BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 10. NO 6. 342–353 doi:10.1093/bfgp/elr027

Microarrays for Pathogen Detection and Analysis

Kevin S. McLoughlin

Advance Access publication date 19 September 2011

Abstract

DNA microarrays have emerged as a viable platform for detection of pathogenic organisms in clinical and environmental samples. These microbial detection arrays occupy a middle ground between low cost, narrowly focused assays such as multiplex PCR and more expensive, broad-spectrum technologies like high-throughput sequencing. While pathogen detection arrays have been used primarily in a research context, several groups are aggressively working to develop arrays for clinical diagnostics, food safety testing, environmental monitoring and biodefense. Statistical algorithms that can analyze data from microbial detection arrays and provide easily interpretable results are absolutely required in order for these efforts to succeed. In this article, we will review the most promising array designs and analysis algorithms that have been developed to date, comparing their strengths and weaknesses for pathogen detection and discovery.

Keywords: *microarrays; pathogens; genomics*

INTRODUCTION

Infectious diseases pose a growing threat to public health, due to increased rates of population growth, international trade and air travel, climate change, bacterial antibiotic resistance and a wide range of other factors. In addition, global conflicts over the past decade have raised concerns that pathogenic agents might be released deliberately by terrorist organizations or other entities. Public and private funding agencies have responded to these concerns by investing heavily in the development of new assays for microbial surveillance and discovery. The majority of these new methods involve direct detection of microbial nucleic acids. Ideally, these methods should be effective both as *detection* assays (for identification of known pathogens) and as *discovery* techniques (for revealing the presence of novel, previously uncharacterized organisms).

Most currently available methods for microbial detection and discovery using nucleic acid samples are based on three technologies. In order of increasing cost, these are the polymerase chain reaction (PCR) [1], oligonucleotide microarrays [2] and DNA sequencing [3]. These platforms have different strengths and weaknesses. While sequencing provides the most in-depth, unbiased information, and is able to reveal completely novel organisms, it can be costly and time-consuming for some applications, particularly when the resources required for data processing and analysis are taken into account. Although multiplex sequencing of bar-coded samples reduces the cost per sample, it also decreases the coverage and thus the sensitivity of the analysis; this may be an issue when the organism of interest has low abundance and the sample has not been treated beforehand to remove host and/or background DNA.

At the other end of the cost spectrum, PCR assays are very fast and sensitive, but have limited capacity for multiplexing [4, 5]. When an assay is required to test for the presence of several organisms simultaneously, many PCR reactions may be needed, erasing any cost benefit. They are also highly specific; this is an advantage for detecting a microbe whose sequence is precisely known, but a great disadvantage for discovery of novel species, or for detecting variant strains of a known species.

Corresponding author. Kevin S. McLoughlin, Global Security, Lawrence Livermore National Laboratory, PO Box 808, L-174, Livermore, CA 94551 USA. Tel.: +925-423-5486; Fax: +925-422-6736; E-mail: mcloughlin2@llnl.gov

Kevin S. McLoughlin is a Computational Biologist in the Pathogen Bioinformatics Group at Lawrence Livermore National Laboratory. His research concerns analysis methods for microbial metagenome sequence and microarray data.

Microarrays occupy a middle ground with respect to cost, processing time, sensitivity, specificity and ability to detect novel organisms. The high-density arrays available at present are able to test for the presence of thousands of different organisms simultaneously, at a cost less than US$ 100 per sample. Arrays can be designed with a combination of high-specificity probes and probes designed against conserved regions, so that they can be used in both detection and discovery modes. While most array designs select probes from fully sequenced genomes in GenBank and other databases, cross-hybridization between probes and similar but non-identical sequences allows detection of novel species, provided that they are closely related to those that were used for probe design. A limitation of microarrays is that, except for so-called universal arrays, probe designs must be updated periodically to include the ever-increasing number of microbial genome sequences being added to GenBank. Nevertheless, for many applications, microarrays offer an ideal balance of capabilities for broad-spectrum microbial surveillance.

A microarray is a miniaturized device containing short (25- to 70-mer) single-stranded DNA oligonucleotide probes (or 'oligos') attached to a solid substrate, as shown in Figure 1. The probes are designed to have sequences complementary to segments of one or more target organism genomes. Oligos may be spotted onto the array by mechanical deposition [2], sprayed on with a modified inkjet printer head [6] or synthesized *in situ* through a series of photocatalyzed reactions [7]. Probes are placed on the array in a rectangular grid of 'features', each containing many copies of the same oligo. The density of features on the array varies between platforms, from 20 000 spots per slide for a typical spotted array, to several million for platforms such as NimbleGen and Affymetrix that use *in situ* synthesized oligos. Arrays may be subdivided with a gasket into subarrays, allowing multiple samples to be tested on one slide. Replicate features, scattered randomly across the array, may be used to allow correction for scratches and spatial effects. On some arrays, negative control probes with random sequences are included, to provide a threshold level for background noise correction.

To analyze a sample with the array, nucleic acids are extracted and converted to cDNA if necessary (e.g. if the target of interest is viral genomic RNA). The DNA is amplified if needed, fragmented and fluorescently labeled. The labeled DNA is incubated on the array surface for several hours, allowing enough time for the DNA fragments to hybridize to complementary or nearly complementary probes, if they exist on the array. The array is then washed to remove unbound DNA and scanned to produce a file of fluorescence intensities for each feature. In the resulting image, bright features will correspond to probes that are complementary to the DNA in the sample.

