BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 19 2007, pages 2550–2557
doi:10.1093/bioinformatics/btm392

Sequence analysis

PhylArray: phylogenetic probe design algorithm for microarray

Cécile Militon[1,†], Sébastien Rimour[1,2,†], Mohieddine Missaoui[1,2], Corinne Biderre[1], Vincent Barra[2], David Hill[2], Anne Moné[1], Geneviève Gagne[1], Harald Meier[3], Eric Peyretaillade[1] and Pierre Peyret[1,*]

[1]Génomique Intégrée des Interactions Microbiennes, Laboratoire de Biologie des Protistes, UMR CNRS 6023, Blaise Pascal University, 24 avenue des Landais, Campus des Cézeaux, [2]LIMOS UMR CNRS 6158, Blaise Pascal University, Clermont-Ferrand II, BP 10125, 63177 Aubière Cedex, France and [3]Lehrstuhl für Rechnertechnik und Rechnerorganisation, Institut für Informatik, Technische Universität München, Germany

Received on April 3, 2007; revised on July 18, 2007; accepted on July 27, 2007

Advance Access publication August 12, 2007

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Microbial diversity is still largely unknown in most environments, such as soils. In order to get access to this microbial 'black-box', the development of powerful tools such as microarrays are necessary. However, the reliability of this approach relies on probe efficiency, in particular sensitivity, specificity and explorative power, in order to obtain an image of the microbial communities that is close to reality.

Results: We propose a new probe design algorithm that is able to select microarray probes targeting SSU rRNA at any phylogenetic level. This original approach, implemented in a program called 'PhylArray', designs a combination of degenerate and non-degenerate probes for each target taxon. Comparative experimental evaluations indicate that probes designed with PhylArray yield a higher sensitivity and specificity than those designed by conventional approaches. Applying the combined PhyArray/GoArrays strategy helps to optimize the hybridization performance of short probes. Finally, hybridizations with environmental targets have shown that the use of the PhylArray strategy can draw attention to even previously unknown bacteria.

Availability: http://fc.isima.fr/~rimour/phylarray/

Contact: pierre.peyret@univ-bpclermont.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microorganisms are present in all environmental habitats, even the most extreme. Despite this extensive distribution, we have relatively little knowledge about these communities. Indeed, environmental studies are often limited by the difficulty to globally evaluate microbial populations and their dynamics in complex environments. For soils, Gans *et al.* (2005) have estimated that 1 g of surface soil might contain more than one million distinct genomes. Moreover, a minority of those microorganisms have now been cultivated and characterized (Dunbar *et al.*, 1999). This underestimation of bacterial diversity, caused by the cultural bias, has forced the development of more suitable investigation methods.

During the last decade, cultivation-independent molecular tools have been developed as an alternative in order to study environmental microbial communities more comprehensively (Amann *et al.*, 1995). These nucleic acid-based methodologies (PCR-based or hybridization methods) usually target the gene encoding, the small subunit ribosomal RNA (SSU rDNA). However, total nucleic acids from complex microbial environments are too rich in information to be easily analyzed by such molecular tools. Therefore, SSU rDNA oligonucleotides microarrays have been developed (Bodrossy and Sessitsch, 2004; Gentry *et al.*, 2006; Loy and Bodrossy, 2006). These high-throughput molecular tools are able to detect up to several thousands of microbial phylotypes simultaneously in a single experiment using species-specific probes immobilized on a solid surface.

The accuracy of such an approach, in terms of a comprehensive exploration of complex environments communities, relies on the efficiency of probes sets. They must be highly sensitive (Peplies *et al.*, 2006) and specifically recognize targeted groups even when the groups are present in low abundance (Gentry *et al.*, 2006). The majority of microarrays used for those studies are based on oligonucleotide probes, which present many design advantages (Ehrenreich, 2006).

Cross-hybridization is the major point that limits the determination of specific probes. In order to evaluate the specificity of a given probe, it is necessary to have a reliable predictor for its hybridization performance. However, the dynamics of probe-target hybridization in a microarray experimental context are very complex and not yet fully understood. Recent work has demonstrated that the use of thermodynamics parameters to assess probe specificity does not show good results (Pozhitkov *et al.*, 2006). In our work, we decided to use sequence similarity to check

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

probe specificity. A study made by Kane *et al*. (2000) on 50-mer oligos shows that a probe must satisfy two conditions to be specific:

(1) The oligonucleotide sequence must not have more than 75% of similarity (among all sequences) with a non-targeted sequence present in the hybridization pool.
(2) The oligonucleotide sequence must not include a stretch of identical sequence greater than 15 contiguous bases.

Several probe design programs generalize these criteria to oligos of any length (Rimour *et al*., 2005). In the remainder of this article, we use these criteria to check probe specificity. Values '75%' and '15 bases' are parameters of the algorithm and may be changed without modifying the bases of our method.

In previous work, we also have proposed a new approach to design oligonucleotides that combines both the specificity and the sensitivity (Rimour *et al*., 2005). In this strategy, named GoArrays, the oligonucleotide sequence is composed of two specific probe sequences (e.g. 25-mers) separated by a short random linker. As the oligonucleotide sequence is therefore quite long (e.g. 55-mers), it keeps the advantages of both short and long oligonucleotides.

