

Sparse Sensing DNA Microarray-Based Biosensor: Is It Feasible?

Mojdeh Mohtashemi
MITRE Corporation, McLean, VA, USA
MIT CSAIL, Cambridge, MA, USA
mojdeh@mitre.org

Haley Smith, Felicia Sutton, David Walburger, and James Diggans
Emerging Technologies
MITRE Corporation, McLean, VA, USA

Abstract—Conventional microarray-based biosensors can only detect a limited number of organisms, and adding sensor capabilities requires re-engineering of reagents and devices to detect the presence of a novel microbial organism. To overcome these limitations, the size of the microarray may need to be prohibitively large, an impractical proposition, cost-wise, using current technology. We hypothesized that a relatively small number of oligomers is sufficient to design a microarray capable of differentiating between the genomic signatures of multiple organisms. To test this hypothesis, we designed a sparse, pseudo-random prototype microarray-based biosensor by generating 12,600 25bp oligomer probes derived from a mathematical model based on random selection of DNA sequences from seven pathogenic prokaryotic genomes. To enable identification of novel organisms, a reference library of pure genomic DNA was generated from three simulant organisms that are known to be phylogenetically distant from the seven base species used to generate the probes. These simulants were combined to produce complex DNA samples meant to mimic the uncertainty and complexity of an unknown environmental genomic background. A mathematical model was then developed to capture the signature of each simulant organism. The model detected the presence of all three simulant organisms in the mixed DNA samples with high accuracy.

Keywords—biosensor; sparse sensing; DNA microarray; VLMC; PLSR; genomic signature

I. INTRODUCTION

Most biosensors can be considered closed systems in that they are built to respond to one or a small number of organisms, and are unable to react in the absence of those organisms. This is true regardless of whether the elements change by natural genetic drift or by intentional engineering of antigens [1]. While an effective approach when the target environment remains static, this framework is not particularly robust or efficient [2], as it requires creation of new physical reagents or sensor capability whenever novel or previously unencountered infectious agents are discovered. Although there have been efforts to design microarrays that are representative of groups or families of organisms, these arrays are sensitive to the presence of specific targets common amongst these groups [3]. An open system would provide data regardless of whether a particular biological event was expected, thus

allowing new microorganisms to be recognized, characterized and managed in short order.

A presumed drawback in the design of an open system for biosensing, however, is that the greater the number of biological species to be detected, the larger the array size required. Thus, to detect the presence of even a few microorganisms, either individually or in combination, the conventional wisdom dictates that the microarray would have to be very large to capture distinct genomic patterns with high degree of specificity, an endeavor that is not cost effective.

It has recently been suggested that many natural phenomena are sparse in that they can be represented in a compressed format using the proper basis [4-9]. Sparsity denotes that, to recover a signal of interest, the number of degrees of freedom needed to approximate the signal may, in principle, be much smaller than the length of the signal [6, 8]. This is the foundation for the new theory of compressive sensing or compressive sampling [6-8].

More recently, Dai, et al. have proposed that DNA microarrays can be designed using the notion of compressive sensing [10]. They used the NIH database of clusters of orthologous groups (COGs) of proteins based on sequences of 66 unicellular organisms to design microarray probes. A limitation of this approach is in the use of COGs to design probe sets rather than the entire DNA sequences of the organisms, thereby limiting the flexibility of the array in detecting novel species that lack certain clustered proteins. Furthermore, their results are based on limited laboratory generated data and experimentations. The key challenge in the design of an open biosensing system is to demonstrate that first, sparsity is an applicable and sensible notion in natural environments and thus sparse sensing can be used to characterize a large number of microorganisms. And second, that a relatively small DNA microarray, if appropriately designed, is capable of capturing the DNA signatures of multiple environmental organisms in a reliable manner.

In this paper, using laboratory data, we provide strong evidence that the underlying genomic imprints of biological organisms may indeed be sparse, and thus a relatively small codebook, or collection of microarray probes, can capture the signature of multiple biological signals succinctly and

differentiate between them when in complex mixtures. We propose and design a prototype nucleic acid-based sensor that makes use of a sparsely generated set of probes paired with mathematical models capable of recognition and classification of a broader array of organisms against a complex background like that in the natural environment.

