c-GAMMA: Comparative Genome Analysis of Molecular Markers

Pierre Peterlongo[1], Jacques Nicolas[1], Dominique Lavenier[2], Raoul Vorc'h[1], and Joël Querellou[3]

[1] Équipe-projet INRIA Symbiose, Campus de Beaulieu, Rennes, France
WWW home page: http://www.irisa.fr/symbiose/
[2] ENS Cachan - IRISA, France
[3] LM2E UMR6197 Ifremer, Centre de Brest, France

Abstract. Discovery of molecular markers for efficient identification of living organisms remains a challenge of high interest. The diversity of species can now be observed in details with low cost genomic sequences produced by new generation of sequencers. A method, called `c-GAMMA`, is proposed. It formalizes the design of new markers for such data. It is based on a series of filters on forbidden pairs of words, followed by an optimization step on the discriminative power of candidate markers.

First results are presented on a set of microbial genomes. The importance of further developments are stressed to face the huge amounts of data that will soon become available in all kingdoms of life.

1 Introduction

The decade started with the complete sequencing of the Haemophilus influenzae genome in 1995 [1]. This period was characterized by the multiplication of sequencing projects for getting a better comprehensive view of the whole tree of life. During this time, an exponential rate of sequencing projects was observed, with a $\times 2$ increasing rate every 20 months [2]. Comparative analyses of complete genomes from Bacteria, Archaea to Human have a huge impact on all aspects of life sciences and is deeply redesigning the evolution theory in the light of genomics [3]. To better understand the driving forces in speciation, the diversity in virulence of pathogens, the diversity in metabolic pathways in various key species, more complete genomes of closely related strains of the same species (or species of the same genus) are needed. This recently triggered a flood of sequencing projects for novel strains of key pathogens (Campylobacter, Haemophilus, Mycobacterium, Streptococcus, etc.), model species (Bacillus, Escherichia), ecological key players (Prochlorococcus, Synechococcus) and species potentially interesting for biotechnology (Pyrococcus, Thermococcus). It appears that for these species the number of sequencing projects is growing exponentially, and the time has come to address specifically comparative genomics at micro-scale evolution (Table 1).

One of the main needs is the design of molecular markers that can achieve a high level of discrimination between different species or strains. The use of molecular markers has become increasingly popular in many fields: phylogenetic reconstruction in microbiology, quality control in food industry, traceability in epizooty and epidemic diseases, barcoding of life, forensics, etc. Each domain of activity has its favourite marker(s) working optimally for a specific purpose. The increasing number of complete genomes of related species available in databases raises the question of rapid determination of additional molecular markers through comparative genomics.

This paper proposes a novel approach to characterize molecular markers within a set of complete genomes of related strains or species targeting PCR (Polymerase Chain Reaction). PCR is one of the most important tools in genetic engineering for amplifying specific DNA fragments defined by flanking pairs of words on both side. These pairs of words are matched by complementary short synthetic nucleotides, called primers. Potential applications include strain quality control, identification, taxonomy and possibly phylogeny.

Phylum	Genus	Species, strains	Genomes projects	Genomes completed
Arthropoda	*Drosophila*		10	10
Euryarchaeota	*Methanococcus*		7	6
	Pyrococcus		4	3
	Thermococcus		5	3
Firmicutes	*Bacillus*	*anthracis*	8	3
	Bacillus	*cereus*	14	4
	Bacillus	other species	13	8
	Clostridium	*botulinum*	7	4
	Clostridium	other species	29	10
	Lactobacillus		12	11
	Staphylococcus	*aureus*	12	12
	Staphylococcus	other species	4	4
	Streptococcus	*pneumoniae*	17	3
	Streptococcus	other species	30	22
Spirochaetes	*Borrelia*	*burgdorferi*	7	1
	Borrelia	other species	4	2
Proteobacteria	*Burkholderia*		45	13
	Campylobacter	*jejuni*	9	4
	Campylobacter	other species	6	2
	Escherichia	*coli*	31	10
	Haemophilus	*influenzae*	13	4
	Haemophilus	other species	4	2
	Pseudomonas		16	11
	Rickettsia		13	10
	Salmonella	*enterica*	23	6
	Shewanella		10	6
	Vibrio	*cholerae*	6	1
	Vibrio	other species	14	6
	Yersinia	*pestis*	13	6
	Yersinia	other species	7	3
Actinobacteria	*Mycobacterium*	*tuberculosis*	5	4
	Mycobacterium	other species	12	12
Tenericutes	*Mycoplasma*		14	13
	Ureaplasma	*urealyticum*	11	1
	Ureaplasma	other species	5	0
Cyanobacteria	*Prochlorococcus*	*marinus*	12	12
	Synechococcus		15	10

Table 1. Number of genome projects related to important prokaryotic genera and species (Source: GOLD http://www.genomesonline.org/ and Microbesonline http://www.microbesonline.org/, modified, april 2009)

2 Identification of genome species markers using PCR

Let us first explain the way markers are used during PCR. Let $\overline{s}$ denotes the reverse complement of word s over a four letter alphabet A, T, C, G and v denotes the marker to be selectively amplified. A DNA double helix corresponds to the hybridization of sequence $x.u.v.w.y$ with sequence $\overline{x.u.v.w.y} = \overline{y}.\overline{w}.\overline{v}.\overline{u}.\overline{x}$. PCR aims at hybridizing of subsequence $u.v.w$ with its complementary strand $\overline{w}.\overline{v}.\overline{u}$, initiated by two short synthetic nucleotides, the primers, which will match $\overline{u}$ and w respectively (x,u,v,w and y are words). Thus, the word v corresponding to the marker itself is produced in the context of two fixed words corresponding to primer sequences.