Obtaining specific and sensitive probes is a big challenge, but the design of explorative oligonucleotides is the greatest one. As described previously, the majority of microorganisms are still unidentified, and not present in public ribosomal databases. Most classical oligonucleotide design software uses therefore incomplete data sets to generate species-specific probes. Thus, only a small fraction of known bacterial communities can be studied with these probes. However, a few design tools try to decrease this bias by allowing the selection of probes targeting higher bacterial taxa. ARB probe design tools (Ludwig *et al*., 2004) and PRIMROSE software (Ashelford *et al*., 2002) can generate this kind of taxon-specific probes. ARB is used in most of the biodiversity studies that use phylogenetic microarrays (Franke-Whittle *et al*., 2005; Loy *et al*., 2005; Sanguin *et al*., 2006). Schliep and Rahmann (2006) use a statistical group-testing approach with non-unique probes to detect targets related by a phylogenetic tree. Their method can detect unknown targets, but it has been validated only on simulations of hybridization experiments.

In this article, we describe a new algorithm, implemented in the program named PhylArray, allowing the generation of efficient probes. Our design strategy is based on the detection of high taxonomic groups and the use of a combination of degenerate and non-degenerate probes to globally monitor known and unknown bacterial communities.

2 PROBE DESIGN STRATEGY

Probe design for microarray experiments is not a trivial computational task. Parameters described previously have to be considered to obtain an efficient probe selection. Designing oligonucleotide probes for bacterial identification is basically the same problem as probe selection for classical gene expression experiments. The only difference is the specificity test. In gene expression experiments, a probe identifying a given gene must be specific among all other gene sequences of the studied organism. When designing oligonucleotides from SSU rRNA, each probe must be specific among all SSU sequences that may be present in the sample during the hybridization step. If the mixture composition is totally unknown, the specificity can only be checked against all known SSU sequences. SSU rRNA sequences can be obtained from various sources: major primary databases (GenBank/EMBL/DDBJ) or curated secondary databases (Cole *et al*., 2005; DeSantis *et al*., 2006; Ludwig *et al*., 2004; Wuyts *et al*., 2004). The Ribosomal Database Project II provides aligned and annotated rRNA gene sequences, along with analysis services. It represents a widely used and good-quality data source for bacterial identification (Cole *et al*., 2005).

Our aim is to develop a probe design algorithm for the selection of microarray oligonucleotides adapted to complex environments studies. The probes must be sensitive enough to detect all microbial community components, even in low abundance, and highly specific in order to recognize only the target groups. Moreover, as the majority of microorganisms from complex environments are still unidentified, we do not wish to use probes targeting only the known species, which would make the microarray unable to identify new species. Such constraints forced us to propose the following guidelines for developing our algorithm:

- The probes must target the genus as well as higher taxonomic units, in order to globally monitor known and unknown bacterial communities belonging to targeted taxonomic units. The polymorphism of the target group must be taken into account. To solve this problem, we decided to design a combination of degenerate and non-degenerate probes.
- The rRNA database must contain high-quality data in order to avoid cross-hybridizations that could be caused by sequences assigned to wrong species (false annotation). We decided to build our own secondary SSU rRNA database.

These points are discussed in the following paragraphs.

2.1 Targeted taxonomic groups and polymorphism questions

It is very hard to identify and differentiate species of the same genus with a specific oligonucleotide microarray strategy because the SSU rRNA variability within some species is too low. Considering Kane's conditions, a probe that targets a species might cross-hybridize with another species of the same genus. This is why our algorithm selects probes that target at least a genus, or a higher taxonomic group (family, order, etc.). In order to take into account the sequence polymorphism within the target group, the probe design software should generate a consensus sequence for the group, using the IUPAC nomenclature. In a microarray experimental context, the spotted probe then is a mixture of all possible sequences which can be generated from the consensus sequence. By following this strategy, some of these sequences targeted may not belong to real rRNAs (Fig. 1a) and could lead to supplementary cross-hybridizations. In our method, the result

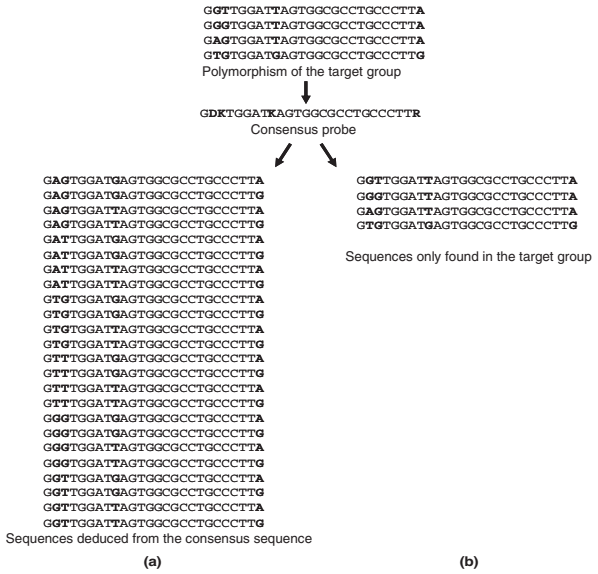

Fig. 1. Schematic representation of probe design to target a group of species. Sequences alignment allows the generation of the consensus sequence with degenerate positions. All sequences deduced from the consensus sequence (**a**) and real sequences used for the alignment step (**b**) are used in the search probe specificity.

given to the user contains the consensus sequence (degenerate oligonucleotide) and all the specific sequences derived from the single rRNA sequences that were used to build the consensus (Fig. 1b). Thus, either the consensus sequence (with all sequence combinations) or each 'real' sequence can be independently spotted on the microarray.