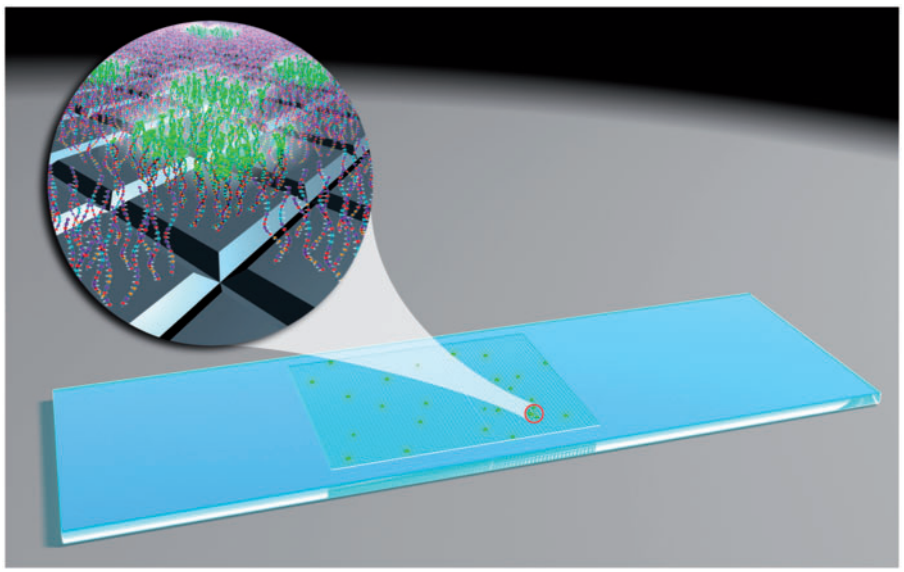

Figure 1: Schematic view of a microarray, showing single-stranded DNA oligo probes attached to substrate, with fluorescently labeled (green) target DNA strands bound to selected oligos.

PATHOGEN DETECTION ARRAY DESIGN

Several groups have applied microarray technology to pathogen detection. Their approaches may be distinguished according to the range of pathogens targeted, the probe design strategy and the array platform used. Each group has also developed analysis algorithms targeting its own array platform, which we will discuss in a subsequent section.

ViroChip

The first microarray designed for detection of a wide range of pathogens was the ViroChip [8]. The initial version of the ViroChip contained 1600 probes derived from the 140 complete viral genomes available in GenBank when the array was designed. Later versions of the array were developed to cover a wider range of viruses as additional genomes were published [9]. The ViroChip is fabricated by mechanically spotting synthesized oligonucleotides on a glass slide. The oligos are 70-mers, usually selected to match sequences common to a taxonomic family, but not found in other families. For some families, oligos were instead selected at the genus level. Since the probes were designed against conserved sequences, the ViroChip could be used to identify novel viruses within the same family as a known, sequenced virus. This capability was used to characterize the virus responsible for the 2003 SARS outbreak as a novel coronavirus [10].

The advantages of the ViroChip platform include the low cost of spotted oligo arrays, and its ability to detect novel viruses within a known family. Its disadvantages result mainly from the limitations of spotted oligo technology: the low density of probes that can be spotted on each slide, the inherent noisiness of the data and the large up-front cost for each array design (because tens of thousands of commercially synthesized oligos must be purchased for each design). The latest version of the ViroChip addresses these issues by using the Agilent ink-jet array platform [11]. As with other target library-based arrays, probe designs must be updated periodically to remain current with novel genomes in GenBank.

Resequencing pathogen microarrays

Another approach to pathogen detection uses resequencing microarrays [12,13]. These arrays contain short probes (25- or 29-mers) tiled along selected genes of the target pathogen species. Four probes are designed for each location in a target gene: one with a perfect match base at the central position of the probe and one for each of the three alternative bases. Hybridization and analysis of these arrays yields a sequence for each target gene homolog present in the sample; the sequence is then matched to a species and strain by comparison against a sequence database, using BLAST [14].

The prototype Respiratory Pathogen Microarray (RPM v1.1) was manufactured as a custom Affymetrix array. It contained probes for several common human respiratory viruses and bacteria, including influenza, adenovirus, coronavirus and rhinoviruses, together with bacteria such as *Bordetella pertussis* and *Streptococcus pneumoniae*. More recent designs based on this approach, such as the resequencing pathogen microarray for tropical and emerging infectious agents (RPM-TEI v1.0) [15], contain probes for a wider variety of pathogens and known toxin genes, focusing on the categories A–C select agents defined by the Centers for Disease Control (CDC).

The use of short oligonucleotide probes, with a large number of oligos per target gene, gives the RPM arrays very high specificity for strain-level identification of target organisms. The disadvantages of the RPM approach are its lower sensitivity (due in part to the use of short oligos), the limited range of organisms that can be covered on a single array (because of the large number of probes required for each target) and its lack of ability to detect novel organisms.

Universal detection array

A sequence-independent 'universal' microarray was described in [16]. Rather than selecting oligos from sequenced microbial genomes, the authors created an array containing 14 283 unique 12-mer and 13-mer probes with randomly generated sequences. Probes were excluded from the array design if they differed at fewer than 4 positions (for 12-mers) or 5 (for 13-mers) from previously selected probes. The authors further refined the probe list by building prototype arrays, and excluding from further analysis probes that did not give reproducible signals when data from replicate hybridizations were compared. Oligos were synthesized on arrays using a photocatalytic process, similar to that used on the NimbleGen platform. Hybridizations were performed at low temperatures (23°C) to compensate for the extremely short probe length.

