AFRL-IF-RS-TR-2005-46
Final Technical Report
February 2005

IDENTIFICATION OF BIOLOGICAL WARFARE (BW) THREAT AGENTS USING DEOXYRIBONUCLEIC ACID (DNA) MICROARRAYS

Science Applications International Corporation

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. IDEN

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2005-46 has been reviewed and is approved for publication

APPROVED: /s/

THOMAS RENZ
Project Engineer

FOR THE DIRECTOR: /s/

JAMES A. COLLINS, Acting Chief
Advanced Computing Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** FEBRUARY 2005	**3. REPORT TYPE AND DATES COVERED** Final Jul 02 – Jul 04

4. TITLE AND SUBTITLE
IDENTIFICATION OF BIOLOGICAL WARFARE (BW) THREAT AGENTS USING DEOXYRIBONUCLEIC ACID (DNA) MICROARRAYS

5. FUNDING NUMBERS
C - F30602-02-C-0136
PE - 61101E
PR - IDEN
TA - TB
WU - 01

6. AUTHOR(S)
Paul R. Schaudies

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Science Applications International Corporation
10260 Campus Point Drive
San Diego California 92121

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Advanced Research Projects Agency	AFRL/IFTC
3701 North Fairfax Drive	26 Electronic Parkway
Arlington Virginia 22203-1714	Rome New York 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2005-46

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Thomas Renz/IFTC/(315) 330-3423/ Thomas.Renz@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

We developed a bioinformatic method for the identification of unique sequences within the large genomes of bacteria. In support of this contract, we have generated unique sequence and microarray data demonstrating species-level discrimination between Bacillus anthracis, vaccinia virus and Yersinia pestis and strain-level discrimination between Escherichia coli 0157:H7 and the non-virulent K-12 strain. By assaying for the presence of: 1) unique sequences at various levels of the phylogenetic tree, 2) virulence genes, 3) antibiotic resistance genes, 4) virulence plasmid sequences and 5) ribosomal genes, we are in a unique position to develop a novel method for the identification and characterization of microorganisms. The application for this technology crosses many fields of research that are important to the Government including: Environmental testing for bioterrorism by the Department of Homeland Security, Forensic analysis by law enforcement agencies, Battlefield testing by the armed forces, Water quality testing by the EPA, Human diagnostics by the CDC and Agricultural testing by the USDA. In addition to the use of this method to identify known pathogens, we believe that our method will also permit the identification of naturally occurring virulent variants of non-pathogenic organisms, which may constitute an emerging threat to public health and the ability to detect genetically engineered pathogens. By utilizing many different organism-specific sequences for each organism assayed, coupled with statistical analysis methods, false positives simply become a phenomenon of past technology. This research represents a significant leap in technology for the identification and characterization of microorganisms.

14. SUBJECT TERMS
Biological Agent Detection, DNA Microarrays, Strain Discrimination of Biological Warfare Agents

15. NUMBER OF PAGES
24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

List of Tables

Acknowledgments

Science Applications International Corporation (SAIC) and Dr. R. Paul Schaudies, the PI on this project, wish to gratefully acknowledge Dr. Sri Kumar of the Defense Advanced Research Projects Agency (DARPA) for his support of this work and the efforts of the many employees of DARPA that made this project successful.

Dr. Schaudies would like to acknowledge SAIC for supplying internal research and development funding, which made the initial development of the unique sequence identification software possible. Without this bioinformatic capability, the work performed under this contract would not be possible.

Dr. Schaudies also wishes to acknowledge the efforts of Dr. Doreen Robinson, Dr. Gregory Eley and Dr. Joseph Vockley for their commitment to this project, without which, the advancements made by SAIC in the area of bioinformatics and microbial analysis would not be possible.

Introduction

Approximately 400 microbial genomes have been sequenced and entered into public domain databases. In spite of this wealth of genomic information, DNA based methods for the identification of human pathogens and the detection of environmental microbes are limited to the identification of a few species-specific genomic sequences from relatively few organisms. Bacterial genomes typically contain between 1 x 10^6 and 1 x 10^7 nucleotides. Their large size is a limiting factor in the ability to identify unique sequences within their genomes because mathematical algorithms that identify unique sequences become increasingly complex as the size of the genome increases (1-3). Unique sequences, identified from viral genomes, which may be more than three orders of magnitude smaller than bacterial genomes, have been used to identify and characterize viruses on a microarray platform (4,5). The search for unique sequences that may be used to identify bacteria is currently limited to highly conserved sequences such as the ribosomal genes (6-9) or to toxin genes that are responsible for the pathogenicity of specific microorganisms (10-12). The large size and number of bacterial genomes make it difficult and costly to manufacture microarrays that contain every possible oligonucleotide from every sequenced organism. We have developed a method that identifies regions of unique genomic sequence from large microbial genomes. These unique sequences may be utilized for the identification and characterization of biological threat agents, genetically manipulated microbes and medically or agriculturally significant bacteria and viruses. Our method utilizes a modified BLAST search algorithm (13,14) to identify unique regions of genomic DNA and RNA. Unique regions can be used as amplification targets in PCR based detection platforms (15-22) or used to generate unique oligonucleotides with specific hybridization characteristics for use on a microarray platform. Accurate strain level identification of an organism is best accomplished by combining information from unique genomic sequence with other strain-specific identifiers such as toxin genes, virulence plasmids, antibiotic resistance genes and ribosomal RNA gene sequences. Although the unique sequence data can be used with any detection platform, utilizing multiple oligonucleotide markers per organism, on a microarray format, significantly reduces the false positive detection rate.

Methods and Procedures

Unique Sequence Identification: Our bioinformatic method utilizes a BLAST search algorithm to identify unique regions of genomic DNA. A database of genomic sequence is created which contains all DNA sequences available in the public domain. This database is stored on a local computer. A single genome is retrieved from the database in FASTA format. All sequences from the genome of interest are removed from the sequence database. The genomic sequence of interest is fragmented and used as the query sequence in a BLAST search against the local database from which it was obtained. The BLAST output file is parsed to identify fragments lacking similarity to other organisms. For this application an E value of e^{-5} is used as the cutoff score for the BLAST search. Scores greater than e^{-5} are considered to be unique, while those with a score lower than e^{-5} are not. Unique regions are fragmented further and the BLAST search is repeated until the search does not identify any more non-unique sequences. Once all genomic sequences have been characterized, amplification primers and oligonucleotides for manufacturing microarrays are identified from the unique genomic fragments using commercially available primer design bioinformatic tools.

Microarray Manufacturing and Hybridization: Microarrays were manufactured by Combimatrix Inc. (Mukitteo, WA). Microarrays were hybridized and washed according to protocols provided with the microarrays. Hybridizations were performed at 45^0C in hybridization buffer provided by Combimatrix, Inc. 5ng of labeled DNA was hybridized to each array.

