Universal Microbial Diagnostics using Random DNA Probes

by

Amirali Aghazadeh Mohandesi

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE Master of Science

APPROVED, THESIS COMMITTEE:

Richard G. Baraniuk, Chair
Victor E. Cameron Professor of Electrical and Computer Engineering and Computer Science, Rice University

Caleb Kemere
Assistant Professor of Electrical and Computer Engineering, Rice University

Rebekah A. Drezek
Professor of Bioengineering and Electrical and Computer Engineering, Rice University

Houston, Texas
April, 2014
ABSTRACT

Universal Microbial Diagnostics using
Random DNA Probes

by

Amirali Aghazadeh Mohandesi

The accurate and efficient identification of microbial organism such as viruses and bacteria has mounting importance in the fields of health care, environmental monitoring, and defense. As an example, sepsis from bacterial infection is currently the 11th leading cause of death in the United States. However, current microbial detection strategies are cost-prohibitive, time-consuming and inevitably use unique sensors that are specific to each species to be detected. In this thesis we present a novel microbial sensing platform capable of both detecting the presence and estimating the concentration of microbial organisms in an infectious sample using a small number of random DNA probes. Our Universal Microbial Diagnostics (UMD) platform leverages the theory of sparse signal recovery (compressive sensing) to stably identify the composition of a sample containing several bacteria from a potentially large library of target bacteria. We experimentally validate UMD in vitro using a set of random sloppy molecular beacons to recover pathogenic bacteria without DNA amplification. We also evaluate the average performance of UMD in silico for genus and species level identification of 38 common human pathogens. A particularly promising property of UMD for health care, environmental monitoring, and defense applications is that a fixed set of random measurement probes are universal in the sense that they
can characterize novel organisms not present in the target library.
Acknowledgments

I would like to thank my advisor, Richard Baraniuk for his guidance, support and encouragement and especially for teaching me to think big and wish for the impossibles. I also want to thank my committee members: Rebekah Drezek, Caleb Kamere, and Joseph Petrosino for all their support and time. I also would like to thank Mona Sheikh for introducing me to the exciting world of genomic sensing and challenging me with her brilliant ideas. Thanks also goes to my experimental collaborators, Adam Lin and Allen Chen, in bioengineering department and Lisa Atkins in Baylor College of Medicine. I also want to thank my office mate, Mr. Lan, for providing the most cheerful working enviroment. Finally, significant thanks are extended to my family who inspire my interest in pursuing science and innovation and friends who helped me make this possible.
# Contents

Abstract ................................................................. ii
List of Illustrations .................................................. vii
List of Tables .................................................................. xiii

1 Introduction ................................................................ 1
  1.1 Signal processing for genomic studies .......................... 1
  1.2 DNA-based Microbial Identification ............................ 1
  1.3 Thesis overview ...................................................... 3

2 Background .................................................................. 4
  2.1 Preliminaries .......................................................... 4
    2.1.1 Notation ......................................................... 4
    2.1.2 Vector Norm ................................................... 4
    2.1.3 Sparse signals ................................................ 5
  2.2 Compressive Sensing ............................................... 6
    2.2.1 Compressive sensing theory ............................... 6
    2.2.2 Compressive sensing for non-negative signals ............ 8
    2.2.3 Greedy recovery algorithms .............................. 9
    2.2.4 Basis pursuit denoising .................................... 11
  2.3 Whole Genomic Sequence Analysis Tools ....................... 13
    2.3.1 DNA Thermodynamic Hybridization Model .......... 16
  2.4 Molecular Beacons ................................................ 19
    2.4.1 Principle of Operation .................................... 19
    2.4.2 Advantages for Microbial Detection ................... 20
3 Universal Microbial Diagnostics

3.1 Universal Microbial Diagnostics Platform
3.1.1 The Platform Overview
3.1.2 Linearity assumption considerations in UMD
3.1.3 Hybridization Affinity Matrix Calculations
3.2 UMD probe design
3.3 Experimental proof of concept
3.3.1 Experimental set up
3.3.2 Probe characteristic curve
3.3.3 Bacteria detection and concentration Estimation in vitro
3.3.4 UMD on most common human pathogens
3.4 Extension by simulation
3.4.1 Genus level single bacteria detection using UMD
3.4.2 Species level single bacteria detection using UMD
3.4.3 Genus level double bacteria detection using UMD
3.4.4 Greedy probe selection

4 Discussion and Conclusion

4.1 Future work

Appendix A

Bibliography
Illustrations

2.1 Molecular Beacons (MBs) [1] Principle of Operation. The probe sequence in the loop binds spontaneously to the target sequence which is exact complement to the probe sequence. Adapted from [2].
3.1 **Schematic of Universal Microbial Diagnostics (UMD) platform.** (A) Genomic DNA is extracted from a bacterial sample and thermal cycled with $M$ random DNA probes. The genome-probe binding is quantified, producing a probe-binding vector $y$; in this study the random probes are in the form of molecular beacons (MBs), and the DNA-probe binding is quantified by the ratio of open/hybridized to closed/non-hybridized MBs. (B) The hybridization binding level of each probe to a potentially large reference database of $N$ bacterial genomes ($B_1, B_2, \ldots, B_N$) is predicted using a thermodynamic model and stored in an $M \times N$ hybridization affinity matrix $\Phi$ (here $N = 10$ for illustration, but it could be very large in practice). (C) Assuming just $K$ bacterial species comprises the sample, the probe-binding vector $y$ is a sparse linear combination of the corresponding $K$ columns of the matrix $\Phi$ weighted by the bacterial concentrations $x$, i.e., $y = \Phi x + n$, where the vector $n$ accounts for noise and modeling errors. When $K$ is small enough and $M$ is large enough, $\Phi$ can be effectively inverted using techniques from compressive sensing (CS), yielding the estimate for the microbial makeup of the sample $x$; in this illustration the $K = 2$ bacteria labeled $B_2$ and $B_7$ are present in the sample.

3.2 The fast approach for modeling of the hybridization affinity of a random probe sequence (red) to a bacterial genome (blue). Hybridization affinity was calculated by counting the total number of significant binding spots of a random probe sequence to the genome under the experimental thermodynamic conditions. All bindings with melting temperature greater than $T_{m_{\text{thresh}}} = 45^\circ C$ were considered significant (see later this chapter for more details on the choice).
3.3 Probe-bacteria hybridization affinity computation process. Both strands of the bacterial genome (blue lines) are randomly segmented into roughly equal length (∼ 700 – 1000 nts) fragments. All of the bacterial fragments and probe sequences along with the experimental conditions are fed into the DNA software [3] to predict all stable probe-bacteria complexes and concentrations. These concentrations, in aggregate, determine the concentration of opened beacons, which is defined as the hybridization affinity of the probes to the bacterial genome.

3.4 DNA sequence structure of three test random DNA probes. Nucleotides in black color are fixed to preserve the hairpin structure, and other nucleotides are randomly chosen.

