Available online at www.sciencedirect.com
ScienceDirect
Journal of Microbiological Methods 70 (2007) 292–300
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Development of a statistically robust quantification method for microorganisms in mixtures using oligonucleotide microarrays

Alex E. Pozhitkov [a,b], Kyle D. Bailey [a], Peter A. Noble [a,*]

[a] Civil and Environmental Engineering, University of Washington, Seattle, 98195, WA, USA
[b] College of Marine Sciences, P.O. Box 7000, University of Southern Mississippi, Ocean Springs MS 39566, USA

Received 25 January 2007; received in revised form 21 April 2007; accepted 1 May 2007
Available online 10 May 2007

Abstract

High-density oligonucleotide arrays can be extremely useful for identifying and quantifying specific targets (i.e., ribosomal RNA of microorganisms) in mixtures. However, current array identification schemes are severely compromised by nonspecific hybridization, resulting in numerous false-positive and false-negative calls, they lack an adequate internal control for assessing the quality of identification, and are dependent on amplification of specific target sequences which introduce biases. We have developed a novel approach for the routine quantification and identification of metabolically active microorganisms in mixed samples. The advantage of our approach over conventional ones is that it avoids designing, optimizing, validating, and selecting oligonucleotide probes for arrays; also, nonspecific hybridization is no longer a problem. The basic principle of the approach is that a fluorescence pattern of a mixed sample is a superposition of the fluorescent patterns for each target. The superposition can be quantitatively deconvoluted in terms of concentrations of each microbe. We demonstrated the utility of our approach by extracting rRNA from three microorganisms, making test mixtures, labeling the rRNA, and hybridizing each test mixture to DNA oligonucleotide (20-mers, $n=346{,}608$) arrays. Comparison of known concentrations of individual targets in mixtures to those estimated by the solution revealed highly consistent results. The goodness-of-fit of the solution revealed that about 90% of the variability in the data could be explained. A new analytical approach for microbial identification and quantification has been presented in this report. Our findings demonstrate that including signal intensity values from all duplexes on the array, which are essentially nonspecific to the target organisms, significantly improved predictions of known microbial targets. To our knowledge, this is the first study to report this phenomenon. In addition, we demonstrate that the method is a self-sufficient analytical procedure since it provides statistical confidence of the quantification.

Keywords: Pattern recognition; Nonspecific hybridizations; Oligonucleotide arrays

1. Introduction

The identification and quantification of specific targets (e.g., nucleic acids of microorganisms) in complex target mixtures is important to many areas of biomedical and engineering science, including health care (e.g., oral cavity microbes and polymicrobial diseases), biological defense (e.g., microbial pathogens in food or transported substances), and environmental monitoring (e.g., biofilms in water distribution systems). High throughput technologies, such as DNA arrays, have significant potential for identifying microbes. Three platforms are currently used for microbial identification: synthetic membrane (Raskin et al., 1994a,b), planar glass (DeSantis et al., 2005; Wilson et al., 2002a, b; Loy et al., 2002; Small et al., 2001, Pozhitkov et al., 2006), and gel-pad arrays on glass slides (Urakawa et al., 2002; Pozhitkov et al., 2005a). All platforms share the common attribute that a sensor detects a signal from target sequences hybridized to immobilized oligonucleotide probes. The intensity of the signal provides a measure of the amount of bound nucleic acid in a sample. Hybridization occurs not only between specific (perfect match) probe–target pairs but also between nonspecific pairs containing mismatches. Therefore, the observed signal intensity from a single array spot might represent a combination of perfect match

* Corresponding author. Tel.: +1 206 685 7583.
E-mail addresses: Alexander.Pozhitkov@usm.edu (A.E. Pozhitkov), kreated@u.washington.edu (K.D. Bailey), panoble@washington.edu (P.A. Noble).

0167-7012/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2007.05.001

and nonspecific targets hybridized to the same probe (Zhang et al., 2005). This situation seriously compromises the quality of data generated from array experiments, affecting microbial identification in complex mixtures.

The most commonly used approach for identifying microorganisms using DNA arrays is to target variants of highly conserved genes, such as those encoding ribosomal RNA (rRNA) (Woese, 2000). In particular, rRNA molecules are ideal for microbial identification because they occur universally in all microorganisms and contain highly conserved as well as highly divergent regions, which can be used to infer phylogenetic relatedness.

Two groups of studies have advocated the implementation of array technology for identifying and/or quantifying microbes: those that use amplified genes (e.g., DeSantis et al., 2005; Wilson et al., 2002a; Loy et al., 2002), and those that use rRNA (e.g., Small et al., 2001; Adamczyk et al., 2003; Chandler et al., 2003). Array studies that have used amplified gene products for hybridization have demonstrated some success in identifying microorganisms in mixed samples (DeSantis et al., 2005; Wilson et al., 2002b; Palmer et al., 2006). Relative quantification of complex microbial populations has also been demonstrated in these samples, however it has been well established that PCR-based methods do not adequately reflect community composition (Suzuki and Giovannoni, 1996). PCR amplification incurs bias and artifacts in the product thus adding uncertainties to the quantity of target microorganisms in mixtures. Amplification of rRNA genes (as well as other conserved genes) is needed because they typically occur in low copy number (e.g., a microbial cell has up to fifteen copies of rRNA genes per genome; Acinas et al., 2004), else they would be difficult to detect. It is also important to note that DNA, being more chemically stable than RNA, is less likely to be degraded in samples, hence amplification products represent both live and dead cells.