Fluorescent Probe Generation: Probe was generated by the random incorporation of Cy3-dCTP during a Klenow labeling reaction. 50ng of genomic or environmental DNA labeled with an Invitrogen Klenow Labeling kit (Invitrogen Inc., Carlsbad, NM) for 2 hours. Labeled DNA was purified through a YM30 column (Millipore Inc., Billerica, MA) as recommended by the manufacturer.

Microarray Analysis: Combimatrix microarrays were scanned on an Axon fluorescent scanner at PMT settings of 500, 750 and 1000. Hybridization intensity at 532nm was determined for each oligonucleotide on the array and visualized with Spotfire visualization software (Spotfire Inc, Somerville, MA). Each oligonucleotide is present in duplicate on the arrays, which are hybridized in triplicate for each sample. Microarrays contain internal hybridization controls used to perform rigorous quality control by Combimatrix after microarray synthesis. The same controls are used to monitor the quality of hybridization.

Results

Whole Genome Amplification

Whole genome amplification was performed in the presence of Cy-3 dCTP with the intent that amplification and fluorescent labeling of the genomic amplicons be performed simultaneously. **Figure 1** is an agarose gel of randomly amplified DNA from *B. subtilis*, *C. perfringens* and *E. coli*. This figure demonstrates that the amplification occurs at the various concentrations cy3-dCTP in each of the organisms tested, however, optimal incorporation of the fluorescent dye occurs at 40uM concentration. This is evident by the difference in color of the high molecular weight smear in the 40uM lanes in the gel, from green to a shade of yellow. This level of incorporation is significant because the ability to detect unique sequence on a microarray is directly proportional to the amount of label incorporated into the amplification product.

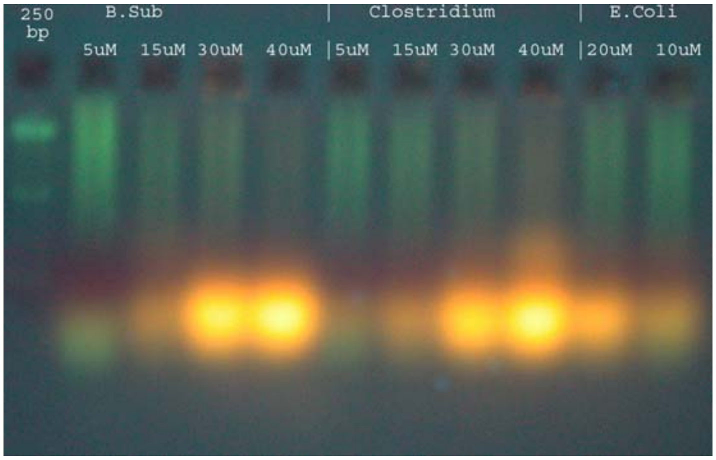


Figure 1. Random amplification and labeling of genomic DNA: Maximum incorporation of cy3- dCTP into the amplification product occurs with a cy3-dCTP concentration of 40uM demonstrated by the color shift in the high molecular weight smear on this agarose gel.

To further demonstrate the ability to randomly amplify genomic material lambda phage DNA was digested with the restriction enzymes to generate the banding pattern seen on the agarose gel on the left in **Figure 2**. This DNA was transferred to a nitrocellulose membrane. A small amount of the digested DNA was used as the template for a random amplification reaction. The random amplification was performed incorporating a biotin label into the amplified product. The amplified product was hybridized to the nitrocellulose membrane containing the digested DNA and detected with a streptavidin conjugate. The presence and relative intensity of each band on the nitrocellulose (right side of **Figure 2**) matches the intensity of each band on the agarose gel on the left. This demonstrates that each of the fragments of the digested DNA was labeled equally indicating that the labeling was random and not dependent on the size of the DNA fragment. This is important when considering that during an amplification reaction, DNA fragments of all sizes are generated. If the amplification were preferential for a particular size range then the amplification would most likely not be random and some of the sequences would be under or over represented in the final amplified product.

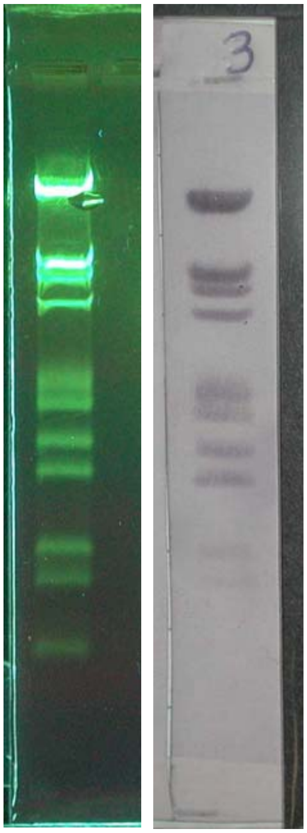


Figure 2. Random amplification of genomic DNA: An agarose gel of restriction enzyme digested Lambda DNA (left) and a southern blot of the same fragmented DNA hybridized with randomly amplified DNA. The relative intensities of the bands on the southern blot match the intensities of the bands on the agarose gel. This indicates that amplification is not dependent on the size of the DNA fragment.

Unique Sequence Analysis

Unique sequence was generated for each of the organisms listed in **Table 1**. The fifth column in that table presents the number of bases of unique sequences available for microarray design. Even though the percent of unique sequence for some of these organisms is below 1%, the size of the genome ensures that sufficient sequence is available to generate enough oligonucleotides for microarray design. An oligonucleotide design software, Oligo 6, was used to identify oligonucleotides that would be useful for probes on a microarray. The length of each oligo is 50 nucleotides long with a melting temperature of 70^{0}C.

Organism	Accession Number	# unique sequences	Average size (bp)	Total unique seuqence (bp)	Size of genome (bp)	% unique
Bacillus Anthracis	NC_003995	91	500	39,000	5,093,554	0.77
Yersinia Pestis	NC_004088	15	1000	15,000	4,600,755	0.33
Brucella Melitensis Chromosome 1	NC_003317	7	160	1,300	2,117,144	0.06
Clostridium perfringens	NC_003366	581	600	360,000	3,031,430	11.88
Escherica coli O157:H7	NC_002695	231	500	130,000	5,498,450	2.36
Escherica coli K12	NC_000913	55	500	27,000	4,639,221	0.58
Vaccinia	NC_001559	8	160	1,400	191,737	0.73
Ebola	NC_003549	4	1000	4,000	18,959	21.10
Eastern Equine Encephalitis Virus	NC_003899	2	1000	2,000	11,675	17.13
Francisella tularensis pOM1 Plasmid	NC_002109	4	600	2400	4,442	54.03

Table 1. Unique sequences generated for 10 organisms: Due to the larger size of the bacterial genomes there is more total unique sequence available for oligonucleotide design in bacteria than in viruses.