3.5 Example of a predicted probe-bacteria fragment binding with many base pair mismatches.

3.6 The predicted number of opened probes when a bacterial genome (E.coli) is exposed to a specific (in the same thermodynamic condition as the experiment) traditional molecular beacon (MB) [1], sloppy MB [4], and MB created according to our design rules. While the nucleotides in the loop were determined randomly, the percentage of opened beacons (hybridization affinity to E.coli) is substantially increased for the random probe MB, thanks to the choice of stem length and configuration of a small number of fixed nucleotides in the loop sequence.

3.7 Experimentally measured FRET ratios to quantify hybridization between bacterial DNA (E. coli, S. aureus, and F. tularensis) and probes 1, 2, and 3.
3.8 Probe characteristic curves for random probes 1, 2, and 3 for determining the experimental relationship between FRET ratio and open beacon concentration (hybridization affinity). The characteristic curve is fit to the measured FRET ratio as a function of the concentration of the probes exact complement for three random probes. Error bars represent the standard error of the mean (SEM) for 3 trials.

3.9 Characteristic curve parameters a, b, and n for three test random probes, and the associated curve fit performance criteria, $R^2$ and root mean square error (RMSE). Parameters were fit to the following curve: $FRET = FRET_0 + \frac{a}{1 + b \log(10^{-6}) - \log(c) - n}$ [5].

3.10 Hybridization affinity between DNA samples and probes, converted from FRET ratio (as in Figure 3.3.1) through the probe characteristic curve fit equations.

3.11 The inner product of the experimentally obtained hybridization affinity (Figure 3.3.2) and center sum normalized predicted (by thermodynamic modeling) hybridization affinity for three bacterial DNA samples ($E. coli$, $S. aureus$, and $F. tularensis$) as a measure of the similarity of the probe measurements to the bacteria in the database. Orthogonal Matching Pursuit (OMP) correctly recovers the highlighted bacterium that was present.

3.12 Comparison of the predicted concentrations of three bacterial DNA ($E. coli$, $S. aureus$, and $F. tularensis$) (gray bars) with the experimentally measured values (black bars).

3.13 The ratio of the hybridization pattern signal to the hybridization pattern noise ranges from 23 to 35 dB.
3.14 **Genus-level single bacterium detection.** Blue lines show the average sensitivity (true positive rate) of the OMP algorithm in recovering 38 species from the most common human infectious genera, simulated by adding i.i.d. Gaussian noise with standard deviation $\sigma = 0.025$ (dashed line) to account for the measurement noise and $\sigma = 0.123$ to account for the modeling error. The probes were randomly selected from a dictionary of 300 probes (100 probes each of GC content 50, 56, 60%). The red lines show the average specificity for the two different noise levels. The error bars show the standard deviation of 100 simulation trails. There exist sets of $M = 3$ and $M = 5$ random probes (among the 300 probes) that enable detection with 100% sensitivity and specificity at these measurement and modeling noise levels. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

3.15 **Species-level single bacterium detection.** The average sensitivity (blue lines) and specificity (red lines) in species level detection of MRSA among 8 *Staphylococcus* species using OMP. There exist many sets of $M = 3$ and $M = 5$ probes (among the 300 probes) that enable detection with 100% sensitivity and specificity at the same measurement and modeling noise levels as above. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

3.16 **Genus-level two bacteria detection.** Basis Pursuit De-Noising (BPDN) was utilized to recover the noise-corrupted sum of all the bacteria pairs ($K = 2$) present in the database. At the above measurement and modeling noise levels, $M = 9$ and $M = 17$ probes achieve sensitivity and specificity of 100%, confirming our intuition that additional probes are required to recover $K = 2$ bacteria. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44
3.17 **Performance and computational complexity comparison of probe selection algorithms.** Figure shows the average sensitivity in detecting 38 human infectious species of different genera using the proposed Greedy Probe Selection (GPS) algorithm (blue line). GPS-selected probes achieve 100% sensitivity with $M = 7$ probes, while on average an $M = 5$ random probe set (see A) achieves the same performance. However, the computational time for selecting probes using GPS grows only linearly with the number of probes (purple solid line), while the time for searching the complete random probe space grows exponentially (purple dashed line).

3.18 A confusion matrix summarizing UMD pathogenic bacterial detection in simulation, using $M = 3$ and $M = 7$ probes selected by GPS.
Tables
Chapter 1

Introduction

1.1 Signal processing for genomic studies

“We live in the post – genomic era, when DNA sequence data is growing exponentially” [6]. The cost for sequencing the whole genome of a bacteria has decreased since 2007 from 10K$ to only 10$ today [7]. The availability of genomic data is opening new doors for signal processing tools and theories to help us better understand the biological phenomenon and engineer biological sensing instruments. In signal processing the Nyquist-Shannon sampling theorem was the starting point for a multitude of signal processing tasks, algorithms and devices. Today, the complete sequencing of genome offers the scope to apply such intelligence to the processing and acquisition of genomic data. This tribute of signal processing in the genomic era will have impacts in a variety of fields vital to human health: personalized learning, medicine, drug development, defense, environmental monitoring and food safety, among others.

1.2 DNA-based Microbial Identification

Among the areas to benefit from the signal processing tools, microbial identification has mounting importance in many critical schemes from health centers to food pro-
cessing plants and battlefields [8, 9, 10]. Sepsis from bacterial infection is currently the 11th leading cause of death in the United States, and the mortality rate of bloodstream infections is high (14-50%) [11, 12, 13]. Conventional strategies for microbial detection are based on microbe-specific genomic and proteomic markers and protocols. For example, polymerase chain reaction (PCR) identifiers, such as the 16S and 23S ribosomal DNA subunits in bacteria. Although these methods show promise as highly specific tools for microbial identification [14, 15], they have limitations in clinical, industrial, or defense settings [16, 17]. There are three main problems that result from the use of the unique identifiers. One, the 16S and 23S genes are short (just 2000-3000 nucleotides) compared to the whole genome length which is in the order of several millions [18], therefore they must undergo PCR process in order for the unique sensors (or probes) to identify them. Two, current molecular tools do not have the capacity to hold all the unique identifiers needed to detect all the different sequenced bacterial species, which means that several different devices each tailored to a different set of bacteria may be needed for a complete detection. Three, current molecular tools have not way to identify unknown or mutated species. Studies of prokaryotic diversity estimated that there are yet millions of undetected bacterial species in the world, of which fewer than 3000 have completely sequenced at last count. By solely focusing on the specific known markers in each bacteria, current molecular tools are severely limited in their detection performance. For example, In the case of a viral epidemic, the detection of a newly mutated species using current PCR methods would
require entirely new capture probes to be manufactured, introducing additional costs and delays. Blood cultures typically require at least 72 hours to produce reliable results [19, 20, 21, 22]. During this waiting period, administration of broad-spectrum antibiotics breeds further threats of bacterial resistance.