II. PROPOSED APPROACH

Our approach consists of three layers of data generation and modeling to: (A) generate a set of probes by training a mathematical model on seven pathogenic sequences, (B) generate a reference library of hybridization patterns for three simulant organisms and for mixed samples, and (C) develop a mathematical framework for validation and identification of distinct presence of the simulant organisms in individual and mixed samples.

A. Probe Design

We utilized variable-length Markov Chains (VLMCs) [11], trained on sequences from seven prokaryotic pathogenic genomes, to generate 25-mer microarray probes. 25-mer sequences had been previously shown as a good trade-off between hybridization specificity and diversity [12]. The seven pathogenic sequences were extracted from GenBank and are listed in Table I. We randomly sampled 500 base pairs from each genome without regard for coding regions. Samples were concatenated end to end to produce a single DNA sequence, and used to train a VLMC model. An initial set of probes consisting of 100,000 unique DNA sequences with a length of 25 were then generated from the trained model. These probes were screened for a melting temperature between 58°C and 68°C and propensity for self-hybridization of ΔG (change in free energy) > -1.1 . Probes with mono-runs of guanine bases longer than three were eliminated as these probes have a propensity to form g-tetrads or pseudo-knots. The remaining probes were ranked by decreasing ΔG for self-hybridization. The top 12,600 probes were selected and 15% were randomly duplicated for quality control purposes resulting in 15,200 total probes sent to Agilent for synthesis on their 8x15k Custom Array format.

Finally, to evaluate the propensity of the VLMC-derived probes for specificity, we generated 12,600, 25-mer probes at random, and aligned both sets of probes against a panel of Gram-positive and -negative prokaryotic organisms using MPI-BLAST [13]. The specificity of each set of probes was evaluated using a metric composed of the number of BLAST hits per 1000 base pair of organism genomic sequence: “Hits/Kilobase”. As seen in Fig. 1, BLAST results for VLMC-trained probes produced at least a two fold increase in BLAST hits against each organism compared with random probes.

B. Microarray Methods and Hybridization

To hybridize against the VLMC-derived probe set and generate data, the purified genomic DNA from 3 different simulant strains, *Bacillus cereus* (BC), *Bacillus subtilis* (BS), and *Pantoea agglomerans* (PA), was fragmented and amplified using a Sigma GenomePlex® Whole Genome

Amplification kit. Amplified DNA was precipitated using sodium acetate and ethanol. DNA was labeled with ULYSIS Alexa Fluor® 546 Nucleic Acid Labeling kit (Invitrogen) and excess dye removed with an Agilent Genomic DNA Purification Module. Samples were then concentrated to 250ng of DNA in 7ul. Labeled sample was prepared for hybridization using an Agilent Oligo aCGH Hybridization kit and loaded onto Agilent 8x15K Custom Arrays. Arrays were hybridized for 16 hours at 42°C, and then washed (Agilent Oligo Wash Buffer Kit) and scanned on a Molecular Devices GenePix 4100A. Analysis was done using Agilent Feature Extraction software v9.5.3.1.

Ten technical replicate arrays were generated for each simulant species, resulting in a total of 30 arrays for training and validation of the detection model (Table II). Next, spike-ins of short oligos was designed to bind to specific probes of the array. The spike-in acted as a positive control for the arrays. Two arrays were run to determine an optimum spike-in concentration for the arrays: 1% and 0.1% of total DNA concentration. Spike-in was then added at a 1% concentration to each single species array. Finally, 8 mixed samples were prepared based on 4 possible combinations of three single genomes (2 arrays per combination) in equal ratio for a total of 250 ng per array (Table II). The mixed samples were labeled as: BC/BS/PA_1, BC/BS/PA_2, BC/BS_1, BC/BS_1, BC/BS_2, BS/PA_1, BS/PA_2, BC/PA_1, and BC/PA_2.

