HiSpOD: probe design for functional DNA microarrays.

Eric Dugat-Bony, Mohieddine Missaoui, Eric Peyretaillade, Corinne Biderre-Petit, Ourdia Bouzid, Christophe Gouinaud, David R.C. Hill, Pierre Peyret

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 5 2011, pages 641–648
doi:10.1093/bioinformatics/btq712

Gene expression

Advance Access publication January 6, 2011

HiSpOD: probe design for functional DNA microarrays

Eric Dugat-Bony[1,2,†], Mohieddine Missaoui[3,4,†], Eric Peyretaillade[2,5], Corinne Biderre-Petit[1,2], Ourdia Bouzid[1,2], Christophe Gouinaud[3,4], David Hill[3,4] and Pierre Peyret[2,5,*]

[1]Clermont Université, Université Blaise Pascal, Laboratoire Microorganismes: Génome et Environnement, BP 10448, F63000, Clermont-Ferrand, [2]UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, [3]Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, [4]UMR CNRS 6158, LIMOS, F-63173 Aubière and [5]Clermont Université, Université d'Auvergne, Laboratoire Microorganismes: Génome et Environnement, BP 10448, F63000, Clermont-Ferrand, France

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The use of DNA microarrays allows the monitoring of the extreme microbial diversity encountered in complex samples like environmental ones as well as that of their functional capacities. However, no probe design software currently available is adapted to easily design efficient and explorative probes for functional gene arrays.

Results: We present a new efficient functional microarray probe design algorithm called HiSpOD (High Specific Oligo Design). This uses individual nucleic sequences or consensus sequences produced by multiple alignments to design highly specific probes. Indeed, to bypass crucial problem of cross-hybridizations, probe specificity is assessed by similarity search against a large formatted database dedicated to microbial communities containing about 10 million coding sequences (CDS). For experimental validation, a microarray targeting genes encoding enzymes involved in chlorinated solvent biodegradation was built. The results obtained from a contaminated environmental sample proved the specificity and the sensitivity of probes designed with the HiSpOD program.

Availability: http://fc.isima.fr/~g2im/hispod/.

Contact: pierre.peyret@univ-bpclermont.fr

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

Received on September 27, 2010; revised on November 23, 2010; accepted on December 15, 2010

1 INTRODUCTION

Microbes mediate almost every conceivable biological process on the planet. It is estimated that about 10^{30} bacterial and archaeal cells are involved in these tremendous metabolic potentialities (Whitman *et al.*, 1998). To access to this wide metabolic diversity, environmental functional gene arrays (FGAs) are high-throughput tools which allow analyzing the presence or the expression of thousands of genes encoding key enzymes involved in many metabolic process in only one experiment. However, designing DNA microarray probes to encompass the full diversity of gene family sequences encountered in nature and not yet identified is still one of the most difficult challenges. In addition, the difficulty to obtain sufficient nucleic acid materials for microarray studies from environmental samples requires the development of highly sensitive probes. Currently, the most complete FGA available to monitor bacterial metabolism in complex ecosystems is GeoChip (He *et al.*, 2007) with a latest version covering ~47 000 gene variants from more than 290 gene categories (He *et al.*, 2008). GeoChip has already demonstrated its power in assessing *in situ* bacterial activity, reinforcing the idea that functional microarray are an efficient approach (Berthrong *et al.*, 2009; Kimes *et al.*, 2010; Leigh *et al.*, 2007; Liang *et al.*, 2009; Mason *et al.*, 2008; Rhee *et al.*, 2004; Van Nostrand *et al.*, 2009; Wang *et al.*, 2009; Yergeau *et al.*, 2007). However, the main problem that must be faced for functional microarray applications is the complexity of the oligonucleotide design step which hinders their development.

Many parameters such as probe length, melting temperature (T_m), %GC, complexity or still potential cross-hybridizations evaluation (percent similarity and identity stretch between probe and non-targeted sequences) must be taken into account to ensure the selection of the oligonucleotides offering the best compromise between specificity and sensitivity (Kane *et al.*, 2000; Pozhitkov *et al.*, 2007). Fortunately, numerous bioinformatics tools, which estimate most of these parameters, have been developed to perform this crucial probe design step (Lemoine *et al.*, 2009). Currently, the most popular software for FGA probe design are OligoWiz (Wernersson and Nielsen, 2005), CommOligo (also named OligoStar) (Li *et al.*, 2005) or YODA (Nordberg, 2005). However, these softwares were developed for use on whole-genome data or on specific sets of genes (Gentry *et al.*, 2006; Lemoine *et al.*, 2009). So, cross-hybridization research is performed only against a reduced set of sequences (Li *et al.*, 2005; Wang and Seed, 2003; Wernersson and Nielsen, 2005). Thus, probe specificity must be checked against databases containing a very large set of sequences while staying within a reasonable time for the design.

A second critical obstacle when dealing with environmental studies is the difficulty to design probes able to cover all the gene variants encoding proteins having a same function (Habe and Omori, 2003). Although Hierarchical Probe Design (Chung *et al.*, 2005) and ProDesign (Feng and Tillier, 2007) softwares propose clustering

*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.

methods to identify conserved regions allowing a probe design for gene families, they are unable to perform the specificity tests. The single promising approach to address this problem is implemented in the PhylArray program (Militon *et al.*, 2007) which consists in the design of degenerate probes targeting multiple sequence variants. This strategy allows the design of probes targeting sequences never described in international databases and therefore considered as explorative probes. Unfortunately, this tool is only dedicated to probe design for phylogenetic microarrays.