The other group of array studies deals with rRNA molecules. Ribosomal RNA occurs at relatively high quantities in metabolically active cells, it varies with growth rate (Neidhardt and Magasanik, 1960; Rosset et al., 1966), and it is rapidly degraded upon cell death. There are fewer array studies dealing with identification and quantification of microbes using rRNA than those based on rRNA genes presumably due to technological difficulties and misconceptions on nucleic acid dissociation processes. Recently, systematic physicochemical studies of nonequilibrium thermal dissociation (gel-pad microarrays) revealed serious issues associated with the technology itself (Pozhitkov et al., 2005a,b, 2007, submitted for publication; Pozhitkov and Noble, 2007a,b) and with the previously-believed notion that dissociation curves are useful for distinguishing between specific and nonspecific hybridization in samples containing multiple microbial species (Pozhitkov et al., 2007). Quantification of rRNA microbial targets in mixtures of targets using microarrays also presents an even greater challenge than identification because the extent of nonspecific hybridization is not known in natural samples, and there has been no effective way of taking it into account.

The goal of our research was to develop a quantitative approach for the routine identification of metabolically active microorganisms in samples containing mixtures of microbial targets using high-density arrays. Our successful approach is based on solving complex fingerprint patterns in terms of individual reference patterns using an over-defined system of equations. The result of the solution is the relative contribution of each reference pattern (i.e., microorganism) to the sample that can be presented in terms of concentration or mass. The solution is determined by the least squares method (Lawson and Hanson, 1974) and the goodness-of-fit is assessed by the R^2 of the solution.

2. Materials and methods

2.1. Nucleic acid preparation, sequencing, and microarray protocols

Nucleic acid was isolated from the log phase cultures of *Porphyromonas gingivalis*, *Streptococcus mutans*, and *Streptococcus macedonicus*. Isolation was performed using the FastRNA BLUE kit (Q-BIOgene, Irvine, CA), following the manufacturers' instructions by bead beating (two times for 30 s each), phenol chloroform extraction, and isopropanol precipitation as previously described (Pozhitkov et al., 2005a). The quality of the total RNA was assessed using Bioanalyzer (Agilent Technologies).

Sequencing of the RNA was accomplished by reverse transcribing the RNA using Super Script kit (Invitrogen) using GM4 (5′-TACCTTGTTACGACT) and amplifying the resultant cDNA using GM3 (5′-AGAGTTTGATCMTGGC) and GM4 primers (Muyzer et al., 1995). Both strands were sequenced at the Genome Sequencing Center at the University of Washington.

Test mixtures were created by dispensing 30 μg of each individual rRNA sample (1 μg/μl) into one tube. The total volume (90 μl) of tube was then split to three tubes at approximately equal fractions.

Ribosomal RNA of individual samples and test mixtures were shipped to NimbleGen for processing. NimbleGen labeled 10 μg of rRNA with biotinylated ddATP via first strand cDNA synthesis, digested the cDNA, perform a terminal transferase end-label reaction, and then purify the biotinylated cDNA fragments using a microcon (Millipore) clean-up column. According to NimbleGen protocol information, 10 μg of rRNA yields approximately 8 μg of cDNA (on average). Four μg of the cDNA of each biotin-labeled cDNA target, or mixture of targets was hybridized with the array. Hybridization conditions were 45 °C, 1 M Na^+. After 16 to 24 h hybridization, the microarray was washed with non-stringent and stringent buffers and the images were recorded. Detection of fingerprints required a post-hybridization stain with the fluorolink cy3-streptavidin. Finally, NimbleGen sent us the microarray data to UW for bioinformatic analysis.

2.2. Analytical approach

The identification and quantification of microorganisms was accomplished according to Record of Invention UW Ref#7346D. The approach involved: (i) making a microarray with a probe set, (ii) obtaining a library of fingerprints for each microorganism, and (iii) developing a numerical solution that provides the relative abundances of each microbial target in a mixture (see Math supplement for explicit details).

The oligonucleotides (20-mers) synthesized on the microarray were based on rRNA sequences of known microorganisms (Table 1). These oligonucleotide probes were generated by tiling along each 16S rRNA sequence using a one base pair shift. The obtained list of probes was then vetted to remove redundancy, and replicated in sufficient number for statistical analysis. The resulting pool of oligonucleotide probes was then synthesized on the microarray surface.

A library of fingerprints was generated by individually hybridizing RNA from one target microbe to an array. By 'fingerprint', we mean a vector of signal intensity values for each probe on the microarray. The values in the vectors (profiles) will have the same order as every individual reference organism.

Hybridizing each mixture to an array also created a library of fingerprints of the mixtures. The fingerprint of the mixture is a superposition of the fingerprints of individual known microorganisms. Therefore, one can solve the complex fingerprint patterns in the mixtures in terms of the individual fingerprints with corresponding weights. The weights are proportional to the concentration of individual microbial targets.

In matrix form, the system can be re-written as:

$$\mathbf{N}\cdot\mathbf{X}=\mathbf{S}$$

where **N** is the matrix of the pre-recorded patterns arranged as column-vectors (design matrix); **X** is the sought vector of the weights; **S** is the vector of intensities of the sample pattern to be solved. This system can be solved analytically by a common method (Teukolsky et al., 1992).

$$\mathbf{X}=(\mathbf{N}^T\cdot\mathbf{N})^{-1}\cdot\mathbf{N}^T\cdot S.$$

To assess the goodness of fit, one will calculate the R^2, which is the squared correlation coefficient between $\mathbf{N}\cdot\mathbf{X}$ and **S** vectors, where **X** was determined as a solution of the system.