Oligos identified with this software were used to manufacture spotted microarrays. The array presented in **Figure 4** contains unique oligonucleotides from vaccinia virus, *E. coli* O157:H7, E. coli K12 and *C. perfringens*. This array was hybridized with an amplicon that corresponded to a region from which unique oligos were identified. The specific pattern of hybridization to the *C. perfringens* oligos in the upper left portion of the array demonstrates the ability to identify genomic DNA by oligonucleotide hybridization. In addition, the missing spots within this cluster of oligos that hybridized to the genomic material acts as a warning that not all oligos will hybridize to the genomic material. This has a great deal to do with the ability of the Oligo 6 program to correctly identify oligonucleotides that have matching melting temperatures and the lack of secondary structure. These data act as a warning that as good as a bioinformatic tool is, laboratory validation is a critical component of the microarray development process. The fact that 13 of the 15 spots accurately hybridized to the array demonstrates that the process of labeling and hybridization to a microarray works.

Unique sequence from the O157:H7 and K12 strains of *E. coli* was identified using SAIC’s unique sequence identifier software, now called FIGUR (Fast Identification of Genomic Unique Regions). This software tool was automated under a contract with the FBI however; the function of the automated software remains identical, in practice, to the software that was originally developed by SAIC. The unique sequence analysis for O157:H7 and K12 was performed with SAIC internal funds however the subsequent steps for the identification and validation of unique oligonucleotides was performed under DARPA funding.

Unique sequence was identified from O157:H7 and K12. **Figure 3** is a graphic representation of the unique sequence available for O157:H7. In this figure, green represents non-unique regions of the genome and yellow represents a mixture of smaller unique and non-unique regions that are below the resolution of the computer screen. These yellow regions, when zoomed in, actually contain red and green bands representing the unique and non-unique sequence from which they were derived. Large blocks of unique sequence would appear to be red on this graphic representation.

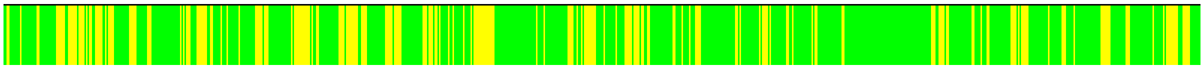


Figure 3. Unique sequence graphic representation for E. coli O157:H7: This graphic represents the genome of *E. coli* O157:H7. Unique sequences are contained within the yellow bands of the genome while non-unique are contained in the green regions. Unique sequences make up 7.55% of the *E. coli* O157:H7 genome. Oligonucleotides generated to the unique regions of the genome are used to manufacture microarrays for the discrimination of this strain from other strains. These unique regions are strain specific.

There was 7.55% unique sequence in the genome of *E. coli* O157:H7 when this genome was compared to the genomic sequence of all other organisms in the NCBI database, to include several other strains of *E. coli*. Although 7.55% is not a large percentage, it is what one might expect when comparing organisms for strain level differences. It has been our experience that when an organism is the single entry in the NCBI for its genus/species, the amount of unique sequence identified can be as high as an order of magnitude greater. 7.55% of the *E. coli* genome represents 425,377 bases of unique sequence in approximately 1000 fragments from which to identify unique oligonucleotides. The software identified a similar amount of unique sequence from the K12 strain.

Unique Oligonucleotide Identification/Array Construction

The 7.55% unique sequences of the genome represent a significant pool of sequence from which to derive unique oligonucleotides for array construction. We identified approximately 150 oligonucleotides that contain 50 nucleotides of unique sequence, with a Tm of 70^0C from the unique genomic sequence available. 150 oligonucleotide probes were identified for K12 and *Clostridium perfringens*. *C. perfringens* will act as our negative control on the microarrays. In addition to the 150 oligonucleotide probes generated from the unique genomic regions of O157:H7, we identified oligonucleotides from the virulence genes contained within the genome of O157:H7. All of these oligonucleotides as well as oligonucleotide probes for 5 other organisms were used to manufacture arrays. The construction of the arrays was preformed by Combimatrix Inc, who has developed a method of *in-situ* oligonucleotide array manufacturing. **Figure 4** is an example of a hybridized array. Each array can contain 902 oligonucleotides in addition to approximately 100 standard QC oligonucleotides present on the array to monitor quality of manufacturing and hybridization.

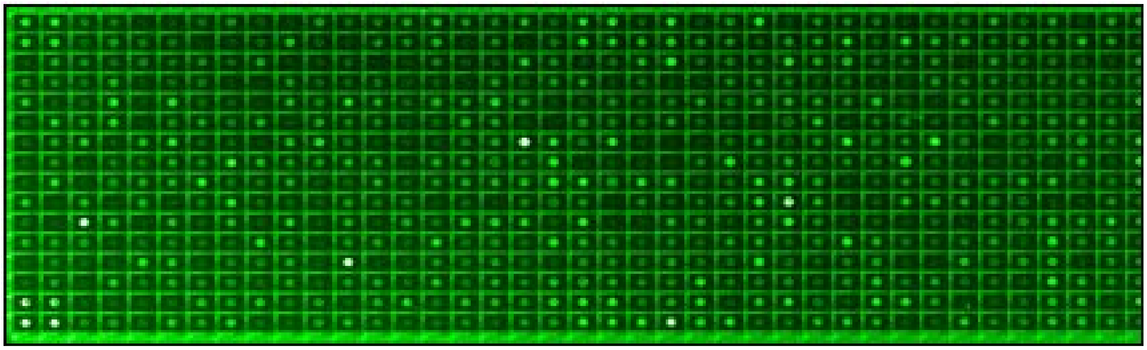


Figure 4. Hybridized microarray. A microarray hybridized with cy-3 d-CTP labeled DNA containing 902 unique oligonucleotides and 100 oligonucleotide standards. Oligonucleotides that are recognized by the probe “light-up” green. The brighter the spot, the more likely that a complimentary piece of DNA is present in the genome. Direct analysis of these data with software designed to detect subtle differences in hybridization intensity yields informative data that demonstrates the quality of hybridization.

Discrimination Between E. coli K12 and O157:H7 by Microarray Analysis.

Microarrays containing *E. coli* K12 and O157:H7 50mer oligonucleotide probes to the genomes of both organisms, oligonucleotide probes specific for virulence genes from O157:H7 and 50mers specific for the unique regions of the *Clostridium perfringens* genome were hybridized with fluorescently labeled DNA purified from *C. perfringens, E. coli* K12 and *E. coli* O157:H7. Fluorescent probe was generated from these genomes by the random incorporation of cy3-dCTP in a klenow reaction. A popular method of displaying microarray data is through the use of a scatter plot. The scatter plots in **Figure 5** represent the comparison on arrays hybridized with O157:H7 and K12, virulence gene hybridization and the hybridization of C. perfringens DNA. The two lobed pattern in panel A is typical of a hybridization to an array that contains oligonucleotides derived from the genomes of two organisms for which DNA sequence is available in the NCBI databases. Microarrays manufactured with oligonucleotide probes and hybridized with genomic DNA from the same organism clearly represent the best-case scenario for the identification of an organism. The scatter plot presented in panel A represents data obtained from an array designed to screen through many oligonucleotide probes in order to identify the most informative probes. The spots on the array that are circled in yellow represent the oligonucleotides that have the greatest differential intensity of hybridization and are thus the most informative oligos for the discrimination between E. coli O157:H7 and K12. The bar graph in panel B presents the most informative oligos for O157:H7 and K12. Panel C presents the hybridization intensity of the virulence genes found in O157:H7. These various colors of spots each represent a different virulence gene. The scatter plot in panel D is the comparison of two *C. perfringens* hybridizations. The blue spots represent ribosomal genes while the green spots represent *C. perfringens* unique genomic regions. The large number of multi-colored spots in the bottom left corner near the origin represent the *E. coli* specific oligonucleotides. There is no cross hybridization between *E. coli* and to the probes for the five other organisms represented on the microarray.