1.3 Thesis overview

The thesis is organized as follows. In Chapter 2 we first introduce the notations that we use throughout the thesis and later describe some background materials in Compressive Sensing (CS) and whole genomic sequence analysis tools and algorithms. Here, we talk about exact theoretical recovery bounds and recovery algorithms (e.g. greedy methods and gradient based methods) for CS. We also give a background on the operation of molecular beacons and contrast them with the newly developed sloppy molecular beacons probes. In Chapter 3 we introduce the idea of Universal Microbial Diagnostics (UMD) platform. We further talk about our proposed methods for calculating the DNA hybridization affinities in UMD using the thermodynamic models. Here we also outline the steps to design a random molecular beacon probe and discuss different design tradeoffs. Next we give an experimental proof of concept for UMD platform in detecting 3 pathogenic bacterial strains and later show UMD’s extensions in detecting more than a single type of bacteria among 38 most common human pathogenic genera using simulations, in silico. Finally, in Chapter 4 we conclude with the summarized implications of our work.
Chapter 2

Background

2.1 Preliminaries

2.1.1 Notation

We provide a brief summary of the notation used in this thesis. Signals are real valued vectors defined over domains that are either continuous or discrete. In general, lower case letters (e.g., $x$, $y$) will be used to denote discrete column signals; upper case boldface letters (e.g., $\Phi$, $D$, $P$) will be used to denote matrices; and unbolded italic letters (e.g, $M$, $N$, $\lambda$) will be used to denote scalar quantities. Calligraphic letters (e.g., $S$, $P$) will be used to denote sets or set operators. Four upper case letters will be used to denote four DNA nucleobases Adenine (A), Cytosine (C), Guanine (G), and Thymine (T).

2.1.2 Vector Norm

In linear algebra, a norm is a function that assigns a strictly real positive number to each vector in a vector space. For example, for a finite-length real signal $x$ of size $N$, which can be viewed as an element of $N$-dimensional vector space $\mathbb{R}^N$, i.e.,
\( x = [x_1, x_2, \ldots, x_N]^T \), well known \( \ell_p \) norms are defined as,

\[
\|x\|_p = \begin{cases} 
(\sum_{i=1}^{N} |x_i|^p)^{1/p}, & p \in (0, \infty); \\
\max_{i=1,2,\ldots,N} |x_i|, & p = \infty; \\
\sum_{i=1}^{N} 1_{x(i)\neq 0}, & p = 0,
\end{cases}
\tag{2.1}
\]

where \( 1 \) denotes the indicator function. We have to note that the \( \ell_p \)-norm does not satisfy the criteria for a well-defined norm when \( p < 1 \). A very common and useful norm is the \( \ell_2 \)-norm which is also commonly known as Euclidian norm. It is easy to see that the \( \ell_2 \)-norm can be written as the square root of the inner product of the vector with itself, i.e., \( \|x\|_2 = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{N} x_i^2} \). The \( \ell_2 \)-norm is particularly useful in characterizing the mean square error (MSE) between pairs of signals; the MSE between \( x_1, x_2 \in \mathbb{R}^N \) is defined as

\[
\frac{1}{N} \|x_1 - x_2\|_2^2.
\tag{2.2}
\]

The sum of absolute differences (SAD) is another common error metric based on the \( \ell_1 \)-norm. Also known as Manhattan distance, SAD between \( x_1, x_2 \in \mathbb{R}^N \) is defined as

\[
\|x_1 - x_2\|_1.
\tag{2.3}
\]

2.1.3 Sparse signals

Given a vector space \( \mathcal{F} \), a dictionary is simply a collection of vectors in \( \mathcal{F} \) (also called atoms) whose linear superpositions can be used to represent signals of interest. For the signals living in \( N \)-dimensional space, any dictionary can be equivalently
represented as a matrix $D \in \mathbb{R}^{N \times D}$, with the columns of $D$ representing the atoms. An *orthonormal basis* (ONB), or simply, a basis is a special type of dictionary with $D = N$ normalized and orthogonal columns. Given a signal $x \in \mathbb{R}^N$ and a basis $\Psi \in \mathbb{R}^{N \times N}$, we can expand $x$ in terms of its basis coefficients $\alpha$, i.e., $x = \Psi^{-1} \alpha = \Psi^T \alpha$.

Common example of orthonormal bases is the *canonical basis* (where $\Psi$ equals the identity matrix and $\alpha = x$). Given a general dictionary $D$, we are often interested in the set of signals that can be written as the linear combination of only a few elements of $D$. Such signals are known as *sparse signals*. Formally, a signal $x \in \mathbb{R}^N$ is said to be $k$-sparse in $D$ if $x = D\alpha$ and no more than $K$ coefficients of $\alpha$ are nonzero. The *support* of $x$, $S(x)$, is defined as the set of indices corresponding to nonzero entries of $\alpha$. The *sparsity* of $x$ is the number of nonzero, or the $\ell_0$-norm, of its dictionary representation $\alpha$ (denotes as $||\alpha||_0$). An elegant feature of a $K$-sparse signal $x \in \mathbb{R}^N$ is that it can be concisely represented using $2K$ pieces of information, corresponding to the *locations* and the *values* of the nonzero coefficients of $\alpha$. In the regime where $K \ll N$, it is often vastly beneficial to store, sense, and process signals entirely in the space of coefficients.

### 2.2 Compressive Sensing

#### 2.2.1 Compressive sensing theory

Compressive Sensing (CS) is an alternative to Shannon-Nyquist sampling theory for the acquisition of sparse or compressible signals in an $N$-dimensional space. Instead
of taking many periodic samples, CS measures inner products with $M \ll N$ random vectors and then recovers the signal via a sparsity-seeking optimization or greedy algorithm. The main result of CS, introduced first by Candes, Romberg, and Tao [23] and Donoho [24], is that robust signal recovery is possible from $M = \mathcal{O}(K \log(N/K))$ measurements. More rigorously, CS obtains a random linear set of measurements, $y = \Phi x$, where $x$ is the $N \times 1$ sparse signal we aim to sense, $y$ is an $M \times 1$ measurement vector, and the measurement matrix $\Phi$ is an $M \times N$ matrix. In the presence of measurement noise, the model becomes, $y = \Phi x + n$ where $n$ is the noise vector. Since $M \ll N$ simply inverting the matrix $\Phi$ is impossible in this case, since it has many more columns than rows. The recovery of $x$ from $y$ is known to be ill-posed using the underdetermined system. Fortunately the sparsity of vector $x$ makes the problem invertible under some assumption imposed by CS theory.

In order to apply the standard compressive sensing theory, the columns of $\Phi$ should satisfy the so-called Restricted Isometry Property (RIP) [25]. It has been shown that a matrix satisfies the RIP if its columns are sufficiently incoherent [26], i.e., when the largest normalized inner product between any two columns of $\Phi$, $\mu = \max_{i,j:i \neq j} \frac{\phi_i^T \phi_j}{\|\phi_i\| \|\phi_j\|}$, known as the coherence, is bounded above by a small constant. More specifically, it has been shown [26] that $\mu < \frac{1}{2K-1}$ is a sufficient condition to exactly recover a $K$-sparse signal with only $M = \mathcal{O}(K \log(N/K))$ measurements in the noise-free scenario (i.e. when $n = 0$). In the presence of noise, the same sufficiency bound holds if the magnitudes of the non-zero elements of $x$ are sufficiently large.
compare to the noise variance [27]. When these conditions hold, we can both detect the presence of bacteria and estimate their concentration using standard CS signal recovery algorithms.

