BIOINFORMATICS
Vol. 24 ECCB 2008, pages i83–i89
doi:10.1093/bioinformatics/btn269

PhyloDetect: a likelihood-based strategy for detecting

Hubert Rehrauer[1,*], Susan Schönmann[2], Leo Eberl[2] and Ralph Schlapbach[1]
[1]Functional Genomics Center Zurich, University/ETH Zurich and [2]Institute of Plant Biology, Department of Microbiology, University of Zurich, Zurich, Switzerland

ABSTRACT

Motivation: Detection and identification of microbes using diagnostic arrays is still subject of ongoing research. Existing significance-based algorithms consider an organism detected even if a significant number of the microarray probes that match the organism are called absent in a hybridization. Further, they do generate redundant results if the target organisms show high sequence similarity and the microarray probes cannot discriminate all of them.

Results: We propose a new analysis strategy that considers organism similarities and calls organisms only present if the probes that match the organism but are absent in a hybridization can be explained by random events. In our strategy, we first identify the groups of target organisms that are actually distinguishable by the array. Subsequently, these organism groups are placed in a hierarchical tree such that groups matching only less specific probes are closer to the tree root, and groups that are discriminated only by few probes are close to each other. Finally, we compute for each group a likelihood score that is based on a hypothesis test with the null hypothesis that the group was actually present in the hybridized sample. We have validated our strategy using datasets from two different array types and implemented it as an easy-to-use web application.

Availability: http://www.fgcz.ethz.ch/PhyloDetect

Contact: Hubert.Rehrauer@fgcz.uzh.ch

Supplementary information: Example data is available at http://www.fgcz.ethz.ch/PhyloDetect

1 INTRODUCTION

Microarrays offering the parallel measurement of thousands of genomic sequences are promising tools to survey microbial ecosystems. They can play an important role in environmental and ecological research that aims at the characterization of natural microbial communities (Bodrossy *et al.*, 2003; DeSantis *et al.*, 2007; Peplies *et al.*, 2004), but can also serve as a diagnostic tool for the identification of pathogens (Burton *et al.*, 2006; Myers *et al.*, 2006, Wilson *et al.*, 2002). Diagnostic microarrays have been demonstrated to be able to detect organisms in different environments: complex microbial communities in the human body (Palmer *et al.*, 2006), Alphaproteobacteria in the rhizosphere (Sanguin *et al.*, 2005), bioremediation bacteria in activated sludge of wastewater treatment plants (Loy *et al.*, 2005), vibrios in marine samples (Marcelino *et al.*, 2006) and pathogenic viruses in humans (Urisman *et al.*, 2005; Wong *et al.*, 2007); although viruses are not considered organisms the approach for their detection is the same. The current status and challenges in the application of diagnostic microarrays has been recently reviewed (Avarre *et al.*, 2007; Sessitsch *et al.*, 2006; Wang *et al.*, 2003).

In this article we present PhyloDetect, a general scheme for the identification and qualitative detection of organisms from hybridizations of taxonomic or diagnostic microarrays (Fig. 1). Our approach is based on no particular assumptions on the nature of the hybridized samples or on the type of microarrays. We only assume that the microarray holds discriminative probes that yield a hybridization signal when hybridized with a matching microorganism and yield no signal when hybridized with a non-matching one. This information can be represented as a probe-organism match matrix. From a hybridization we extract the outcome in form of a presence vector that specifies which probes yielded a presence signal in the hybridization. Using the match matrix and the hybridization outcome the organism detection is done in three steps: grouping of non-distinguishable organisms, arrangement of organism groups in a hierarchical tree and likelihood computation. Organisms are non-distinguishable by the array if they have identical entries in the match matrix. The organization of organism groups in the organism tree is done such that the more specific organism groups that have more matching probes are placed at the leaves and the more general groups having fewer matching probes are placed in the nodes closer to the root. Each node is characterized by the organism group located at the node and the set of matching probes for this group. The hierarchy is built in such a way that the matching probes at each node are given by the matching probes at the parent node plus additional, more specific probes. This hierarchical structure has the advantage that matching more general organism group may be detected even if some of the highly specific probes fail. It is most beneficial if the probes are designed according to a phylogenetic tree with many specific probes for individual organisms and additional general probes matching higher taxa, according to the so-called multiple nested probe concept (Behr *et al.*, 2000; Loy *et al.*, 2002; Militon *et al.*, 2007). Finally, the actual identification is done on the basis of hypothesis tests that are run for each node of the organism tree and takes as null hypothesis the assumption that an organism contained in that node was present. Those nodes (and the assigned organisms) that are not compatible with this presence hypothesis at a specified significance level are declared as absent.