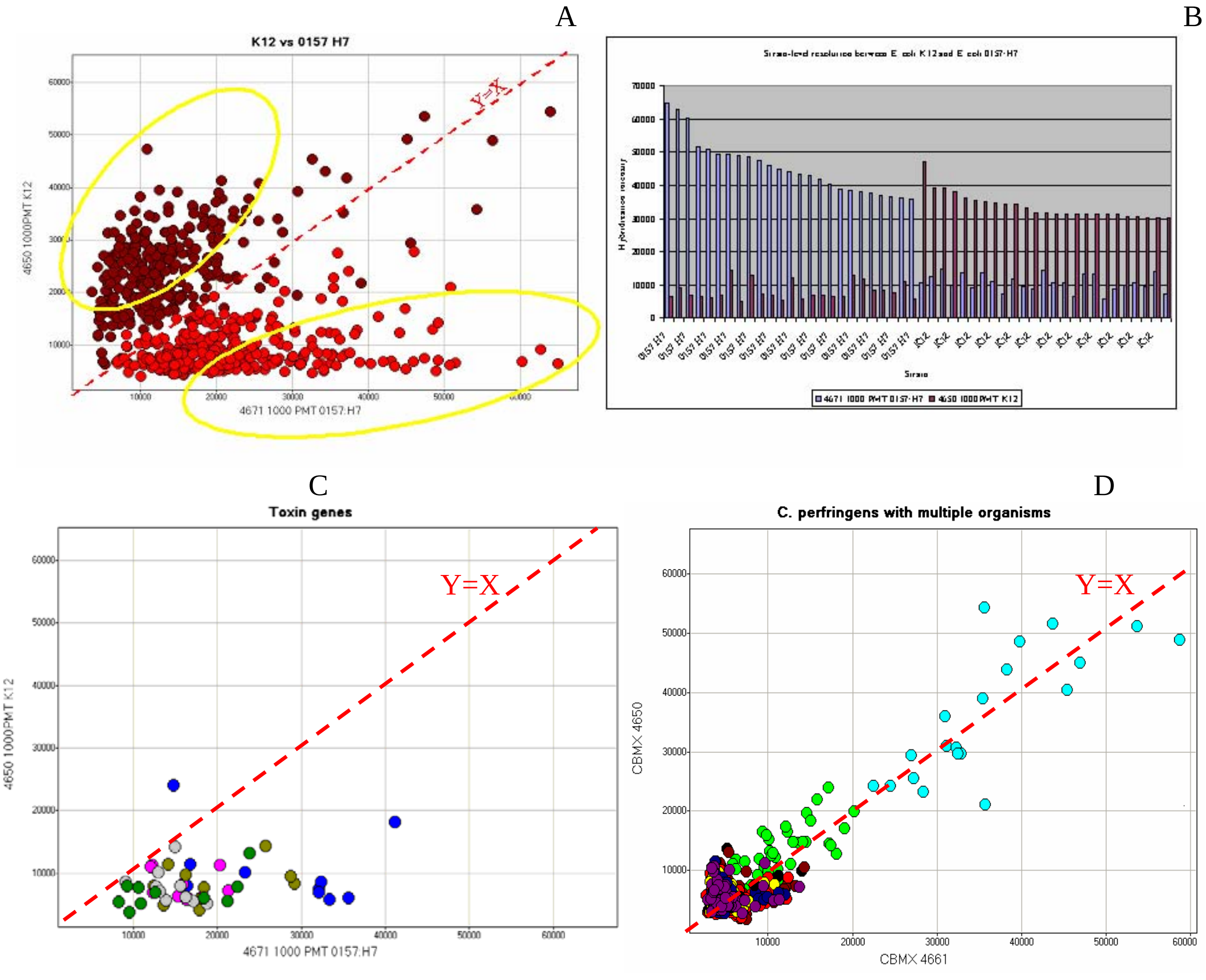


Figure 5. E. coli strain resolution and Clostridium perfringens ribosomal RNA gene detection: Panel A. is a scatter plot of intensity data for two identical high-density screening arrays, one hybridized with *E. coli* O157:H7, the other with *E. coli* K12. *E. coli* O157:H7 sequences are red. *E. coli* K12 sequences are brown. The yellow circles are placed around the spots representing oligonucleotides that would be used on a second-generation array. The bar graph in panel B. presents the hybridization data for the 25 oligonucleotides (located within the yellow highlighted area of panel A.) that have the greatest differential of hybridization between the two organisms. Panel C. presents the virulence genes from a hybridization with *E. coli* O157:H7 DNA vs. *E. coli* K12. Panel D. presents a scatter plot of an identical array hybridized with *Clostridium perfringens* DNA. The light blue spots represent the hybridization intensity for the ribosomal genes for *Clostridium perfringens*. The greens spots represent the hybridization intensity for *C. perfringens* chromosomal specific sequences. The remaining spots near the origin represent the intensity of hybridization of *E. coli* specific-oligonucleotides illustrated in panel A. and C. to the *C. perfringens* Cy-3 dCTP labeled probe.

Array Specificity and Complex Backgrounds

Arrays containing probes for multiple organisms can be used to determine the sensitivity and specificity of the microarray platform. The array utilized in the above experiment contains oligonucleotide specific to numerous organisms. **Table 2** outlines the array design. This array, when hybridized with either *E. coli* strain or with a probe to *C. perfringens* did not hybridize to the oligos designed to other organisms (**Figure 5**).

E. coli genomic sequences:
151 unique sequences for *E. coli K12*
152 unique sequences for *E. coli 0157:H7*

E. coli Toxin Genes:
5 unique sequences each:

- AidA1
- EivE
- Shiga 2 sub A
- Shiga 1 sub A
- Tox B

Other organisms:
25 *B. anthracis* unique sequences
15 *B. melitensis* unique sequences
21 *C. perfringens* unique sequences
10 *C. perfringens* 16S unique sequences
27 *Vaccinia* unique sequences
25 *Y. pestis* unique sequences

451 total sequences, in duplicate.
902 spots on array

Table 2. E. coli Array Design: This array contained *E. coli* strain specific oligonucleotides, toxin specific oligonucleotides and oligonucleotides to five other organisms. The lack of cross hybridization of genomic probes for *E. coli* O157:H7, K12 and *C. perfringens* demonstrates the specificity of each oligonucleotide for the organism from which the sequence was generated.