Most of the specific sequences that are used as molecular markers come from ubiquitous components of the cell with limited nucleic material such as ribosomes. One of the main resources concerns 16S rRNA and can be found on various dedicated websites including the Ribosomal Database project [4]. The last release (April 3, 2009) reports 836 814 16S rRNA annotated and aligned sequences. They mostly come from uncultured microbes, as a result of the standard investigation of microbial diversity by molecular methods and more recently by metagenomics.

The limits of 16s rRNA for species identification have been reached to handle this high number of species. Firstly, there is no linear relationship between 16S rRNA similarity and DNA-DNA hybridization. A consensus was reached, specifying that 16S rRNA similarity levels fewer than 97% between two strains is equivalent to DNA-DNA hybridization level fewer than 70%, and

discriminates two different species. However, many different species display 16S rRNA sequences similarity within the range of 98-99%, and in those cases, 16S rRNA cannot be used to establish a strain as a novel species. Other variable molecular markers are frequently used in addition to 16S rRNA like housekeeping genes [5]. The major drawback is that none of these additional markers are universal. Secondly, in phylogeny reconstruction, 16S rRNA cannot solve all the problems and for some taxonomic groups, tree topologies are uncertain. The help of additional sequences is required and the current trend is to use a set of sequences corresponding to the concatenation of ribosomal proteins.

Another widely used molecular marker for Eukarya "barcoding" is the 648-bp region of the cytochrome oxidase I (COI). DNA barcoding employs a minimum of 500-bp sequence of COI to help species identification and discovery in large assemblages of life [6]. Although well adapted to species identification, COI sequences cannot be used in phylogeny nor in ecotypes identification. Here again, additional molecular markers need to be found in the set of complete genomes currently available for various tasks ranging from quality control in laboratory collections of pico and micro-eukaryotes to traceability of pathogens, of pests in the environment, etc.

Biologists need help both for choosing markers on a less restrictive set of sequences and for choosing the primers that will select these markers. Since most authors consider short sequences, they rely generally on multiple alignments for a subset of species of interest, so that conserved and non conserved regions become directly visible. Targets are then defined as conserved regions delimiting some highly variable regions and potential primers are then checked with the whole database in order to prune solutions matching elsewhere in the sequence. Nice environments have been developed in this context [7]. However, the task of finding suitable markers is not fully automated and does not scale to many species or long sequences due to the multiple alignment step.

In [8], A. Pozhitkov and D. Tautz propose a program for simply finding one probe discriminating a given set of sequences from others. Although the algorithm could be widely improved, they use the interesting idea of building a crude index of words of fixed size in order to speed the search for common patterns.

Finding the best primer pairs where each primer is a substring of a given sequence is by itself a complex multicriteria task that has been well described in [9]. It extends beyond string matching and the use of edit distances since it involves criteria including the proximity between primer melting temperatures, minimization of hybridization effects between forward and reverse primers, and avoidance of hybridization of primers with themselves. It may be solved efficiently using dynamic programming schemes that use extensions of approximate string matching equations.

The large scale design of primers has been tackled in another context: the observation of gene expression for a given organism. The issue is to produce for a subset of genes, or the complete set of genes of some genomes, a set of markers that identify each gene. The technique used in such a case, microarrays, involves an array of spots, each being attached to a primer. The main objective, in this context, is to find a set of primers working at a same temperature (called probes), each one recognizing a unique gene in the given set. Combining suffix-tree indexing, dynamic programming and ad'hoc filters, Kaderali and Schliep [10] showed that it is possible to identify organisms but that this technique requires long probes for identifying many species. A recent review of long primers (size greater than 40) identification tools is available in [11]. Producing a microarray with many long probes remains an expensive operation. One of the interests of working on species instead of gene expression is that the primer-gene association does not needs to be bijective. The issue becomes the choice of a minimal set of primers, a problem easily reduced to the minimum set covering problem [12].

The present study shows yet another variation of primer design problem. The idea of working on whole genomes is kept but restricted to PCR as a low cost identification technique. The genome-based marker design problem consists in determining within a set of genomes, (i) primer pairs conserved over these genomes (ii) usable for PCR amplification for genomes differentiation (iii), associated to at least one homologous flanking region (iv) that can be used for diverse objectives: speciation, strain and species rapid identification, taxonomy, search of variable regions and contextual gene analysis.

To the best of our knowledge, this problem has never been stated before. The closest study we are aware of in terms of constraints to be solved, corresponds to a very different application: the study of the variability of individual genomes in terms of deletions or translocations that can occur in mutants and pathogenic states like cancer. A recent clever experimental protocol, PAMP [13], uses multiplex PCR to selectively amplify the variations observed in pathogenic cells. Authors have developed an optimization technique to design primer sets based on simulated annealing and integer programming [14]. This technique can processes sequences up to one Mbp. Although the setting is different, it shares an interesting characteristic with our approach, namely the comparison of ordered pairs of primers.