TABLE I. PATHOGENIC SEQUENCES USED TO GENERATE PROBES

Species	Pathogenicity	GenBank ID
Bacillus anthracis (Ames strain)	Anthrax	NC_003997
Yersinia pestis (CO92)	Bubonic plague	NC_003143
Francisella tularensis (Schu 4)	Tularemia	NC_006570
Brucella suis	Brucellosis	NC_004310
Burkholderia mallei	Glanders	NC_006348
Burkholderia pseudomallei	Melioidosis	NC_006350
Escherichia coli O157 H7 str. Sakai	Hemolytic uremic syndrome	NC_002695

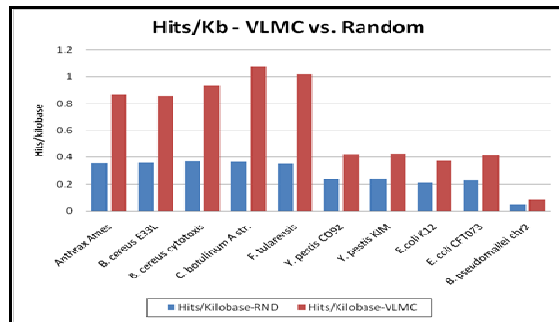


Figure 1. Specificity in Hits/Kilobase of the VLMC trained vs. random probes against a panel of gram negative and positive prokaryotic organisms

TABLE II. EXPERIMENTAL DESIGN

Genomic DNA	# Arrays	gDNA
<i>B. subtilis</i>	10	250 ng
<i>B. cereus</i>	10	250 ng
<i>P. agglomerans</i>	10	250 ng
<i>B. subtilis</i> / <i>B. cereus</i>	2	125 ng/species
<i>B. subtilis</i> / <i>P. agglomerans</i>	2	125 ng/species
<i>B. cereus</i> / <i>P. agglomerans</i>	2	125 ng/species
<i>B. cereus</i> / <i>B. subtilis</i> / <i>P. agglomerans</i>	2	84 ng/species
Oligo spike-ins	2	2.5 ng and 25 ng

C. Detection Model

A multivariate mathematical model using partial least squares regression (PLSR) was developed to effectively capture the signature of each simulant organism. Briefly, given a number of predictor, or independent, variables, PLSR iteratively finds the best fit for one or more response by achieving correlation between the two [14-16]. By constructing new predictor variables, or latent variables, as linear combinations of the original variables, PLSR seeks to maximize correlation between the response and predictor variables while capturing and explaining most of the variation within the predictor variables.

Here, the predictor variables are the 30 single species arrays hybridized against the set of probes using the three simulant organisms (ten arrays each), and the response variables are the 8 mixed samples hybridized against the probe set using 4 possible combinations (two arrays each). There were 12,600 hybridization measurements made on all variables, resulting in a 12,600×30 matrix of observations on the predictor variables (single species), and a 12,600×8 matrix of observations on the predicted variables (mixed species).

The first two latent variables from the PLSR model achieved maximum correlation with the response variables while together they captured most of the variation in the predictor variables (>80%). Thus, the signature of each simulant organism and its contribution to each test sample was derived from the corresponding regression coefficient derived from the PLSR model based on the first two latent variables. The goodness of fit of the model for each test sample was determined using the R^2 statistic which is the normalized value of the total squared error explained by the model. Finally, to determine which probes are critical in differentiating between patterns of hybridization, the contributing value of each probe to the goodness of fit was assessed using the Hotelling's T^2 statistic, a multivariate measure of variation in each row of observations per probe.

III. RESULTS

The PLSR model was first validated using single species arrays by iterative leave-one-out cross validation. Briefly, every time one array (from the set of 30 single species arrays) was randomly selected as a test sample and excluded from training data. The model was then trained on the remaining 29 arrays and two oligo spike-in arrays, and tested on the array that was left out. The experiment was repeated 200 times and

the average value of results was reported. As illustrated in Fig. 2, all three simulant organisms were classified with high degree of specificity (mean(R^2) = 0.96, CI = 0.95). The percentage of contribution as depicted on the y-axis represents the specificity or amount of contribution of each organism to the test sample as explained by the model. Recall from the previous section that short oligo spike-ins were added to each single species sample, which the model identifies in the validation step.