The ability of a complex background to hybridize to oligonucleotides designed to detect specific organisms was tested by hybridizing arrays with fluorescently labeled DNA from *B. anthracis, Y. pestis, B. melitensis*, vaccinia virus, cow, pig, chicken and humans (**Figure 6**). The hybridization of these eight genomes including four microbial genomes and four large complex genomes, demonstrates specificity for the organisms of interest and no significant background hybridization.

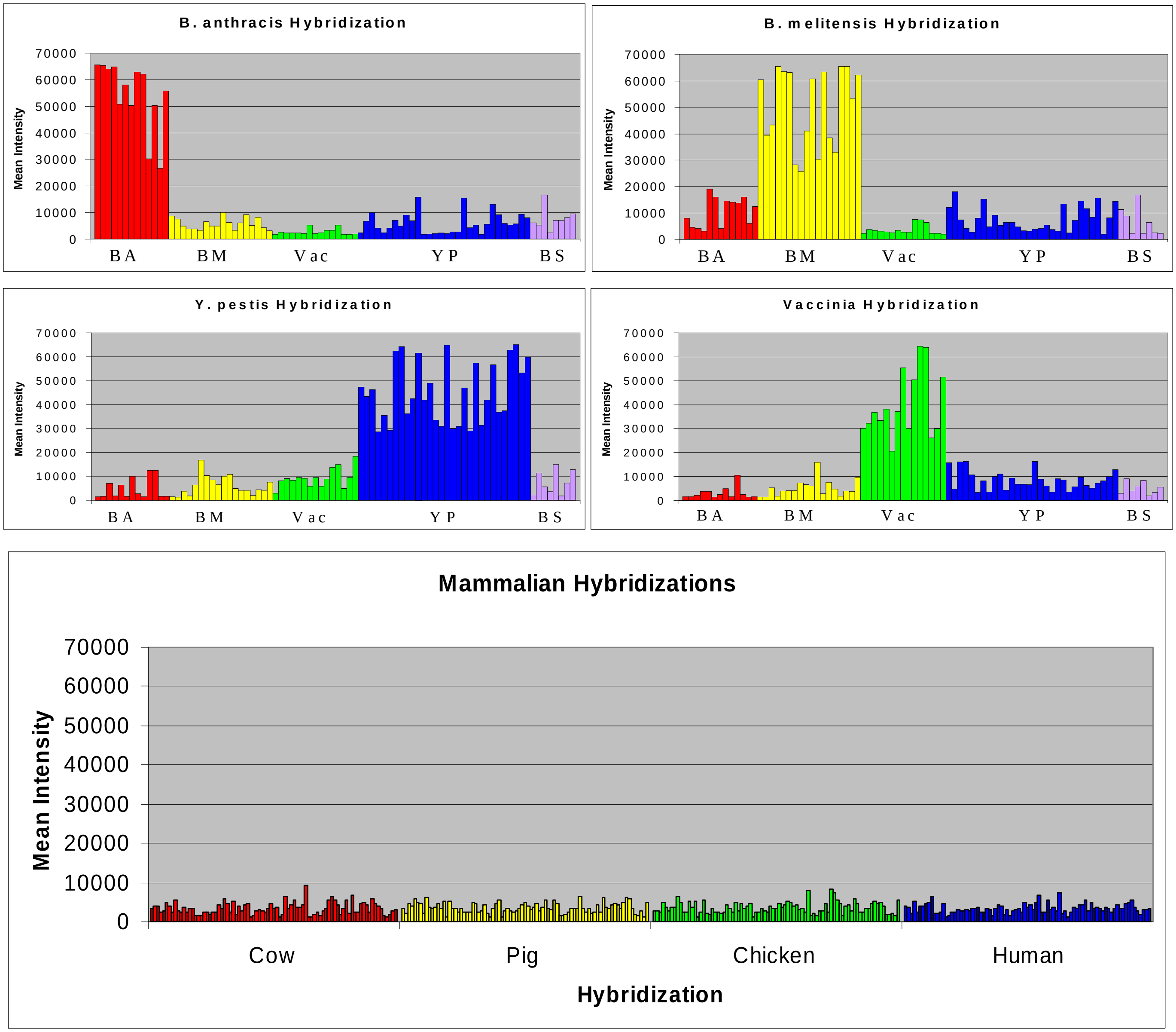


Figure 6. Organisms specific hybridization: To demonstrate that the oligonucleotide probes on the array were organism-specific and did not cross hybridize with mammalian DNA generating a background hybridization pattern, four microbial genomes and cow, pig, chicken and human DNA was Cy-3 dCTP labeled by direct incorporation and hybridized to identical arrays containing these oligonucleotides. The organism used as the hybridization probe is identified at the top of the four smaller panels above. The identity of the oligos is provided by the initials BA (*Bacillus anthracis*), BM (*Brucella melitensis*), Vac (vaccinia virus), YP (*Yersinia pestis*) and BS (*Bacillus subtilis*). The graph on mammalian hybridization presents the hybridization data for the hybridization of these four genomes against the organism specific oligos presented in the upper four graphs. There was no cross hybridization between the oligos and the mammalian genomes.

Hybridizations performed with a human probe and a probe containing vaccinia virus DNA in the presence of a human background is presented in **Figure 7**. The microarray was able to detect the presence of the vaccinia virus the vastly more complex background of the human genome.