Compared to algorithms for the analysis of differential expression from microarrays, the field of organism detection has received only little attention. Only three algorithms for significance-based organism detection with diagnostic microarrays can be found in the literature: E-predict (Urisman *et al.*, 2005), the Pathogen Detection Algorithm (PDA) (Wong *et al.*, 2007) and DetectiV (Watson

[*] To whom correspondence should be addressed.

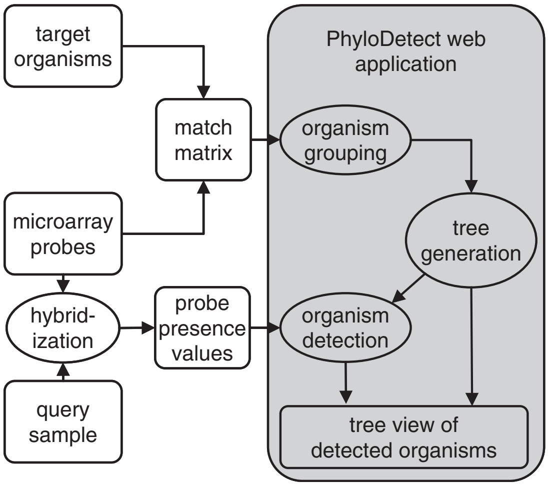


Fig. 1. Schematic workflow for the organism detection. PhyloDetect uses the probe-organism match matrix to determine organism groups and their hierarchical structure. The probes that give a present signal in a microarray hybridization are then used as input for the organism detection. The result is displayed in a tree view.

et al., 2007). These algorithms rely on hypothesis tests with the null hypothesis that the specific organism was *not* present in the hybridized sample. Further they have been designed and validated for the detection of viral pathogens using microarrays with many (up to 10^2) matching probes for each virus. However, in environmental applications where organisms are detected based on their ribosomal RNA, such high number of probes specific for individual organisms are not available. In comparison to these algorithms, the PhyloDetect approach is designed to work also well with fewer probes and is unique by representing the target organisms in a tree structure. This organism tree is especially useful for applications that aim at a high resolution of the microbial organisms where it is not possible to find large numbers of discriminating probes. Additionally, PhyloDetect is the only analysis tool for taxonomic microarrays that is available as a web application and therefore readily usable by the scientific community.

2 METHODS

The PhyloDetect strategy for organism detection consists of the following parts:

(1) Group the non-distinguishable microbial organisms.

(2) Given the probes on a microarray, arrange all detectable and distinguishable microbial groups in an organism tree.

(3) Using the probe presence values of a microarray hybridization, run a hypothesis test for each node of the tree and compute the likelihood that the signal was generated by one of the microbial organisms assigned to the tree node.

The definition of the organism groups and their hierarchy is determined from probes of the diagnostic microarray and has to be computed only once for a given array type. The actual organism identification in the third step is done for every sample which is hybridized to an individual slide of this array type.

As input, PhyloDetect requires a match matrix m_{ij} with $i=1\ldots N$and $j=1\ldots M$, where N is the number of discriminating probes and M is the number of target organisms. We require $m_{ij}=1$ if probe i matches the organism j, and $m_{ij}=0$ if not. This match matrix can either be directly derived from BLAST searches of the probes against the genomes of the target organisms, but ideally the matrix also considers the predicted binding energies of the found matches (Urisman *et al.*, 2005). Alternatively; the weighted mismatch score implemented in the ARB software (Ludwig *et al.*, 2004) may be used. Or, the match matrix may even be determined empirically from a large set of test hybridizations (Loy *et al.*, 2002; 2005).

The second input required by PhyloDetect is the hybridization outcome. This has to be in the form of a binary vector of presence values $o_i, i=1\ldots N$ that specifies whether probe i was present ($o_i=1$) or absent ($o_i=0$). Such presence values for probes may be computed in different ways from the hybridization signals. Frequently probes are called present if the intensity is higher than the median background signal plus twice the SD of the background (Peplies *et al.*, 2004; Sanguin *et al.*, 2005). Alternatively, normalization against a positive control with subsequent thresholding against an empirically determined value may be used (Bodrossy *et al.*, 2003; Loy *et al.*, 2002).

2.1 Microbial group definition

Target organisms are clustered into one group if they cannot be distinguished by the set of probes on the microarray. Non-distinguishable organisms may occur if their genomic sequence is very similar, or if after the design of the microarray new organisms are discovered and added to the set of target organisms. In both cases the array may not have discriminating probes for the target organisms. Non-distinguishable organisms do lead to identical detection scores thus generating redundancy in the detection result, which is eliminated if they are grouped together and if the detection works on the organism groups. A present value for an organism group in hybridization has to be interpreted as: 'A group member or a mixture of group members was contained in the hybridized sample'. PhyloDetect groups organisms together if their columns in the match matrix are identical. Specifically, organism j and organism j' are non-distinguishable if the match matrix columns m_{ij} and $m_{ij'}$ are identical for each probe i.