Finally, despite the wide range of available programs, few provide access to oligonucleotide design through a web interface. This solution is the easiest one since it does not require any local software installation. Unfortunately, only six softwares offer this possibility and among them four (OligoFaktory, Osprey, PROBEmer and ROSO) allow the user to implement its own biological models for a tailor-made probe design.

Therefore, many improvements in the probe design technology are still necessary and the HiSpOD (High Specific Oligo Design) program was developed to meet these challenges. In addition to take into account the crucial parameters for the design of effective probes, we combine:

- the possibility to produce specific probes and also degenerate probes from consensus sequences, and thus with an explorative power, in order to detect a broad spectrum of variants from gene families;
- the use of a large formatted database dedicated to microbial communities [all coding DNA sequences (CDSs) from Prokaryotes (PRO), Fungi (FUN) and Environmental (ENV) taxonomic divisions of the EMBL databank] to check specificity of selected probes and limit cross-hybridizations;
- program parallelization in order to optimize the speed of the design;
- the oligonucleotide design access through a web interface.

This user-friendly bioinformatics tool is dedicated to probe design for studying microbial communities. It can also be applied for all research areas after adaptation of the database used to search for cross-hybridizations. The algorithm and its features are discussed and demonstrated with the analyses of a microarray built from gene sequences encoding enzymes involved in chlorinated solvent [tetrachloroethene (PCE) and trichloroethene (TCE)] biodegradation. The functional DNA microarrays developed in this study will be particularly useful to monitor *in situ* bacterial capacities on contaminated environments.

2 COMPLETE CDS DATABASE BUILDING

A new database was developed in order to assess the potential cross-hybridizations of probes designed with HiSpOD by using non-target sequences present in environmental samples. All annotated transcript sequences in all classes of EMBL databank release 97 of PRO, FUN and ENV taxonomic divisions (ftp.ebi.ac.uk/pub/databases/embl/release) were extracted using a Perl script named '*buildDatabase.pl*'. Briefly, the script downloads compressed files containing EMBL entries from the EMBL databank, extracts the CDS from each EMBL entry and their associated 5′ and 3′ untranslated regions (UTR) fixed by default as the 300 upstream and downstream bases to the CDS. Indeed, UTRs were considered to avoid cross-hybridizations because they are components of mRNA sequences. To obtain a suitable and curated database, a filtering step is carried out to ensure that sequences with bad quality were rejected if:

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

(2) At least one stretch of 10 consecutive *N* is detected.

The data extracted are then saved into a relational database implemented under Oracle and named EnvExBase. Finally, the whole database contains 9 129 323 sequences (about 10 GB). Proteins obtained from each original CDS were also saved in a separate table in order to allow rapid accession of their specific information like source organisms and biological functions of encoded proteins.

3 DESIGN ALGORITHM

HiSpOD software allows oligonucleotide probe design for FGAs in the context of microbial ecology but could also be adapted to various biological models. It needs an input file in FASTA format containing the target sequences. Two kinds of sequence are accepted by the program: specific nucleic acid sequence and degenerate nucleic acid sequence [consensus sequence resulting from multiple alignments and based on IUB/IUPAC code (International Union of Biochemistry and Molecular Biology/International Union of Pure and Applied Chemistry)]. The program workflow consists in five consecutive steps (Fig. 1):

(1) Oligonucleotide probe design. For each input sequence, the probe length is defined by the user. Along the sequence, the algorithm extracts all probes by incrementing the constant defined probe size in a window. First, in the case of degenerated sequences, all potential oligonucleotides generated with a maximal degeneracy higher than the threshold fixed by the user are removed. Then, all oligonucleotides or all combinations resulting from degenerate probes are filtered according to the melting temperature (minimum and maximum T_m determined by the formula (A) in which [Na^+] is fixed by default at 0.5 M) and the complexity stretch (maximum authorized number of identical consecutive bases).

$$Tm = 79.8 + 18.5 \times \log_{10}([Na^+]) + 58.4 \times \frac{yG+zC}{wA+xT+yG+zC} + 11.8 \times \frac{(yG+zC)^2}{(wA+xT+yG+zC)^2} - \frac{920}{wA+xT+yG+zC} \quad \text{(A)}$$

(2) Similarity search. Oligonucleotides which successfully pass the first step are tested for similarity search against all the sequences from EnvExBase database using BLASTN program (Altschul *et al.*, 1990). The BLAST program is tuned for this task as follows: the 'DUST' filter, used to mask low complexity regions from nucleic acid sequences, is inactive (−F F), the search strand is set to the positive one (−S 1), the expected-value (−e) is a parameter chosen by the user (1000 by default), the word size is set to the minimum length (−W 7) and the number of reported results is fixed to 500 (−b 500).

