions between the combinations of two among the $2n$ primers.

Conditions for Multiplex PCR

We detail in this section a model of compatibility between primers and its requirements for classical PCR; we think that it applies also to Long Accurate PCR (LA-PCR) (Barnes 1994a; 1994b) which permits amplification of very long segments of several thousand bases.

We speak of *locus* amplification when considering the amplification of a single segment. Only one amplification is allowed inside a given *locus*, and to each *locus* amplification correspond a *forward* and a *reverse* primer. We define a *subprimer* as a subsequence of length σ of a primer: in practice, σ will take value 4

[*] Pierre Nicodème, INRIA-Rocquencourt, BP105, 78153 - Le Chesnay Cedex - France. Tel: 33-(0)139635443 - Fax: 33-(0)139635659

```
5' ..CTGACACAACTGTGTTCACTAGCAA......AAGGTGAACGTGGATGAAGTTGGTG.. 3'
                               3'<<-TTCCACTTGCACCTACTTCAAC 5'
                                    ****    reverse primer
        forward primer    ****
      5'ACACAACTGTGTTCACTAGCAA->> 3'
3' ..GACTGTGTTGACACAAGTGATCGTT......TTCCACTTGCACCTACTTCAACCAC.. 5'
```

Figure 1: Primers for DNA polymerase: 3'-subprimers are marked with *'s.

or 5. A *3'-subprimer* is the subprimer ending a primer at its 3' extremity (primers are always read in the direction 5' ⇒ 3').

Figure 1 shows the synthesis initiated by the forward primer 5'-ACACA...AGCAA-3' on the 3'-5' strand of a segment of DNA.

The consistency requirements on a set of primers are the following:

1. **Locus** amplification **requirements**:
(a) *Pairing-distance.* The distance between a forward primer and a reverse primer (pairing-distance) must be in the range of 150-450 bases (the minimum and maximum values given here are indicative).
(b) *Non-palindromicity.* The primers satisfy the conditions of non-palindromicity preventing self-homology.
(c) *Reverse-complementarity.* The 3'-subprimers must not be reverse complementary to any of the subprimers (subprimers as 3'-subprimers are assumed to be of length σ bases).
2. **Multi-locus** amplification or **experiment requirements**:
(a) *Reverse-complementarity.* Condition 1-c is extended to all the primers in the experiment.
(b) *GC-AT composition.* The temperatures of denaturation, or the GC/AT percentage in the primers of a multi-*locus* PCR amplification must belong to a limited range of values (for instance, $48\% - 52\%$).
(c) *Electrophoresis distance.* The difference of lengths between any two segments amplified in the same multi-*locus* PCR amplification must be 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; if not, the problem of differentiating the amplified segments has to be handled in a different way.

Note that subprimers (including the 3'-subprimers) may be identical between different *loci*, or inside a *locus.* In fact this situation is combinatorially unavoidable and experimentally sound.

The algorithm: MULTIPCR

We propose an approximate algorithm of high efficiency in practical computations; this algorithm is likely to be almost optimal.
We will use $\rho(s)$ to denote the reverse complementary of a string s: for instance, $\rho(\texttt{ATCG}) = \texttt{CGAT}$.

The algorithm is as follows:
1. Sort the set of *loci* in increasing order of the number of candidate pairs of primers. 2. Process the ordered set of*loci*, *locus* after *locus*:
2.1 For each *locus*, try each possible pair of primers with respect to conditions 1. given above, including the distance condition.
Let π_1 and π_2 be the two 3'-subprimers in a pair; compute $\rho(\pi_1)$ and $\rho(\pi_2)$. Check that (i) none of these two strings occurs as substrings of the set P of already chosen primers, (ii) π_1 and π_2 do not conflict, symmetrically, with the subprimers of set P and (iii) $\rho(\pi_1)$ does not occur in π_2 and $\rho(\pi_2)$ does not occur in π_1: if this is the case keep the pair π_1 and π_2 as a candidate for the locus.
2.2 Select as the representative (for the current locus) the pair minimizing the number of different selected 3'-subprimers and, in case of equality, the number of different internal subprimers of length σ in the set P. Add it to the set P of selected primers.

The *loci* providing no compatible pair with the pairs of the *loci* already selected for the current experiment are left apart and processed for another experiment.

Complexity. The complexity C of our algorithm is $C = O(Knl)$ with $K \approx E\tau_f k_r \approx 6E$, where E is the number of experiments, n is the number of *loci*, l is the average length of a *locus*, $\tau_f \approx 1/7$ is the probability of getting a forward primer at a position, and $k_r \approx 40$ is a constant giving the approximate number of acceptable reverse primers for a forward primer. (See the discussion for the numerical values of τ_f and k_r).

Software Implementation

The *MULTIPCR* program implements the algorithm described in the preceding section.