3 Model

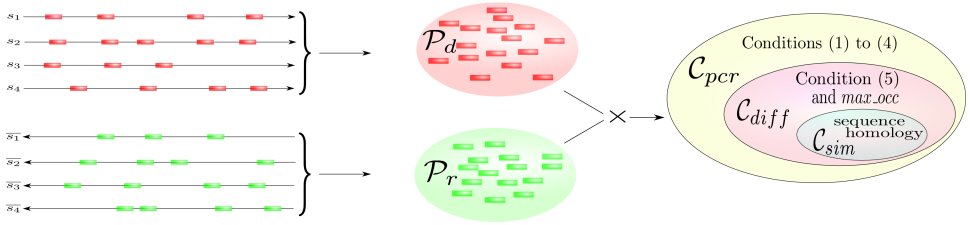

Fig. 1. Overview of the model. A set $\mathcal{S}$ of four sequences guides the design of two primers (red and green rectangles) sets $\mathcal{P}_d$ and $\mathcal{P}_r$, whose pairs generate sets $\mathcal{C}_{pcr}$, $\mathcal{C}_{diff}$ and $\mathcal{C}_{sim}$.

We propose a generic formalization and a model for designing primer pairs for PCR amplification for finding markers able to differentiate genomes. The model relies on four steps (see Figure 1 for an overview):

1. Given a set of sequences and primer parameters, detection of oligonucleotides that could be primers and that theoretically hybridize on each of these sequences on direct or reverse-complementary strand. This detection is based on physico-chemical properties of DNA fragments (Section 3.1);
2. From this set of possible primers, selection of all pairs that respect location properties on each sequence. These properties derive from PCR amplification technical constraints (Section 3.2);
3. Selection of primer pairs that define fragments considered as molecular markers; they permit to differentiate sequences from each other, using a simple length criterion (Section 3.3);
4. Selection of all pairs that further define flanking regions (fragments on the left and right hand sides of primer pairs) sharing homology or being highly variable (Section 3.4).

3.1 Primers characteristics

The primary goal of a primer is to hybridize a complementary strand of a DNA sequence at a well defined position. Optimal primers are produced on the basis of an hybridization model taking into account various criteria:

1. **G+C content**: the G+C percentage of a primer is framed between a minimum and a maximum threshold value, typically between 40% and 60%.
2. **Melting temperature**: the melting temperature of a primer must be in bounded interval. The computation is based on the nearest neighbour method [15]. The melting temperature calculation also takes into account the concentration of nucleotides and the concentration of salt.

3. **Repeats**: primers containing large trains of identical nucleotides or dinucleotides are eliminated.
4. **Hairpin loops**: primers must not include hairpin loops. The size of the stem or the size of the loop must be lower than a predefined value.
5. **Self-complementarity**: a primer must not self hybridize during PCR. Thus, primers that form a duplex with their complementary strand are removed.
6. **Thermodynamic stability at primer ends**: The Gibbs Free Energy (ΔG: in units of kcal/mole) values computed on the 5' and 3' ends of the primers are bounded. The ΔG value determines the strength of the hybridization and triggers the decision of considering the position as a potential hybridization site.

If a nucleic sequence possesses all these qualities, it can be considered as a successful primer. The next question is: what are the conditions for this primer to hybridize with a DNA sequence? In other words, given this primer and any portion of the genome, can they hybridize together? The answer is brought by the calculation of the thermodynamic stability between the two strands. The nearest neighbour method proposed by SantaLucia [15] is used to compute ΔG along the two oligonucleotide sequences, with special care in the 5' and 3' extremities of the primer.

3.2 primer pairs for PCR amplification

Interesting primer pairs are those defining a fragment that may be amplified by PCR. Their hybridization positions (called PCPP, for Primers Couple PCR Positions) must respect some distances characteristics and some repartition conditions over the hybridization locations of each of these primers. Figure 2 gives an example of PCPP.

For a non-ambiguous characterization of PCR results and for a given a primer pair, two PCPP could not start or end at the same position. This avoids amplification of alternative sizes of fragment at a same position.

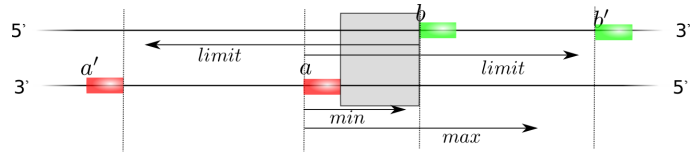

Fig. 2. A portion of two DNA strands is shown. On each strand, a primer has two hybridization locations at position a' and a on the direct strand (red rectangles on the bottom line) and position b and b' on the reverse complementary one (green rectangles on the top line). This pair shows hybridization sites (a and b) which respect conditions (1) to (3). The shaded area corresponds to a putative molecular marker. If condition (4) is also respected, (a, b) is a PCPP.