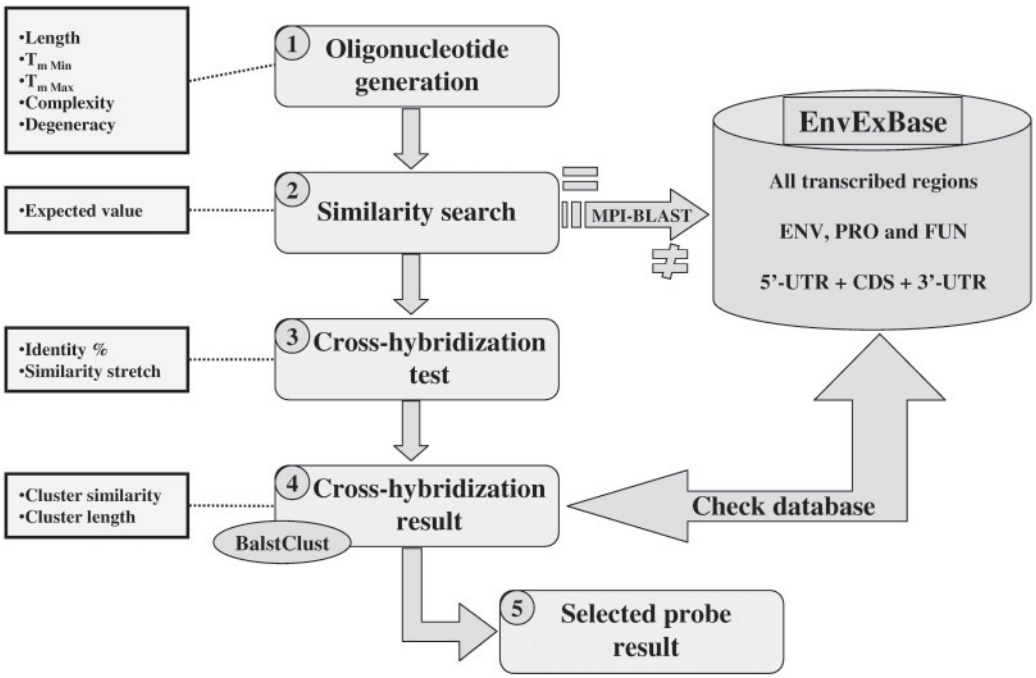

Fig. 1. HiSpOD design workflow. HiSpOD program is composed of five steps and each parameter chosen by the user are indicated on the left.

(3) Cross-hybridization test. Kane's criteria (Kane *et al.*, 2000), e.g. percent similarity and identity stretch, defined by the user (default values: 85 and 15, respectively), are checked for all positive results from previous BLASTN analysis to identify potential cross-hybridization.

(4) Cross-hybridization result. All sequences identified to potentially cross-hybridize with designed probes are clustered using BLASTClust tool provided by the BLAST program. The options (−S) which gives the minimum percentage of similarity between sequences that should be grouped together and (−L) which gives the minimum percentage of length of the implicated sequences in the alignment that should be considered, are also fixed by the user (default values: 90 and 0, respectively).

(5) Selected probe result. The final result provided by our software is stored within two output files: the first file with the extension '.probes' included all filtered oligonucleotides sorted by the number of cross-hybridizations. The position on the input sequence, the length and the sequence of the oligonucleotide are also mentioned. The second file '.result' details the non-targeted sequences that may induce potential cross-hybridizations with their accession numbers, source organism and function of the encoded protein.

4 IMPLEMENTATION

HiSpOD is written in Perl 5 scripting language and uses a BioPerl module for better visibility and to simplify the source code. A parallel implementation allows probe finding to be done in parallel on computing cluster architecture. The parallelism is introduced in Step (3) (cross-hybridization test) which is the most time-consuming task of the process. For this step, HiSpOD integrates a distributed and parallel implementation of the BLAST program (Darling *et al.*, 2003) which uses the MPI library (Message Passing Interface) and runs on parallel architectures.

HiSpOD software is installed on a symmetric multiprocessor (SMP) machine with 8 AMD Opteron Dual-Core processors at 1.8 GHz, 128 GB RAM and running Linux CentOS 64 bits. For large data storage, a SAN system disk with 2TB is used.

The application is freely accessible from a web interface developed in PHP/MySql at http://fc.isima.fr/~g2im/hispod/. The design results are made available on the server. HiSpOD is started via the interface using the chosen data entered by the user. The computing is made on the symetric multiprocessor (SMP) machine after a remote connection.

5 EXPERIMENTAL VALIDATION

The efficiency of HiSpOD was then assessed by the design of probes targeting genes encoding enzymes involved in the chlorinated solvent biodegradation.

5.1 Oligonucleotide probe set design

Probes were designed from 54 nucleic sequences corresponding to 21 genes coding for enzymes involved in chlorinated solvent biodegradation pathways (15 involved in anaerobic pathways and 6 in aerobic ones), which were downloaded from the NCBI database (Supplementary Table S1). Fourteen genes were treated, starting with individual nucleotide sequences (*amoB*, *todC1*, *todC2*, *bmoC*, *mmoC*, *xamoA*, *bvcA*, *prdA*, *rdhA1*, *rdhA2*, *pceA* from *Dehalococcoides ethenogenes* 195, *pceA* from *Sufluropirillum multivorans*, *pceA* from *Dehalobacter restrictus* PER-K23 and *pceA* from *Shewanella sediminis* HAW-EB3), 5 genes were treated at the gene family level (*pceA* from *Geobacter*, *rdhA1A*, *rdhA1B*, *rdhA2A*

Table 1. Input parameters for HiSpOD probe design

Probe characteristics				Probe specificity			
Probe length (mers)	T_m range (°C)	Single nucleotide stretch length (complexity)	Maximal degeneracy	Maximal consecutive match with non-target	Maximal identity with non-target (%)	BLASTClust option (−S) (%)	BLASTClust option (−L) (%)
50	64–79	10	1 or 32	15	75	90	0

(−*L*) = 0 indicates that sequence length is not considered for the clustering.