2.2 Organism tree algorithm

The organism grouping yields the microbial groups $g_k, k=1\ldots K$, where K is the total number of distinguishable organism groups. For each group g_k we denote S_k as the set of probes that matches the microbial group g_k.

We define the organism tree as a tree of nodes where each node ν is given by a set of probes denoted by S_ν, and the corresponding organism group g_k with $S_k == S_\nu$. The tree is computed recursively with the following algorithm:

Initialization:

Initialize the set of remaining groups $G_{rem}$with all groups $g_k, k=1\ldots K$ and define a root node ν_{root} that has the empty set as the set of probes. Use the root node as current node ν_{curr}.

Recursion:

(1) If G_{rem} contains a group g_k with $S_k == S_{\nu_{curr}}$, remove g_k from G_{rem} and assign the group to the current node ν_{curr}, else generate a new empty group and assign it to the current node ν_{curr}

(2) While G_{rem} contains microbial groups, do:

   (a) Select the probe i that matches most frequently in the set of groups G_{rem} and is not contained in S_ν

   (b) Add to the current node a new child node γ with the set of probes of the child node being the union of probe i and the set of probes at the current node: $S_\gamma = i \cup S_{\nu_{curr}}$

   (c) Compute the new set of remaining organism groups G'_{rem} as the set of all groups g_k in G_{rem} where $S_\gamma \subseteq S_k$ and remove G'_{rem} from G_{rem}

   (d) Continue at Step 1, with γ as the current node and G'_{rem} as the current set of remaining groups

This recursion algorithm generates a tree of organisms groups and matching probe sets with the following properties:

- The probe set of a child node is always a superset of the probe set of the parent node.
- A node may have an empty organism group attached. However, there will always be descendents with non-empty organism groups. If in a detection such an empty node is called present, but none, if its descendant nodes is called present, then this points to the potential presence of an unknown organism.
- Probes assigned to nodes at higher positions in the tree are more general and match many organism groups.
- Non-detectable organisms, i.e. organisms for which no matching probe exists are assigned to the tree root.
- The tree is not a phylogenetic tree! It rather depicts similarities between organisms as ‘seen’ by the microarray. However, if the microarray probes are chosen to closely match monophyletic groups, the tree will be very close to a phylogenetic tree.

Ties may occur in Step 2a of the recursion algorithm, because there may be several probes having the maximal match frequency in the remaining organism groups. This situation is treated as follows:

- The probe that comes first in the match matrix is selected.
- If one or more of the other probes with the maximal match frequency have the same match signature in the remaining organism groups, then these probes are also selected.

2.3 Organism identification

Finally, we compute from a hybridization result, the organism groups that were present in the hybridized sample. For each organism group we run a hypothesis test, with the null hypothesis that the organism group was present and the alternative hypothesis that the group was absent. We consider all groups that have a P-value larger than α as present and below α as absent. This implies that the organism detection rate for samples containing individual organisms is in the ideal case $1-\alpha$. When computing the hypothesis test for group g_k, PhyloDetect uses only the outcomes of those probes that are matching probes for the group g_k, i.e. only the probes $i \in S_k$. This is because under the null hypothesis (g_k was in the sample) we can only formulate our expectations about probes matching g_k. For the other probes we cannot make any statements, since the null hypothesis leaves it open whether other organisms were also contained in the sample or not.

Under the assumption that the probes discriminate perfectly and that a given organism is present in the sample, all matching probes for that organism will yield a present signal. However, using real world arrays and hybridizations, some of the matching probes will give an absent signal because they have inadequate hybridization properties (Pozhitkov *et al.*, 2006). We account for that by assuming a probe-independent false negative rate *fnr* that reflects that a probe may give an absent signal despite the fact that its matching counterpart is contained in the sample. With these assumptions the number of matching probes of the organism that give an absent signal follows the binomial distribution, so that we can formulate the hypothesis test with the null hypothesis that the given organism was present. Specifically, let g_k be the considered organism group, and let r be the number of probes with absent signals among the set of matching probes S_k. Then the likelihood of observing r or more absent signals is given by