2.2 SSU rRNA database re-build

All SSU sequences of taxonomic division PRO (prokaryotic sequences) were obtained from the EMBL database. The latter were classified by organism (OC line in EMBL entry). This classification is made only for the purpose of pre-filtering and might be modified by the next step of the database build. Candidate sequences were rejected from further analysis if:

(1) The percentage of *N* (unknown base) in the sequence is >10%.

(2) The sequence contains a stretch of 10 consecutive *N*.

(3) The sequence is too short (lmax < 400, where lmax is the length of the longest sequence of the considered taxonomic unit).

The next step is the database curation. It aims to reject sequences that are assigned to the wrong organism, or which present chimeric anomalies. We use the K-means algorithm. For each taxon, the sequences are first partitioned in two clusters ($K = 2$). The distance between two sequences is derived from sequence similarity. When K-means has finished iterating for finding homogenous group of sequences, the well-annotated sequences are gathered in the same cluster. The other sequences are rejected.

3 DESIGN ALGORITHM

Input parameters for our algorithm are the name of the target taxon (T), the desired probe length (l), the specificity threshold (s) (Kane's criteria), the maximum number of degenerate bases in the oligonucleotide sequence (*xdeg*), and the sequence database used for the specificity test. The target taxon could be a group located at any level of the phylogenetic tree, for example *Micrococcus* (Genus) or *Micrococcaceae* (Family). The specificity threshold is used to determine if the probe may hybridize with a non-target sequence. Thus, the user can modify Kane's criteria described earlier in this article (75% similarity, 15 identical contiguous bases) if it is too restrictive in the experimental conditions. The used database must contain all sequences which could be present in the sample during the hybridization step. The default database used has been described in the previous section with all SSU rRNA from the PRO division of EMBL database filtered using our algorithm. The obtained secondary database is composed of 25 110 sequences belonging to 1900 genera.

The design algorithm consists of four consecutive steps:

(1) Extraction and filtering. All sequences of the taxon T are extracted from the chosen rRNA database. We use NCBI Taxonomy (Wheeler *et al.*, 2000) stored in a relational database to facilitate the selection of sequences at different taxonomic levels.

(2) Multiple sequence alignment. The sequences obtained in step (1) are aligned using the ClustalW algorithm (Thompson *et al.*, 1994).

(3) Search for a consensus sequence. A 'consensus sequence' using the IUPAC code is created from the alignment. The aim is not only to obtain a sequence that represents the group polymorphism, but also to remove possible sequencing errors. In each column of the alignment, the bases are replaced by a single consensus base. If numerous unspecified bases (*N*) or gaps (–) are observed in the alignment the following procedure is applied:

- If the number of '*N*' or '−' characters in a column is ≥ 50% of the total number of characters, a '−' is inserted in the consensus sequence.
- Otherwise, a consensus base is created, which corresponds to the bases present in the column. In this case, '*N*' and '−' characters are assumed to be sequencing errors.

Figure 1 in Supplementary material shows an example of this step.

(4) Search for specific probes. In the last step, specific probes are searched along the consensus sequence. The algorithm first tries to find a subsequence with less than *xdeg* degenerate bases (maximum number of degenerate bases specified by the user) incrementing a window of length l along the consensus sequence. Then, the program checks the specificity of this subsequence, which is the critical step of our algorithm.

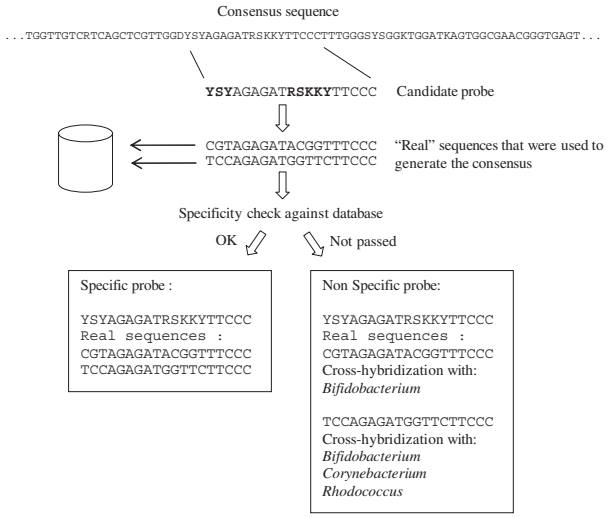

Fig. 2. Schematic illustration of the last step of the algorithm. A candidate probe is extracted from the consensus sequence. Only the sequences that were used to generate a consensus are tested for specificity. Even if the probes are not specific, the results are stored with possible cross-hybridizations.

In order to check the specificity of the potential probe, the program does not generate all base combinations from a consensus sequence, which would involve more checking than necessary. It gets only the sequences that were used to generate the consensus and checks their specificity. This process is illustrated in Figure 2. If the tested sequences are specific, the probe is considered to be specific for the target taxon. The user result consists of the degenerate probe and the non-degenerate sequences that represent the taxon. Even if the probe is not specific, the results are stored including all targets of potential cross-hybridizations. The user can specify a maximum number of cross-hybridizations for the probes to be stored in the result file, so that he can list all possible probes or only the most specific ones.