Uncertainties associated with each solution will be found in accordance with Teukolsky et al. (1992). First, one has to determine the unscaled covariance matrix based on the reference fingerprints (alone). The non-scaled covariance matrix for the vector of solutions (**X**) can be computed as follows:

$$C=(\mathbf{N}^T\cdot\mathbf{N})^{-1}.$$

Second, the matrix is then scaled by a factor (sf) based on the deviation between the fingerprint and the patterns found solutions as follows:

$$\mathrm{sf}=\frac{(\mathbf{NX}-\mathbf{S})^T\cdot(\mathbf{NX}-\mathbf{S})}{r-p}$$

where **X** is the solution of the mentioned above equation, r — number of rows in **N** and p — number of columns in **N** (Lawson and Hanson, 1974).

Diagonal elements of the scaled covariance matrix are in fact variances (squared standard deviations) of each solution in the vector **X**. Therefore, these diagonal elements can be used as a measure of an error associated with each solution. Values in the covariance matrix as well as solution are only meaningful if there is a good correlation between **S** and $\mathbf{N}\cdot\mathbf{X}$, where **X** is the computed solution. Correlating **S** and $\mathbf{N}\cdot\mathbf{X}$ determines how well the pattern of the unknown sample can be explained in terms of individual components. If the correlation is poor, the design matrix **N** is not adequate to the sample and the solution as well as covariance matrix will be meaningless.

Finally, the determined weights (elements of **X**) can be converted to concentrations or masses of individual organisms provided that total concentration or mass of the sample is measured beforehand. Mathcad (Math Soft Inc.) was used to solve numeric equations.

Table 1
16S rRNA sequences that were used to generate oligonucleotide probes on arrays

GI number	Microbial species	Number of perfect match probes
45491	*Propionibacterium freudenreichii*	1040
45659	*Peptococcus niger*	1374
173681	*Actinobacillus actinomycetemcomitans*	1337
174375	*Escherichia coli*	1522
175621	*Peptostreptococcus anaerobius*	1288
176044	*Streptococcus bovis*	1451
176047	*Streptococcus salivarius*	1430
294287	*Porphyromonas endodontalis*	1296
294288	*Porphyromonas gingivalis*	1343
294420	*Prevotella denticola*	1243
576603	*Staphylococcus aureus*	1481
829087	*Acinetobacter baumannii*	1440
853707	*Actinomyces odontolyticus*	1255
1834295	*Abiotrophia defectiva*	1272
2183315	*Streptococcus gordonii*	1493
2580430	*Brevundimonas diminuta*	1336
3712666	*Treponema denticola*	1300
4490387	*Fusobacterium nucleatum* subsp. *nucleatum*	1440
5578753	*Enterococcus faecalis*	1430
5578899	*Streptococcus mutans*	1493
10946530	*Bacteroides* cf. forsythus oral clone BU063	1465
15011532	*Butyrivibrio fibrisolvens*	1518
23477251	*Streptococcus macedonicus*	1377
23821283	*Ralstonia eutropha*	1440

GI number refers to the unique sequence number in GenBank.

3. Results

We developed a Microsoft Access database for the fingerprint signal intensities, which is publicly available at http://faculty.washington.edu/pozhit/default.htm. The database also contains queries that can be used to analyze the data.

3.1. Evaluation of mismatch (MM) probes as controls for nonspecific hybridization

We followed the approach of DeSantis et al. (2005) because they identified microorganisms in complex mixtures using high-density arrays. It is important to emphasize that we did not include the PCR amplification step, or, for that matter, deal with rRNA genes as done in their study. Rather, total rRNA targets

were used because we did not want to introduce PCR amplification bias into the microarray results. Their approach addresses the nonspecific hybridization issue by including a specifically designed probe that contains a mismatch in the middle (Wilson et al., 2002a,b), so we generated three different mismatch (MM) probes for each perfect match probe (PM), which served as controls for nonspecific hybridization. A tiling array of 20-mers was designed to target 16S rRNA molecules of 24 microbes (Table 1) common to the human oral cavity.

According to DeSantis et al. (2005), the PM–MM pairs were assigned to the organisms to perform identification. A PM duplex was required to pass the following criteria: (i) the intensity of the PM duplex had to be at least 2.5 times greater than that of the MM duplex, and (ii) the difference in intensity between a PM duplex and an MM duplex had to be 25 times greater than the noise of background probes. DeSantis et al. (2005) determined noise of background probes as those probes having the lowest 2% of all signal intensities. We did not use this measurement in our study because we had multiple random probes that served as controls in each array experiment, and we were not convinced that the arbitrary 2% threshold adequately reflected noise. We used the average intensity of these probes as background noise. If all three MM duplexes of the PM–MM pairs passed, the hybridization was considered to be positive.

When a single target *P. gingivalis* was hybridized to the array, multiple triplets of PM–MM pairs correctly passed the DeSantis et al. (2005) criteria for the *P. gingivalis* target. However, not only *P. gingivalis* specific probes passed the criteria, but also probes specific to other microorganisms. Table 2 shows the top-five microorganisms having an overwhelming number of probe pairs that passed the criteria. In other words, all microorganisms shown on the right side of Table 2 were falsely identified to be present in the sample using the DeSantis et al. (2005) approach. Note that these results are not based on mixtures but rather a single target. Similar results were obtained for the other two microbial targets when they were also individually hybridized to the microarray (data not shown). These results show that, although the DeSantis et al. (2005) approach might be suitable for identifying microbial targets when they are amplified by PCR (which increases the concentration of specific targets over nonspecific ones, reducing nonspecific signal), total RNA target was not suitable for this purpose.