In the following, the set $\mathcal{C}_{pcr}$ of primer pairs defining at least one PCPP on each sequence is defined. Given a set of sequences $\mathcal{S}$, we dispose of two primers sets: $\mathcal{P}_d$ containing primers that hybridize on the direct strand of each sequence and $\mathcal{P}_r$ that contains primers hybridizing the reverse complementary strand of each sequence. $pos(s, p)$ is then defined as the set of positions where the primer $p \in \mathcal{P}_d$ hybridizes on the sequence $s \in \mathcal{S}$. $pos(\overline{s}, p')$ is defined as the set of positions where the primer $p' \in \mathcal{P}_r$ hybridizes on the sequence s reverse-complemented. For the sake of clarity, all positions are reported on the direct strand.

$\mathcal{C}_{pcr}$ is defined as the set of pairs $c = (p, p')$ of primers from $\mathcal{P}_d \times \mathcal{P}_r$ such that for each $s \in \mathcal{S}$:

$$\exists a \in pos(s,p) \text{ and } \exists b \in pos(\overline{s},p')\, min \leq b - a \leq max \quad (1)$$

Moreover, the conditions of uniqueness for fragments starting or ending at a given position can be expressed as follows:

$$\forall\, a' \neq a \in pos(s,p)\, a' < b \Rightarrow b - a' \geq limit \quad (2)$$

$$\forall\ b' \neq b \in\ pos(\overline{s}, p')\ a < b' \Rightarrow b' - a \geq limit \tag{3}$$

Conditions (1) ensures that the pair of primers defines at least a fragment of length in $[min +$ primers length$, max]$. Conditions (2) and (3) ensure that the selected pair of primers defines non-ambiguous fragments at given positions. Figure 2 represents hybridization locations respecting conditions from (1) to (3).

In order to get rid of amplification of alternative sizes of fragment at a same position, $\mathcal{C}_{pcr}$ does not contain pairs of primers with hybridization sites respecting condition (1) and not respecting condition (2) and (3). Formally, $\forall (p, p') \in \mathcal{C}_{pcr}$ and $\forall s \in \mathcal{S}$

$$\forall (a, b) \in pos(s, p) \times pos(\overline{s}, p')\ min \leq b - a \leq max \Rightarrow (2) \wedge (3) \tag{4}$$

3.3 Primer pairs for sequence differentiation

Primer pairs in $\mathcal{C}_{pcr}$ are potential candidates for PCR amplification. Let $\mathcal{C}_{diff}$ be a subset of $\mathcal{C}_{pcr}$ containing all pairs of primers defining inner fragments whose length enable to differentiate sequences.

To do so, $lengths(s, c)$ ($s \in \mathcal{S}$ and $c \in \mathcal{C}_{pcr}$) is defined as the set of lengths of inner fragments defined by PCPP of c on the sequence s. Additionally, $\mathcal{C}_{diff}$ is defined as the subset of pairs c from $\mathcal{C}_{pcr}$ such that $\forall s, s' \in \mathcal{S}, s \neq s', \exists l \in lengths(s, c)$ such that, $\forall l' \in\ lengths(s', c)$

$$\max\left(\frac{l}{l'}, \frac{l'}{l}\right) \geq \delta, \text{ with } \delta \text{ a fixed parameter.} \tag{5}$$

Informally, condition (5) ensures that for each pairs of primers $c \in \mathcal{C}_{diff}$ and that for each couple of sequences $s, s' \in \mathcal{S}$, at least one of the fragments defined by any PCPP of c on s has a length different enough from all fragments defined by PCPP of c on sequence s'. This property enables selected fragments to differentiate sequences from each other with a simple length-based test.

Moreover, in order to provide readable PCR results by clearly distinguish amplified fragments, an additional parameter *max_occ.* is applied. In $\mathcal{C}_{diff}$, couples c whose number of hybridization sites is bigger than *max_occ.* are removed. Formally, $\forall c \in \mathcal{C}_{diff}, \forall s \in \mathcal{S} : |lengths(s, c)| \leq$ *max_occ.*, with $|\mathcal{E}|$ denoting the cardinality of the set $\mathcal{E}$ (this notation is used in the rest of the paper).

3.4 Sequence similarity / variability

At last, sequence composition of fragments defined by PCPP is taken into consideration. Given a PCPP, the internal region (red area on Figure 3) and the two flanking regions (yellow fragments on Figure 3) are considered. Depending on the application, one may want these areas to be homologous or variable.

Bearing in mind that any combination of searched homology is possible, $\mathcal{C}_{sim}$ (see Figure 1) is defined as the subset of pairs of primers from $\mathcal{C}_{diff}$, such that there exists at least on PCPP for these pairs with variable centre fragment and at least one homologous flanking region. Each fragment is considered both on the direct and on the reversed strand.

4 Methods

This section presents the methods for finding potential primers (Section 4.1). Then, respectively, sections 4.2, 4.3 and 4.4 show how previously defined sets $\mathcal{C}_{pcr}$, $\mathcal{C}_{diff}$ and $\mathcal{C}_{sim}$ are detected.

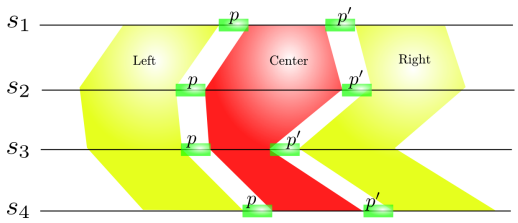

Fig. 3. A pair of primers $c = (p, p') \in \mathcal{C}_{sim}$ has one PCPP on each of four genomes. For the sake of clarity only one strand is represented. Inner fragments defined by this PCPP (red) present for instance high variability while left or right flanking regions (yellow) present for example high similarity.