The specificity of all possible sequences which can be generated from the consensus sequence is also tested, and the result can be added to the result file (program option). This gives information about the global specificity of the degenerate probe, including sequences that are not present in databases.

4 IMPLEMENTATION

The design algorithm is implemented in a program called PhylArray. The program is written in Perl, and it uses ClustalW for multiple sequence alignment and BLAST (with the following parameters: Word size $W = 7$, Low-complexity Filtering F = false, Expectation value $E = 1000$) for checking the specificity. A parallel implementation allows probe finding to be done in parallel on a computing cluster architecture. The parallelism is introduced in step (2) (multiple alignment) and step (4) (specificity checking) of the design algorithm. These are the most time-consuming tasks of the process. PhylArray has been tested on a cluster with 15 computing nodes: a 'master' computer (management node) and 14 worker nodes. Each node is equipped with two processors (Xeon 2.67 GHz with hyperthreading) and 2 GB of RAM.

For the parallelization of multiple sequence alignment, we used ClustalW-MPI (Li, 2003). It is a distributed and parallel implementation of ClustalW, which uses the MPI library (Message Passing Interface) and runs on parallel architectures (a computing cluster in our case).

The fourth step of our algorithm (search for a specific probe within the consensus sequence) has been parallelized using the fact that the specificity tests of the different probes extracted from the consensus sequence are totally independent. To partition the computation on p machines, we only need to split the consensus sequence in p parts. The databases used for specificity checking are sent to each computing node.

For the implementation on the computing cluster, we use OpenPBS (http://www.openpbs.org/) to submit the jobs to the computing nodes. It is a flexible batch queuing system, which is easy to use when submitting independent jobs to a computing cluster.

PhylArray is available via a Web interface at http://fc.isima.fr/~rimour/phylarray/. It is necessary to register to obtain a login and password. The user can access the PhylArray interface and submit jobs. This part of the application consists of a Web server (Apache) and is written in XHTML and PHP. Perl scripts are used to communicate with the master node of the cluster. All information concerning user management and submitted jobs are stored in a relational database (MySQL). Figure 2 in Supplementary Material, presents a schematic overview of the PhylArray architecture.

5 EXPERIMENTAL VALIDATION

In order to evaluate the efficiency of probes selected with PhylArray, we compared them to oligonucleotides generated by another design program (PRIMROSE) and retrieved from a probe database (ARB Probe Library), both in common use for the design of probes (Freitag *et al.*, 2005). The GoArrays strategy, mentioned in the Introduction section of this article, was used in order to increase the efficiency of short oligonucleotide probes. We focused on the design of oligonucleotides targeted to SSU rRNA of the following genera: *Staphylococcus*, *Micrococcus*, *Aeromonas*, *Pseudomonas*, *Streptomyces*, *Rhodococcus* and *Bifidobacterium*. Two hundred and sixty-five oligonucleotides targeting these groups were selected and spotted on a prototype microarray in order to evaluate them with nucleic acid extracts from pure bacterial cultures.

5.1 Probe selection

The selection of 10 probes was performed using PRIMROSE and the supplied SSU_Prok database including archaeal and bacterial sequences. As advised, 15–20 SSU rRNA sequences belonging to the targeted genus have been autonomously selected in the database. The probe length was set to 40 or 50 nt and the number of degenerate bases has been increased to 10 bases. Fourteen probes have been selected from the ARB Probe Library (20 bases long). PhylArray has been used to design 172 long probes (50-mers) and 34 short probes (25-mers). In order to increase the efficiency of small probes, the

GoArrays strategy has been used on short probes. Thirty-one long oligonucleotides (56-mers) were created by combining short probes proposed by PhylArray, and 12 oligonucleotides (46-mers) were designed by concatenating short probes selected from the ARB Probe Library. Oligonucleotide sequences and characteristics are available in the Supplementary Material (Tables 1–3).

5.2 Experimental procedure

For microarray production, oligonucleotides probes were synthesized with a 5′ amino linker modification and spotted on Corning® GAPS II Coated Slides by Eurogentec. Each oligonucleotide was spotted in triplicate. Total RNA was extracted from pure cultures of *Staphylococcus xylosus* (DSM20266), *Enterococcus faecalis* (DSM20478), *Micrococcus antarcticus* (JCM11467), *Micrococcus lylae* (DSM20315), *Nesterenkonia sandarakina* (DSM15664) and *Aeromonas species* (laboratory strain) using the RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. Total RNAs from a polluted soil were obtained using a modified protocol originally described by Fleming *et al*. (1998). Nucleic acid extraction is followed by a purification using phenol/chloroform and with the RNeasy Mini kit.