2.2.2 Compressive sensing for non-negative signals

In a special class of problems where both the hybridization matrix $\Phi$ and the sparse concentration vector $\mathbf{x}$ are non-negative a more optimistic recovery bound has been proven to hold. In particular [28] defines an alternative notion of incoherence that improves the recovery guaranty. To state the new recovery bound it is essential to introduce the alternative notion for coherence known as one-sided coherence. For an arbitrary matrix $\Phi$, the one-sided coherence is defined as $\rho(\Phi) = \max_{i \neq j} \frac{\phi_i^T \phi_j}{\|\phi_i\|_2 \|\phi_j\|_2}$.

It is shown in [28] that, for a non-negative matrix $\Phi$, if a non-negative $K$-sparse solution exists such that $\rho(PD) < \frac{1}{2K-1}$, then the solution is unique and CS recovery algorithms can find it. Here $D$ is defined as the column $\ell_1$ (sum) normalized matrix of $\Phi$, and $P$ is column mean subtraction operator [28]. While the one-sided coherence recovery bound is based on the better-conditioned matrix $PD$, it remains pessimistic, and better recovery performance is typically achieved empirically in practice.
2.2.3 Greedy recovery algorithms

The concentration vector $\mathbf{x}$ is recovered from the measurement vector $\mathbf{y}$ via a sparsity-penalized optimization of the form,

$$\min_{\mathbf{x}} ||\mathbf{x}||_0, \text{ subject to } ||\mathbf{y} - \mathbf{\Phi x}||_2 < \sigma.$$ 

Here $||\mathbf{x}||_0$, known as $\ell_0$ norm, counts the number of non-zero values in the vector $\mathbf{x}$, and $\sigma$ estimates the noise standard deviation. While this optimization problem has exponential complexity, a variety of different greedy algorithms have been developed to solve it approximately. Orthogonal Matching Pursuit (OMP) [29] is an iterative greedy algorithm that, at each step, selects the column of $\mathbf{\Phi}$ that is most correlated with a residual vector and then updates the residual by projecting the measurements vector $\mathbf{y}$ onto the linear subspace spanned by the columns that already been selected. The primary advantages of OMP are its simplicity and fast convergence. Moreover, if the sparsity level $K$ (the number of bacteria is known, then the algorithm can use it as the stopping criterion.

In non-negative CS setting we can leverage the fact that both the hybridization affinity matrix $\mathbf{\Phi}$ and the sparse concentration vector $\mathbf{x}$ are non-negative to improve the performance of OMP. We utilize a variant of OMP described in [28] that is adapted to recover non-negative sparse solutions from non-negative sensing matrix $\mathbf{\Phi}$. Instead of directly working on $\mathbf{\Phi}$, this algorithm operates on the canonical matrix $\mathbf{PD}$, where $\mathbf{D}$ is the column $\ell_1$ (sum) normalized version of matrix $\mathbf{\Phi}$ defined as:

$$\mathbf{D} = \mathbf{\Phi} \mathbf{W}^{-1},$$

where $\mathbf{W}$ is an $M \times M$ diagonal matrix containing the column sums of

...
the hybridization matrix $\Phi$ and $P$ is the pre-conditioner matrix and can be chosen as any invertible $N \times N$ matrix. In the case of positive matrix $D$ an efficient pre-conditioning can be obtained just by subtracting the weighted mean of each column of $D$: 

$$PD = (I - \frac{1-\epsilon}{N}E)D,$$

where $E$ is $N \times N$ matrix of ones, $I$ is the identity matrix, and $0 < \epsilon < 1$ is a weighing constant to make the $P$ matrix invertible. Working with the preconditioned matrix $PD$ does not change the solution of the problem [28]; however it significantly improves the OMP algorithm behavior and performance guarantees. This OMP algorithm variant from [28] is constructed as follows:
Algorithm 1: Non-negative OMP

Data: $D, y$

Result: $x^i$

Initialization;

The temporary solution: $x_i = 0$; The temporary residual: $r^i = y$; The temporary solution support $S^i = support\{x^i\} = \emptyset$

while $||r||_2^2 < T$ do

Sweep: $\forall j \in [1, 2, \ldots, N]$: find $\epsilon(j) = \min_{x \geq 0} ||D_j x_j - r||_2^2$;

Update support: find $j_0$ such that $\forall j \in S^i$, $\epsilon(j_0) \leq \epsilon(j)$, and update $S^i = S^{i-1} \cup \{j_0\}$;

Update solution: compute $x^i: x^i = \min_x ||Dx - y||_2^2$ subject to $support\{x^i\} = S^i$ and $x \geq 0$;

Update residual: compute $r^i = y - Dx^i$;

end

2.2.4 Basis pursuit denoising

While greedy algorithms such as OMP attempt to solve the exact $\ell_0$ norm penalized problem, another important class of CS recovery algorithm relax the problem by replacing the $\ell_0$ norm with the more tractable $\ell_1$ norm:

$$\min_x ||x||_1, \text{ subject to} ||y - \Phi x||_2 < \sigma.$$ 

Here $||x||_1$ sums the absolute value of the entries in the vector $x$. The $\ell_1$ norm penalized optimization problem above is convex and reduces to a linear program
known as Basis Pursuit Denoising (BPDN) [25]. Surprisingly the $\ell_1$ norm penalized problem, under the same sufficient conditions on mutual coherence state above, can exactly recover a $K$-sparse signal from the noisy measurements $y$ with $M = \mathcal{O}(K \log(N/K))$ measurements.

In this thesis we are interested in solving the BPDN problem with an additional non-negativity constraint ($x \geq 0$). This problem can be efficiently solved using gradient-projection algorithm with Nestrov acceleration as described in [30, 31]. Starting from an initial guess for $x_0$, one step in taken in the direction of negative gradient, i.e., $x_k = x_{k-1} - t_k \Delta f(x_{k-1})$, (where $t_k$ is the step size and can at each step be determined by the Nestrov rule as described in [30]) and then $x_k$ is projected back to the non-negative orthant after zeroing out the negative entries of $x$ (in every iteration) until converge.

For very large scale problems where $\Phi$’s dimension is big ($M, N > 1000$), we can improve the performance of the $\ell_1$ penalized non-negative regression solver by casting the problem as a non-negative least squares problem. The non-negative least squares problem can be efficiently solved using fast active-set based algorithms such as [32]. These algorithms avoid the gradient calculation step in every iteration, which is very time-demanding in large scale problems. To formulate the non-negative least squares problem lets consider the optimization problem,

$$\min_x ||x||_1, \text{ subject to } ||y - \Phi x||_2 < \sigma, \text{ and } x \geq 0.$$ 

By replacing the objective function with the squared norm $||x||_2^2$ and bringing the
constraint into the cost function using a Lagrangian multiplier $\lambda > 0$, we have,

$$\min_{x} ||x||_1, \text{ subject to } ||y - \Phi x||_2 < \sigma, \text{ and } x \geq 0.$$ 

Since the $\ell_1$ norm of a non-negative valued vector $x$ is $||x|| = \sum_{i=1}^{N} x_i$, we can reformulate the above problems as a non-negative least squares problem

$$\min_{x} ||\tilde{y} - \tilde{\Phi} x||_2, \text{ subject to } x \geq 0,$$

that can be solved using standard methods (e.g., the \texttt{lsqlin} command in Matlab).