4.1 Methods for primers detection ($\mathcal{P}_d$ and $\mathcal{P}_r$)

Given a set $\mathcal{S}$ of n sequences, ideally all potential primers that may hybridize at least once on each sequence should be generated. Such an approach is actually unfeasible by enumerating all the primer configurations. In this case study, considering primers of length 25 would lead to test 4^{25} elements, which is unrealistic.

Instead, the following approach is used. To search common primers of length l, all l-mers of each sequence $s \in \mathcal{S}$ are first considered. In addition, to extend the space search, these l-mers are degenerated in their middle. Practically, 2 nucleotides are modified, leading to generate up to 4^2 l-mers per position.

After this stage, a huge set of l-mers are considered as potential primers. Only those respecting conditions presented in Section 3.1 are selected. More precisely, the selection of primers is achieved through a pipeline of filters. Each stage of the pipeline eliminates the candidates which do not fit specific criteria. For efficiency purpose, the most stringent criteria are first taken into consideration. The implementation is based on a series of functions which does not present algorithmic challenges and are not detailed here.

After this process, a new set of l-mers considered as putative primers is available. From this set only those that hybridize on all different sequences are selected. The whole set of primers is thus checked against $\mathcal{S}$. In that way, a list of hybridizing primers is associated to each sequence. The intersection of these lists, results in the set of primers that hybridize at least once on every sequences.

To speed-up the hybridization test, the sequences are first indexed with a seed-based index technique. The length of the seeds are set to 6, meaning that a primer hybridization will be reported only if the primer and the genome share at least 6 common nucleotides (or, more exactly, two complementary 6-nt words). In that case, a ΔG value is computed as presented in section 3.1. Depending of the ΔG value, the primer is added or not to the primer list associated with the genome.

4.2 Methods for detection of primer pairs for PCR amplification ($\mathcal{C}_{pcr}$)

From this point, two sets of potential primers are available: $\mathcal{P}_d$ and $\mathcal{P}_r$ that hybridize respectively at least once on each sequence s and once on each reverse complementary sequence $\overline{s}$.

In order to verify conditions (1) to (4), all possible primer pairs $(p, p'_j) \in \mathcal{P}_d \times \mathcal{P}_r$ are checked. On each sequence s, the ordered hybridization locations $pos(s, p)$ and $pos(\overline{s}, p')$ are available from previous steps.

In a few strokes, the algorithm works as follows: positions over $pos(s, p)$ and $pos(\overline{s}, p')$ are read conjointly as long as condition (1) is not fulfilled. In case a pair of hybridization positions $(a, b) \in pos(s, p) \times pos(\overline{s}, p')$ respecting these three conditions is found, then previous positions a' on $pos(s, p)$ (resp. next position b' on $pos(\overline{s}, p')$) is checked in order to validate that condition (2) (resp. (3)) is respected. In case of success, the pair (p, p'_j) is tagged as a potential pair for PCR, otherwise the pair is rejected (condition (4)) and the reading of its positions is stopped. All pairs

of primers respecting conditions (1) to (4) are stored is the set $\mathcal{C}_{pcr}$.

For a pair of primers $(p, p'_j) \in \mathcal{P}_d \times \mathcal{P}_r$, this approach reads all positions in $pos(s, p)$ and in $pos(\overline{s}, p')$ leading to a complexity in $O(|pos(s, p)| + |pos(\overline{s}, p')|)$ that is $O(N)$ with N the total length of the input genomes. As this computation is done for each possible pair of primers, the overall time complexity of this step is in $O(|\mathcal{P}_d| \times |\mathcal{P}_r| \times N)$ that is $O(N^3)$. In practice, the time complexity is much lower, as confirmed by experimental tests described in Section 5.

4.3 Methods for detecting primer pairs for sequence differentiation ($\mathcal{C}_{diff}$)

Finding the subset $\mathcal{C}_{diff}$ from the set $\mathcal{C}_{pcr}$ is straightforward. For each primer pair $c \in \mathcal{C}_{pcr}$ and each sequence $s \in \mathcal{S}$, $lengths(s, c)$ is known (see Section 3.3). Trivially, for each primer pair c in $\mathcal{C}_{pcr}$ and for each couple of sequences $s, s' \in \mathcal{S}, s \neq s'$, c is conserved in $\mathcal{C}_{diff}$ if there exists $l \in lengths(s, c)$ that is different enough from all $l' \in lengths(s', c)$ so that condition (5) is respected. Simultaneously, it is trivial to conserve in $\mathcal{C}_{diff}$ only primer pairs for which the number of occurrences of PCPP on each sequence respects the *max_occ.* parameter.

This checking is done in $O(|lengths(s, c)| \times |lengths(s', c)|)$ for each couple of sequences s, s' and each primer pair $c \in \mathcal{C}_{pcr}$. Thus, for each primer pair, this checking is done in $O(n^2 \times |lengths(s, c)| \times |lengths(s', c)|)$ leading to an overall time complexity of $O(|\mathcal{C}_{pcr}| \times n^2 \times |lengths(s, c)| \times |lengths(s', c)|)$. Note that in practice n, $|lengths(s, c)|$ and $|lengths(s', c)|$ are negligible with regard to $|\mathcal{C}_{pcr}|$.