Table 2. Total number of selected probes and time consuming for the design for individual gene sequences and degenerate consensus sequences (gene family)

Data observed	Individual gene sequences	Degenerate consensus sequences
Design time (min)	3256	4014
Average design time per sequence (min)	203.5	573
No. of probes selected	140	54 (155[a])
Average No. of probes per sequence	9	8 (22[a])

[a]Number of probes after non-degenerate probe creation from each degenerate ones.
No, number.

and *rdhA2E*) and 2 were treated at both the individual and the family levels (*tceA* and *vcrA*). In order to target gene families, multiple sequences were aligned by using ClustalW program (Thompson *et al.*, 1994) to determine degenerate consensus sequences before probe design with HiSpOD. For each job launched on HiSpOD program (23 entries), probes of 50mers in length were designed. Table 1 summarizes the parameter values used for the probe design.

T_m range was determined to obtain a probe set with a GC content between 40 and 60%. Thus, ∼5 days (121 h) were necessary for the complete design with an average of 3.5 h for an individual sequence and of 9.5 h for a degenerate sequence. Based on the specificity (no or few cross-hybridizations) and the location of the probe along the target sequence, multiple probes were then selected per gene (Table 2 and Supplementary Fig. S1).

In total, 295 oligonucleotide probes were designed. These were used to elaborate a functional DNA microarray.

5.2 Experimental procedures

5.2.1 Microarray building Microarrays were produced with the NimbleGen high-density array synthesis technology (Roche NimbleGen, Madison, Wisconsin, USA). Each oligonucleotide was synthesized *in situ* in triplicate. Probes were randomly distributed across the array in order to minimize spatial effect. In addition, the microarray also reported 8863 random probes (20–56mers) used as a metric of technical background noise and 11 573 other control probes (positive and negative) which allowed quality control of oligonucleotide synthesis and hybridization conditions.

5.2.2 Synthetic antisense RNA target preparation and labelling The complete *mmoC* (S81887), *vcrA* (AY322364) and *tceA* (AF228507) gene sequences associated to SP6 promoter sequence at their 3′ end were synthesized by Biomatik Corporation (Cambridge, Ontario, Canada) and cloned into the pGH vector. Antisense RNAs (aRNAs) were produced by *in vitro* transcription using the MEGAscript SP6 Kit (Ambion, Austin, Texas, USA) and then labelled with Cy3-ULS as described in the Kreatech ULS labelling procedure (Kreatech Diagnostics, Amsterdam, The Netherlands). The aRNAs were quantified by spectrophotometry (at 260 nm) as well as dye incorporation (at 550 nm) with a Nanodrop ND-1000 (NanoDrop Technologies, Wilmington, North Carolina, USA) and RNA integrity was controlled with a Model 2100 Bioanalyzer using the RNA 6000 Nano kit (Agilent Technologies, Santa Clara, California, USA).

5.2.3 Environmental target preparation and labelling Total RNAs were extracted according a modified protocol (Vetriani *et al.*, 2003) from an industrial site contaminated with both tetrachloroethene (PCE) and TCE) and biostimulated during 31 months before sample collection by SITA Remediation Company (Lyon, France). In this protocol, freeze-thaw cycles were replaced by beat betting 1 min at 30 Hz with 1 g of glass beads of ≤106 μm in diameter (Sigma-Aldrich, Saint Louis, Missouri, USA). Co-extracted DNA was removed by digestion with 4U of DNase I (DNA-free, Ambion, Austin, Texas, USA) at 37°C for 35 min. Using the MicrobExpress kit (Ambion, Austin, Texas, USA), 8–9 μg of total RNA mixture were subjected to mRNA enrichment protocol. In a second step, antisense amino-allyl dUTP marked RNA (aRNA) was obtained by amplification with the MessageAmp II-Bacteria kit (Ambion, Austin, Texas, USA) and labelled with Cy3 fluorescent dye (GE Healthcare, Chalfont St Giles, UK) following the manufacturer's instructions. RNA integrity was controlled with a Model 2100 Bioanalyzer using the RNA 6000 Nano kit (Agilent Technologies, Santa Clara, California, USA).

5.2.4 Microarray hybridization A microarray was hybridized with 10^{11} copies of each labelled Cy3-aRNA obtained from synthetic genes and another one with 6 μg of Cy3-aRNA obtained from the contaminated groundwater. Targets were dried in a SpeedVac (Thermo Fisher Scientific, Villebon sur Yvette, France) and resuspended in the NimbleGen Hybridization Buffer (Roche NimbleGen, Madison, Wisconsin, USA). Then, targets were hybridized on the microarray at 42°C for 72 h using a 4-bay NimbleGen Hybridization system (Roche NimbleGen, Madison, Wisconsin, USA). Array washes were performed as

recommended by NimbleGen and slides were scanned at 2 μm resolution using the InnoScan® 900 and Mapix® software (Innopsys, Carbonne, France). Finally, pixel intensities were extracted using NimbleScan™ v2.5 software (Roche NimbleGen, Madison, Wisconsin, USA) and pair reports containing signal intensity data for every spot on the array linked to its corresponding probe identifier were generated.