To obtain labeled targets, the SSU rRNA fraction (125 ng) was reverse transcribed (2 h at 42°C) in a 20 μl final volume using dNTPs from Invitrogen (0.25 mM), RNasin+ from Promega (1U), SuperScriptIII (100 U) and its associated buffer (1X) from Invitrogen, DTT (0.1 M) and 0.625 μM of the following primer (the bold part of the oligonucleotides allows the formation of a T7 promoter): R1406-T7: 5′-**TAATACGACTCACTATAGGTA CT**ACGGGCGGTGWGTRCAA-3′

The second strand was created using all the neo-synthesized heteroduplex, dNTPs (0.3 μM), Ribonuclease H (1 U), *Escherichia coli* DNA Polymerase I (20 U) and its associated buffer (1X), *E.coli* Ligase (5 U) (all these products are provided by Invitrogen) in 100 μl final volume. *In vitro* transcription reaction allows the incorporation of amino allyl UTPs (final concentration 5 mM) for the indirect labeling of the targets and was conducted using MEGAscript kit (Ambion) according to the manufacturer's instructions. Then, the labeling of the amplified SSU rRNA was done by coupling the amino-modified aRNAs to the fluorescent dye Cy3 or Cy5 (0.5 mM) by incubation of the aRNAs with the succinimidyl ester-derivitized reactive free dye (Cy3 or Cy5 mono-reactive NHS-ester) in a coupling buffer (1X) from Ambion. This reaction was performed in the obscurity for 1 h at room temperature and was stopped with the addition of 1.3 M of hydroxylamine.

Hybridizations were carried out in a 25 μl final volume (17 μl of DigEasy buffer from Boehringer, 500 ng of labeled aRNAs, 1.4 μg of salmon sperm DNA and 6 nM of doubly labeling -Cy5/Cy3- reference oligonucleotide) at 37°C for 2 h in a TrayMix^{S1} hybridization chamber (Biotray; http://www.biotray.fr/). After hybridization, the microarrays were washed two times at room temperature during 5 min with the following solutions (solution1: 0.2X SSC, 0.1% SDS; solution2: 0.2X SSC). The slides were then scanned on the Affymetrix 428 Array scanner to detect Cy3 and Cy5 fluorescence.

Raw data analysis was carried out using a tool of the TM4 software suite: TIGR Spotfinder 3.1.0 (Saeed *et al*., 2003). Spot segmentation was done with the Otsu method using 10–30 px for the searched diameter. A median for each spot triplicate was calculated both in Cy5 and Cy3.

5.3 Probe efficiency analysis

5.3.1. Probe sensitivity Figure 3 shows a section of the microarray with probes targeting *Staphylococcus* and the signals obtained after hybridization with labeled SSU rRNA extracted from the bacterial strain *S.xylosus*. These targets hybridize with all oligonucleotides targeting this genus (Fig. 3a). Nevertheless, significant differences in signal intensities can be observed according to the methods used for probes selection (Fig. 3c). Probes designed with PhylArray are more sensitive than PRIMROSE and ARB ones (except, as expected for highly degenerate oligonucleotides like SStaphd3 showing a degeneracy of 1536). For long probes, signals of PhylArray oligonucleotides [from 9520 to 51 379 FU (Fluorescence Unit)] are higher than the PRIMROSE ones (6423 FU).

This observation is the same for short PhylArray and ARB probes with respectively 7731–17 572 FU and 405–946 FU. Moreover, hybridization analysis also shows that the use of the GoArrays strategy on short probes on average trebles their signal (e.g. GoArbStaph increase the signal by 2.3–5.4-fold).

5.3.2 Probe specificity Microarray analysis (Fig. 4) shows that probes generated with PRIMROSE, the ARB Probe Library and PhylArray (50-mers probes) are not strictly specific. Actually, all PRIMROSE probes selected to target *Pseudomonas*, *Streptomyces* and *Rhodococcus* genera also recognize *S.xylosus* targets. The SSU rRNA of *S.xylosus* does also cross-react with all probes from the ARB Probe Library targeting *Micrococcus* and *Aeromonas* and 50% of probes designed PhylArray (50-mers) that target *Micrococcus* and *Aeromonas*. The most important rate of cross-hybridization with *S.xylosus* targets is obtained for PRIMROSE probes (78%).

Oligonucleotides selected with PhylArray (50-mers) and ARB have lower rates of false-positive hybridizations (respectively, 8 and 1.5%). Moreover, false-positive intensities of ARB probes are clearly weaker and disappeared when the GoArrays strategy was used. PhylArray (25-mers) associated or not with the GoArrays strategy allowed the generation of more specific probes.

Similar results have been obtained with targets extracted from others reference bacteria: *M.antarcticus*, *E.faecalis*, *M.lylae*, *A.species*. and *N.sandarakina* (results not shown). PRIMROSE probes always show highest cross-hybridization rates (57–90%). PhylArray (50-mers) oligonucleotides have lower cross-hybridization percentages (0.05–46%) but quite higher compared to ARB probes (0–42%) and PhylArray short ones (3–28%). Depending on the tested bacterial strain, PhylArray short probes and ARB probes are more specific.

Hybridization analyses have also shown that the use of the GoArrays strategy on short probes often allows the removal of false-positive signals (*A.species*, *N.sandarakina*, *M.antarcticus*, *M.lylae*).