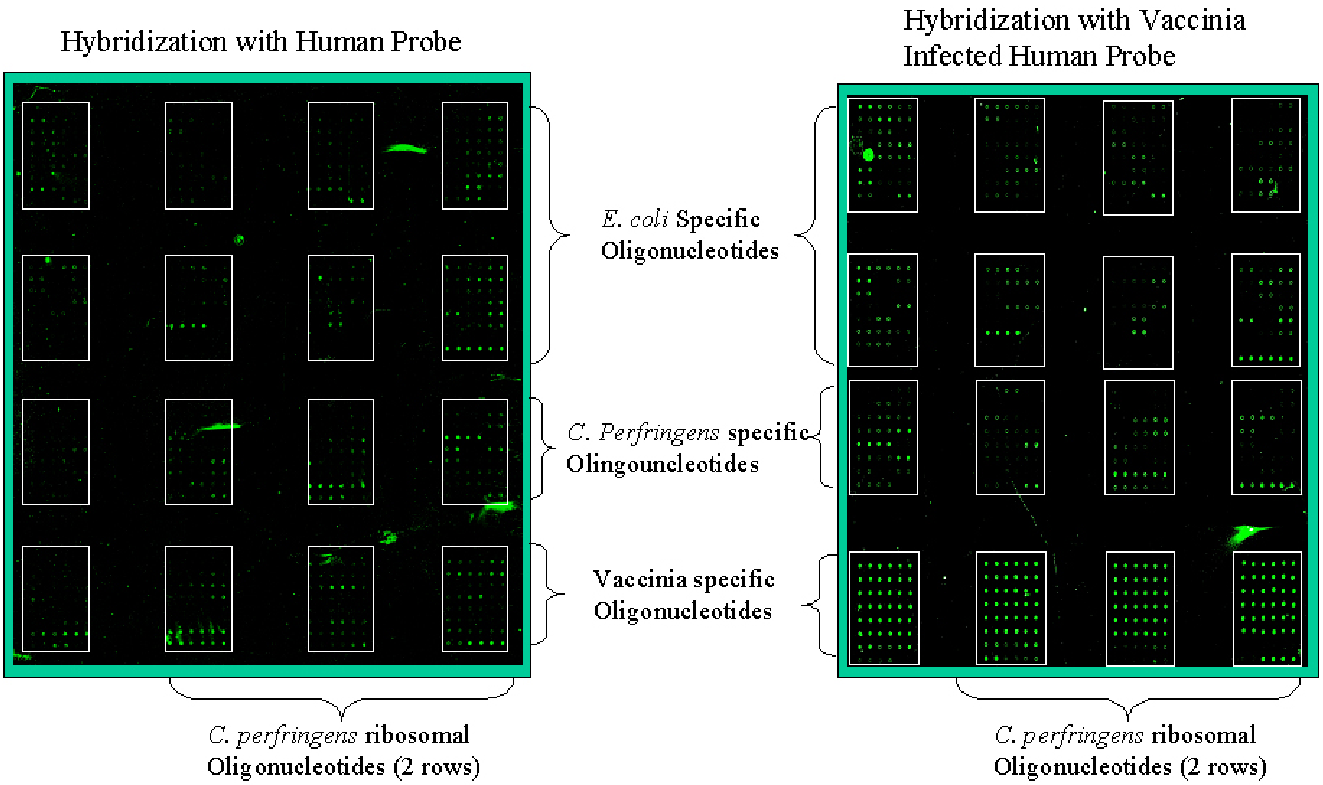

Hybridization with Human Probe
Hybridization with Vaccinia Infected Human Probe
E. coli Specific Oligonucleotides
C. Perfringens specific Olingouncleotides
Vaccinia specific Oligonucleotides
C. perfringens ribosomal Oligonucleotides (2 rows)
C. perfringens ribosomal Oligonucleotides (2 rows)

Figure 7. Hybridization of human and vaccinia infected human probes to oligonucleotide microarrays: Specific hybridization of vaccinia cy3 labeled probe to the array and the lack of hybridization of cy3 labeled human probe to the array indicates that this method may be used for developing a diagnostic that identifies species and strains of organisms in animals and humans.

Microarray data was generated for nine organisms. The identity of these organisms can be see in **Figure 9** where organism specific oligonucleotides for *B. anthracis, E. coli, B. melitensis* and vaccinia virus are presented. An example of the hybridization data that resulted in the graphs in **Figure 9** is presented in **Figure 8**. In these scatter plots, data for two hybridizations is presented, one on the X-axis, the other on the Y-axis. Spots that are located in the upper left or lower right quadrants of the scatter plot represent oligos that had the greatest differential of hybridization. Those that fall on the line Y=X which runs diagonally up and to the right from the origin are non-informative for the discrimination between the organisms on the scatter plot.

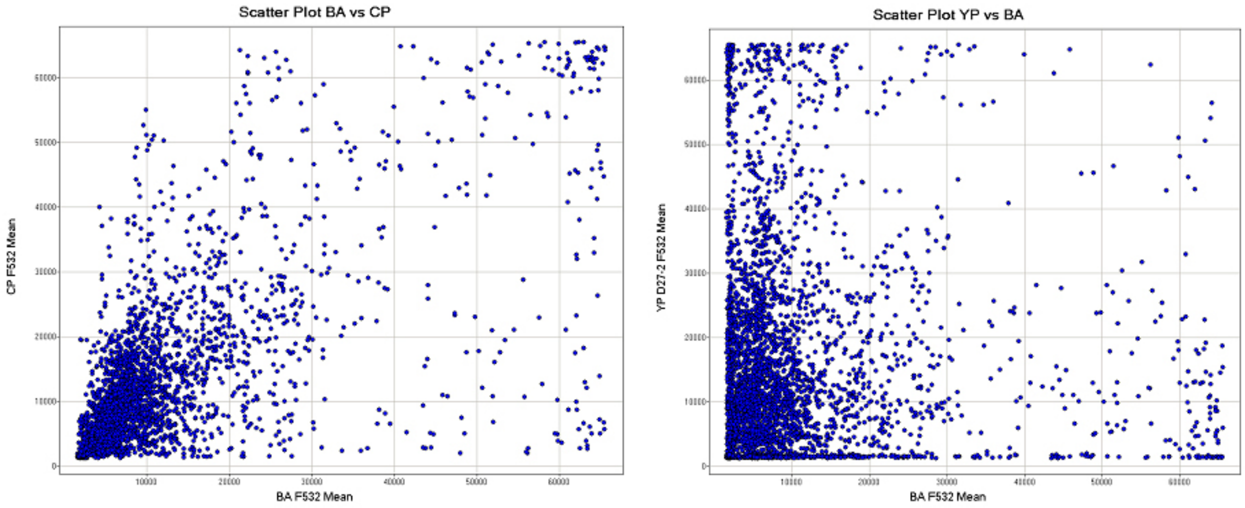


Figure 8. Scatter plot of microarray analysis: Hybridization intensities for 3840 oligonucleotides for *C. perfringens* (CP), *B. anthracis* (BA) and *Y. pestis* (YP) are presented as an example of how intensity of hybridization can be used to discriminate between organisms. The closer to the axis and farther away a spot is from the origin, the more specific the oligonucleotide is for the organism on that axis. Spots that fall on the line Y=X do not discriminate between organisms.