5.2.5 Data normalization and statistical analysis The background noise was determined using random probes present on the microarrays with the method described in Supplementary Material 1 and was defined by two parameters: the background median intensity ($B_{position}$) and its dispersion ($B_{dispersion}$). Finally, a modified signal-to-noise ratio named SNR′ and based on the formula SNR′ = (probe signal intensity $-B_{position}$)/$B_{dispersion}$ was calculated in order to normalize our data. As suggested in the study of He and Zhou (2008), positive hybridization was considered significant for probes having a SNR′ > 5.

5.2.6 Gene isolation from the contaminated groundwater Genomic DNA (gDNA) was extracted from the industrial contaminated site using a modified protocol originally described by Vetriani *et al.* (2003). In this protocol, freeze-thaw cycles were replaced by beat betting 1 min at 30 Hz with 1 g of glass beads of ≤106 μm in diameter (Sigma-Aldrich, Saint Louis, Missouri, USA). gDNA concentrations were measured using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, North Carolina, USA). PCR reactions were carried out to isolate gene fragments for *vcrA*, *bvcA, pceA* and *todC1* (primer list and PCR conditions in Supplementary Table S2). Amplicons of correct size were then cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, California, USA) and sequenced by MWG DNA sequencing services (Ebersberg, Germany).

5.2.7 Quantitative RT-PCR In order to validate the semi-quantitative aspect of microarray results, qRT-PCR assays were performed. RT reactions were carried out in duplicate from 100 ng of total RNA extracted from the contaminated environment using *vcrA, bvcA* and *pceA* (0.625 μM of each primer) reverse primer mix (Supplementary Table S2). Control for DNA contamination evaluation was the same reaction without reverse transcriptase. Then, quantitative reactions were performed for each gene independently. For each of the two RT replicates, the quantitative reaction was achieved twice, thus leading to four measurements for each gene. For the gDNA contamination control, the quantitative reaction was completed four times. Finally, three replicates for each point of the standard curve (serially diluted cDNA) were measured. The reaction was performed in a final volume of 20 μl containing 5 μl of cDNA, 10 μl of 2X MESA Green qPCR for SYBR assay mixture (Eurogentec, Liege, Belgium) and the corresponding primer sets described in Supplementary Material (Supplementary Table S3) at 0.2 μM final concentration each. qPCR was realized in the Mastercycler Realplex (Eppendorf, Hamburg, Germany) starting with 5 min of denaturation at 95°C followed by 40 cycles consisting of denaturation at 95°C for 15 s, annealing at 59°C for 15 s and elongation at 68°C for 30 s. Data analysis was achieved with *realplex* software version 1.5 (Eppendorf, Hamburg, Germany). The quantification of transcripts in the RNA sample was determined after subtraction of the gene copy number obtained for the gDNA contamination control.

Table 3. Probe sensitivity and specificity assessment after microarray hybridization with three aRNA targets (*mmoC*, *vcrA* and *tceA*)

Targeted gene	No. of positive probes/No. of total probes	Average SNR'
mmoC	5/5	1522.5
vcrA	38/38	1624.5
tceA	9/9	1709.4
Others	0/243	< 5

5.2.8 Accession numbers Nucleotide sequences without PCR primer sequences were deposited in the GenBank database under accession numbers HM140426 and HM140427 for *vcrA*, HM140428 and HM140429 for *bvcA* and HM140430 and HM140431 for *pceA*. The microarray data discussed in this publication are available at the GEO web site (http://www.ncbi.nlm.nih.gov/geo/) under accession number: GSE21492.

5.3 Probe efficiency analysis

5.3.1 Probe sensitivity and specificity A first experiment was performed to test the ability of probes to reliably identify genes involved in chlorinated solvent degradation. Thus, three synthetic genes [*mmoC* (S81887), *vcrA* (AY322364) and *tceA* (AF228507)] were used to produce aRNA by *in vitro* transcription. Then, 10^{11} copies of each labelled aRNA were hybridized to the array. Whatever the gene tested, the results (GSE21492) showed that all corresponding probes gave a strong positive hybridization signal (Table 3), which is in agreement with a great sensitivity of probes designed with HiSpOD. These probes are also highly specific because no hybridization signals (SNR′ > 5) were obtained with probes targeting the other genes (21 genes are targeted by the microarray).

5.3.2 Industrial polluted site analysis Microarray hybridization with labelled antisense mRNA extracted from a PCE/TCE contaminated groundwater, allowed a full identification of expressed genes on the environmental sample. Four genes were detected (average intensity of all specific probes with SNR′ > 5): three encoded reductive dehalogenases involved in the anaerobic dehalorespiration pathway (*vcrA* and *bvcA* from *Dehalococcoides* and *pceA* from *Sulfurospirillum*) and one implicated in the aerobic TCE biodegradation pathway (*todC1* from *Pseudomonas*). Positive hybridization signals were obtained for all regions of the three genes involved in the anaerobic dehalorespiration pathway. In contrast, only one out of the two targeted regions for *todC1* was detected. This finding suggested that hybridization of the probes designed to target the 3′ extremity of *todC1* gene resulted from cross-hybridization and did not reflect the presence of transcripts on the contaminated site. In addition, PCR and sequencing experiments allowed the isolation of gene fragments for *vcrA* (HM140426 and HM140427), *bvcA* (HM140428 and HM140429) and *pceA* (HM140430 and HM140431) but not for *todC1*, confirming the results obtained with the DNA microarray.