Here $\tilde{y} = \begin{bmatrix} \sqrt{\lambda} y \\ 0 \end{bmatrix}$, and $\tilde{\Phi} = \begin{bmatrix} \sqrt{\lambda} \Phi \\ 1 \ldots 1 \end{bmatrix}$, where the last column penalizes the $\ell_1$ norm. In the compressive sensing regime where the number of measurements $M < N$ is less than the number of columns in $\Phi$, this method implicitly assumes an additional $\ell_2$ penalty on vector $x$ in calculating the pseudo inverse of the nonnegative least squares problem, therefore seeks a slightly different solution from the classical BPDN problem solved using fast gradient based thresholding methods described earlier.

### 2.3 Whole Genomic Sequence Analysis Tools

The genome of an organism refers to its collective genetic material consisting of DNA molecules amino acid molecules in the nucleus of every cell. Incredibly, DNA naturally lends itself to a discrete structural description of its subunits, with respect to an ordering that is of great biological consequence. Each molecule of DNA is comprised of a linear sequence of discrete nucleotides a sequence of nucleic bases against a sugar-
phosphate backbone. The permutation of nucleotides in the genome is a blueprint for instructions governing every aspect of an organism’s physical and mental being, so in fact this is a very useful representation to understand. It is no coincidence that the discrete nucleotide sequence of the genome has become synonymous with the genome itself. The transcriptome and proteome of an organism (comprised of its RNA and expressed proteins, respectively) also have similar discrete signal representations.

Traditionally, deriving a complete, discrete sample representation of the genome has been a challenge due to its minuscule size and the problems associated with making such molecular-level measurements. Instead, for decades there have existed several technologies to determine shorter, partial representations of the genome by indirect means, i.e. by observing the hybridization of short sections of DNA strands with their complements, and then inferring the original DNA strand composition from it. Some such technologies are Southern blots, microarrays, and PCR. In fact, the Southern blot, named after its inventor, Edwin Southern, was so influential that it inspired later techniques named Northern, Eastern, Western and Southwestern blots for RNA and protein detection.

Dramatic change occurred with the advent of the first sequencing method, Sanger sequencing. Today, next-generation sequencing (NGS) technology, buoyed by rapid advances in optics, nanotechnology and computer science, has made it possible to derive a complete, high-fidelity sampling of the genome. Instead of observing and measuring the whole-sequence hybridization of short DNA sections, we record base-
by-base hybridization of DNA sections, using fast, redundant measurements, followed by a computational phase involving the alignment and assembly of those sections. The genomics world is at an exciting stage where several types of sequencing technologies have not only been developed, but are being further perfected every day while providing us with freely flowing genomic data. For the first time, the genetic material of every organism has the potential to become a known quantity. Unlike analog-to-digital signal converters, technologies to sense genetic material are expensive. Sequencing may be the gold standard for the rigorous assessment of genomic material, but its cost is still extremely prohibitive for practical non-research use. Until the price of sequencing falls from bench-levels to bedside-levels, it is critical to focus on other partial sampling methods like microarrays, molecular beacons and microfluidic devices that are readily available and less expensive for practical purposes.

Until then, we can take advantage of the fact that an organisms DNA is almost unvarying over its lifetime, so it only needs to be sequenced once to be known. We can leverage our knowledge of fully-sequenced genomes to make informed pronouncements on the data that partial-sampling based technologies provide us. The technology that we focus on re-engineering is molecular beacons. We discuss DNA hybridization models in this section before delving into the molecular beacon technologies in the next section.
2.3.1 DNA Thermodynamic Hybridization Model

Practically all present-day genomic sampling tools are based on complementary hybridization. The DNA molecule of interest binds with its complement; this binding is measured, and the original DNA molecules nucleic acid composition is inferred from it. Tools like molecular beacons use the complementary binding between probe and target DNA strands, and then measure emitted fluorescence of each bound section. Therefore it is especially important to establish a model to predict the binding strength between a probe and the target is intended for, in order to design them appropriately. Sequencing-by-synthesis technologies, as the vast majority of NGS technologies are, also depend on complementary hybridization. However they use single base binding and fluorescence to ascertain the DNA sequence one base at a time, so there is no probe design component and hence no critical dependence on a hybridization model. Single base binding is essentially binary—either it is bound and fluoresces, or not. There are several factors that influence the hybridization affinity of a DNA fragment with another (e.g. Smith-Waterman, etc.). We discuss a thermodynamic based approach here: the SantaLucia Nucleic Acid Thermodynamic model, which is based on the empirical and theoretical predictions of the bound and unbound free energies of any pair of nucleotides [28].

Our goal is to estimate the probe-target hybridization intensities at each probe spot, given probe and target DNA sequences. We aim to use a thermodynamic model for DNA-DNA hybridization to give us a physically realistic model for hybridization
affinities. For instance an A-T bond formation is a double covalent bond between hydrogen atoms, which releases relatively less free energy since it is relatively less stable than a C-G bond, which is a triple covalent bond between corresponding hydrogen atoms on each base. More complicated to quantify are the free energies when there are base mismatches.

Unlike many other models, this model, also incorporates thermodynamic parameters for mis-hybridizations between two DNA sequences (accredited to SantaLucia [28]). The frequent occurrence of base mismatches in probe-target hybridizations during the probe design phase makes it critical that we estimate affinities in the face of these mis-hybridizations as accurately as possible. Using Smith-Waterman sequence alignments to do so is inexact for cases of mismatches, since the algorithm is highly dependent on the type of penalty that gaps are assigned by the alignment scoring matrix in use. There is no consensus on which penalty values most accurately reflect hybridization reality. The SantaLucia thermodynamic model is commercially available from the company DNA Software, Inc. We purchased two component software packages: (1) ThermoBlast, which performs fast alignments of sequences against large genome databases to discover thermodynamically stable hybridizations, and (2) Visual OMP DE, which can simulate hybridization experiments with detailed solution conditions through scripts launched from the command line, and generate results for melting temperatures (Tm), Gibbs free energy (∆G), and the percentage-based concentration of each resultant species post-experiment. There is also the capabil-
ity to visually generate the secondary structure for each monomer, homodimer and heterodimer species formed from the constituent probes and target fragments. We follow a two-step procedure in determining the hybridization intensities of a given set of probes against a given dictionary of potential targets. First, we run each probe against each double stranded target genome, using ThermoBlast, to discover all possible target fragments at which there is significant alignment. Second, we use Visual OMP DE to run a script to simulate each hybridization experiment between a single probe and a single target, using these target fragments to simulate the target. Every experiment simulation script contains information on: the probe sequence, the target fragments from ThermoBlast it was shown to hybridize with, and conditions for the experiment. Experimental conditions included probe and target concentrations, assay temperature, hybridization cocktail composition (Mg++, Glycerol, DMSO, Formamide, TMAC, Betaine concentrations), and pH. This procedure is repeated for each probe-target pair. The hybridization results of each experiment in Visual OMP DE produces a data file, which is parsed for the concentrations of resultant species at the end of the hybridization. In each data file we use the percentage of probe-target heterodimer structures formed, i.e. the percentage of target fragments that are bound to probes, as an estimate for the hybridization affinity of that probe-target pair.
2.4 Molecular Beacons