By hybridizing genomic DNA from several bacterial species, the authors demonstrated that their array produced reproducible patterns of probe intensities (or 'signatures') that distinguished between one species and another. Probe intensities were not correlated either with occurrences of the oligo sequence in the target bacterial genome or with predicted free energies of hybridization. Thus, identification of an unknown target using this array relies on comparison of the observed intensity pattern to a compendium of signatures acquired by hybridizing known targets to the array. The primary advantage of the universal array approach is that the probes are not specific to genomes that have already been sequenced, so that the array design does not need to be updated as new genomes become available; this is a unique characteristic of the universal array approach. The principal disadvantage is the lack of any means of predicting signatures for organisms of known sequence; these must be obtained experimentally for every species of interest. The large number of experiments required makes this approach impracticable for broad-spectrum detection, especially for agents that must be handled using BSL-3 and BSL-4 procedures.

GreeneChip

The 'GreeneChip' arrays represent a broader-spectrum approach to pathogen detection [17, 18]. These are high-density oligonucleotide arrays, fabricated using the Agilent inkjet system. GreeneChipVr version 1.0 contains 9477 probes for viruses infecting vertebrates. GreeneChipPm v1.0 is a panmicrobial array design, containing all of the GreeneChipVr probes, together with probes for several thousand pathogenic bacteria, fungi and protozoa, comprising a total of 29 495 60-mer oligos. Viral probes were designed to target a minimum of three genomic regions for each family or genus of virus. Typically, one highly conserved region was chosen, along with two or more variable regions. Probe sequences were selected so that every vertebrate virus in the ICTV database (International Committee on Taxonomy of Viruses) or in GenBank was represented by at least one probe, with five or fewer mismatches. Bacterial, fungal and protozoan probes were selected by a similar strategy, except that the target sequences were only chosen from the 16S ribosomal RNA (rRNA) genes of bacteria and 18S rRNAs of fungi and protozoa.

When they were tested with virus-infected cell cultures and clinical samples from virally infected patients, the GreeneChip arrays correctly identified the virus at the species level. Performance with bacterial samples was poorer, due to the choice of 16S rRNA as the target gene; probes for these targets tended to cross-hybridize across taxa, so that some bacteria could only be identified at family or class resolution. The sensitivity of these arrays was comparable to that of the ViroChip series, due to the use of long (60-mer) oligos.

Lawrence Livermore microbial detection array

The most comprehensive pathogen detection arrays reported to date were designed by a team at Lawrence Livermore National Laboratory [19, 20]. Initial versions of the Lawrence Livermore Microbial Detection Array (LLMDA) contained target probes for all bacteria and viruses (pathogenic and otherwise) for which full genome sequences were available. More recent versions also include probes for pathogenic fungi and protozoa. Probe lengths on a single array vary between 50 and 65 nt and are adjusted so that all probes on the array have roughly equivalent affinities for their complementary target DNA molecules.

As on the GreeneChip, probes are selected from target genomes by one of two strategies. 'Discovery' probes match genome regions that are unique to a taxonomic family or subfamily, but are shared by the species within that family. By targeting sequences that evolve more slowly within families, the discovery probes are optimized for detection of novel species within a known family. 'Census' probes target highly variable regions that are unique to an individual species or strain. They are optimized for forensic use, to identify the specific strain of organism in a sample as precisely as possible.

The LLMDA designs are primarily deployed on the NimbleGen platform and have also been prototyped with Agilent inkjet technology. NimbleGen arrays are fabricated by photocatalytic synthesis, using a digital micromirror device to control the addition of nucleotides to each oligo. This technology supports higher probe densities than either spotted oligo arrays or the Agilent inkjet platform, with up to 4.2 million features per array. Thus, the LLMDA is able to represent each sequenced microbial genome by a large number of diverse probes, varying from 10 to 50 depending on the array format and the range of microbes targeted. A probe is counted toward the minimum

representation if it has an alignment to the target genome with at least 85% identity and with a 29-mer perfect match subsequence. Bacterial probes are selected from full genome sequences, rather than from 16S rRNA genes as in the GreeneChip, enabling strain-level discrimination for most bacteria. The LLMDA was shown to have high sensitivity and specificity when tested against a variety of previously characterized viral and bacterial cultures and was also used successfully to detect adventitious porcine circovirus DNA in a pediatric rotavirus vaccine [21].

The advantages of the LLMDA are its broad coverage of bacteria and viruses, as well as of eukaryotic pathogens. Its disadvantages include its cost (compared to spotted arrays such as the ViroChip), and the need (unlike with 'universal' arrays) to design new probes periodically to incorporate new genomes deposited in GenBank. The cost differential can be mitigated by exploiting the multiplex NimbleGen array formats, in which 3, 4, 12 or 24 samples can be hybridized to a single array.

PATHOGEN DETECTION ARRAY ANALYSIS

General issues

Much of the initial work on microarray data analysis focused on the use of these arrays to measure gene expression [22], that is, to infer changes in messenger RNA (mRNA) concentration in cells or tissues, resulting from changing experimental conditions, by hybridizing labeled copies of mRNAs to arrays containing probes for specific gene transcripts. Algorithms for expression data analysis had to deal with the fact that, in most experiments, the true mRNA concentrations were unknown. Therefore, most work on expression analysis was aimed at background correction [23], normalization [24] and estimation of concentration ratios (fold changes) between different conditions [25].