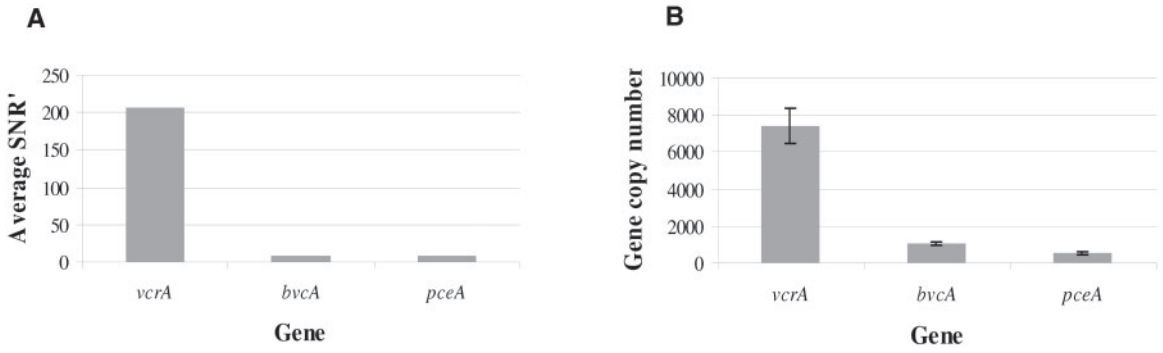

Fig. 2. Gene expression profile identified in a contaminated groundwater obtained with (**A**) microarray analysis and (**B**) qRT-PCR assays. For the three genes *vcrA*, *bvcA* and *pceA*, graphs indicate either the average ratio SNR′ obtained for all probes targeting the same gene (A) or the gene copy number per nanogram of total RNA (B).

5.3.3 Gene expression quantification SNR′ levels obtained for *vcrA* gene (average SNR′ = 205.6 for the individual gene design and average SNR′ = 220.6 for the gene family design) were much higher than those obtained for the last two genes *bvcA* and *pceA* (average SNR′ = 7.3 and 8.2, respectively), which could suggest a higher abundance of *vcrA* transcripts in the sample (Fig. 2A). To confirm these results, transcript quantification by qRT-PCR was performed on the same groundwater sample and the measurement of the expression level showed a higher number of transcripts for *vcrA* (7415 ± 949 copies/ng of RNA) than for both *bvcA* and *pceA* (1072 ± 86 and 559 ± 87 copies/ng of RNA, respectively) in the RNA sample (Fig. 2B).

These data support the applicability of our FGA for gene identification and as a semi-quantitative tool for gene expression evaluation with the advantage to survey numerous genes at the same time.

6 DISCUSSION

HiSpOD takes into account a greater number of criteria to select more efficient probes (Table 4) compared with most popular softwares used for oligonucleotide probe design for FGAs (Li *et al.*, 2005; Nordberg, 2005; Wernersson and Nielsen, 2005), The entire software suite is available through web services. To produce the DNA probes, users can enter their sequences of interest and set the required parameters.

Furthermore, our software can be applied to both, individual nucleotide gene sequences and degenerated consensus sequences, produced by multiple alignments of gene sequences belonging to a same gene family. Therefore, the user has the possibility to identify specific probes either at the gene or at the gene family level. For this last case, design methods previously reported are based on sequence clustering (Chung *et al.*, 2005; Feng and Tillier, 2007) or on the use of mismatch probes (Jaing *et al.*, 2008), which constrain the selection of probes in the most conserved regions. In contrast, HiSpOD allows considering regions generating the least number of cross-hybridizations even if they are more variable within the gene family. Because all sequence combinations for each degenerate probe are re-created for the DNA microarray synthesis, some of them could detect new gene sequences not yet available in international databases. Furthermore, as the number of probes per array is no longer a constraint, given the development of very high density microarrays (several million probes proposed by NimbleGen), designing microarray to encompass the full diversity of sequence variants encountered in the environment or in specific metabolic process is now feasible. By this aspect, the HiSpOD method is then a pioneer explorative approach to the design of probes dedicated to FGAs. Today, the only probe design software described to allow an exploratory study of environmental samples using degenerate probes is PhylArray (Militon *et al.*, 2007), but it is dedicated to phylogenetic oligonucleotide microarrays (POAs).

Cross-hybridizations with non-target sequences can be a significant contribution to the hybridization signal, potentially introducing substantial error. According to He *et al.* (2007), they represent one of the most critical limits of DNA microarray approach when dealing with complex environmental samples of unknown composition. However, over the past two decades, new sequencing technology developments and the trend of declining sequencing costs, allowed an exponential growth of public DNA sequence data. Data analysis has revealed an extraordinary gene diversity and a huge nucleic sequence polymorphism for all investigated environmental microbial habitats especially by metagenomics approach (Kottmann *et al.*, 2010). This extraordinary resource can now be mined for DNA microarrays applications to assess the potential cross-hybridizations. Currently, HiSpOD is the most suitable tool to take into account this wide diversity to limit cross-hybridations with a test of specificity conducted against EnvExBase, a large and representative database of all prokaryotic, fungal and environmental sequences. This database includes all known potential CDSs and their putative UTRs. However, as the HiSpOD algorithm extracts all probes by incrementing the constant defined probe size in a window along the sequence, the time required for evaluating their specificity can be quite long (several hours per gene) especially for highly degenerate oligonucleotide probes. So, HiSpOD is greatly dependent on the performance of MPI-BLAST program used in the specificity test step. Currently, we are focusing our efforts to increase the program performance by deploying it on a larger cluster infrastructure and on a computing grid. Given the exponential increase in the number of sequences in databases, the DNA microarrays will be upgraded regularly with novel probes allowing the targeting of either new variants of the targeted genes or genes newly discovered. In addition, EnvExBase component of the HiSpOD software will be updated regularly. An