The model was then trained on all 30 single species arrays and two oligo spike-in arrays, and tested on 8 mixed samples. As depicted in Fig. 3, the signatures of individual organisms contributing to each mixture (true positive) was captured in all 8 samples (mean(R^2) = 0.76, CI = 0.95). In two BCPA samples (the last two stacked bars in Fig. 3), however, the signature of the third organism, BS, was incorrectly detected at small concentration (false positives). While this may be due to in-vitro hybridization errors, more experiments are needed to investigate the cause.

To determine the contribution of each probe to the process of capturing the genomic signature of each organism, and thus guide future probe design based on sparse sensing, probe values were assessed using the Hotelling's T^2 statistic for each mixed sample in the model. Because this may result in a different ordering of probes for different mixed samples, the average value of each probe was calculated and probes were sorted in descending order of their average T^2 statistic. The PLSR model was then run iteratively, each time adding the next top 200 probes to the model until all 12,600 probes were included. The average value of the R^2 statistic was then recorded as a function of number of sorted probes included in the model. Fig. 4 illustrates the cumulative R^2 curve as more probes are included in the model. Note that over 70% of the R^2 value is achieved using the top 7,000 probes, while the remaining 5,200 probes contribute less than 6% to the average goodness of fit. It is also important to note the plateau effect of the cumulative R^2 curve, implying that the fit cannot be further improved upon using the remainder of the probe set and after about 11,500 probes are included in the model.

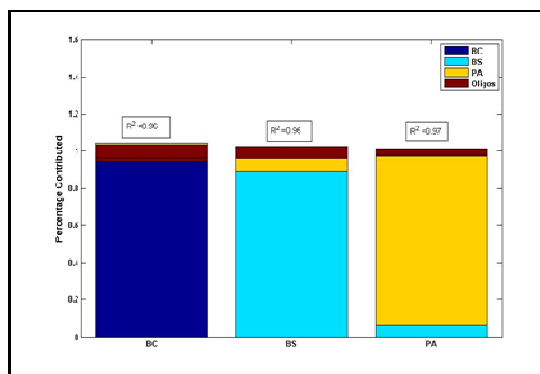


Figure 2. Validation of the PLSR model using single species arrays by iterative leave-one-out. All three simulant organisms were classified correctly with high R^2 values.

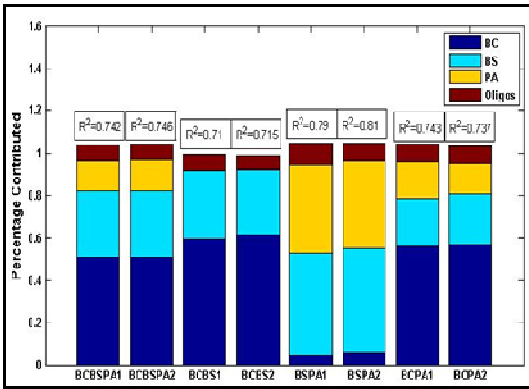


Figure 3. The PLSR model was tested on 8 mixed genomic DNA.

IV. CONCLUDING REMARKS

We hypothesized that the underlying genomic imprints of biological organisms are sparse, and thus can be represented in a compressed format using a relatively small DNA microarray as a codebook. We then designed a prototype DNA microarray using known sequences of seven pathogenic species and generating probes by simulating VLMCs. We tested this hypothesis on three simulant organisms and their mixed samples by laboratory experimentation, and developing a mathematical model to analyze the resulting data. We provided strong evidence that a relatively small set of randomly generated probes, paired with a mathematical model, was capable of capturing the signature of each organism and detecting its presence in mixed samples under a controlled environment.

Two observations are worthy of note here. First, the observed plateau effect of the cumulative R^2 curve in Fig. 4 indicates that the majority of the generated probes are valuable for decoding the hybridization patterns in mixed samples, and only about 1,000 probes with lowest T^2 values are not capable of improving upon the goodness of fit. Second, nearly 73% of the cumulative R^2 value in Fig. 4 is achieved using the top 7,000 probes, while the rest of the probes contribute only about 3% to the R^2 value. These observations, together with the results of Fig. 3, substantiate the validity of the conjecture that sparsity may be a common phenomenon in the biological domain, and must be exploited toward the efficient design of biosensors.