Detection array analysis offers the opportunity to understand microarray behavior in much greater detail, because the samples analyzed are produced from genomic DNA. Sequences are known for many microbial genomes, and standard laboratory techniques exist to measure the concentration of DNA in a sample. Therefore, one can design experiments in which sample DNA molecules with known degrees of similarity to probe sequences are present, at a wide range of known concentrations. The wealth of information available in these experiments makes it possible to develop detection algorithms based on models, in which the probe signal given the presence of a target organism at some concentration is predicted from the probe and target genome sequences. After fitting model parameters from experiments with known samples, one can solve the inverse problem to find the targets that best explain the observed array data for an unknown sample.

Nevertheless, detection arrays present many of the same analysis issues as other types of microarrays. Probe signals must be corrected for background fluorescence of the array substrate [23] and have additional noise contributions due to transient hybridization with noncomplementary or partially matching DNA molecules [26]. Probe and target DNAs may form hairpins or other secondary structures that prevent hybridization between expected partners [27, 28] or enhance hybridization between unexpected probe–target pairs. Chemical saturation, in which most or all of the oligos in a probe feature are bound by target DNAs, creates a nonlinear relationship between target concentration and probe intensity [29]. Another source of nonlinearity is optical saturation [30], which occurs when the scanner converts the analog probe intensity to a 16-bit digital value. All intensities greater than some threshold are reported as the maximum value (65 535); thus, if the scanner photomultiplier tube gain is set too high, a substantial amount of information about the true probe intensities may be lost.

GreeneLAMP

Many current algorithms for detection array analysis follow purely empirical approaches, without trying to model the physical processes underlying hybridization, washing and scanning. The algorithm developed for GreeneChip analysis, 'log-transformed analysis of microarrays using *P*-values' (GreeneLAMP) [17], is one such approach to the species identification problem. The GreeneLAMP algorithm makes several key assumptions about array experiments:

- probe intensities are log-normally distributed;
- probe intensities represent independent measurements of target genome concentrations;
- the number of probes for any species having positive signals is limited, on the order of 100 or fewer.

When pairs of probe sequences are 95% or more identical, the independence assumption is clearly violated. In this case, the algorithm clusters probes into equivalence groups and pools their signals in an unspecified manner. Probes are associated with target taxa using BLAST; the score threshold for association is not specified, and there appears to be no attempt to differentiate between probes with strong and weak similarity.

To analyze an array experiment, the GreeneLAMP software first subtracts background levels from the probe intensities, for probes >2 SD above the mean. The background levels are derived from matched control samples when they are available and from random 60-mer control probes on the same array otherwise. The software then centers the log intensities, divides them by the SD to form Z-scores and computes tail probabilities (*P*-values) under the log-normality assumption. It then categorizes the probes as positive or negative according whether the *P*-values exceed a fixed threshold: 0.1 for arrays with matched controls and 0.023 otherwise.

Finally, *P*-values are computed for each taxon using the QFAST algorithm to combine the individual *P*-values for positive probes associated with the taxon [31]. This step depends crucially on the independence assumption. The product of the n *P*-values is used as a test statistic; its tail probability assuming independence of the *P*-values can shown to be:

$$P\left(\prod_i p_i > p\right) = 1 - p\sum_{k=0}^{n}\frac{(-\log p)^k}{k!}$$

The candidate taxa are then ranked by this combined *P*-value.

As mentioned in our discussion of the GreeneChip design, the GreeneLAMP algorithm was moderately successful in the analysis of viral samples, providing correct identification at the species level. Since the algorithm has only been applied to GreeneChip data, it is difficult to assess its performance independently from that of the chip design. The failure of the GreeneChip platform to precisely identify bacteria can be partially explained by the cross-reactivity of ribosomal RNA probes. However, an algorithm design that accounted for the greater affinity of probes for highly similar target sequences might have been able to overcome the limitations of the array design. A more severe limitation of GreeneLAMP is its inability to deal with complex mixtures, such as those found in clinical and environmental samples. Since the output of the algorithm is a single ranked list of taxa, there is no means to identify a combination of taxa that if present would best explain the observed intensity data.

E-Predict

Another empirically motivated method is the E-Predict algorithm [32], which was developed for analyzing ViroChip arrays. E-Predict computes a 'theoretical hybridization energy profile' for each complete viral genome, by using BLAST to align probes to the genome sequence, and then computing a predicted hybridization free energy for each probe having a significant alignment. Free energies, which provide a measure of affinity for a probe to bind to a target genome fragment, are computed using a nearest-neighbor stacking energy method [33], and are then scaled to produce a vector within the unit hypercube (using quadratic normalization by default):

$$\Delta G_{ij}^{(\mathrm{norm})} = \frac{\Delta G_{ij}^2}{\sum_{k=1}^{n}\Delta G_{kj}^2}$$

Here, n is the number of probes and ΔGij is the raw free energy for probe i hybridizing to target j. Probes with no BLAST hit to the target genome are assigned zero free energies. To identify the target hybridized to an array, E-Predict by default normalizes the probe intensities γ_i to sum to 1:

Alternative methods (sum, quadratic and unit vector) may be used to normalize both intensities and free energies. The normalized intensity vector is then compared to the normalized energy vector for each target in a database, using one of several similarity metrics: dot products, centered and uncentered Pearson correlation coefficients, Spearman rank correlations, or a value based on Euclidean distance. The target having the highest similarity score is then predicted to be present in the array sample.