$$\sum_{s=r}^{|S_k|} \binom{|S_k|}{s} fnr^s (1-fnr)^{|S_k|-s},$$

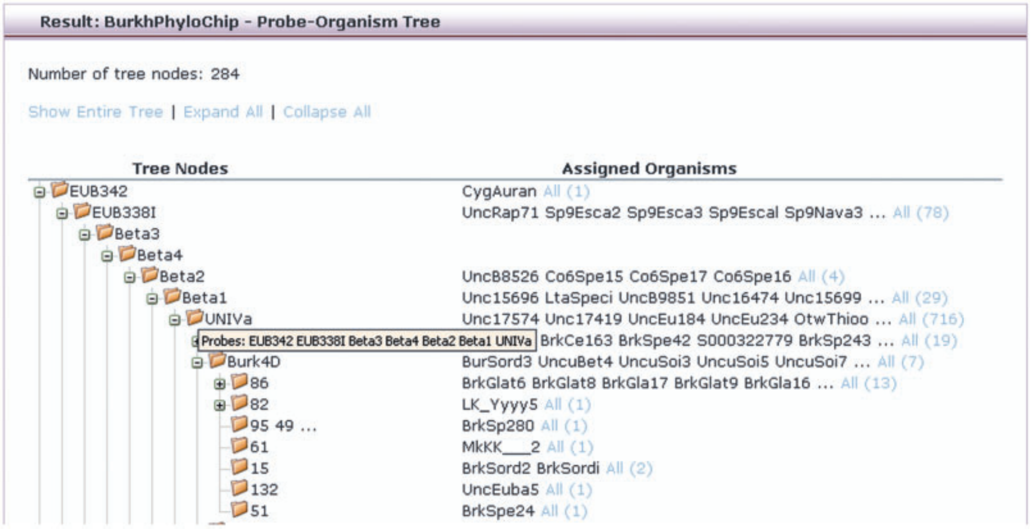
which is the complement of the cumulative binomial distribution. If for an organism group, this likelihood is below α, then we can rule out—at significance level α—that a member of group g_k was present in the hybridized sample.

In case the microarray contains probes matching more than one organism group this type of testing leads to the false detection of organism groups if their matching set of probes has a large overlap with the matching probe set of the actually hybridized group. In order to prevent against this type of false detection, the test is run twice for each group g_k: once with all probes in S_k, and once with only the four most discriminative probes. Only if both tests call the group g_k present, the group is called present.

3 RESULTS

We demonstrate the performance of the PhyloDetect approach on the example of a diagnostic taxonomic microarray that targets and discriminates bacteria of the genera *Burkholderia* and *Pandoraea*. These bacteria are ubiquitously distributed in nature and have been isolated from soil, water, the rhizospheres of various plants, industrial settings, the hospital environment and from infected humans. Some species have an enormous biotechnological potential and have been used for bioremediation of recalcitrant xenobiotics, plant growth promotion and biocontrol purposes. At the same time, however, some species have emerged as important pathogens of plants, animals and humans; for a review see Coenye and Vandamme (2003). This array was developed to identify these bacteria as pure cultures but also in mixtures in the environment in a high-throughput manner. It contains 131 oligonucleotide probes with lengths of 15–18 nt matching the 16S rRNAs of *Burkholderia* and *Pandoraea* species. The probe design of the array is according to the nested probe approach and contains next to the highly specific probes for individual strains also probes common to many *Burkholderia* and *Pandoraea* bacteria. A paper discussing the application of this array for the monitoring of these bacteria in soil is currently in preparation. We used a dataset of 95 hybridizations covering established type strains as well as newly sequenced clones for the evaluation of the PhyloDetect approach. The dataset is available as example dataset of the PhyloDetect web application. In a hybridization, a probe was considered as present if its signal was at least 10% of the signal of the positive control probes. The performance of the organism detection depends only slightly on the choice of this presence threshold; this is because the majority of the probes give either a clear present or a clear absent signal. For example, if the presence threshold is increased by 50% the number of probes that declared as present in the hybridizations decreases only by 12% and leaves still enough present probes to detect the organisms. The same is true when choosing different hybridization preprocessing or normalization steps. In particular, we considered different methods of subtracting hybridization backgrounds (no subtraction, chip-specific or probe-specific background) and normalization steps (negative controls, negative controls considering spread of the controls) and did observe only minor changes in the overall false-positive and false-negative rates of individual probes. With the chosen presence threshold of 10%, the empirically determined false negative rate of individual probes is as low as 5%.