2.4.1 Principle of Operation

Molecular Beacons (MBs) [33] are single-stranded, free-floating oligonucleotide hybridization probes in a stem-loop structure with two fluorescent dyes, Cy3 and Cy5, bound to the 5′ and 3′ ends. When a target sequence presents a more stable binding with the loop section than the binding between its two stem ends, MBs open and report the presence of a target molecule through the fluorescence resonance energy transfer (FRET) between the Cy3 and Cy5 molecules (see Figure 2.4.1). The loop portion of a conventional MB is a probe sequence that is complementary to a predetermined sequence in a target nucleic acid. The stem is formed by the annealing of two complementary arm sequences that are on either side of the probe sequence. The stem keeps the fluorescent dyes in close proximity to each other, causing the fluorescence of the fluorophore to be quenched by FRET. When the probe encounters a target molecule, it forms a hybrid that is longer and more stable than the hybrid formed by the arm sequences. Since nucleic acid double helix is relatively rigid, formation of a probe-target hybrid precludes the simultaneous existence of a hybrid formed by arm sequences. Therefore, the probes undergo a conformational change that forces the arm sequences apart and causes the fluorophore and quencher to move away and fluoresces when illuminated by ultraviolet light.
Figure 2.1: Molecular Beacons (MBs) [1] Principle of Operation. The probe sequence in the loop binds spontaneously to the target sequence which is exact complement to the probe sequence. Adapted from [2].

2.4.2 Advantages for Microbial Detection

There are many advantages to using MBs for bacterial detection. First, they are free-floating rather than immobilized to a surface (as in microarray) and therefore have a great probability of making contact with a target molecule, resulting in faster reaction kinetics. Second, the probe-target hybrids need not be isolated from their surplus probes prior to analysis, making them more amenable for field use. Another advantage of MBs is their relatively high signal-to-background ratio, providing higher sensitivity of detection [34]. Upon hybridization to its target, a well designed MB can generate as high as 200-fold fluorescence enhancement under optimal conditions [1].
This provides the MBs with a significant advantage over other fluorescent probes in ultra-sensitive analysis.

2.4.3 Sloppy Molecular Beacons

A traditional MB only reacts to a target sequence that is compete complement to the loop section of the probe. This require upwards of 100 different species-specific probes in many PCR assay tubes for a comprehensive screening procedure. A different approach was introduced in [4] as a screening technique for bacterial identification that requires only a single gene amplification assay containing a set of four differently colored molecular beacon probes of a unique design. After the completion of amplification, the molecular beacons in the reaction tube are hybridized to the amplicons at a relatively low temperature. The temperature is then slowly raised to determine, from the consequent changes in fluorescence intensity of each molecular beacon in the set, the temperature at which each of the four probe-target hybrids melts apart (Tm). The resulting set of four Tm values uniquely identifies the species present in the sample.

The key feature that enables this design to react to more that one target organism is the non specificity introduced in the probes. Unlike sequence-specific molecular beacons [1], which possess short probe sequences (18 to 26 nucleotides long) that form probe-target hybrids only with perfectly complementary, or nearly perfectly complementary, target sequences, these sloppy molecular beacon probes are designed
to form probe-target hybrids with the amplicons generated from all of the species that we wish to identify. In order to enable this property, the probe sequences in these molecular beacons are about 40 nucleotides long. Consequently, they form probe-target hybrids even if the duplexes possess a substantial number of mismatched base pairs.

Sloppy MBs are among the first novel practical approaches that enable non specific sensing of a group of bacteria. The design, however, is based on probes that are meticulously designed to target the 16S rDNA in different bacteria genomes and therefore inevitably have the limitation of other gene (marker)-specific approaches. The most important of them is probably that they can not be used to identify unknown bacterial species whose 16S is not available (sequenced) yet. The applicability of SMBs to other microbial organisms (such as viruses) is also not investigated.
Chapter 3

Universal Microbial Diagnostics

3.1 Universal Microbial Diagnostics Platform

3.1.1 The Platform Overview

In UMD (Figure 3.1.1 A), the genomic DNA of an infectious sample is extracted and exposed to a small number $M$ of DNA probes, which hybridize to the genomic DNA at various locations; this hybridization is experimentally quantified, producing a probe-binding (or hybridization) vector $y$ whose entries correspond to the hybridization binding level of each probe with the microbial sample. *A priori*, the hybridization binding level of each probe to a reference database of $N$ bacterial genomes is obtained and stored in an $M \times N$ hybridization affinity matrix $\Phi$ (Figure 3.1.1 B). The hybridization affinity matrix can be measured either experimentally *in vitro* or predicted computationally *in silico*. To speed up the probe design and prove the concept of UMD, here we predict the affinity matrix using a thermodynamic model *in silico*. To compute the entry $\phi_{ij}$ in the matrix $\Phi$, the hybridization binding level of probe $i$ to genome $j$ (Figure 3.1.1 B), we start by cutting the genome up into 700-1000 nucleotide fragments. The fragment-probe mixture is then fed into a thermodynamics-based hybridization model [3]. This model predicts all possible stable probe-bacteria
fragment bindings along with their resulting concentrations for a given set of experimental conditions (Figure 3.1.3). The overall hybridization affinity $\phi_{ij}$ is computed by summing the concentrations of all predicted, stable probe-fragment bindings for a unit concentration of bacterial genome.

With an excess concentration of probes as compared to sample DNA, the probe-binding vector $\mathbf{y}$ consists of a linear combination of the predicted hybridization affinities of the species in the reference genome database (the columns of the matrix $\mathbf{\Phi}$, with $\phi_i$ denoting the $i^{th}$ column) weighted by their concentrations $\mathbf{x}$; i.e., $\mathbf{y} = \mathbf{\Phi x} + \mathbf{n}$, where the vector $\mathbf{n}$ accounts for noise and modeling errors (Figure 3.1.1 C).