P-values are associated with scores by comparing them to an empirical probability distribution derived from 1009 microarray experiments, which was found to be approximately log normal. The authors assume that the underlying null distribution is exactly log normal and estimate its parameters by iteratively trimming the highest score values for each virus until the remaining logged scores show the least

deviation from normality, according to a Shapiro–Wilk test. The mean and variance are computed from the remaining untrimmed values.

To be useful for analyzing clinical and environmental samples, a detection algorithm must be able to identify multiple organisms within a sample. This problem is addressed with an iterative version of E-Predict. After identifying the most likely target as described above, E-Predict sets the intensities of the probes matching that target to zero, renormalizes the intensity vector, recomputes the similarity scores and repeats the identification process.

At first glance, E-Predict appears to be motivated by a thermodynamic model, since it uses free energies to represent probe–target similarity or affinity. However, the authors do not present a physical justification for their choices of normalization and scoring functions; these were instead chosen because they provided the best separation in between-family comparisons and the least separation within families, for a particular test dataset. Therefore, one might be concerned that the algorithm might not generalize well to a wider range of datasets. Nevertheless, E-Predict has been successfully applied to identify or characterize viruses in thousands of ViroChip experiments.

VIPR

VIPR (Viral Identification using a PRobabilistic algorithm) [34] is a technique developed for analysis of viral diagnostic microarrays. It is essentially a naïve Bayes classifier, based on the assumption that probe intensities follow a log-normal distribution with one of two sets of parameters for each probe, according to whether it is predicted to bind the target in the sample (is 'on') or not ('off'); i.e. the log intensities Y_i are distributed as follows:

$$Y_i|\text{on} \sim N(\mu_{i,\text{on}},\sigma^2_{i,\text{on}}); \quad Y_i|off \sim N(\mu_{i,\text{off}},\sigma^2_{i,\text{off}}).$$

The binding predictions are obtained by calculating free energies ΔG with a nearest-neighbor approach, and treating probes with ΔG below a fixed threshold as 'on'. To give accurate results, the VIPR model must be trained using data from positive control arrays to estimate the parameters of the 'on' and 'off' intensity distributions for each probe. Priors for the latent variables in the model (the on/off states) are derived by considering the fraction of probes $P_{\text{pred}}(\text{on})$ predicted to bind a target T, along with the numbers of targets $n(\text{state})$ sharing the same state for a given probe, and applying Bayes' rule:

$$P_{\text{marg}}(\text{on}) = \frac{P(T|\text{on})P_{\text{pred}}(\text{on})}{\sum_{\text{state=on,off}} P(T|\text{state})P_{\text{pred}}(\text{state})}$$

$$P_{\text{marg}}(\text{off}) = 1 - P_{\text{marg}}(\text{on}).$$

Here $P(T|\text{state})$ is assumed to be uniform, i.e. equal to $1/n(\text{state})$.

Given the prior probabilities, the on/off distribution parameters for each probe, and the observed log intensities γ_i, VIPR computes posterior probabilities for each probe i:

$$P(\text{on}|Y_i = \gamma_i) = \frac{P(Y_i = \gamma_i|\text{on})P_{\text{marg}}(\text{on})}{\sum_{\text{state=on,off}} P(Y_i = \gamma_i|\text{state})P_{\text{marg}}(\text{state})}$$

$$P(off|Y_i = \gamma_i) = 1 - P(\text{on}|Y_i = \gamma_i).$$

Finally, VIPR combines these probabilities to calculate a posterior likelihood for each target in a list of candidate viruses (assuming conditional independence of the probes):

$$L(T) = \prod_{i:i\text{ binds }T} P(\text{on}|Y_i = \gamma_i) \prod_{i:i\text{ does not bind }T} P(\text{off}|Y_i = \gamma_i)$$

When compared to E-Predict and other published algorithms, VIPR had greater accuracy in identifying viruses hybridized to a custom hemorrhagic fever virus array. Like GreeneLAMP, VIPR is not designed to deal with complex samples, where a mixture of targets might be present. Also, the requirement that parameters be fitted to data from arrays hybridized to each candidate target limits its usefulness for broad-spectrum microbial detection arrays. However, its ability to 'learn' from additional training data means that VIPR may be more accurate than other algorithms when applied to specialized diagnostic arrays, designed to test for a limited range of species.

DetectiV

DetectiV [35] is a software package, written in the R language [36], which provides simple visualization, normalization and significance testing functions for detection array data. Unlike most of the methods discussed here, it is not tightly coupled to any particular array platform, and runs in any computing environment that supports the R language, including Mac OS X, Unix/Linux and Windows. To normalize probe intensities, DetectiV divides them by a reference intensity, which may be either the mean of a

set of designated control probes, the global median intensity for the array or the intensity of the corresponding probe on a reference array; the logarithm of the intensity ratio is then reported for each probe. Significance testing is performed by selecting groups of probes sharing common family, species, or other annotations, and computing a one-sample *t*-test, with the null hypothesis that the log intensity ratio for each group is zero.

The DetectiV software leaves interpretation of the log ratio and *t*-test results to the user; it does not correct *P*-values for multiple testing, nor does it define any threshold values for 'detection' of a particular species. Typically a user will rank families or species by *P*-value, and examine the log intensity ratios for the top n groups to decide which species are most likely to be present. DetectiV may thus be regarded as a useful package for exploratory data analysis, rather than a rigorous statistical tool. Nevertheless, the authors found that, when applied to two ViroChip data sets used in the E-Predict study [32], DetectiV gave better prediction performance than E-Predict.