For the organism identification we considered ∼50 000 16S rRNA target sequences (>1200 bp) available from the ARB Project (Ludwig *et al.*, 2004) at http://www.arb-home.de and from the ribosomal database project (Cole *et al.*, 2005) release 9.27 plus additional sequences of environmental samples, not published before. Since all hybridized samples were amplified using primers that specifically match 16S rRNA gene sequences from *Burkholderia* and *Pandoraea*, we could reduce the number of potential target

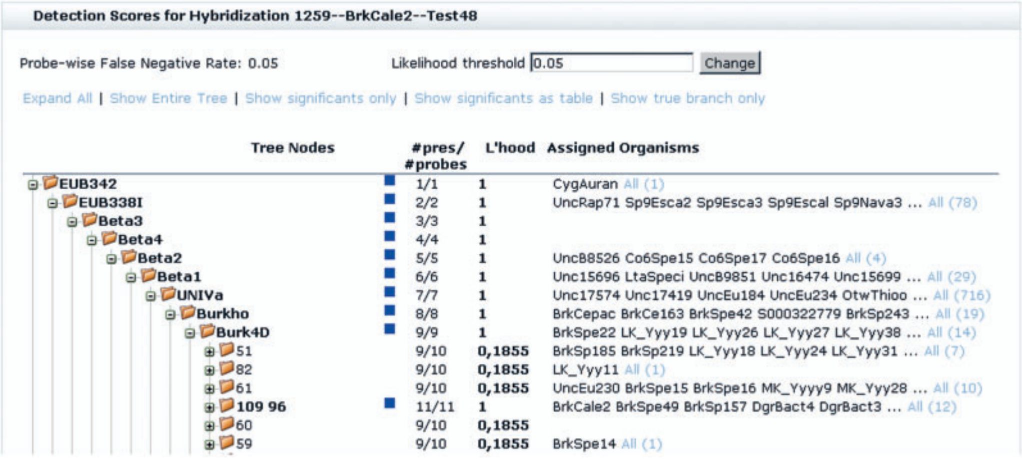


Result: BurkhPhyloChip - Probe-Organism Tree

Number of tree nodes: 284

Tree Nodes	Assigned Organisms
EUB342	CygAuran All (1)
EUB338I	UncRap71 Sp9Esca2 Sp9Esca3 Sp9Escal Sp9Nava3 ... All (78)
Beta3	
Beta4	
Beta2	UncB8526 Co6Spe15 Co6Spe17 Co6Spe16 All (4)
Beta1	Unc15696 LtaSpeci UncB9851 Unc16474 Unc15699 ... All (29)
UNIVa	Unc17574 Unc17419 UncEu184 UncEu234 OtwThioo ... All (716)
Probes: EUB342 EUB338I Beta3 Beta4 Beta2 Beta1 UNIVa	BrkCe163 BrkSpe42 S000322779 BrkSp243 ... All (19)
Burk4D	BurSord3 UncuBet4 UncuSoi3 UncuSoi5 UncuSoi7 ... All (7)
86	BrkGlat6 BrkGlat8 BrkGla17 BrkGlat9 BrkGla16 ... All (13)
82	LK_Yyyy5 All (1)
95 49 ...	BrkSp280 All (1)
61	MkKK___2 All (1)
15	BrkSord2 BrkSordi All (2)
132	UncEuba5 All (1)
51	BrkSpe24 All (1)

Fig. 2. Section of the Phylodetect organism tree showing the hierarchical arrangement of target organisms of the Burkholderia phylochip. The tree nodes are labeled according to the characteristic probes for each node. All probes matching the entire branch down to that node are shown as tooltip (yellow box) upon mouse over. The organisms matching all probes of a node but no other probes are listed to the right.



Detection Scores for Hybridization 1259--BrkCale2--Test48

Probe-wise False Negative Rate: 0.05

Likelihood threshold 0.05 Change

Expand All | Show Entire Tree | Show significants only | Show significants as table | Show true branch only

Tree Nodes		#pres/ #probes	L'hood	Assigned Organisms
EUB342	■	1/1	1	CygAuran All (1)
EUB338I	■	2/2	1	UncRap71 Sp9Esca2 Sp9Esca3 Sp9Escal Sp9Nava3 ... All (78)
Beta3	■	3/3	1	
Beta4	■	4/4	1	
Beta2	■	5/5	1	UncB8526 Co6Spe15 Co6Spe17 Co6Spe16 All (4)
Beta1	■	6/6	1	Unc15696 LtaSpeci UncB9851 Unc16474 Unc15699 ... All (29)
UNIVa	■	7/7	1	Unc17574 Unc17419 UncEu184 UncEu234 OtwThioo ... All (716)
Burkho	■	8/8	1	BrkCepac BrkCe163 BrkSpe42 S000322779 BrkSp243 ... All (19)
Burk4D	■	9/9	1	BrkSpe22 LK_Yyy19 LK_Yyy26 LK_Yyy27 LK_Yyy38 ... All (14)
51		9/10	**0,1855**	BrkSp185 BrkSp219 LK_Yyy18 LK_Yyy24 LK_Yyy31 ... All (7)
82		9/10	**0,1855**	LK_Yyy11 All (1)
61		9/10	**0,1855**	UncEu230 BrkSpe15 BrkSpe16 MK_Yyyy9 MK_Yyy28 ... All (10)
109 96	■	11/11	1	BrkCale2 BrkSpe49 BrkSp157 DgrBact4 DgrBact3 ... All (12)
60		9/10	**0,1855**	
59		9/10	**0,1855**	BrkSpe14 All (1)