4.4 Methods for detection of primer pairs taking into account sequences similarity and variability ($\mathcal{C}_{sim}$)

Pairs of primers from $\mathcal{C}_{diff}$ that define fragments respecting conditions exposed in Section 3.4 (see also Figure 3) are selected to be included in $\mathcal{C}_{sim}$. Knowing the large amount of work previously done for finding multiple local alignments, we decided to not develop our own algorithm. In this framework, we used `MEME` [16] which provides an e-value estimation enabling used as a formal criterion for creating set $\mathcal{C}_{sim}$.

As stated earlier, this step is highly tunable depending on the biological applications. In this framework, the method is the following: for each primer pair $c \in \mathcal{C}_{diff}$, `MEME` is applied on all combination of PCPP of c on the set of genomes. The primer pair c is stored in $\mathcal{C}_{sim}$ if one of the `MEME` results provides both:

- an e-value bigger than a fixed threshold for center fragments alignments,
- an e-value bellow another threshold for flanking regions alignments.

5 Results

The method has been implemented in the `c-GAMMA` tool acting as a pipeline of programs. As a preliminary test `c-GAMMA` was applied on a set $\mathcal{S}$ of height Thermococcales genomes (source GOLD database http://www.genomesonline.org/) of total length $N \approx 16$ Mb. Thermoccocales were chosen due to their high interest in biotechnology. Species belonging to this family display thermostable hydrolases of interest. It is therefore important to find molecular markers that can help to identify strains within Thermococcus and Pyrococcus species and insure quality control. The goal of our study was to design couples of primers defining molecular markers both identifiable by PCR and by a sequence homology criterion.

All experiments were computed using a PC Intel® dual core 2.40 GHz running under Linux Fedora with 2 GBytes memory.

5.1 Primer detection results ($\mathcal{P}_d$ and $\mathcal{P}_r$)

The method exposed in Section 4.1 was applied for generating primers of length 25 that hybridize at least once on each genome (direct and reverse complementary strand) in $\mathcal{S}$. The primers generation was done by testing all 25-mers presents on each genome (direct and reverse complementary strands) and degenerating two central positions on of each of them. Thus ≈ 512 million of 25-mers were tested. Each of these 25-mers was selected for further analysis if classical parameters for PCR amplification were respected. This method generated 2 803 510 primers on the direct strand and 2 796 747 on the reverse complementary strand.

Then only primers that hybridized at least once on each sequence (direct and reverse complementary strands) were conserved. This step conserved 62 247 primers on the direct strands (set $\mathcal{P}_d$) defining 6 309 356 hybridization sites. On the reverse complementary strands, 62 764 primers were conserved (set $\mathcal{P}_i$) having a total of 6 295 992 hybridization sites. Note that, in average, a primer hybridization site is found each ≈ 2.38 positions on each strand and that each primer has ≈ 100 hybridization sites.

This step is the most time consuming. It was performed in less than six hours.

5.2 Primer pairs for PCR amplification and sequence differentiation results ($\mathcal{C}_{pcr}$ and $\mathcal{C}_{diff}$)

For creating the sets $\mathcal{C}_{pcr}$ and $\mathcal{C}_{diff}$ from $\mathcal{P}_d$ and $\mathcal{P}_i$, methods presented in Section 4.2 were sequentially applied. For defining $\mathcal{C}_{pcr}$ the parameters were the following: $min = 200$, $max = 2000$ and $limit = 3500$. These parameters facilitate standard PCR procedure used by most diagnostic laboratories.

First, set $\mathcal{C}_{pcr}$ containing pairs of primers that respect conditions (1) to (4) is selected. This was done on all possible primer pairs in $|\mathcal{P}_d| \times |\mathcal{P}_i|$ (≈ 3.9 billion pairs in this experimentation). This computation took less than four hours and provides 63877 couples.

(a)

δ	$max\_occ.$	$\lvert\mathcal{C}_{diff}\rvert$
1	10	63872
1	9	63865
1	8	63782
1	7	63518
1	6	63193
1	5	62018
1	4	59050
1	3	53218
1	2	42187
1	1	18122

(b)

δ	$max\_occ.$	$\lvert\mathcal{C}_{diff}\rvert$
1.01	2	1149
1.02	2	301
1.03	2	180
1.04	2	107
1.05	2	71
1.06	2	56
1.07	2	37
1.08	2	11
1.09	2	11
1.10	2	11

(c)

δ	$max\_occ.$	$\lvert\mathcal{C}_{diff}\rvert$
1.01	1	137
1.02	1	68
1.03	1	41
1.04	1	36
1.05	**1**	**24**
1.06	1	23
1.07	1	17
1.08	1	0

Table 2. Quantitative results while varying parameters for finding $\mathcal{C}_{diff}$ from $\mathcal{C}_{pcr}$. $\mathcal{C}_{pcr}$ contained initially 63877 primer pairs. Tests **(a)** make variation over the maximal number of occurrences ($max\_occ.$) of PCPP of each couple on each genome. Tests **(b)** (resp. **(c)**) make variation over the parameter δ (see Section 3.3) using at most 2 (resp. 1) occurrences of PCPP of each couple on each genome.