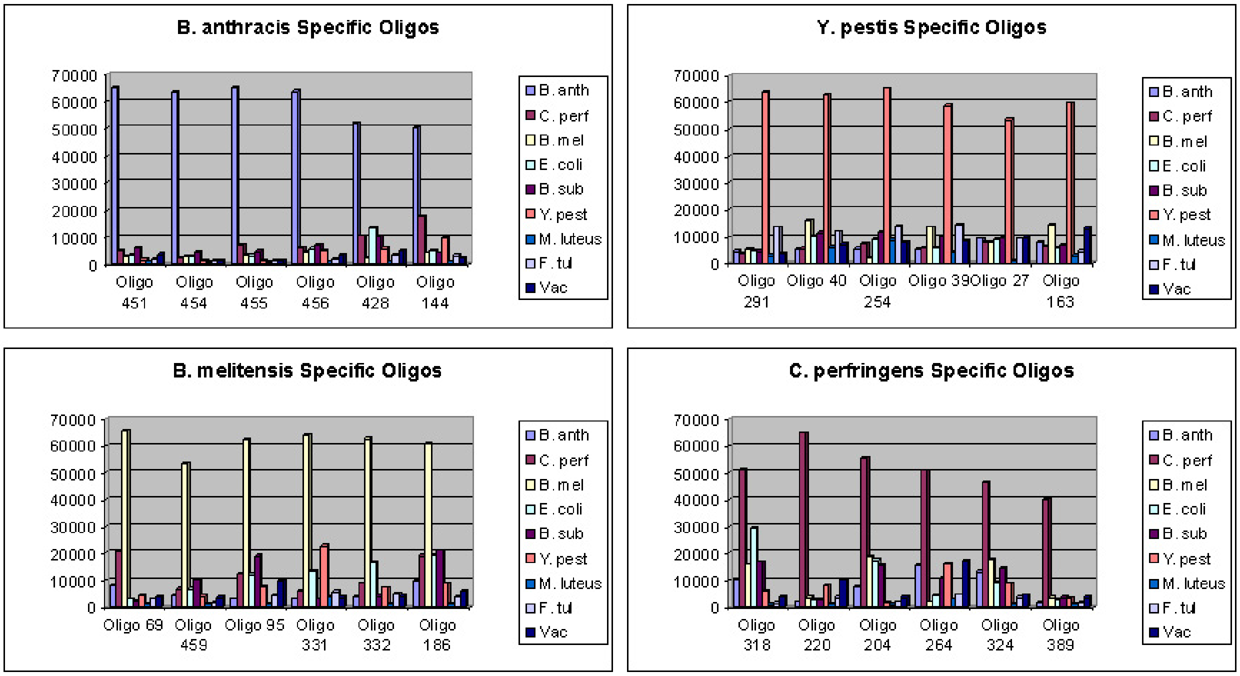


Figure 9. Organism specific oligonucleotides identified from scatter plots: Six oligonucleotides per organism are used to demonstrate the ability to identify unique oligonucleotides using this method. The intensity of hybridization is presented for nine organisms. Oligonucleotides are on the x-axis. The intensity of the hybridization of each oligonucleotide for nine organisms is presented by the different colored bars.

Discussion

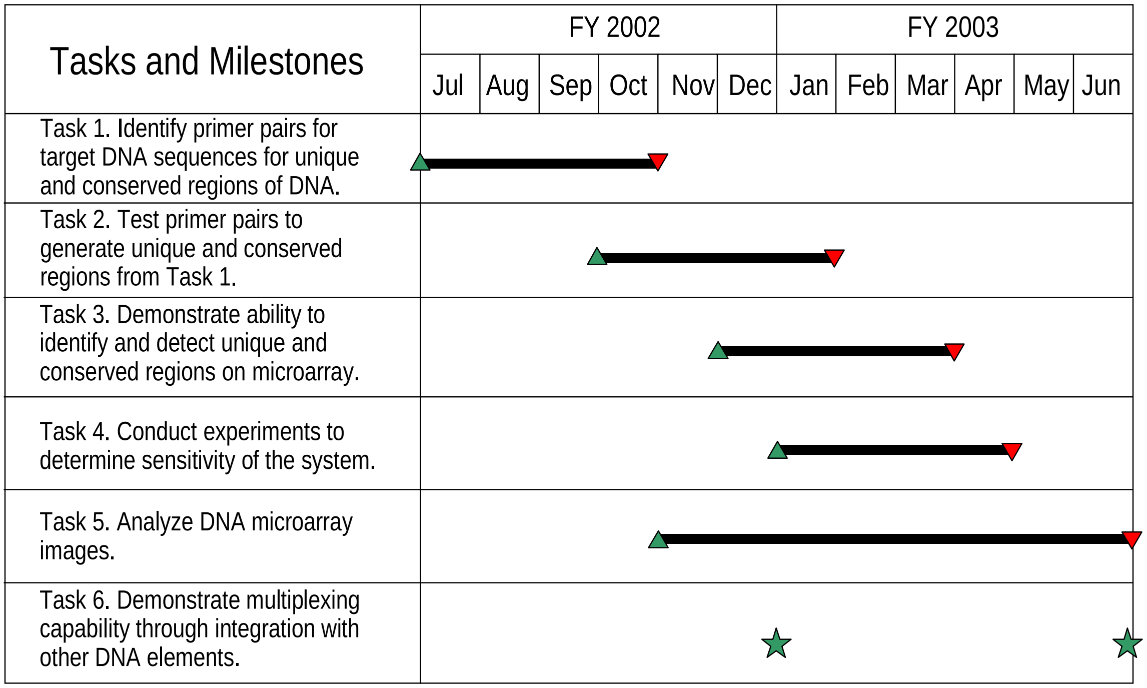
The use of microarrays to identify microorganisms is not a new idea. Microarrays are currently being developed to discriminate between organisms based on 16S ribosomal RNA gene sequences and based on unique sequence identified from the small genomes of viruses. The problem with using 16S ribosomal RNA sequences for microarray analysis is that: 1) It ignores greater than 99% of sequence contained in the genome of an organism, 2) 16S ribosomal RNA gene sequences are very similar between organisms and they cannot typically be used to discriminate between strains of organisms and 3) They cannot be used to identify the presence of virulence genes and antibiotic resistance genes. Alternatively, the use of unique sequences from viruses works well, however, the bioinformatic methods used to identify unique sequences from the small genomes of viruses do not work on larger bacterial genomes, which can be three orders of magnitude larger than viral genomes.

We have developed a bioinformatic method that identifies unique sequences from large bacterial genomes. We have utilized this method to identify unique sequences from *Bacillus anthracis*, *Yersinia pestis,* Vaccinia virus and *Escherichia coli*, the four organisms that are the subject of this contract. Microarrays were used as a detection platform to validate the unique sequences identified by our bioinformatic method. The unique sequences generated by our bioinformatic tool are not, however, platform dependant. These unique sequences will work with other detection platforms such as real time PCR, the current gold standard of microbial detection. As a demonstration of our ability to discriminate between organisms, we identified unique sequences that discriminated between organisms at the level of genus, species and strain. The data generated under this contract clearly show that strain level resolution between the genomes *of E. coli* K12 and *E. coli* O157:H7 is possible, as is species level discrimination between *B. anthracis* and *B. subtilis* and genus level discrimination between organisms such as *B. anthracis* and *Y. pestis*. As a demonstration of the ability to detect pathogenic strains of an organism, we included virulence genes on the microarrays and clearly demonstrated that these genes track with the unique genomic sequences identified for the O157:H7 strain of *E. coli*. Further data is provided that demonstrates that the unique sequences, identified from the genomes of these organisms, do not cross react with the complex genomes of humans and agriculturally important species. This suggests that this method may be well suited to human diagnostic and agricultural applications.