Fig. 3. Visualization of a detection result. The figure shows the result of a hybridization of the 16S rRNA gene from BrkCale2. The truly hybridized organism group and all parent groups are indicated by blue squares. Eleven out of the 11 probes that match BrkCale2 gave present signals in the hybridization and the likelihood for the correct organism group is 1. However, since PhyloDetect assumes that multiple organisms may have been present in the hybridized sample, it cannot be excluded that one of the parents was also contained in the sample. Additionally, we allow for false negative probe signals such that a neighboring node with only one discriminating probe may also have been in the sample. Therefore the likelihoods for those nodes are also well above 0.05.

organisms from ∼50 000 to 2400 having sequence matches to the primers. For the 131 probes and the 2400 organisms the matching information was obtained from the ARB software (Ludwig *et al.*, 2004). A probe was considered as a match to an organism if the alignment was perfect. For this array, PhyloDetect identifies 284 discriminable organism groups and arranges them in a tree. In Figure 2 we show a small section of the tree. The folders represent the nodes of the tree and are labeled with the probes specific to that node, i.e. with the probes that match at this node but not at any of the parents. The entire set of probes that matches at this node is displayed as tooltip when the mouse is pointed at it. The organism group associated with a node is displayed to the right of each node with the first few representatives being directly displayed by their identifier. As an example, the result of a hybridization of the 16S rRNA gene amplicon of *Burkholderia caledonica* (BrkCale2) is shown in Figure 3. After processing the hybridization results, PhyloDetect displays the tree with all probes, which gave positive signals indicated in bold. Additionally, the group's likelihood of generating the observed result is shown next to each organism group. All groups for which the likelihood is above the chosen likelihood threshold of 5% are printed in bold. Figure 3 shows that the group that contains BrkCale2 was correctly identified. Furthermore, all parents of this node are also significant, because they are characterized by a subset of the probes matching BrkCale2.

Since PhyloDetect assumes generally that more than one organism may have been present in the hybridized sample the presence of parent groups cannot be excluded. The results only allow the conclusion that the hybridized sample contained either exclusively a member of the BrkCale2 group or a mixture of the BrkCale2 group plus any organism or combination of organisms from the parents of the BrkCale2 group. Siblings or descendants of the BrkCale2 group may also have been present in the sample but here the likelihood is lower because the respective specific probes were not declared as present. In the entire dataset of 95 hybridizations of single bacterial strains, PhyloDetect found in 88% of the cases the correct organism group as being present. The number of false positives was rather high and led to a false discovery rate of 70%. This is because the array aims at resolving the bacterial strains at the sub-species level where only single or few discriminating probes can be found. Considering only the organisms with at least two discriminating probes relative to the actually hybridized strain, the false discovery rate drops down to 31%, showing that PhyloDetect mainly leaves highly similar organisms belonging to the same species undiscriminated.

In order to demonstrate the generality of the PhyloDetect approach, we analyzed also the dataset associated with the publication of the E-predict virus detection algorithm. The dataset is publicly available from the NCBI GEO Database (Edgar *et al.*, 2002) under the accession GSEE2228. The study used the MegaViro array (Wang *et al.*, 2003) that contains ∼11 000 oligonucleotides targeting ∼1000 viruses. For the analysis of this dataset with PhyloDetect we used the energy profile matrix (Urisman *et al.*, 2005, Supplementary Material file 5) to build the match matrix that is required as input information for PhyloDetect. We considered a probe as a match for a virus if the reported binding energy was above 80. From the hybridization signals we considered only those probes as present where the background corrected and normalized intensity was above 200. The generated match matrix and the hybridization outcomes are available on the PhyloDetect web site (http://www.fgcz.ethz.ch/PhyloDetect). For the 18 nasal lavage samples that contained infections by the Influenza A and/or the Human respiratory syncytial (RSV) virus, we computed the detection scores using PhyloDetect (see Table 1 for the results). In all 18 samples the correct virus was identified, for some samples two additional closely related viruses, Bovine respiratory syncytial virus and Human metapneumovirus, were detected. Only in one sample (GEO Accession GSM40816) an unrelated virus (*Lymantria dispar* nucleopolyhedrovirus; highlighted in bold in Table 1) was identified because 3 of 6 of the matching probes were found as present. This same virus was also found present by the E-predict software which suggests that this is rather an artifact of the dataset than a failure of the detection algorithm. Overall, this example demonstrates that the PhyloDetect approach is readily applied and does perform well on this virus detection array.