The two key goals of the UMD platform are to (i) detect the presence and (ii) estimate the concentrations $\mathbf{x}$ of a potentially large number $N$ of reference microbial genomes in an infectious sample given only a small number $M \ll N$ of probe-binding vector measurements $\mathbf{y}$. Simply inverting the matrix $\mathbf{\Phi}$ is impossible in this case, since it has many more columns than rows. Fortunately, it is reasonable to assume that only a small number $K$ of microbial genomes will be present in a given sample, in which case the concentration vector $\mathbf{x}$ is sparse, with $K$ nonzero and $N - K$ zero (or close to zero) entries; when $K < M$, one can hope to invert $\mathbf{\phi}$ to estimate the $K$ nonzero concentrations. More rigorously (see background for detailed statements), when the columns of $\mathbf{\phi}$ are sufficiently incoherent and when $M = \mathcal{O}(K\log(N/K))$, we can apply the theory of CS to recover the concentrations $\mathbf{x}$ from the measurements.
Figure 3.1: Schematic of Universal Microbial Diagnostics (UMD) platform. (A) Genomic DNA is extracted from a bacterial sample and thermal cycled with $M$ random DNA probes. The genome-probe binding is quantified, producing a probe-binding vector $y$; in this study the random probes are in the form of molecular beacons (MBs), and the DNA-probe binding is quantified by the ratio of open/hybridized to closed/non-hybridized MBs. (B) The hybridization binding level of each probe to a potentially large reference database of $N$ bacterial genomes (B1, B2, ..., BN) is predicted using a thermodynamic model and stored in an $M \times N$ hybridization affinity matrix $\Phi$ (here $N = 10$ for illustration, but it could be very large in practice). (C) Assuming just $K$ bacterial species comprises the sample, the probe-binding vector $y$ is a sparse linear combination of the corresponding $K$ columns of the matrix $\Phi$ weighted by the bacterial concentrations $x$, i.e., $y = \Phi x + n$, where the vector $n$ accounts for noise and modeling errors. When $K$ is small enough and $M$ is large enough, $\Phi$ can be effectively inverted using techniques from compressive sensing (CS), yielding the estimate for the microbial makeup of the sample $x$; in this illustration the $K = 2$ bacteria labeled B2 and B7 are present in the sample.
\( \mathbf{y} \) via a sparsity-penalized optimization of the form

\[
\min_{\mathbf{x}} ||\mathbf{x}||_0, \text{ subject to } ||\mathbf{y} - \Phi \mathbf{x}||_2 < \sigma
\]

Here \( ||\mathbf{x}||_0 \) counts the number of non-zero values in the vector \( \mathbf{x} \), and \( \sigma \) bounds the energy of noise vector \( \mathbf{n} \). Since \( M = \mathcal{O}(K \log(N/K)) \) scales only logarithmically in \( N \), the UMD platform has the potential to identify and estimate the concentrations of a large number \( N \) of potential microbial genomes using only a small number \( M \) of measurement probes. What remains is to ensure that the columns of \( \Phi \) are incoherent. We do this by using DNA probes whose sequences are generated via a random permutation of nucleotides. In this chapter we experimentally validate the incoherence of a small, fixed set of randomly selected probes in vitro and measure the average performance of many sets of randomly selected probes for pathogenic detection \textit{in silico} at the noise level observed in the experiment.

### 3.1.2 Linearity assumption considerations in UMD

In the UMD platform the probe concentration \((1 \times 10^{-6} M)\) is in far excess of the target organism concentrations \((\sim 1 \times 10^{-10} M)\), therefore we are able to linearly combine the hybridization affinity signatures that we measure for individual targets using the hybridization model. Due to the flooding of excessive number of probes, each target fragment has its choice of binding/not binding to the probes, and thus we can safely sum together multiple target interactions of the same probe, assuming them to be independent.
3.1.3 Hybridization Affinity Matrix Calculations

A comprehensive thermodynamic model by SantaLucia et al. [3] was utilized to predict the hybridization of probes to bacterial genomes. The frequent occurrence of base-pair mismatches in hybridization of random probes to DNA sequences made it critical that we estimate the affinities as accurately as possible. The SantaLucia model incorporates thermodynamic parameters for mis-hybridizations between two DNA sequences. We purchased two software packages: ThermoBlast DE, which performs fast alignment of sequences against large genome databases to discover thermodynamically stable hybridizations, and Visual OMP DE, which simulates hybridization experiments with detailed solution conditions and generates results for melting temperature (Tm), Gibbs free energy (∆G), and the percentage-based concentration of each resultant species post-experiment. The secondary structure of each monomer, homodimer, and heterodimer species formed from the constituent probes and target fragments can also be visualized.

We developed two procedures to determine the hybridization affinity of a given set of random probes against a given dictionary of potential bacterial targets: the fast method and the parameter-free method.

In the fast method, we ran random probe sequences against the double-stranded target genome using ThermoBlast to discover all possible locations at which there was significant alignment (threshold at Tm = 45°C, see later in this chapter for more information about this choice). The hybridization affinity was reported as the number
of significant alignments per molecule of the genome (Figure 3.1.3).

Figure 3.2: The fast approach for modeling of the hybridization affinity of a random probe sequence (red) to a bacterial genome (blue). Hybridization affinity was calculated by counting the total number of significant binding spots of a random probe sequence to the genome under the experimental thermodynamic conditions. All bindings with melting temperature greater than $T_{m_{\text{thresh}}} = 45^\circ C$ were considered significant (see later this chapter for more details on the choice).

In the parameter-free method, however, we first randomly cut double strands of the target genome into fragments of fixed length (700-1000 bps) and then used Visual OMP DE to simulate the hybridization between the probe and the target genome, using the target fragments. Every simulation contained information on the probe sequence, the target fragment sequences, and conditions for the experiment, including probe concentration (1 $\mu$M), target concentration (500 ng/$\mu$L), assay temperature ($4^\circ C$), hybridization buffer composition (4 mM Mg++, 50mM Na+, Glycerol, DMSO, Formamide, TMAC, Betaine concentrations), and pH (=8). This procedure was repeated for each probe-target genome pair. The hybridization result of each experiment in Visual OMP DE was parsed for the concentration of resultant species at the end of the hybridization. We used the percentage of probe-target heterodimer structures formed, i.e., the percentage of probes that are bound to target fragments, as an estimate for the hybridization affinity of the probe to each target (Figure 3.1.3).
Figure 3.3: Probe-bacteria hybridization affinity computation process. Both strands of the bacterial genome (blue lines) are randomly segmented into roughly equal length (∼700 – 1000 nts) fragments. All of the bacterial fragments and probe sequences along with the experimental conditions are fed into the DNA software [3] to predict all stable probe-bacteria complexes and concentrations. These concentrations, in aggregate, determine the concentration of opened beacons, which is defined as the hybridization affinity of the probes to the bacterial genome.

3.2 UMD probe design

This universal sensing strategy can take any physical embodiment (e.g., quantitative real time PCR or DNA microarrays [35]) for the detection of any DNA sequence (bacterial, viral, or fungal). To test and validate the concept, we recovered pathogenic bacteria using random probes in the form of mismatch-tolerant Sloppy Molecular Beacons (Sloppy MBs) [4].

In a conventional MB for bacterial detection, the loop sequence is designed to target specific regions (e.g., 16S rDNA) within a single bacterium [14] or multiple bacteria [4]. In the MB probes for UMD, the loop sequence is selected as a random sequence of length 38 nts, and the 4 nt-long stem sequence is consistent across all
probes (see Figure 3.2), although other choices might be utilized (different design tradeoffs are discussed in [36]).