The power of this technique is obvious. By assaying for the presence of: 1) unique sequences at various levels of the phylogenetic tree, 2) virulence genes, 3) antibiotic resistance genes, 4) virulence plasmid sequences and 5) ribosomal genes, we are in a unique position to develop a novel method for the identification and characterization of the microorganisms. The application for this technology crosses many fields of research that are important to the Government: Environmental testing for bioterrorism by Homeland Security, Battlefield testing by the armed forces, Water quality testing by the Environmental Protection Agency, Human diagnostics by the Centers for Disease Control and Agricultural testing by the United States Department of Agriculture, to name a few. In addition to the obvious use of this method to identify known pathogens, we believe that our method will also permit the identification of naturally occurring virulent variants of non-pathogenic organisms, representing the ability to identify emerging threats to public health. Another less obvious application is the ability to detect the intentional manipulation of microbial genomes by bioterrorists, designed to place a pathogen in proximity to the public. Genetic manipulation of genomes could place a virulence gene within the genome of a non-pathogenic organism, making it undetectable by current methods of pathogen detection. In addition, by utilizing many different organism specific sequences for each organism assayed, coupled with statistical analysis methods, false positives simply become a phenomenon of past technology.

Gantt chart

Schedule for Identification of BW Threat Agents Using DNA Microarrays



Task 1. Identify primer pairs for target DNA sequences for unique and conserved regions of DNA.

Task 2. Test primer pairs to generate unique and conserved regions from Task 1.

Task 3. Demonstrate ability to identify and detect unique and conserved regions on microarray.

Task 4. Conduct experiments to determine sensitivity of the system.

Task 5. Analyze DNA microarray images.

Task 6. Demonstrate multiplexing capability through integration with other DNA elements.

Conclusions

We have successfully demonstrated the ability to identify unique sequences from the genomes of microorganisms and the ability to detect these unique sequences by microarray analysis. We present compelling evidence that these methods can be utilized to develop a microarray based detection method for the identification of organisms from different genera as well as closely related strains of bacteria. We also demonstrate the ability to detect the presence of virulence genes in pathogenic bacteria and cross hybridization to complex genomes. While these unique sequences were verified as informative for specific organisms by microarray analysis, it is almost certain that these unique sequences would be ideal targets for PCR based analysis methods. Utilizing these sequences on a microarray platform would provide a capability to identify all organisms of interest simultaneously, the ability to detect natural differences within genomes that lead to the evolution of emerging threat organisms and the intentional manipulation of genomes for the purpose of introducing a pathogen into the environment while avoiding classical detection methodologies.

References

1. Fixed-parameter algorithms for closest string and related problems, Jens Grann, Rolf Niedermeier and Peter Rossmanith. Algorithmica (2003) 37:25-42

2. Selecting signature oligonucleotides to identify organisms using DNA arrays, Lars Kaderali and Alexander Schliep, Bioinformatics (2002) Vol 18 no 10: 1340-1349

3. Comparative genomic tools applied to bioterrorism defense, Tom Slezak, Tom Kuczmarski, Linda Ott, Clinton Torres, Dan Medeiros, Jason Smith, Brian Truitt, Nisha Mulakeen, Marisa Lam, Elizabeth Vitalis, Adam Zemla, Carol Ecale Zhou and Shea Gardner, Briefings in Bioinformatics, (2003) Vol 4 No. 2: 1-17

4. Rapid Development of Nucleic Acid diagnostics, J. Pathrick Fitch, Shea N. Gardner, Thomas A Kuczmarski, Stefan Kurtz, Rich Myers, Linda L. Ott, Thomas R. Slezak, Elizabeth A. Vitalis, Adam T. Zemla and Paula M. McCready (2002) Proceedings of the IEAA. 90, No.11:170-1720

5. Microarray-based detection and genotyping of viral pathogens, David Wang, Laurent Coscoy, Maxine Zylberberg, Pedro C. Avila, Homer A. Boushey, Don Ganem and Joseph L. DeRisi, (2002) PNAS, Vol 99. No24 15687-15692

6. Genetic variation in 16S-23S rDNA internal transcribed spacer regions and the possible use of this genetic variation for molecular diagnosis of Bacteroides species. Kuwahara T, Norimatsu I, Nakayama H, Akimoto S, Kataoka K, Arimochi H, Ohnishi Y, Microbiol Immunol. 2001;45(3):191-9.

7. Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. Reyes-Lopez MA, Mendez-Tonorio A, Maldonado-Rodriguez R, Doktycz MJ, Fleming JT, Beattie KL, Nucleic Acids Res. 2003 Jan 15;31(2):779-89.

8. Comprehensive detection of bacterial populations by PCR amplification of the 16S-23S rRNA spacer region. Gonzalez N, Romero J, Espejo RT, J Microbiol Methods. 2003 Oct;55(1):91-7.

9. Development of a PCR-based method for specific identification of genotypic markers of shiga toxin-producing Escherichia coli strains. Osek J, Dacko J. J Vet Med B Infect Dis Vet Public Health. 2001 Dec;48(10):771-8.

10 Simultaneous detection of two verotoxin genes using dual-label time-resolved fluorescence immunoassay with duplex PCR. Watanabe K, Arakawa H, Maeda M. Luminescence. 2002 Mar-Apr;17(2):123-9.

11. PCR detection of Clostridium perfringens producing different toxins in faeces of goats. Uzal FA, Plumb JJ, Blackall LL, Kelly WR. Lett Appl Microbiol. 1997 Nov;25(5):339-44.

12. PCR amplification on a microarray of gel-immobilized oligonucleotides: detection of bacterial toxin- and drug-resistant genes and their mutations. Strizhkov BN, Drobyshev AL, Mikhailovich VM, Mirzabekov AD, Biotechniques. 2000 Oct;29(4):844-8, 850-2, 854 passim.

13. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

14. Zhang, J. & Madden, T.L. (1997) "PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation." Genome Res. 7:649-656.

15. Rigby PW, Dieckmann M, Rhodes C, Berg P. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I.
J Mol Biol. 1977 Jun 15;113(1):237-51

16. Ziebell KA, Read SC, Johnson RP, Gyles CL. Evaluation of PCR and PCR-RFLP protocols for identifying Shiga toxins.
Res Microbiol. 2002 Jun;153(5):289-300.

17. Wittwer CT, Herrmann MG, Gundry CN, Elenitoba-Johnson KS. (2001)
Real-Time Multiplex PCR Assays.
Methods. 2001 Dec;25(4):430-442

18. Freeman, WM, Walker, SJ, and Vrana, KE (1999)
Quantitative RT-PCR: pitfalls and potential. Biotechniques 26, 112-122.

19. Svanvik N., A. Stålberg, U. Sehlstedt, R. Sjöback & M. Kubista. (2000)
Detection of PCR Products in Real-time Using Light-up Probes.
Anal. Biochem. 287, 179-182 (2000).

20. Souazé, F, Ntodou-Thomé, A, Tran, CY, Rostene, W, and Forgez, P (1996)
Quantitative RT-PCR: Limits and accuracy. BioTechniques 21, 280-285.

21 Kainz P. (2000)
The PCR plateau phase - towards an understanding of its limitations.
Biochim Biophys Acta 2000 Nov 15;1494(1-2):23-7

22. Giulietti A, Overbergh L, Valckx D, Decallonne B, Bouillon R, Mathieu C. (2001)
An overview of real-time quantitative PCR: applications to quantify cytokine gene expression.
Methods 2001 Dec;25(4):386-401