Table 4. Comparison of parameters used by HiSpOD and by three popular softwares

Software	Probe length	T_m	GC%	Complexity	% Similarity	Identity stretch	Degeneracy	Cross-hybridization clustering	Web interface
OligoWiz	Yes	Yes	No	No	Yes	Yes	No	No	No
CommOligo	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
YODA	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
HiSpOD	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

improvement of the database structure, consisting in a split into multiple sub-databases according to particular ecosystems (soil, marine, freshwater, atmosphere, etc.), is also under development. It will provide an enhanced usability of the program, with a faster speed and an elimination of unneeded cross-hybridization information. Currently, no known software offers the opportunity to test for specificity on such a large database. Furthermore, no tested software (i.e. OligoWiz, CommOligo and Yoda) is able to integrate this database to perform the specificity test with it. This is probably due to an inadequate performance optimization of these algorithms which systematically crashed when using a large database. In addition, the clustering of cross-hybridization sequences using BLASTClust tool is another original approach proposed by our software to define gene sequence families that may induce potential cross-hybridizations and so to facilitate selection of the most appropriate probes.

Probe sensitivity is another crucial parameter, particularly for environmental studies, where biomass can be low. So, design process allowing the selection of oligonucleotide probes need to be optimized. Generally, microarray-based hybridization presents a low detection limit. For example, 50mer probe FGA detection limit was estimated to be in the range of 5–10 ng of gDNA in the absence of background DNA and was about 10-fold lower in its presence (Rhee *et al.*, 2004). In this last case, target sequences were extrapolated to represent genomic material from <5% of the total community of an environment (Gentry *et al.*, 2006). For mRNA, no detection limit is currently assessed in environmental sample. It remains hard to solve the sensitivity issue despite the existence of some protocols bringing those limits down (Gao *et al.*, 2007; Wu *et al.*, 2006). In this study, we demonstrated the ability of probes designed with HiSpOD to detect mRNA retrieved from environmental samples with a high sensitivity (e.g. 559 ± 87 copies/ng of RNA for *pceA*) (Fig. 2). Furthermore, expression profiles obtained were comparable with those of qRT-PCR assays suggesting the semi-quantitative aspect of the microarray elaborated using this software. We demonstrate that the microarray developed in this study provides a powerful tool to monitor chlorinated ethene biodegradation capabilities in complex environments. This tool will be now used for a wider study incorporating several industrial polluted sites where a more or less complete degradation of chlorinated solvents was observed.

However, certain limitations still remain. Indeed, numerous authors have highlighted the problem of signal variability obtained with probes targeting different regions of the same gene (Bruun *et al.*, 2007; Held *et al.*, 2006), certainly when linked with probe thermodynamic properties and/or target secondary structure. The function of these parameters is still difficult to evaluate and to visualize in the microarray results. Although their calculation is time consuming, they will be taken into account in the next version of HiSpOD.

In this study, the new software was shown to be particularly accurate to design probes dedicated to microbial ecology. The experimental validation of the oligonucleotide probe set designated using HiSpOD was performed with mRNA extracted from a PCE/TCE contaminated aquifer where pollutant biodegradation was observed (Site description and chemical analysis in Supplementary Material 2 and Table S4). These pollutants undergo biodegradation through two different pathways: (i) use as an electron acceptor (reductive dechlorination) (Futagami *et al.*, 2008) and (ii) co-metabolism where degradation of the chlorinated solvent is simply fortuitous and provides no benefit to the microorganism (Arp *et al.*, 2001), the first being the major process for the natural biodegradation of chlorinated solvents. In this study, the most recognized gene sequences involved in these two metabolic processes where extracted from international databases and were used for probe design. The microarray presented here is, to our knowledge the most complete high-throughput molecular tool describing these pathways with 295 probes covering 54 sequence variants of 21 genes and in addition with an explorative side. By comparison, GeoChip mentioned 35 probes targeting genes characteristics of PCE/TCE biodegradation pathways (He *et al.*, 2007). In the context of industrial site rehabilitation, it could be also applied as a diagnostic tool to detect microbial activities before, during and after bioremediation treatment. Moreover, this explorative high-throughput tool could be used to monitor the distribution of these catabolic pathways along non-contaminated ecosystems in order to better understand the origin of these fascinating microbial functions.

7 CONCLUSION

In summary, we present a novel probe design software called HiSpOD dedicated to FGA developments. HiSpOD allows probe design at both specific gene and gene family levels through an original approach based on degenerated probe determination. Moreover, specificity test was conducted against a large formatted database composed of all known CDS retrieved from the taxonomic divisions PRO, FUN and ENV of the EMBL databank in order to avoid cross-hybridization events. Finally, the user-friendly web interface developed simplifies greatly the user access to the program.