PhyloDetect

PhyloDetect [37] is one of the few analysis methods that deals explicitly with the genomic similarity between taxonomically related organisms, and the consequent tendency of some probes to cross-hybridize to multiple organisms. Given a 'match matrix' $M = [m_{ij}]$, in which $m_{ij} = 1$ if probe i matches target j and 0 if not, PhyloDetect groups targets into a nested hierarchy, based on the similarity of their column vectors in the match matrix. Targets that are indistinguishable (because their match vectors are identical) are collapsed. This grouping is done once for each array design and candidate target set. To analyze an array, PhyloDetect reduces the probe intensities to binary indicators (e.g. by thresholding against the median +2 SD of the background intensities), and performs a series of hypothesis tests, one for each group in the hierarchy. Interestingly, the null hypothesis in each test is that an organism in the group is present; the alternative is that no organism in the group is present. The test statistic is based on the number of probes matching the group that have zero indicators, and a probe-independent false negative rate γ. If there are n probes matching the group, the likelihood of observing r or more probes below the detection threshold is the complement of the cumulative binomial distribution,

$$P(m \geq r) = \sum_{k=r}^{n} \binom{n}{k} \gamma^k (1-\gamma)^{n-k}.$$

This likelihood score is compared against a significance threshold α, and the group is predicted to be absent (at significance level α) if the likelihood is below α. The test is repeated for every group at every level in the hierarchy, and the scores are displayed in a tree structure format.

PhyloDetect is designed to work with data for any detection array, provided that one can construct a match matrix for its probes against a list of candidate targets; in fact, it is available as a web application provided by the authors. However, this implementation does not scale well for high-density microarrays or large candidate target sets, because the matrix must be instantiated in memory. In addition, the false negative rate parameter must be chosen carefully for each array design. The greatest strength of PhyloDetect is that its results can be easily interpreted, in the common situation where the sample contains an organism related but not identical to one or more of the candidate targets.

Composite Likelihood Maximization

The CLiMax algorithm (composite likelihood maximization) is based on a biophysical model of probe-target hybridization. It was developed to analyze data from LLMDA arrays [19], but could in principle be applied to other whole-genome arrays such as the ViroChip. CLiMax models the likelihood of the observed pattern of probe intensities as a function of the set of targets present in the sample, and follows a greedy maximization procedure to identify a locally optimal set of targets that best explains the observed intensities. The likelihood is estimated as a product of probe-specific factors, treating the probe intensities as conditionally independent variables, given the sample composition (thus, it is a composite likelihood approximation).

The factors in the CLiMax model are conditional probabilities that the probe intensity will exceed a certain threshold, given the presence of a particular target at some minimal concentration. The threshold value is determined from the distribution of negative control probe intensities on the array, and is used to distinguish target-specific signals from signals due to

nonspecific hybridization. Essentially, CLiMax reduces the probe intensities to binary variables, which are described by a logistic regression model. The apparent loss of information implied here is mitigated by the fact that, in most experiments, probe intensities fall into two distinct clusters, one near the optical saturation limit and the other near the background fluorescence level. The covariates in the logistic regression model are derived from BLAST alignments of probe sequences to targets in a database of microbial genomes: the BLAST score, the alignment start position within the probe sequence and a melting temperature computed for the aligned DNA duplex. Coefficients for the model predictors are fitted to data from hybridizations to targets with known genome sequences. The CLiMax logistic regression model also includes a term to account for the observation that probes with low sequence complexity are more likely than others to hybridize nonspecifically to dissimilar targets. It measures sequence complexity by the entropy of the frequency distribution of trimers in the probe sequence.

If X_j is a binary variable indicating whether target j is present in a sample, and Y_i is an indicator representing whether the intensity of probe i is above the detection threshold, the likelihood of X_j given the observed data $Y = \{Y_i\}$ is

$$L(X_j; Y) = \prod_{i:Y_i=1} P(\gamma_i = 1|X_j) \prod_{i:Y_i=0} P(\gamma_i = 0|X_j),$$

where

$$\text{logit}\big[P(Y_i = 1|X_j)\big] = a_0 + a_1 S_i + X_j(a_2 T_{ij} + a_3 B_{ij} + a_4 Q_{ij}),$$

provided that, at most, one target genome is present in the sample. Here, S_i is the probe sequence entropy, T_{ij} is the melting temperature for a duplex of probe i and target j, B_{ij} is the BLAST score, Q_{ij} is the position of the aligned target on the probe sequence and a_0 through a_4 are the fitted model coefficients. When multiple targets may be present, CLiMax uses the following approximation to estimate the probe detection probabilities:

$$P(Y_i = 1|X) = 1 - \prod_{j:X_j=1} P(Y_i = 0|X_j = 1).$$

To identify a set of targets that best explains the observed probe intensities, CLiMax starts with an empty target set, and then iterates through a list of candidates, at each iteration choosing the target which, when added to the present set, yields the largest increase in the likelihood score. Iterations continue until a preset maximum or until no additional increase in the likelihood is possible. Log-odds scores for each predicted target are presented graphically, as two values: its actual log likelihood contribution given the previously selected targets, and the contribution it would have yielded if it were chosen first (assuming no other targets present). An example of CLiMax results for a typical LLMDA experiment is shown in Figure 2. In the figure, viral strains annotated in red boldface are those with positive log-odds scores given the previously selected targets; the yellow vertical dashed line corresponds to a zero score. The darker shaded portion of each bar represents the log-odds score given the previously selected targets; the full length of the bar is the score obtained if the target were selected first. The four human papillomavirus strains listed first have scores substantially above zero, indicating their likely presence in the sample; these were in fact confirmed by an independent assay.