Furthermore, detection lacks for *M.lylae* have been observed using PRIMROSE probes. Oligonucleotides from the ARB

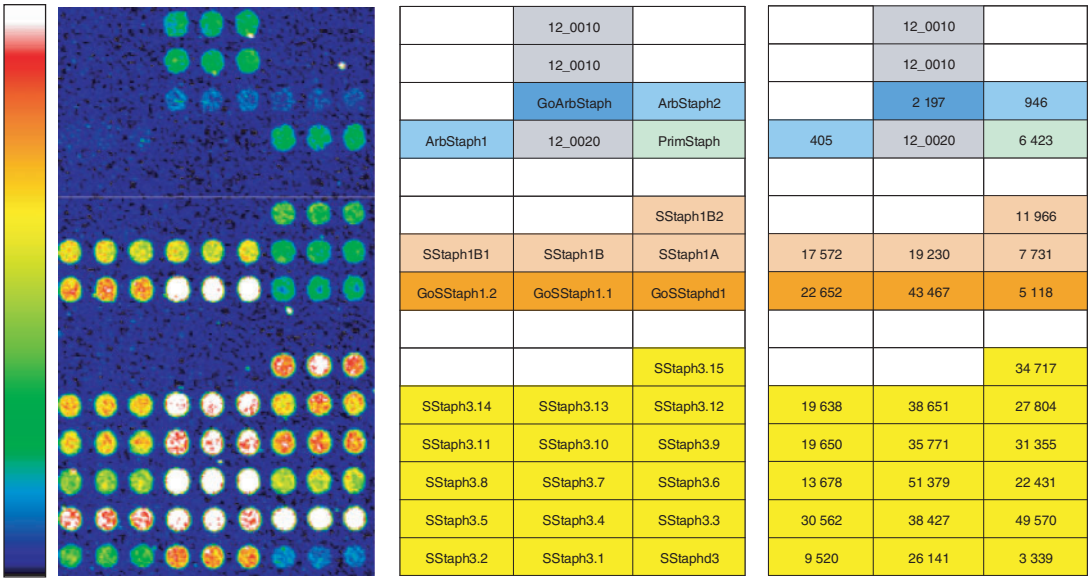

Fig. 3. Sensitivity of designed *Staphylococcus* probes. (**a**) Microarray image obtained after hybridization of labeled 16S rRNAs of *S.xylosus*. (**b**) Location of probes targeting *Staphylococcus* and control probes on the microarray. Probes were designed with the following software: PhylArray (50-mers: yellow and 25-mers: light orange), PRIMROSE (green), ARB probes (light blue) and the GoArrays strategy applied on Arb probes (dark blue) and PhylArray probes (dark orange). (**c**) Measurements of signal intensities (Fluorescence Units) with the TIGR Spotfinder program. 12_0010 is the positive control oligonucleotide (the complementary oligonucleotide, doubly labeled, is added to the hybridization mix). 12_0020 is the negative control oligonucleotide. All probes are spotted in triplicate.

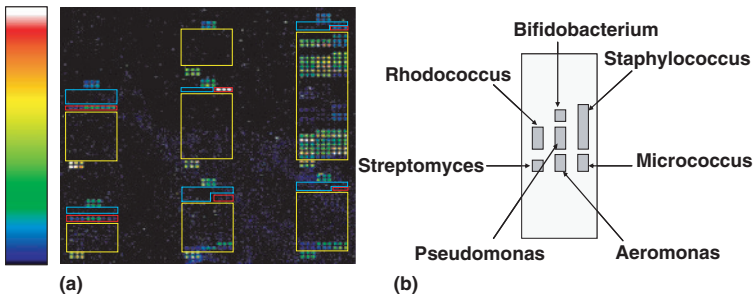

Fig. 4. Specificity evaluation of the probe design. Probes were designed with standard methods as PRIMROSE (red), ARB Probe Library (blue) and with a new method named PhylArray (yellow) (**a**). These oligonucleotides have been spotted on the microarray (**b**) and have been hybridized with 16S rRNA from *S.xylosus* labeled with Cy5.

probe library did not detect *M.lylae* and *A.species*. PhylArray is the only design method allowing a sufficient recognition level for all the tested strains.

5.3.3 Explorative power Figure 5 shows an element of the microarray image (*Aeromonas*-related part) obtained after hybridization with labeled SSU rRNA extracted from a polluted soil.

The fluorescence of the degenerate probe (SAerod2) is higher than non-degenerate probes fluorescence (SAero2.1, SAero2.2, SAero2.3 and SAero2.4). Thus unknown species belonging or close to the genus *Aeromonas* could be present in this complex environment.

6 DISCUSSION

In this article, we have shown that conventional approaches to design microarray probes do not always allow the generation of efficient biosensors for studying complex environments accurately.

The first critical point is the sensitivity of probes. In fact it is important to monitor all representatives of a bacterial community, even the low abundant ones, which also have an impact on the functionality of an ecosystem. In this study, the comparison of sensitivities between probes designed with PhylArray, PRIMROSE and those chosen in the ARB Probe Library has shown a better efficiency of longer probes. Fluorescence intensities of PhylArray probes are up to 20-fold higher than those obtained with PRIMROSE or selected from the ARB Probe Library (respectively compared to PhylArray 50- and 25-mers). It is well known that the presence and the positions of mismatches in the probe or target sequences can affect the signal intensity value (Urakawa *et al.*, 2003). In spite of the fact that ARB probes (20-mers) are smaller than PhylArray probes (25-and 50-mers), the sensitivity differences

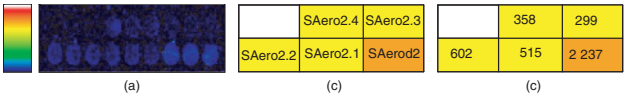

Fig. 5. Potential explorative power of PhylArray probes. Degenerate probes are designed by the PhylArray program allowing potential recognition of all species of the targeted genus, even the unknown fraction. (**a**) The microarray image shows hybridization of *Aeromonas*-targeting probes with bacterial targets extracted from a polluted soil. (**b**) Location of the degenerate probe SAerod2 and the specific probes SAero2.1–2.4. (**c**) Signal intensities (Fluorescence Units) are determined with TIGR Spotfinder 3.1.