ACKNOWLEDGEMENTS

We are grateful to our undergraduate bioinformatics students Nicolas Parisot (IUT Génie Biologique option bio-infomatique, Aurillac) and Mathieu Roudel (Master Bioinformatique, Université Blaise Pascal, Clermont-Ferrand) for the development of statistical and annotation

scripts used in HiSpOD, Pascale Gouinaud for her assistance and Dr Biron David for reviewing the English version of the manuscript. We thank SITA Remediation society for sample collection.

Funding: 'Agence De l'Environnement et de la Maîtrise de l'Energie' (ADEME, France) (grant ID 2598); INSTRUIRE, PREVOIR and LifeGrid programs financed by the regional council of Auvergne (France), the FEDER European commission and the 'Centre national de la Recherche Scientifique' (CNRS, France); the grant ANR-07-ECOT-005-05 for the program PRECODD Evasol from 'Agence Nationale de la Recherche' (ANR, France).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arp,D.J. *et al*. (2001) Molecular and cellular fundamentals of aerobic cometabolism of trichloroethylene. *Biodegradation*, **12**, 81–103.

Berthrong,S.T. *et al*. (2009) Afforestation alters the soil functional gene composition and biogeochemical processes in South American grasslands. *Appl. Environ. Microbiol.*, **15**, 6240–6248.

Bruun,G.M. *et al*. (2007) Improving comparability between microarray probe signals by thermodynamic intensity correction. *Nucleic Acids Res.*, **35**, e48.

Chung,W.-H. *et al*. (2005) Design of long oligonucleotide probes for functional gene detection in a microbial community. *Bioinformatics*, **21**, 4092–4100.

Darling,A. *et al*. (2003) The design, implementation, and evaluation of mpiBLAST. In *4th International Conference on Linux Clusters: The HPC Revolution 2003.* San Jose, California.

Feng,S. and Tillier,E.R.M. (2007) A fast and flexible approach to oligonucleotide probe design for genomes and gene families. *Bioinformatics*, **23**, 1195–1202.

Futagami,T. *et al*. (2008) Biochemical and genetic bases of dehalorespiration. *Chem. Rec.*, **8**, 1–12.

Gao,H. *et al*. (2007) Microarray-Based analysis of microbial community RNAs by whole-community RNA amplification. *Appl. Environ. Microbiol.*, **73**, 563–571.

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

Habe,H. and Omori,T. (2003) Genetics of polycyclic aromatic hydrocarbon metabolism in diverse aerobic bacteria. *Biosci. Biotechnol. Biochem.*, **67**, 225–243.

He,Z. *et al*. (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J.*, **1**, 67–77.

He,Z. and Zhou,J. (2008) Empirical evaluation of a new method for calculating signal-to-noise ratio for microarray data analysis. *Appl. Environ. Microbiol.*, **74**, 2957–2966.

He,Z.L. *et al*. (2008) Development and application of functional gene arrays for microbial community analysis. *Trans. Nonferrous Met. Soc. China*, **18**, 1319–1327.

Held,G.A. (2006) Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Res.*, **34**, e70.

Jaing,C. *et al*. (2008) A functional gene array for detection of bacterial virulence elements. *PLoS ONE*, **3**, e2163.

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

Kimes,N.E. *et al*. (2010) Microbial functional structure of "Montastraea faveolata", an important Caribbean reef-building coral, differs between healthy and yellow-band diseased colonies. *Environ. Microbiol.*, **12**, 541–556.

Kottmann,R. *et al*. (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res.*, **38**, D391–D395.

Leigh,M.B. *et al*. (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *ISME J.*, **1**, 134–148.

Lemoine,S. *et al*. (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge, *Nucleic Acids Res.*, **37**, 1726–1739.

Li,X. *et al*. (2005) Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Res.*, **33**, 6114–6123.

Liang,Y. *et al*. (2009) Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field. *FEMS Microbiol. Ecol.*, **70**, 324–333.

Mason,O.U. *et al*. (2008) Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. *ISME J.*, **3**, 231–242.

Militon,C. *et al*. (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics*, **23**, 2550–2557.

Nordberg,E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.

Pozhitkov,A.E. *et al*. (2007) Oligonucleotide microarrays: widely applied poorly understood. *Brief. Funct. Genomic Proteomic*, **6**, 141–148.

Rhee,S.-K. *et al*. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **70**, 4303–4317.

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.

Van Nostrand,J.D. *et al*. (2009) GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. *Environ. Microbiol.*, **11**, 2611–2626.

Vetriani,C., *et al*. (2003) Fingerprinting microbial assemblages from the oxic/anoxic chemocline of the black sea. *Appl. Environ. Microbiol.*, **69**, 6481–6488.

Wang,F. *et al*. (2009) GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. *Proc. Natl Acad. Sci. USA*, **106**, 4840–4845.

Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.

Wernersson,R. and Nielsen,H.B. (2005) OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res.*, **33**, W611–W615.

Whitman,W.B. *et al*. (1998) Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA*, **95**, 6578–6583.

Wu,L. *et al*. (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl. Environ. Microbiol.*, **72**, 4931–4941.

Yergeau,E. *et al*. (2007) Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. *ISME J.*, **1**, 163–179.