3.2. Nonspecific hybridization

We examined the extent of nonspecific hybridization to help explain why the DeSantis et al. (2005) approach yielded numerous false-positive results. Table 3 shows that only 14 (212 out of 1493) to 19% (262 out of 1343) of the specific probe pairs passed the hybridization criteria, while 80 to 86% did not. This fact should indicate that the criteria are very strict and that nonspecifically hybridized probes should have a very low probability of passing the criteria. However, many probe pairs passed the criteria that were nonspecific to the target (Table 3). We were surprised at the extent of nonspecific probe pairs passing the criteria because only one target was hybridized to one array (not mixtures of targets). Apparently, the target hybridized to probes having various mismatches as well. Some of the MM probes yielded higher signal intensities to the target than those of corresponding PM probes (Fig. 1), which is consistent with a previous study (i.e., Naef and Magnasco, 2003). Although the PM and MM intensities display a general linear trend, the high variability of the points negates their predictive power (Fig. 2). These findings confound our ability to interpret microarray results because, at present, there is no acceptable way to accurately predict signal intensities of probes even when the extent of nonspecific hybridizations (as represented by MM duplexes) is known.

Table 2
Top five hits for the number of false-positive PM–MM pairs passing criteria for the identification of *Porphyromonas gingivalis*

Microbial species	Number of false-positive probe pairs
Ralstonia eutropha	174
Acinetobacter baumannii strains 1 and 2	173
Fusobacterium nucleatum subsp. *Nucleatum*	159
Burkholderia sp. oral strain C37KA	159
Fusobacterium naviforme	157

Table 3
Number of specific and nonspecific PM–MM probe pairs passing (+) DeSantis et al. (2005) hybridization criteria

	Specific probe pairs passing [a]	Total possible specific probe pairs [b]	Nonspecific probe pairs passing
P. gingivalis	262	1343	1528
S. macedonicus	213	1377	3294
S. mutans	212	1493	2871

[a] "specific" means that PM probe in the PM–MM pair was perfectly matched to the target; "nonspecific" means that the PM probe in the PM–MM pair was not matched to the target and the central 16 nucleotides of the PM probe did not match anywhere in the target.
[b] Derived from Table 1.

3.3. A new approach: pattern analysis

Given a set of three reference fingerprints and a fingerprint of the mixture (see Methods), a solution was found using least squares method. Since each probe on the microarray was replicated four times, we sought for the solution of the system of

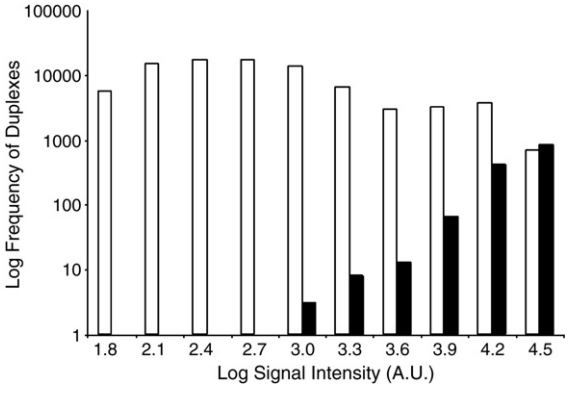

Fig. 1. Distribution of average signal intensity values (midpoints shown) for perfect match (black; $n=1343$) and all mismatched duplexes (white; $n=86{,}652$) when single target rRNA from *P. gingivalis* was hybridized to one microarray.

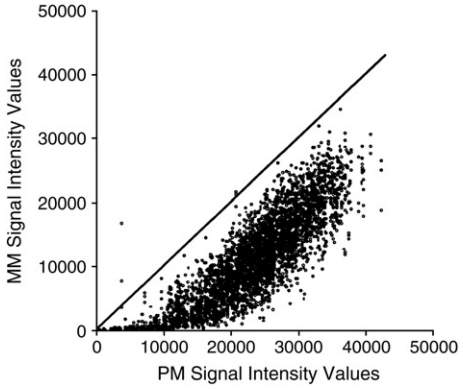

Fig. 2. Distribution of average signal intensity values for perfect match (PM) and designed mismatched (MM) duplexes using target rRNA from *P. gingivalis* hybridized to one microarray. The line represents SI(PM)=SI(MM).

equations in two ways: (i) by treating replicates as separate signal intensities, and (ii) by averaging four replicates and using the average as one signal intensity. The R^2 of the solution for the averaged sample was about 2 to 4% higher than that of treating replicates as individual cases, so we used averaged signal intensities for the analysis below.

The solution yielded high Pearson correlation coefficients and accounted for ~86 to 88% of the variability in the data based on R^2 (Table 4). Since the concentration of rRNA in the mixture is known from spectrophotometric readings, one can calculate the relative concentrations of each target in solution (Table 5). Note that each target microbe in the mixtures had a concentration that was close to 1.3 μg, and that these concentrations were more or less consistent in all three mixtures (Table 5). Subtle variations in concentrations of the mixtures are likely due to pipetting errors and/or noise in spectrophotometric measurements that occurred when preparing the original stock solutions.