could also be somewhat explained by the localization of the hybridization sites on the targeted molecule (SSU rRNA). Fuchs *et al.* (1998) have demonstrated that probe accessibility is variable according to the binding region. If we localize probe hybridization sites on the drawn SSU rRNA accessibility map, we can observe that binding regions of PhylArray probes (SStaph3.x; position 50, SStaph1.x; position 650) could be more accessible compared to ARB (ArbStaph; positions 100 and 200) and PRIMROSE (PrimStaph; position 300) probes. Other parameters could influence the hybridization efficiency. Several investigations have demonstrated the potential implication of thermodynamic properties of nucleic acids in the target-probe duplex formation and dissociation (e.g. secondary structures of SSU rRNA, intra- and inter-self structure of probes) that could be used to predict the probe efficiency (Gentry *et al.*, 2006). However, Pozhitkov *et al.* (2006) have recently shown that these thermodynamic parameters are only weakly correlated with probe efficiency. Our results demonstrated that probe sensitivity is influenced by the probe length, the perfect match between probes and targets, the location of the mismatches (Table 4 in Supplementary Material), the accessibility of the target and other complex thermodynamic parameters not yet well understood.

The second critical point is the specificity. The design of probes only recognizing a defined taxon is a real challenge due to the high conservation of the SSU rDNA biomarker within members of the bacterial domain and the presence of unknown bacteria in the studied samples (Gentry *et al.*, 2006). We have observed specificity differences between different long probes (PRIMROSE versus PhylArray 50-mers) as well as between short ones (ARB Probes Library versus PhylArray 25-mers). Long oligonucleotide probes designed with PRIMROSE show a significantly high cross-hybridization rate, which is based on their high complementarity to non-targeted bacterial sequences. The sequences targeted by the PRIMROSE probes specific for *Streptomyces*, *Pseudomonas* and *Rhodococcus* and the sequence of *S.xylosus* SSU rRNA have a high similarity of up to 94%. This value exceeds by far Kane's criteria (Kane *et al.*, 2000) and is probably responsible for the observed cross-hybridization with *S.xylosus*. PhylArray 50-mer probes targeting *Streptomyces*, *Pseudomonas* and *Rhodococcus* are more specific because their target sequences are dissimilar enough to not be able to bind to *S.xylosus* SSU rRNA. The reason for not finding these specific regions with PRIMROSE could be in its degeneracy setting limitation of 10 degenerate nucleotides per probe. This constraint restricts the search for probes to more conserved regions which could create cross-hybridizations with others taxons. PhylArray identifies probes in less conserved regions that are more specific for a given taxon. The limitation of the degeneracy rate in PRIMROSE has been certainly done in order to reduce computing constraints. In fact, PRIMROSE generates all the combinations of oligonucleotides from the consensus sequence and checks the specificity of each one. Thus, this process is time-and space-consuming if the consensus sequence is highly degenerated. The originality of our method relies on the specificity test of PhylArray probes which avoids problems due to a high degeneracy. In order to perform this test, the algorithm does not generate all possible sequences from the degenerate probe but only those used to create the consensus. Thus, the set of non-degenerate probes we obtain is highly specific for the target taxon. Another restrictive issue is the homology threshold used to define a potential cross-hybridization with a non-targeted sequence. The PRIMROSE user can modify this setting while varying the number of tolerated mismatches (up to 7 mismatches). However, for long probes (50-mers), this threshold corresponding to 86% of similarity, is insufficient because cross-hybridization events can occur even at 75% as described by Kane (Kane *et al.*, 2000). PhylArray uses by default Kane's criteria: 75% similarity (1) and 15 identical contiguous bases (2), but if it is too restrictive the user can set individual parameter values. Compared to PRIMROSE and PhylArray long probes, short oligonucleotide probes designed with PhylArray or selected from the ARB Probe Library have shown smaller cross-hybridization rates (respectively, 3–28% and 0–42%). These results are in accordance with Kane's criteria: the longer the oligonucleotide, the easier condition (1) is satisfied. However, to satisfy condition (2) the situation is more complex since as the length increases, the probability of finding a stretch of 15 identical bases might also increase. These observed cross-hybridizations are also explained by their similarities with non-targeted bacterial sequences. These false-positive signals seem to be unavoidable in a context of probes designed to target the SSU rRNA of closely related bacteria. However, we have shown a successful alternative by applying the GoArrays strategy. By concatenating suitable short oligonucleotide probes, we could minimize or even avoid false-positive signals in most cases. PhylArray 25-mer probes as well as probes selected from the ARB Probe Library showed high specificity. This could be explained by greater destabilization of non-perfect probe-target duplexes due to the constraint of loop formation on the target during the hybridization with the probe. It is furthermore important to mention that the signal intensities increased significantly using this approach.