The primary weaknesses of the CLiMax approach are its assumption (shared by VIPR and other likelihood-based algorithms) that probe intensities are conditionally independent, and its greedy algorithm for selecting multiple targets. The independence assumption is violated when probes are designed against a large set of related strains that are similar but not identical. When this happens, and the sample contains DNA with a similar sequence originating from a genome that is not part of the target database, a large number of probes may cross-hybridize to it. Since the probes contribute additively to the log-odds score, a false-positive target prediction may result. These errors can be recognized by examining the locations of the positive probes in each predicted target genome and eliminating targets for which only probes in a narrow subregion are detected.

The greedy target selection algorithm may lead to erroneous predictions in some cases, when the sample contains a species for which the target database contains both complete genomes and single replicons. In this case, the algorithm will preferentially allocate log-odds contributions from positive probes to complete genomes, since they have greater numbers of probes, even though the probes may match a single-replicon sequence more closely.

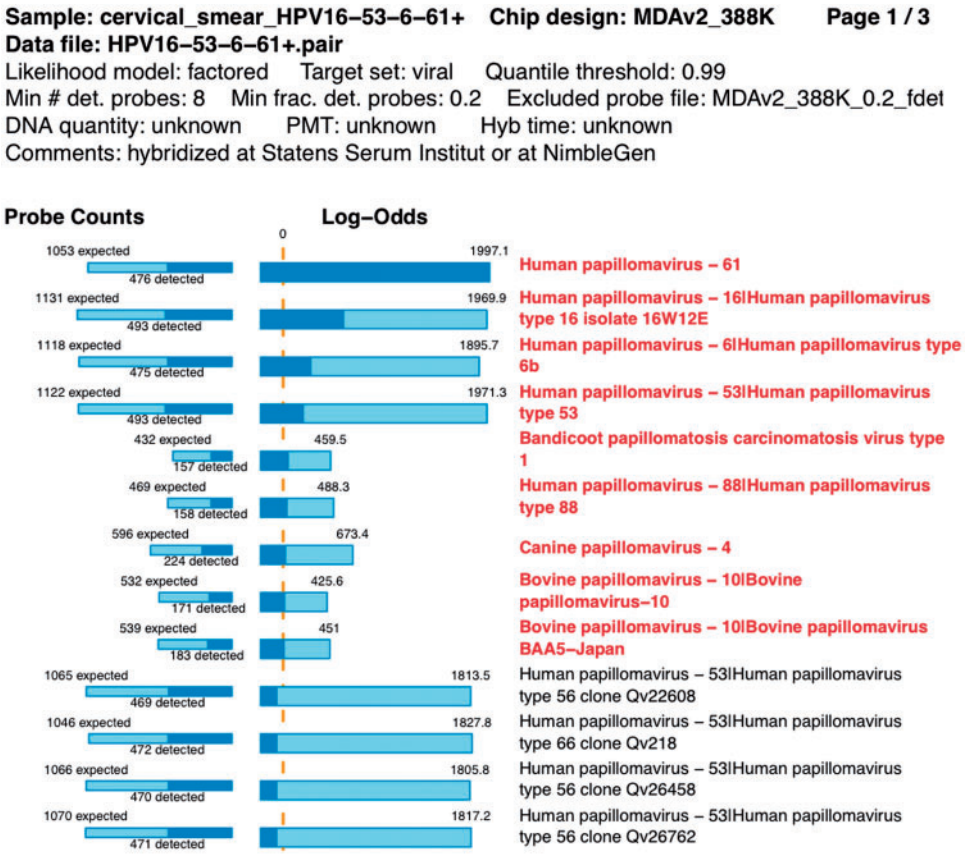

Figure 2: Results of LLMDA analysis of a cervical smear sample, indicating the presence of four types of human papillomavirus (16, 53, 6 and 61), confirmed by an independent assay.

CONCLUSION AND FUTURE PERSPECTIVES

Several promising approaches to microbial detection array design and analysis have been tested during the past decade. The array platforms vary widely in terms of fabrication cost, range of organisms targeted, sensitivity and specificity of detection. An essential component of any pathogen detection platform is an analysis algorithm that can make sense of the noisy data produced with current array technology and yield easily interpretable results. Analysis and visualization software will become especially important as microbial detection arrays move from the research environment to widespread medical, industrial and military use. Most of the analysis algorithms described in this review can be adapted to handle data from a variety of array types, and each algorithm has a unique set of strengths. An ideal analysis and visualization tool would combine the best features of the current approaches.

Further advances in array technology will also facilitate their broader use for pathogen detection. These will include higher probe densities; automated techniques for sample preparation, DNA/RNA extraction and amplification; faster hybridization times; label-free methods to detect probe–target binding; and efficient analysis algorithms that run on smart phone-size computers. Ultimately, these advances will lead to detection arrays that function as components of mobile, point-of-care devices, which can deliver results in less than an hour rather than overnight.

Improved techniques for metagenomic sequence analysis will also lead to better array designs, by allowing probes to be designed for species that do not grow in isolation under laboratory culture. Besides being able to identify a wider range of organisms, these newer designs will have better specificity, as probes that cross-hybridize to previously unknown species can be excluded.

Finally, application of detection microarrays in food safety and medical settings will require the development of tools to assess the quality of array hybridizations and the resulting data. In addition,

microarray technology will need to evolve in order to achieve the level of reproducibility needed for diagnostic use. For this to occur, we will have to improve our understanding of the complex biophysics underlying array hybridization, currently an area of active research.