To assess whether the prerecorded patterns are linearly independent of one another, we conducted *in silico* cross-validation tests. In order to discriminate microbial targets in mixtures, it is necessary that the patterns show linear independence. Cross-validation tests provide an indication of the linear independence of the patterns because one can compare the predicted outputs of targets (i.e., the mass of individual targets) as a function of each input pattern. Any overlap in the predicted outputs would indicate that the patterns were not independent. The results showed absolute discrimination (Table 6), without any apparent overlap. We were surprised at these findings because *S. macedonicus* and *S. mutans* are phylogenetically related (i.e., same genus *Streptococcus*), and therefore one would expect to have at least some overlap in their fingerprint patterns. However, the solution yielded perfect R^2-values, indicating absolute discrimination.

To investigate the robustness of the solution, we randomly shuffled the order of signal intensities of one of the reference patterns and added the resulting pattern to the reference list, simulating a fingerprint of a fourth target. The approach correctly identified the absence of the fourth target (Tables 7 and 8), without compromising the identification of the other targets as implied by the goodness-of-fit of the R^2. The small negative value of the shuffled sample indicates that the concentration was zero.

We also investigated the robustness of the solution by randomly shuffling the order of the signal intensity values in the pattern of the mixtures, imitating a pattern that is not comprised from the reference patterns, i.e. simulating a pattern of unknown microorganisms. The model did not converge to a solution as indicated by the R^2 of zero (data not shown). These results indicate that mixtures, containing significant amounts of targets whose patterns are not presented in the reference library, do not yield false-positive identifications.

3.4. Contribution of duplex sets to predicting the targets

In the previous experiments, signal intensities of all duplexes on the microarray were used to predict concentrations of known microbes in mixtures, regardless if they were perfect match duplexes, or duplexes with single or multiple mismatches. To determine if excluding some duplexes from the analysis would significantly improve predictions of target concentrations (as anticipated), we reanalyzed the data by: (i) excluding designed mismatch probes (MM), and (ii) excluding all probes that were not perfect matches to the three microorganisms in the mixtures. The R^2 of the solutions are presented in Table 9. Removing the

Table 4
Solutions and standard deviations of the solution for 1:1:1 mixtures

Target	Relative amounts by mixture (±SD)		
	#1	#2	#3
P. gingivalis	0.776±0.002	0.846±0.002	0.694±0.002
S. macedonicus	0.404±0.003	0.416±0.003	0.367±0.003
S. mutans	0.526±0.003	0.607±0.004	0.440±0.003
R^2	0.88	0.87	0.86
r	0.94	0.93	0.93

The goodness-of-fit of the solution is reflected in the R^2-values.

Table 5
Predicted quantities (μg) of specific targets in three 1:1:1 mixtures

Target	Amount by mixture		
	#1	#2	#3
P. gingivalis	1.8	1.8	1.8
S. macedonicus	0.9	0.9	1.0
S. mutans	1.2	1.3	1.2

Each mixture should contain approximately 4 μg of target.

Table 6
Cross-validation of model using individual reference patterns as inputs (4 μg)

	P. gingivalis	*S. macedonicus*	*S. mutans*
P. gingivalis	4	0[a]	0
S. macedonicus	0	4	0
S. mutans	0	0	4

The predicted quantity (output) of each target based on the model is shown by row. The R^2-value for all rows was 1.

[a] Values were below detection limits.

Table 7
Solutions and standard deviations for 1:1:1 mixtures

Target	Relative amounts by mixture (±SD)		
	#1	#2	#3
P. gingivalis	0.777±0.002	0.849±0.002	0.699±0.002
S. macedonicus	0.406±0.003	0.420±0.003	0.373±0.003
S. mutans	0.527±0.003	0.610±0.004	0.445±0.003
Shuffled *P. gingivalis*	−0.017±0.002	−0.038±0.002	−0.057±0.002
R^2	0.88	0.87	0.86
r	0.94	0.93	0.93

The goodness-of-fit of the solution is reflected in the R^2-values.

designed MM probes and thus decreasing the number of duplexes by 75% only marginally decreased the R^2. Subsequent removal of duplexes having multiple mismatches to the target organisms resulted in a sudden decrease in R^2 to approx. 0.53. To further investigate the reasons for such phenomenon, we studied the distribution of signal intensities within the three analyzed sets of duplexes. The frequencies of signal intensity values for the different data sets revealed two different distributions. Signal intensity values of all duplexes on the array, and all duplexes excluding those that are designed MMs, had a weakly bimodal distribution, while the distribution of perfect match duplexes (only) was skewed toward high signal intensities (Fig. 3). The observed distribution of all duplexes and all but designed MM duplexes indicates that nonspecific hybridization occurs frequently on the array, even at high intensities. Paradoxically, this nonspecific hybridization remarkably improves microbial identification.

Our findings demonstrate that including signal intensity values from all duplexes, i.e., PM and MM duplexes, significantly improved predictions of known microbes. To our knowledge, this is the first study to report this phenomenon.

4. Discussion

Almost two decades has past since oligonucleotide arrays were first purposed as a way to identify target sequences simultaneously (Dramanac et al., 1989; Southern, 1988). Yet, a robust, quantitative, analytical approach for identifying rRNA targets in mixtures has not been developed. The reason for this situation is a high level of nonspecific hybridization and a lack of adequate probe-design strategies to confront this problem. A combined molecular–analytical tool that provides reliable and robust identification of microorganisms is highly desired in the field of microbiology because current methods for identifying microbes, such as *in situ* hybridization (Amann, 1995) membrane dot blot (Raskin et al., 1994a,b), and quantitative PCR (Lim et al., 2001), are labor intensive, and not suitable for simultaneous identification of multiple microbes in a single experiment. This study addressed the problem of nonspecific hybridization, and describes a new approach that successfully quantified known microorganisms in mixtures of targets with statistical confidence.