Different software programs are available predicting which pair of primers to choose inside a given *locus*.

We use the program *PRIMER*, of S.E. Lincoln, M.J. Daly, and E.S. Lander (Lincoln, Daly, & Lander 1991) as a preparation step in our program. *PRIMER* is a two-step program; *step-1* selects candidates for forward and reverse 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.

Both *PRIMER* and *MULTIPCR* are written in the C language and implemented on the *Unix System V* system. The package "*PRIMER+MULTIPCR*" is available by anonymous ftp at **`ftp://ftp.infobiogen.fr /pub/logiciels/unix/bio`** and portable on SUN, DEC and Silicon Graphics workstations.

Experimental results and NP-completeness

Table 1 shows the results obtained when processing 248 *loci* of Genbank; as a typical *locus*, the *locus* HUMDYSGP of the gene of dystrophy is 2202 base pairs long. The program *Primer* produces for this gene 343 forward and 339 reverse primers. Therefore, on the average, a primer starts each seventh position, and the pairing-distance condition implies that there are approximately 40 admissible reverse primers for each forward primer.

It is clear that simultaneous amplification of 214 *loci*, as proposed in Table 1, is biologically unrealistic, but a multiplex of about 10 to 20 PCR looks reasonable.

Let us now comment about the NP-completeness of the problem. G. Robins and *al.* (Robins *et al.* 1993) have proved the NP-completeness of the minimisation problem for the multigene primer selection (a somehow different problem) by reduction to the *Minimum Set Cover* problem.

We reduce our compatibility problem to the *Multiple Choice Matching* problem which proves NP-completeness (Nicodeme 1993). Apparently there is no connexion between the two reductions.

Analysis of the MULTIPCR algorithm

Evaluating the limit probability of rejection of a *locus*. This section presents a simplified analysis of the model, when a steady state is reached; it gives insight into the optimality of our algorithm. We denote by λ the number of amplified *loci*, $\varpi^\star$ the number of different 3'subprimers, ϖ the number of internal subprimers at a current state of the algorithm. The figures in Table 1 show these values when it is no longer possible to add primers to the current experiment.

We make some empirical observations on Table 1: we see that, with an electrophoresis distance $\delta = 5$,

σ	δ	ν	λ	$\varpi^\star$	ϖ
4	1	1	214	34	222
		2	32	27	223
		3	2	4	48
	3	1	84	31	224
		2	81	30	226
		3	73	34	222
		4	10	17	168
	5	1	48	31	223
		2	45	29	223
		3	49	32	218
		4	49	30	219
		5	47	31	219
		6	10	19	171
no possible amplification			0		
number of *loci* processed			248		

Table 1: Submitting 248 *loci* to *MULTIPCR*, compatibility check length is 4 bases (model of 256 subprimers).
The columns indicate respectively: σ checking length, δ electrophoresis distance, ν experiment number, λ number of amplified *loci*, $\varpi^\star$ number of different 3'subprimers, ϖ number of internal subprimers.

$\varpi + \varpi^\star$, the number of subprimers and 3'-subprimers involved in a sequence of experiments, is almost equal to the total number of possible subprimers (256), and that there are less than 50 *loci* by experiment.

If we can reduce the electrophoresis distance we are able to discriminate more *loci* in the first experiments. If s is the number of subprimers inside the primers (in practice, if $\sigma = 4$, we have $s = 17$ for primers of length 20); denoting by $\pi_{1,w,b}$ the probability that a primer is accepted by a system of S subprimers ($S = \varpi + \varpi^\star$), we have

$$\pi_{1,\varpi^\star,\varpi} = \frac{\varpi^\star}{S}\left(1 - \frac{\varpi^\star}{S}\right)^s \quad \text{and} \quad \pi_{1,30,226} \approx 0.014 \qquad (1)$$

The probability π_{11} of compatibility of two primers (thrown inside an empty system of urns), is

$$\pi_{11} = \frac{1}{S}\left(1 - \frac{1}{S}\right)^{2s} + \left(1 - \frac{1}{S}\right)\left(1 - \frac{2}{S}\right)^{2s} \approx 0.766 \qquad (2)$$

The *MULTIPCR* algorithm considers a number F of candidate forward primers for a *locus*, and, for each forward primer, a number R of candidate reverse primers at an acceptable pairing-distance (in the range 150-450 bp.) of this forward primer. Depending on the *loci* length, F is in the range 100-500, while R stays close to 50.