Key Points

- Microarrays are emerging as a cost-effective, broad-spectrum platform for detection of pathogens in clinical and environmental samples, foods, pharmaceuticals and other products.
- A variety of array designs have been developed by researchers, some focused only on viruses and others targeting the full range of known bacterial, viral and fungal pathogens.
- Statistical analysis methods coupled to the features of each array design are a key element of each array platform.
- Application of microbial detection arrays in the clinic will require quality control procedures for array manufacture and processing, as well as analysis and data visualization methods that provide easily interpretable results.

Acknowledgements

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

FUNDING

This work was supported by Laboratory Directed Research and Development (grant number 08-SI-002) from Lawrence Livermore National Laboratory, and by the National Biodefense Analysis and Countermeasures Center (award number L164212/F0901). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the National Biodefense Analysis and Countermeasures Center (NBACC), Department of Homeland Security (DHS), or Battelle National Biodefense Institute (BNBI).

References

1. Mullis K, Faloona F, Scharf S, *et al*. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 1986;**51**(Pt 1):263–73.
2. Schena M, Shalon D, Brown P, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.
3. Sanger F, Nicklen S, Coulson A. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; **74**:5463–7.
4. Bej A, Mahbubani M, Miller R, *et al*. Multiplex PCR amplification and immobilized capture probes for detection of bacterial pathogens and indicators in water. *Mol Cell Probes* 1990;**4**:353–65.
5. Vandenvelde C, Verstraete M, Van Beers D. Fast multiplex polymerase chain reaction on boiled clinical samples for rapid viral diagnosis. *J Virol Methods* 1990;**30**:215–27.
6. Hughes T, Mao M, Jones A, *et al*. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001;**19**:342–7.
7. Pease A, Solas D, Fodor S, *et al*. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;**91**:5022–6.
8. Wang D, Coscoy L, Zylberberg M, *et al*. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 2002;**99**:15687–92.
9. Wang D, Urisman A, Liu Y-T, *et al*. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* 2003; **1**:E2.
10. Ksiazek T, Erdman D, Goldsmith C, *et al*. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1953–66.
11. Chen E, Miller S, DeRisi J, *et al*. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. *J Vis Exp* 2011. http://www.jove.com/details.php?id=2536 (25 August 2011, date last accessed).
12. Lin B, Wang Z, Vora G, *et al*. Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res* 2006;**16**:527–35.
13. Malanoski A, Lin B, Stenger D, *et al*. Automated identification of multiple micro-organisms from resequencing DNA microarrays. *Nucleic Acids Res* 2006;**34**:5300–11.
14. Altschul S, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
15. Leski T, Lin B, Malanoski A, *et al*. Testing and validation of high density resequencing microarray for broad range biothreat agents detection. *PLoS ONE* 2009;**4**:e6569.
16. Belosludtsev Y, Bowerman D, Luebke K, *et al*. Organism identification using a genome sequence-independent universal microarray probe set. *BioTechniques* 2004;**37**:654–8, 660.
17. Palacios G, Quan P-I, Jabado O, *et al*. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis* 2007;**13**:73–81.
18. Quan P-I, Palacios G, Jabado O, *et al*. Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *J Clin Microbiol* 2007;**45**:2359–64.
19. Gardner S, Jaing C, McLoughlin K, *et al*. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* 2010;**11**:668.
20. Jaing C, Gardner S, McLoughlin K, *et al*. A functional gene array for detection of bacterial virulence elements. *PLoS ONE* 2008;**3**:e2163.
21. Victoria J, Wang C, Jones M, *et al*. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J Virol* 2010;**84**:6033–40.
22. Smyth G, Yang Y, Speed T. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* 2003;**224**:111–36.
23. Kooperberg C, Fazzio T, Delrow J, *et al*. Improved background correction for spotted DNA microarrays. *J Comput Biol* 2002;**9**:55–66.
24. Bolstad B, Irizarry R, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**:185–93.

25. Irizarry R, Bolstad B, Collin F, *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;**31**:e15.
26. Zhang L, Miles M, Aldape K. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol* 2003;**21**:818–21.
27. Ratushna V, Weller J, Gibas C. Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* 2005;**6**:31.
28. Gibas C. The biophysics of DNA microarrays. *Handbook of Physics in Medicine and Biology*. Boca Raton, Florida: CRC Press, 2005, 42–1.
29. Burden C, Pittelkow Y, Wilson S. Statistical analysis of adsorption models for oligonucleotide microarrays. *Stat Appl Genet Mol Biol* 2004;**3**:Article 35.
30. Dodd L, Korn E, McShane L, *et al*. Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics* 2004; **20**:2685–93.
31. Bailey T, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 1998;**14**:48–54.
32. Urisman A, Fischer K, Chiu C, *et al*. E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* 2005; **6**:R78.
33. SantaLucia J, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 2004;**33**:415–40.
34. Allred A, Wu G, Wulan T, *et al*. VIPR: a probabilistic algorithm for analysis of microbial detection microarrays. *BMC Bioinformatics* 2010;**11**:384.
35. Watson M, Dukes J, Abu-Median A.-B, *et al*. DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol* 2007;**8**: R190.
36. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* 2011. http://www.R-project.org (29 April 2011, date last accessed).
37. Rehrauer H, Schönmann S, Eberl L, *et al*. PhyloDetect: a likelihood-based strategy for detecting microorganisms with diagnostic microarrays. *Bioinformatics* 2008;**24**:i83–9.