For obtaining $\mathcal{C}_{diff}$ from $\mathcal{C}_{pcr}$, a set of tests using several distinct parameters was performed. Each test was computed in less than 30 seconds. Results are shown Table 2.

This experiment shows that the $max\_occ.$ parameter (table **(a)**) has a strong influence and that most of the primer pairs define between 1 and 5 PCPP per genome. However, even by constraining exactly one occurrence per genome (last line of **(a)**), one notices that still 18122 pairs respect the parameters.

Moreover, these results show that, fortunately, even while applying very stringent parameters, some primer pairs are found. For instance, while asking for a minimal fragment length difference

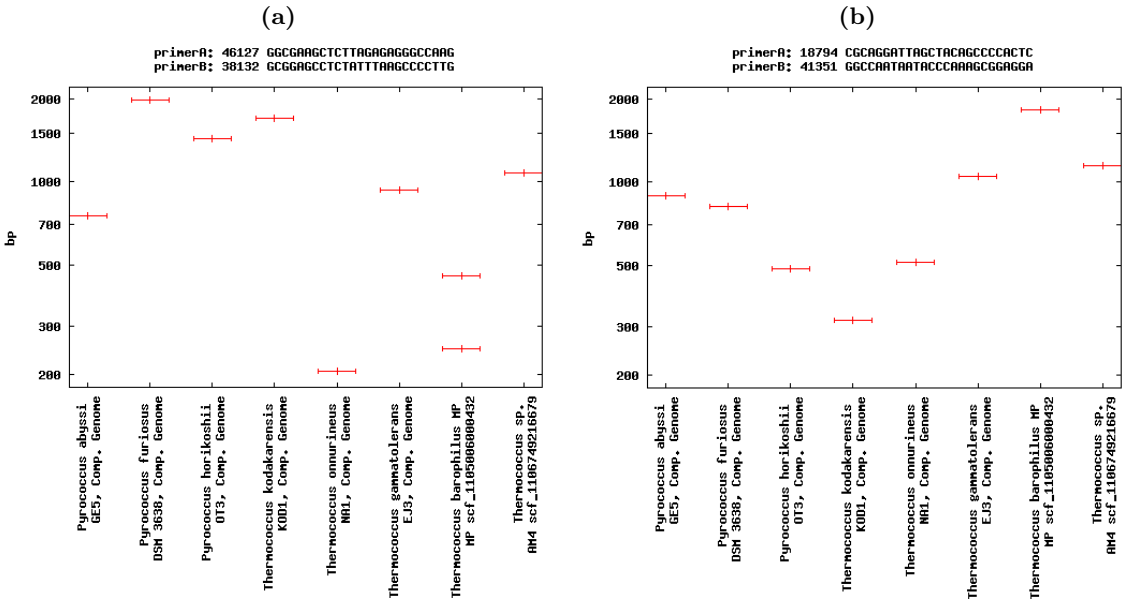

Fig. 4. (a) A randomly chosen theoretical PCR obtained on the studied set of genomes using a pair of primers respecting conditions $min = 200$, $max = 2000$, $limit = 3500$, $\delta = 1.10$ and $max\_occ. = 2$. **(b)** theoretical PCR obtained on a primer pair respecting conditions $min = 200$, $max = 2000$, $limit = 3500$, $\delta = 1.05$ and $max\_occ. = 1$ and defining a variable marker region and a homologous flanking region between set of genomes.

of $\delta = 10\%$ and at most 2 fragments occurrences on each genome (last line of **(b)** of Table 2), 11 primer pairs are found. Figure 4**(a)** shows the theoretical PCR result that would be obtained on the studied set of genomes thanks to a randomly chosen primer pair respecting such conditions. It is worth mentioning that, as expected, this single PCR result clearly permit to distinguish strains from each other, as for any primer pair respecting the required parameters.

5.3 Detection of primer pairs taking into account sequence similarity and variability results ($\mathcal{C}_{sim}$)

The goal here is to show that an approach involving similarity criterion in addition to lengths attributes provides realistic biological results. Thus, we show results of an experimentation run on the set of 24 primer pairs generating one PCPP on each sequences with at least $\delta = 5\%$ of difference of length between them (bold faced line of **(c)** of Table 2.). For each of these primer pairs PCPP, `MEME` was applied both on central fragment and on the two flanking areas (over 1000 bp). We selected primers pairs for which the best alignment had an e-value higher than 1 for the central fragment and lower than 10^{-1} for any of the flanking regions.

Among this 24 primer pairs, one of them gave satisfying results. Indeed, the couple of primers (CGCAGGATTAGCTACAGCCCCACTC, GGCCAATAATACCCAAAGCGGAGGA), having exactly one PCPP on each genome (see Figure 4**(b)**) define a central fragment highly variable (best local alignment found has an e-value equal to $4.2e{+}5$) and has a left area containing an homologous motif (shown on Figure 5) of length 98 with a e-value of $1.3e^{-2}$.