When considering a *locus*, the asymptotical distributions of the number of accepted forward primers f and accepted reverse primers r are both binomials; the small value of $\pi_{1,30,226} \approx 0.014$ allows us to apply the Poisson approximation to these binomial distributions, with respective parameters $\phi = F\pi_{1,w,b}$ and $\rho = R\pi_{11}\pi_{1,w,b}$ (with $w + b = U$). The probability Π of rejection of a *locus*, at the stationary state, is then

$$\Pi(\phi(F), \rho(R)) = e^{-\phi+\phi e^{-\rho}}, \qquad (3)$$

probability for which some values are given for $R = 50$ in the table below:

F	250	300	350	400	450
Π	0.230	0.171	0.128	0.095	0.071

Considering the results obtained for 248 *loci* (Table 1), with an average number of 300 to 350 forward primers per *locus*, we see that our algorithm is quasi-optimal.

As a final quantitative remark, we emphasize the fact that our approximation algorithm reduces drastically the number of configurations to be checked and therefore is able to treat in a few minutes a problem with $n = 200$ *loci*, $m = 10$ and $k = 20$, for which the number of possible confugurations is

$$p_{200,20} = \frac{200!}{10!(20!)^{10}} \approx 3 \times 10^{185}.$$

Conclusion

We showed that the theoretical problem of grouping a maximum number of *loci* that satisfy constraints allowing PCR multiplex is NP-complete. We explored a model of compatibility between primers that allows multiplexing and provided the biologists with an efficient and simple tool to choose the pairs of primers in the general case. This model considers that the crucial point is to avoid reverse complementarity of the 3' end of the primers to subsequences of any primer in the mixture used for the multiplex PCR experiment. Considering the traditional PCR method, where the length of the amplified segments is limited to some 500 bp., our *MULTIPCR* approach encounters serious limitations when applied to highly polymorphic *loci*. The method is also not applicable when only one pair of primer is available for each *locus*, which is the case when considering *primers dictionary.* However, the recent emergence of *Long Accurate PCR* (Barnes 1994b; 1994a), which allows amplification of long segments (comprising as much as 35-kb) offers excellent perspectives to the application of our method to genotyping and to DNA physical mapping.

Acknowledgements. We are grateful to Gilles Thomas of Institut Curie, Paris and to Philippe Dessen of Infobiogen, Villejuif, for frequent discussions and advice concerning the biological subjects addressed in this paper. We also thank Marie-France Sagot for reading this paper.

References

Barnes, W. M. 1994a. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci. USA* 91:5695–5699.

Barnes, W. M. 1994b. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proc. Natl. Acad. Sci. USA* 91:2216–2220.

Garey, M. R., and Johnson, D. S. 1979. *Computers and Intractability, A Guide to the Theory of NP-completeness.* W. H. Freeman and Company. Appendix.

Grebinski, V., and Kucherov, G. 1996. Reconstructing a Hamiltonian circuit by querying the graph: Application to DNA physical mapping. INRIA-Lorraine and CRIN/CNRS, Campus Scientifique, 615 rue du Jardin Botanique, BP 101, 54602 Villers-lès-Nancy, France.

Harrison, J. K.; Pearson, W. R.; and Lynch, K. R. 1991. Molecular characterization of alpha-1 and alpha-2 adrenoreceptors. *Trends Pharm. Sci.* 12:62–67.

Lincoln, S. E.; Daly, M. J.; and Lander, E. S. 1991. *PRIMER: A Computer Program for Automatically Selecting PCR Primers.* MIT Center for Genome Research and Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142.

Nicodeme, P. 1993. A computer support for genotyping by multiplex PCR. Technical Report LIX/RR/93/09, LIX, École Polytechnique, 91128, Palaiseau Cedex, France.

Olschwang, S.; Delaitre, O.; Melot, T.; Peter, M.; Schmitt, A.; Frelat, G.; and Thomas, G. 1989. Description and use of a simple laboratory-made automat for *in vitro* DNA amplification. *Methods in Molecular and Cellular Biology* 1(3):121–127.

Pearson, W. R.; Robins, G.; Wrege, D. E.; and Zhang, T. 1995. A new approach to primer selection in polymerase chain reaction experiments. In *Third International Conference on Intelligent Systems for Molecular Biology*, 285–291. AAAI Press.

Robins, G.; Wrege, D. E.; Zhang, T.; and Pearson, W. R. 1993. On the primer selection problem in polymerase chain reaction experiments. Technical Report CS-93-68, Department of Computer Science, University of Virginia, Charlottesville, Virginia 22903, U.S.A.

Sedgewick, R., and Flajolet, P. 1996. *An Introduction to Analysis of Algorithm.* Addison-Wesley.

Sorokin, A.; Lapidus, A.; Capuano, V.; Galleron, N.; Putjic, P.; and Ehrlich, S. D. 1996. A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial mapping and sequencing. *Genome Research* 6:448–453.

Watson, J. D.; Witkowski, J.; Gilman, M.; and Zoller, M. 1992. *Recombinant DNA.* Scientific American Books. second edition.