Future direction includes improvement to the design of microarray probes as guided by the analysis and experimentation in Fig. 4, expansion of the reference library to encompass additional test organisms, and performing environmental testing by external air sampling and genomic recovery against a complex environmental background.

ACKNOWLEDGMENT

This work was fully funded by the MITRE Corporation.

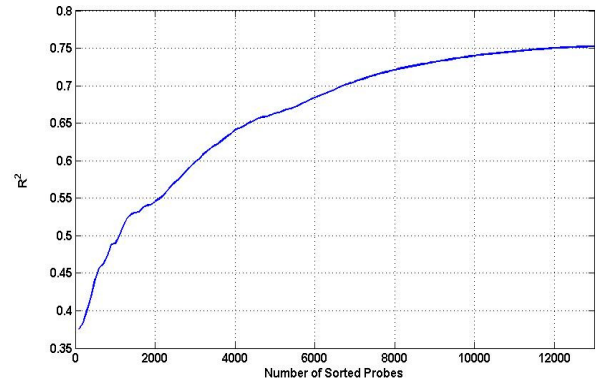


Figure 4. Goodness of fit of the PLSR model based on number of sorted probes with decreasing value of T^2 statistic. The value of each probe is averaged over all mixed samples.

REFERENCES

- [1] A. P. Pomerantsev, N. A. Staritsin, Yu. V. Mockov and L. I. Marinin, "Expression of cereolysine AB genes in Bacillus anthracis vaccine strain ensures protection against experimental hemolytic anthrax infection," *Vaccine*, **15**, 1846-50, 1997.
- [2] A. Sabelnikov, V. Zhukov, & R. Kemp, "Probability of real-time detection versus probability of infection for aerosolized biowarfare agents: A model study," *Biosens Bioelectron*, vol. 21, pp. 2070-7, 2006.
- [3] D. Wang, L. Coscoy, M. Zylberberg, P. C. Avila, H. A. Boushey, D. Ganem, et al., "Microarray-based detection and genotyping of viral pathogens," *Proc Natl Acad Sci*, vol. 99(24): pp. 15687-92, 2002.
- [4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction," *IEEE Trans. On Info. Theory*, vol. 52(2), 489-509, 2006.
- [5] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59(8), 1208-1223, 2006.
- [6] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Info. Theory*, vol. 52(4), 1289-1306, 2006.
- [7] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, 2008.
- [8] E. J. Candès and M. B. Wakin, "Introduction to compressing sampling," *IEEE Signal Processing Magazine*, 2008.
- [9] R. Berinde and P. Indyk, "Sparse recovery using sparse random matrices," unpublished.
- [10] W. Dai, M. A. Sheikh, O. Milenkovic and R. G. Baraniuk, "Compressive sensing DNA microarrays," *EURASIP Journal on Bioinformatics and Systems Biology*, pp. 1-12, 2009. doi:10.1155/2009/162824
- [11] M. Mächler, "Variable Length Markov Chains: methodology, computing, and software," *Journal of Computational and Graphical Statistics*, vol. 13, pp. 435-455, 2004.
- [12] S. Rimour, D. Hill, C. Milton and P. Peyret, "GoArrays: highly dynamic and efficient microarray probe design," *Bioinformatics*, vol. 21(7), pp. 1094-103, 2005.
- [13] A. Darling, L. Carey and W. Feng, "The design, implementation, and evaluation of mpiBLAST," *4th International Conference on Linux Clusters*, 2003.
- [14] S. Wold, "Nonlinear PLS modeling II: spline inner relation (SPL_PLS)," *Chemom. Intell. Lab. Syst.*, vol 14, 1992.
- [15] P. Geladi and B. R. Kowalski, "PLS tutorial," *Anal. Chim. Acta.*, vol. 185(1), 1986.
- [16] A. Lorber, L. E. Wangen and B. R. Kowalski, "A theoretical foundation for the PLS algorithm," *J. Chemometrics*, vol. 1(19), 1987.