It is of course important for probes to avoid cross-hybridizations but it is also essential to recognize the target taxon. Hybridization analyses have highlighted a recognition issue for probes designed by PRIMROSE or selected from the ARB Probe Library in order to target *Micrococcus*. These oligonucleotides did not detect strains of *M.lylae* and *M.antarcticus*. This lack could be due to the difficulty to design probes targeting higher phylogenetic levels than the species. This kind of design implies that the polymorphism of the targeted group is taken into account in order to recognize all its components. For polymorph taxa (e.g. *Micrococcus*),

it is necessary to use highly degenerate oligonucleotides to cover all the group diversity. The ARB Probe Library and PRIMROSE do not allow the design of such probes because only 0 and 10 degenerate bases are permissible, respectively. On the contrary, PhylArray allows the generation of highly degenerate probes in order to cover the polymorphism of all taxa. This could be the reason why no recognition problem has been detected for PhylArray probes.

The last critical point is the explorative challenge. Available design programs only generate probes for the known microbial fraction. It does not allow a global view of microbial communities and is mostly insufficient for comprehensive biological interpretations. Another specificity of our algorithm is to ensure an explorative process in the study of bacterial communities. The combined application of degenerate and non-degenerate probes can highlight the presence of bacteria which are not referenced in sequence databases if the spot composed of degenerate probes is the only one that give a hybridization signal. Hybridizations with environmental targets have suggested this potentiality (see results for SAerod2, SAero2.1, SAero2.2, SAero2.3 and SAero2.4). Unknown species belonging or close to the genus *Aeromonas* could be present in this complex environment. Biological experiments will help us to characterize these strains in order to validate this potential explorative power.

7 CONCLUSION

In summary, we present here a new probe design software tool called PhylArray, used to generate oligonucleotide probes for microarray analysis, allowing the targeting of the genus or higher taxonomic levels. PhylArray produces specific and sensitive probes which cover the polymorphism of the targeted taxon owing to the use of a high level of degeneracy. Moreover, the combined application of highly degenerate probes and associated non-degenerate probes even allows the exploration of the unknown part of bacterial communities. This exciting possibility could help us to create a better understanding of how microbial communities are functioning.

ACKNOWLEDGEMENT

We are grateful to Dr. Wyatt Paul for reviewing the English version of the manuscript. This work was financially supported by the CNRS, the FEDER, the ACI Microbiologie, the ACI Non Pollution-Dépollution Oxygen project, the ANR RSAmaturation, the ANR ECCO Metanox project, the regional council of Auvergne (France) in the INSTRUIRE, PREVOIR and PRAI e-nnovergne LifeGrid projects. C.M. was supported by a grant from 'Ministère de l'éducation, de la recherche et de la technologie' and M.M., S.R. and C.B by grants from ''Auvergne Council''.

Conflict of Interest: none declared.

REFERENCES

Amann,R.I. *et al.* (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, **59**, 143–169.

Ashelford,K.E. *et al.* (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res.*, **30**, 3481–3489.

Bodrossy,L. and Sessitsch,A. (2004) Oligonucleotide microarrays in microbial diagnostics. *Curr. Opin. Microbiol.*, **7**, 245–254.

Cole,J.R. *et al.* (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.*, **33**, D294–D296.

DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.

Dunbar,J. *et al.* (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Appl. Environ. Microbiol.*, **65**, 1662–1669.

Ehrenreich,A. (2006) DNA microarray technology for the microbiologist: an overview. *Appl. Microbiol. Biotechnol.*, **73**, 255–273.

Fleming,J.T. *et al.* (1998) Optimization of differential display of prokaryotic mRNA: application to pure culture and soil microcosms. *Appl. Environ. Microbiol.*, **64**, 3698–3706.

Franke-Whittle,I.H. *et al.* (2005) Design and application of an oligonucleotide microarray for the investigation of compost microbial communities. *J. Microbiol. Methods*, **62**, 37–56.

Freitag,T.E. *et al.* (2005) Influence of inorganic nitrogen management regime on the diversity of nitrite-oxidizing bacteria in agricultural grassland soils. *Appl. Environ. Microbiol.*, **71**, 8323–8334.

Fuchs,B.M. *et al.* (1998) Flow cytometric analysis of the in situ accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **64**, 4973–4982.

Gans,J. *et al.* (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, **309**, 1387–1390.

Gentry,T.J. *et al.* (2006) Microarray applications in microbial ecology research. *Microb. Ecol.*, **52**, 159–175.

Kane,M.D. *et al.* (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.

Li,K.B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics*, **19**, 1585–1586.

Loy,A. and Bodrossy,L. (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clin. Chim. Acta*, **363**, 106–119.

Loy,A. *et al.* (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order ''Rhodocyclales''. *Appl. Environ. Microbiol.*, **71**, 1373–1386.

Ludwig,W. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.

Peplies,J. *et al.* (2006) A DNA microarray platform based on direct detection of rRNA for characterization of freshwater sediment-related prokaryotic communities. *Appl. Environ. Microbiol.*, **72**, 4829–4838.

Pozhitkov,A. *et al.* (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res.*, **34**, e66.

Rimour,S. *et al.* (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, **21**, 1094–1103.

Saeed,A.I. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Sanguin,H. *et al.* (2006) Development and validation of a prototype 16S rRNA-based taxonomic microarray for Alphaproteobacteria. *Environ. Microbiol.*, **8**, 289–307.

Schliep,A. and Rahmann,S. (2006) Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree. *Bioinformatics*, **22**, e424–e430.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Urakawa,H. *et al.* (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2848–2856.

Wheeler,D.L. *et al.* (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.

Wuyts,J. *et al.* (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.