4 DISCUSSION

PhyloDetect represents a universal approach to analyze the hybridization results of diagnostic microarrays in order to identify the organisms present in a sample. PhyloDetect works equally well for small arrays, like the *Burkholderia* phylochip with only few probes per targeted organism, and for large arrays like the MegaViro array (Wang *et al.*, 2003) with many probes per targeted organism. While the detection of distantly related organisms is less challenging because there are usually many discriminating probes available, PhyloDetect does also perform well for closely related organisms with some or many probes matching more than one organism. In the latter case, the user benefits from the hierarchical structure of the result that arranges highly similar organisms close to each other in an organism tree. PhyloDetect intrinsically supports the detection of multiple organisms in a sample since the detection score is computed for each organism independently.

There are three other published algorithms that tackle the problem of statistical detection of organisms from diagnostic microarrays: the E-predict algorithm (Urisman *et al.*, 2005), the PDA (Wong *et al.*, 2007) and the DetectiV approach (Watson *et al.*, 2007). We compared PhyloDetect with the E-predict algorithm on the MegaViro dataset where both algorithms perform comparably. A comparison of DetectiV and E-predict on the same dataset (Watson *et al.*, 2007) shows a slight superiority of DetectiV over E-predict. We have no comparison data for the PDA algorithm because the definition of the 'recognition signature probe' that is part of the algorithm is not sufficiently described (Wong *et al.*, 2007).

E-predict, PDA and DetectiV are primarily designed for virus detection arrays and have been validated in the respective studies. In order to detect a given virus, these approaches perform a hypothesis test with the null hypothesis that the respective virus is *not* present. The difficulty hereby is that the null distribution for this null hypothesis is not properly defined. As a workaround the authors of E-Predict and PDA derive the null distributions empirically by averaging over many (up to thousands) of hybridizations. However this null distribution is not appropriate in cases where it is desired to detect organism x but the sample actually contains a closely related organism x' that has many matching probes in common with x. Here, the null hypothesis for organism x is true, namely organism x was not present in the sample, but the observed distribution is nevertheless very different from the empirically derived null distribution. In the DetectiV approach a one-sample t-test is performed and the probe intensities are compared against a fixed value that is empirically determined from the same array data. PhyloDetect avoids the problem of the unspecific null hypothesis by formulating the hypothesis test in the opposite direction. We assume as null hypothesis that the respective organism is *present* and we only call the organism present if the signal of the matching probes is compatible with this null hypothesis, i.e. the probes that match the organism but are absent can be explained by the statistical model. The E-Predict and PDA algorithms have been designed for arrays with many probes per target organism and the assumption that a large portion of the probes may not work and always give a negative signal. This makes them robust in situations where the data is compromised by noise or by low concentrations of the target organisms. While PhyloDetect is not specifically built for these situations it can accommodate them by setting a larger expected false negative rate, as we did for the analysis of the nasal lavage samples ($fnr = 10\%$). In addition, both algorithms, E-Predict and PDA, make specific assumptions on how the probe-organism match score is computed, while PhyloDetect leaves it open. Additionally, these algorithms do not perform a binarization of the hybridization signals but exploit directly the signals. However, we did not observe any performance improvement when working with the hybridization signals instead of the present calls. Instead by working with binary presence scores we are independent from assumptions about signal distributions and noise characteristics of the different array platforms

Table 1. Detection scores for the E-predict dataset (GEO accession GSE2228)