6 Conclusion

This paper proposes a generic model to efficiently (1) detect primers on a set of genomes, and (2) define suitable molecular markers for genomes differentiation. The differentiation occurs at

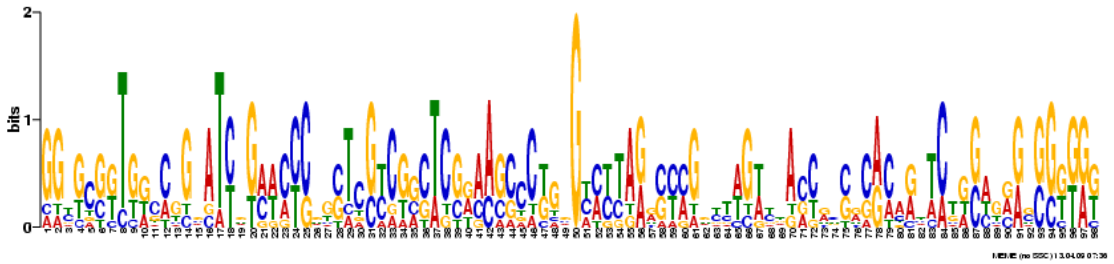

Fig. 5. Motif found by `MEME` on left flanking region of couple of primers CGCAGGATTAGCTACAGCC-CCACTC and GGCCAATAATACCCAAAGCGGAGGA.

two levels, a simple length criterion, and a more precise criterion on flanking regions homologies and/or variability.

The model is fully implemented within a bioinformatics pipeline called `c-GAMMA`. Applied on a set of height bacterial genomes (16 Mb), c-GAMMA designed primers for the detection of molecular markers in 12 hours on a standard work station, making possible the genomes differentiation both using length and homology criterion.

These encouraging preliminary results open the way to other experimentations on the huge source of data produced by next generation sequencing machines.

Moreover, methods proposed in this framework mark a step over molecular markers detection. They are highly suitable for further enhancements such as:

- improving the primers generation by producing all oligonucleotides that may hybridize on a genome fragment. Generation is currently achieved through a simple degeneration scheme on the middle part of the fragments. Such an approach will provide more suitable results. However, it will dramatically increase the number of possible primer pairs ($|\mathcal{P}_d| \times |\mathcal{P}_r|$) and will raise computational issues for finding hybridization sites;
- instead of considering only one primer pair on each sequence, the model may be improved by considering simultaneously several primer pairs to perform multiplex PCR in order to efficiently differentiate close species.

References

1. Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., al.: Whole-genome random sequencing and assembly of haemophilus influenzae rd. Science **269**(5223) (1995) 496–512
2. Koonin, E., Wolf, Y.: Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucl. Acids Res. **36**(21) (2008) 6688–6719
3. Koonin, E.: Darwinian evolution in the light of genomics. Nucl. Acids Res. **37**(4) (2009) 1011–1034
4. Cole, J., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R., Kulam-Syed-Mohideen, A., McGarrell, D., Marsh, T., Garrity, G., Tiedje, J.: The ribosomal database project: improved alignments and new tools for rrna analysis. Nucl Acids Res **37**(suppl 1) (2009) D141–145
5. Stackebrandt, E., Frederiksen, W., Garrity, G., Grimont, P., Kampfer, P., Maiden, M., Nesme, X., Rossello-Mora, R., Swings, J., Truper, H., Vauterin, L., Ward, A., Whitman, W.: Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. **52**(3) (2002) 1043–1047
6. Ratnasingham, S., Hebert, P.: Bold: the barcode of life data system. Mol. Ecol. Notes (2007)
7. Ludwig, W., Strunk, O., Westram, R., Lothar Richter, H.M., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Anton W. Ginhart, O.G., Grumann, S., Hermann, S., Jost, R., Andreas Konig, T.L., Lubmann, R., May, M., Nonhoff, B., Boris Reichel, R.S., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H.: Arb: a software environment for sequence data. Nuc. Acids Res. **32**(4) (2004) 1363–1371

8. Pozhitkov, A., Tautz, D.: An algorithm and program for finding sequence specific oligonucleotide probes for species identification. BMC Bioinformatics **3**(9) (2002)
9. Kampke, T., Kieninger, M., Mecklenburg, M.: Efficient primer design algorithms. Bioinformatics **17**(3) (2001) 214–225
10. Kaderali, L., Schliep, A.: Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics **18**(10) (2002) 1340–1349
11. Lemoine, S., Combes, F., Le Crom, S.: An evaluation of custom microarray applications: the oligonucleotide design challenge. Nuc. Acids Res. **37**(6) (2009) 1726–1739
12. Wang, J., Li, K., Sung, W.: G-primer: greedy algorithm for selecting minimal primer set. Bioinformatics **20**(15) (2004) 2473–2475
13. Liu, Y., Carson, D.: A novel approach for determining cancer genomic breakpoints in the presence of normal dna. PLoS ONE **2**(4) (2007)
14. Bashir, A., Liu, Y.T., Raphael, B.J., Carson, D., Bafna, V.: Optimization of primer design for the detection of variable genomic lesions in cancer. Bioinformatics **23**(21) (2007) 2807–2815
15. SantaLucia, J.J.: A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A **95**(4) (1998) 1460–1465
16. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (1994) 28–36