Most array studies dealing with microbial identification are based on low-density (small-scale) arrays, which typically contain from ten to several hundred oligonucleotide probes attached to a solid support (see review of Loy and Bodrossy, 2006). The development and implementation of these arrays require a significant investment in terms of time and resources, and given that they contain few probes, they have limited statistical power for making identifications, particularly in the case of mixed samples. To date, no study has provided a statistically robust approach for microbial identification in mixed samples using low-density arrays.

High-density (large-scale) arrays offer significant advantages over low-density arrays because they contain hundreds of thousands of oligonucleotide probes making it possible to conduct rigorous statistical analysis. Moreover, probe design, optimization and evaluation steps may not be necessary, in contrast to low-density arrays, since one can create a tiling array, where probes overlap and cover the entire target sequence. In addition to this, no spotting and immobilization steps are required since the probes are directly synthesized on the array surface. As a consequence, array-to-array variability is very low

Table 8
Predicted quantities (μg) of specific targets including a shuffled artificial target in three mixtures

Target	Amount by mixture		
	#1	#2	#3
P. gingivalis	1.8	1.8	1.9
S. macedonicus	1.0	0.9	1.0
S. mutans	1.2	1.3	1.2
Shuffled *P. gingivalis*	0	−0.1	−0.2
R^2	0.88	0.87	0.86
r	0.94	0.93	0.93

Table 9
R^2-values of predicted quantities of specific targets using three different sets of probes and three mixtures

Mixture	R^2-values (*n*=number of duplexes)		
	All array (*n*=86652)	All array minus designed MMs (*n*=21948)	Perfect match probes (*n*=3296)
1	0.88	0.86	0.56
2	0.87	0.84	0.50
3	0.86	0.84	0.54

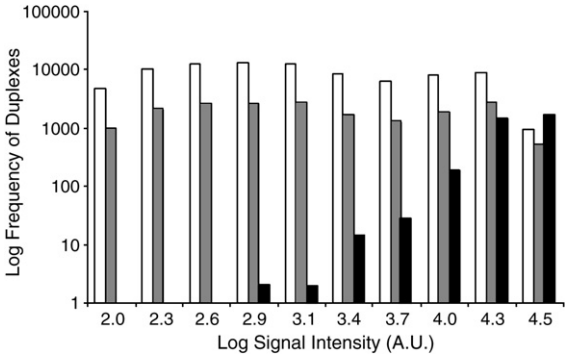

Fig. 3. Distribution of average signal intensity values (midpoints are shown) for all mismatch duplexes on the microarray (white; *n*=83,356), for all mismatch duplexes excluding those that are formed with designed mismatch probes (gray *n*=18,652), and for only perfectly match duplexes (black *n*=3296) for mixture 1 containing rRNA from three microbial targets.

for high-density arrays (R^2 range from 0.95 to 0.98 for NimbleGen arrays, this study), which drastically improves the fidelity of array readout over low-density arrays (R^2 range from 0.45 to 0.99 for spotted arrays, Pozhitkov, pers. comm.). Given the above mention advantages, we chose high-density arrays for our microbial identification and quantification study.

In contrast to previous array studies that used PCR methods to amplify rRNA genes, we used rRNA as the target for our experiments. The custom-designed high-density array contained tiled 20-mer oligonucleotide probes (in quadruplicate), representing 16S rRNA sequences of 24 oral cavity microorganisms. Given that DeSantis et al. (2005) have identified microbes in complex mixtures, we evaluated their approach for identifying microbes by hybridizing individual rRNA targets to one array. We anticipated that the approach should identify the hybridized target in each experiment.

4.1. Assessment of DeSantis et al. (2005) approach for identifying pure targets

DeSantis et al. (2005) reported a modification of a microbial identification scheme originally proposed by Wilson et al. (2002b) that involved *ad hoc* criteria determining occurrence of hybridization between the target and perfect-match (PM) probes. According to this scheme, identification occurs when all PM probes assigned to the operational taxonomic unit (OTU) of the microorganism in question, hybridize to the target, as inferred by signal intensity values. We applied this scheme by assigning PM–MM probe pairs to the organisms and requiring that the signal intensity criteria developed by DeSantis et al. (2005) to be satisfied. When total rRNA from *P. gingivalis* was hybridized to an array we found that the number of PM–MM pairs that satisfied the criteria was only 262 out of a possible 1343 specific pairs (Table 3, middle column). One could conclude from this result that the criteria were very strict since 80% of the specific pairs did not pass. According to Wilson et al. (2002b), the criteria can be 'loosened': not every probe-pair of the OTU was required to pass the criteria for the identification to take place. Yet, despite the strict criteria, many nonspecific PM–MM pairs to *P. gingivalis* passed the criteria (Table 3, right column), which resulted in false identification of unrelated microorganisms (Table 2). Further 'loosening' of the criteria increased the occurrence of false-positive identifications (data not shown).