DFA Result	PhyloDetect result				
	No. of present probes	No. of match probes	*P*-value	Taxon IDs	Names
Influenza A, RSV	10	12	0.989	11250	Human RSV
	6	10	0.828	11320; 183764	Influenza A virus
	7	15	0.291	12814	RSV
Influenza A	4	10	0.377	11320; 183764	Influenza A virus
Influenza A	5	10	0.623	11320; 183764	Influenza A virus
Influenza A	5	10	0.623	11320; 183764	Influenza A virus
Influenza A	6	10	0.828	11320; 183764	Influenza A virus
Influenza A	8	10	0.989	11320; 183764	Influenza A virus
Influenza A	8	10	0.989	11320; 183764	Influenza A virus
Influenza A	9	10	0.999	11320; 183764	Influenza A virus
RSV	15	15	1.000	12814	RSV
	12	12	1.000	11250	Human RSV
	10	16	0.895	11246	Bovine RSV
	4	7	0.773	162145	Human metapneumovirus
RSV	3	6	0.656	10449	**Lymantria dispar** nucleo. virus
	5	12	0.387	11250	Human RSV
	6	15	0.304	12814	RSV
	5	16	0.105	11246	Bovine RSV
RSV	15	15	1.000	12814	RSV
	11	12	0.999	11250	Human RSV
	6	7	0.992	162145	Human metapneumovirus
	11	16	0.962	11246	Bovine RSV
RSV	10	15	0.941	12814	RSV
	6	12	0.613	11250	Human RSV
RSV	13	15	0.998	12814	RSV
	10	12	0.989	11250	Human RSV
	4	7	0.773	162145	Human metapneumovirus
	6	16	0.227	11246	Bovine RSV
RSV	11	15	0.982	12814	RSV
	5	7	0.938	162145	Human metapneumovirus
	7	12	0.806	11250	Human RSV
	5	16	0.105	11246	Bovine RSV
RSV	12	15	0.996	12814	RSV
	5	7	0.938	162145	Human metapneumovirus
	8	12	0.927	11250	Human RSV
	9	16	0.773	11246	Bovine RSV
	4	38	0.000	147712	Human rhinovirus B
RSV	10	12	0.997	11250	Human RSV
	9	15	0.849	12814	RSV
	5	16	0.105	11246	Bovine RSV
RSV	12	12	1.000	11250	Human RSV
	11	16	0.962	11246	Bovine RSV
	11	15	0.954	12814	RSV
	4	7	0.773	162145	Human metapneumovirus
RSV	11	15	0.982	12814	RSV
	6	12	0.613	11250	Human RSV
	3	7	0.500	162145	Human metapneumovirus
	4	16	0.038	11246	Bovine RSV

The samples represent nasal lavage samples where viruses have been identified by DFA. For all samples except one the related viruses were correctly affiliated. The probably incorrectly identified virus Lymantria dispar was also found by the E-predict algorithm and do not need an extensive training dataset to teach the algorithm, as is the case for the E-predict algorithm. Altogether our approach is versatile and general, and readily applied to various kinds of diagnostic microarrays.

ACKNOWLEDGEMENTS

We thank all co-workers at the Functional Genomics Center Zurich for making this study possible, with special mention of Dr Jens Sobek and Catherine Aquino for generating the *Burkholderia* phylochip microarrays. Further we would like to thank Dr Alexander Loy from the University Vienna for fruitful discussions.

Conflict of Interest: none declared.

REFERENCES

Avarre,J.C. *et al.* (2007) Hybridization of genomic DNA to microarrays: a challenge for the analysis of environmental samples. *J. Microbiol. Methods*, **69**, 242–248.

Behr,T. *et al.* (2000) A nested array of rRNA targeted probes for the detection and identification of enterococci by reverse hybridization. *Syst. Appl. Microbiol.,* **23**, 563–572.

Bodrossy,L. *et al.* (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environ. Microbiol.,* **5**, 566–582.

Burton,J.E. *et al.* (2006) Differential identification of Bacillus anthracis from environmental Bacillus species using microarray analysis. *J. Appl. Microbiol.*, **101**, 754–763.

Coenye,T. and Vandamme,P. (2003) Diversity and significance of Burkholderia species occupying diverse ecological niches. *Environ. Microbiol.*, **5**, 719–729.

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

DeSantis,T.Z. *et al.* (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb. Ecol.*, **53**, 371–383.

Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Loy,A. *et al.* (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.

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

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

Marcelino,L.A. *et al.* (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proc. Natl Acad. Sci. USA*, **103**, 13629–13634.

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

Myers,K.M. *et al.* (2006) Molecular identification of Yersinia enterocolitica isolated from pasteurized whole milk using DNA microarray chip hybridization. *Mol. Cell. Probes*, **20**, 71–80.

Palmer,C. *et al.* (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Res.*, **34**, e5.

Peplies,J. *et al.* (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environ. Microbiol.*, **6**, 638–645.

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

Sanguin,H. *et al.* (2005) Development and validation of a prototype 16S rRNA-based taxonomic microarray for Alphaproteobacteria. *Environ. Microbiol.*, **8**, 289–307.

Sessitsch,A. *et al.* (2006) Diagnostic microbial microarrays in soil ecology. *New Phytol.*, **171**, 719–736.

Urisman,A. *et al.* (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol.*, **6**, R78.

Wong,C.W. *et al.* (2007) Optimization and clinical validation of a pathogen detection microarray. *Genome Biol.*, **8**, R93.

Wang,D. *et al.* (2003) Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.*, **1**, E2.

Watson,A. *et al.* (2007) DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol.*, **8**, R190.

Wilson,W.J. *et al.* (2002) Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. *Mol. Cell. Probes*, **16**, 119–127.