4.2. MM probes as controls for nonspecific hybridization

Our lack of success with the DeSantis et al. (2005) approach might also be attributed to the designed MM probes on the array that do not adequately account for nonspecific hybridization. The idea behind using PM–MM pairs is to include specially designed mismatch (MM) probes in the probe set, that are identical to a perfect match (PM) probe except that they have a single internal mismatch to the target. It is assumed that the intensity of the MM probe provides a measure of nonspecific binding to the target. However, one could conclude from Fig. 2 that MM probes are not useful for assessment of the nonspecific hybridization since, as Wu and Irizarry (2004) pointed out, MM probes are detecting signal from specific as well as nonspecific binding.

4.3. New approach to identify and quantify microbes in mixtures

Nonspecific hybridization appears to be inherent to all microarray experiments (Fig. 1), and so far, there has not been proposed any effective method to avoid or account for this phenomenon. Bearing this in mind, we devised an alternative approach that utilizes nonspecific hybridization as a source of information and is similar to another (patent pending), which is based on pyrosequencing (Pozhitkov et al., 2005b). Briefly, in this approach, the entire set of probes and their corresponding signal intensities are viewed as a whole and considered to be a "fingerprint". The fingerprint of each organism that is suspected to be present in the mixture is recorded into a library of fingerprints (Fig. 4). A pattern of an unknown mixture is then solved in terms of relative contributions of each fingerprint from the library of fingerprint patterns, and the goodness-of-fit for the solution was calculated (i.e., R^2). These contributions can be easily converted into mass or concentration units when the total amount of rRNA in the mixture is known.

In this study, approximately 4 μg of labeled target was hybridized to the arrays to make fingerprints. Clearly, smaller amounts could have been used. For example, we have successfully obtained highly reproducible fingerprints using 3 ng of target (1 nM) with high-density Febit (Geniom) arrays (not shown), which is similar to NimbleGen arrays except having fewer probes. Inter-array variation in signal intensities for 8 arrays hybridized to the same rRNA target (n=3160 probes, in duplicate) ranged from 2 to 8% (average=4%; data not shown). Studies aimed at resolving the limitations of our approach in terms of determining a feasible detection limit with complexity of mixtures of targets are currently ongoing.

Possible applications of the approach include identification and quantification of harmful microorganisms (pathogens) in cases where microbial populations of 'normal' samples are more or less consistent, such as food produce and health-related samples (e.g., yogurt, pasteurized milk, blood, or urine). The

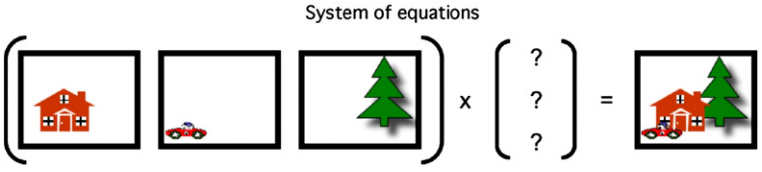

Fig. 4. Cartoon showing the system of equations, the unknown concentrations (?), and the mixture.

fingerprint of the 'normal' sample must be recorded and placed into the library of prerecorded patterns of the pathogens. Using our approach with the library of patterns, one can identify and quantify possible contamination of the sample under investigation. Again, the R^2 of the solution will be a final criterion for quality of identification, preventing the occurrence of false-negative and false-positive results.

This study provides proof-of-principle of a new approach. We recognize that additional experiments are needed to fully validate the approach within the context of more complex mixed samples and also environmental samples. Given that array sensors typically can detect signal in the range of four orders of magnitude, it is reasonable to suggest that we should be able to detect microbial targets on the scale of at least three orders of magnitude as done in the study of Palmer et al. (2006). In the case of environmental samples, it will be necessary to consider the effects of rRNA extraction and dye labeling methods, as they might affect fingerprints of both single targets and target mixtures. Also, interactions among RNA targets may occur in the case of a complex mixture, thus biasing quantification. If this would be proven to occur, one would have to determine target–target specific interaction factors that may be deduced from analyzing of a set of model mixtures. Further optimization of the approach might also be necessary since we showed that 75% reduction of the probe number (removal of all designed mismatch probes) did not significantly alter R^2 of the solution (see Table 9). These findings suggest that all probes of our current version of the microarray are not needed to make precise identifications. Finally, it is important to recognize that our approach cannot be used to identify unknown microbial targets — it will only work for microbial targets whose patterns have been prerecorded.

5. Conclusion

A new high-density microarray-based microbial identification and quantification method has been presented in this report. Our method is quantitative, and does not rely on PCR amplification, or probe design and probe validation. Our findings demonstrate that including signal intensity values from all duplexes on the array, which are essentially nonspecific to the target organisms, significantly improved predictions of known microbial targets. To our knowledge, this is the first study to report this phenomenon. In addition, we have demonstrated that the method is a self-sufficient analytical procedure since it provides statistical confidence of identification and quantification.

Acknowledgements

We thank Sergey Stolyar and Marius Brouwer for their critical comments on earlier versions of the manuscript. Supplemental information based on the Febit (Geniom) experiments conducted by Barbara Kleinhenz, Georg Nies, and Diethard Tautz at the Institute for Genetics at the University of Cologne, Germany. This work was supported by grants 1U01DE014955-01 from NIH/NIDCR and R-82945801 from EPA-CEER-GOM to P.A.N.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mimet.2007.05.001.

References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., Polz, M.F., 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J. Bacteriol. 186, 2629–2635.

Adamczyk, J., Hesselsoe, M., Iversen, N., Horn, M., Lehner, A., Nielsen, P.H., Schloter, M., Roslev, P., Wagner, M., 2003. The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. Appl. Environ. Microbiol. 69, 6875–6887.

Amann, R., 1995. Fluorescently labelled, rRNA-targeted oligonucleotide probes in the study of microbial ecology. Mol. Ecol. 4, 543–554.

Chandler, D.P., Newton, G.J., Small, J.A., Day, D.S., 2003. Sequence vs structure for the direct detection of 16S rRNA on planar oligonucleotide microarrays. Appl. Environ. Microbiol. 70, 2621–2631.

DeSantis, T.Z., Stone, C.E., Murray, S.R., Moberg, J.P., Andersen, G.L., 2005. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. FEMS Microbiol. Lett. 245, 271–278.

Dramanac, R., Labat, I., Brukner, I., Crkvenjakov, R., 1989. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 4, 114–128.

Lawson, C.L., Hanson, R.J., 1974. Solving Least Squares Problems. Prentice-Hall.

Lim, E.L., Tomita, A.V., Thilly, W.G., Polz, M.F., 2001. Combination of competitive quantitative PCR and constant-denaturant capillary electrophoresis for high-resolution detection and enumeration of microbial cells. Appl. Environ. Microbiol. 67, 3897–3903.

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

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., Schleifer, K.-H., Wagner, M., 2002. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Appl. Environ. Microbiol. 68, 5064–5081.

Muyzer, G., Teske, A., Wirsen, C.O., Jannasch, H.W., 1995. Phylogenetic relationships of Thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. Arch. Microbiol. 164, 165–172.

Naef, F., Magnasco, M.O., 2003. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys. Rev., E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics 68 Art. No. 011906 Part 1.

Neidhardt, F.C., Magasanik, B., 1960. Studies on the role of ribonucleic acid in the growth of bacteria. Biochim. Biophys. Acta 42, 99–116.

Palmer, C., Bik, E.M., Eisen, M.B., Eckburg, P.B., Sana, T.R., Wolber, P.K., Relman, D.A., Brown, P.O., 2006. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. 34, e5.

Pozhitkov, A., Noble, P.A., 2007a. Comment on discrimination of shifts in soil microbial communities using nonequilibrium thermal dissociation and gel pad array technology. Environ. Sci. Technol. 41, 1797–1798.

Pozhitkov, A., Noble, P.A., 2007b. Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. Nucleic Acids Res. 35, e70.

Pozhitkov, A., Chernov, B., Yershov, G., Noble, P.A., 2005a. Evaluation of gel-pad oligonucleotide microarray technology using artificial neural networks. Appl. Environ. Microbiol. 71, 8663–8676.

Pozhitkov, A., Stemshorn, K., Tautz, D., 2005b. An algorithm for the determination and quantification of components of nucleic acid mixtures based on single sequencing a reactions. BMC Bioinformatics 6, 281.

Pozhitkov, A., Noble, P.A., Domazet-Loso, T., Staehler, P., Beier, M., Tautz, D., 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.

Pozhitkov, A.E., Stedtfeld, R.D., Hashsham, S.A., Noble, P.A., 2007. Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. Nucleic Acids Res. 35, e70.

Pozhitkov, A., Stedtfeld, R.G., Hashsham, S.A., Noble, P.A., submitted for publication. Effects of nucleic acid concentration on nonequilibrium thermal dissociation curves.

Raskin, L., Poulsen, L.K., Noguera, D.R., Rittman, B.E., Stahl, D.A., 1994a. Quantification of methanogenic groups in anaerobic biological reactors by oligonucleotide probe hybridization. Appl. Environ. Microbiol. 60, 1241–1248.

Raskin, L., Stromley, J.M., Rittman, B.E., Stahl, D.A., 1994b. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl. Environ. Microbiol. 60, 1232–1240.

Rosset, R., Julien, J., Monier, R., 1966. Ribonucleic acid composition of bacteria as a function of growth rate. J. Mol. Biol. 18, 308–320.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., Chandler, D.P., 2001. Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Appl. Environ. Microbiol. 67, 4708–4716.

Southern, E.M., 1988. Method and apparatus for analysing polynucleotide sequences. European Patent 373203B1.

Suzuki, M.T., Giovannoni, S.J., 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. 62, 625–630.

Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical recipes in C. The Art of Scientific Computing. W. H. Press.

Urakawa, H., Noble, P.A., ElFantroussi, S., Kelly, J.J., Stahl, D.A., 2002. Single-base-pair discrimination of terminal mismatches by using oligonucleotide microarrays and neural network analyses. Appl. Environ. Microbiol. 68, 235–244.

Wilson, W.J., Strout, C.L., DeSantis, T.Z., Stilwell, J.L., Carrano, A.V., Andersen, G.L., 2002a. Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. Mol. Cell. Probes 16, 119–127.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kucmarski, T.A., Andersen, G.L., 2002b. High-density microarray of small-subunit ribosomal DNA probes. Appl. Environ. Microbiol. 68, 2535–2541.

Woese, C.R., 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. U. S. A. 97, 8392–8396.

Wu, Z., Irizarry, R.A., 2004. Stochastic models insprired by hybridization tehory for short oligonucleotide microarrays. In: Proceedings of RECOMB 2004. San Diego, CA.

Zhang, Y., Hammer, D.A., Graves, D.J., 2005. Competitive hybridization kinetics reveals unexpected behavior patterns. Biophys. J. 89, 2950–2959.