![DNA sequence structure of three test random DNA probes. Nucleotides in black color are fixed to preserve the hairpin structure, and other nucleotides are randomly chosen.](image)

**Figure 3.4** DNA sequence structure of three test random DNA probes. Nucleotides in black color are fixed to preserve the hairpin structure, and other nucleotides are randomly chosen.

In the design of molecular beacons (MBs) [1] for random DNA probes, the length and GC content (ratio of G+C to other nucleotides) of the probe loop and stem sequences were considered to strike a balance between two factors: fluorescence signal level and probe stability. Signal intensity was especially important in our detection scheme, since no DNA amplification method (such as PCR) was utilized. Similarly to sloppy molecular beacons (sloppy MBs) [4], we selected the random probes loop sequence to be longer than traditional MBs. In addition, we made the stem sequence one nucleotide shorter to introduce additional sloppiness. The unusually long loop and short stem enable our random probes to form hybrids with several base pair mismatches across the entire bacterial genome and compensate for the lower signal
intensity in the absence of DNA amplification methods such as PCR. Figure 3.2 shows an example of a predicted probe-bacteria fragment binding with many base pair mismatches. Clearly both the loop and arm sequences contribute in forming the probe-bacteria complex.

Figure 3.5: Example of a predicted probe-bacteria fragment binding with many base pair mismatches.

The random probes were designed to maintain the MBs signature hairpin structure over a wide range of temperatures (4 – 50°C). One potential approach to ensure this is to generate random MBs and then filter out the probes with undesired secondary structures or melting temperatures. However, this would be extremely time consuming and might even require visual verification of the probes. We therefore developed a new MB structure that avoids probe self-folding altogether (Figure 3.2). This was carried out using the following heuristic. We limited the number of Cyto-
sine (C) molecules in the loop sequence, positioned them maximally far apart from each other, and fixed their neighboring nucleotides to avoid potential G-C pairing that could fold the probe. We also fixed the nucleotides next to the stem sequences to avoid the random generation of probes with longer stems (longer stems make the probe less sloppy). All the other nucleotides were randomly selected with uniform probability, as depicted by the colored letters in Figure 3.2. Figure 3.2 shows the gain in hybridization affinity obtained using this probe design technique in comparison with traditional MBs and sloppy MBs in binding to the E.coli genome. By no means is this the only method to generate random probes for a UMD platform; any method that produces probes with stable hairpin structure and uniform melting temperature while providing the required signal intensity can be utilized.

3.3 Experimental proof of concept

3.3.1 Experimental set up

To prove the UMD concept *in vitro*, we mixed three UMD MBs (GC-contents 50, 56.5, and 60.8%, identical melting temperature of 40°C, and concentration 1µM) with genomic DNA from each of three human infectious bacterial strains: *E. coli*, *S. aureus* and *F. tularensis* (*Escherichia coli* str. K12 substr. MG1655, Methicillin-resistant *Staphylococcus aureus* str. USA300, *Francisella tularensis* subsp. holarctica str. LVS). For each MB-bacterial species pair, equal volumes of probe and bacterial DNA were combined and subjected to a thermal cycling process of denaturing (95°C) and
Figure 3.6: The predicted number of opened probes when a bacterial genome (*E. coli*) is exposed to a specific (in the same thermodynamic condition as the experiment) traditional molecular beacon (MB) [1], sloppy MB [4], and MB created according to our design rules. While the nucleotides in the loop were determined randomly, the percentage of opened beacons (hybridization affinity to *E. coli*) is substantially increased for the random probe MB, thanks to the choice of stem length and configuration of a small number of fixed nucleotides in the loop sequence.

binding/cooling to 4°C overnight. The fluorescence intensities of Cy3 and Cy5 were measured with a fluorometer, and the FRET ratios were calculated by computing Cy5 intensity over total fluorescence intensity (Cy3 + Cy5). The FRET ratios from binding of *E. coli*, *S. aureus*, and *F. tularensis* to the MBs were [0.55, 0.66, 0.69], [0.51, 0.65, 0.71], and [0.51, 0.60, 0.72], respectively (Figure 3.3.1). All experiments were performed in triplicate, and the results shown here average over the trials with
the error bars representing the standard error of the mean.

Figure 3.7: Experimentally measured FRET ratios to quantify hybridization between bacterial DNA (E. coli, S. aureus, and F. tularensis) and probes 1, 2, and 3.

3.3.2 Probe characteristic curve

In order to estimate the bacterial concentrations in physical units, we translated the FRET ratio of each bacterium-MB pair into an opened MB concentration or hybridization affinity, represented in units of molarity. For this, we experimentally obtained and fit FRET ratios for each of the three molecular beacons as a function of the concentration of their exact probe complements, using an optimization method described by Jeričević et al. [5] (Figure 3.3.2). The $R^2$ values for the fit ranged from 0.9517 to 0.9889, suggesting a satisfactory fit, Figure 3.3.2. Therefore, based on the fit equations, the hybridization affinities corresponding to the FRET ratios for E. coli, S. aureus, and F. tularensis were calculated to be: [1.30, 1.31, 1.54], [1.70, 1.52, 1.29]
and [1.72, 2.28, 0.91] (×10^{-7}M), respectively. We refer to these measurements as the measured affinity vectors and show them in Figure 3.3.2.

Figure 3.8: Probe characteristic curves for random probes 1, 2, and 3 for determining the experimental relationship between FRET ratio and open beacon concentration (hybridization affinity). The characteristic curve is fit to the measured FRET ratio as a function of the concentration of the probes exact complement for three random probes. Error bars represent the standard error of the mean (SEM) for 3 trials.

<table>
<thead>
<tr>
<th></th>
<th>Random probe 1</th>
<th>Random probe 2</th>
<th>Random probe 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.5970</td>
<td>0.6400</td>
<td>0.5833</td>
</tr>
<tr>
<td>b</td>
<td>0.4500</td>
<td>0.3100</td>
<td>0.0482</td>
</tr>
<tr>
<td>n</td>
<td>1.9</td>
<td>1.6</td>
<td>4.0</td>
</tr>
<tr>
<td>FRET_0</td>
<td>0.1657</td>
<td>0.1976</td>
<td>0.1638</td>
</tr>
<tr>
<td>R^2</td>
<td>0.9517</td>
<td>0.9722</td>
<td>0.9889</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.1587</td>
<td>0.1274</td>
<td>0.0918</td>
</tr>
</tbody>
</table>

Figure 3.9: Characteristic curve parameters a, b, and n for three test random probes, and the associated curve fit performance criteria, $R^2$ and root mean square error (RMSE). Parameters were fit to the following curve: $FRET = FRET_0 + \frac{a}{1+b(\log(10^{-6})-\log(c))^{-n}}$ [5].
Figure 3.10: Hybridization affinity between DNA samples and probes, converted from FRET ratio (as in Figure 3.3.1) through the probe characteristic curve fit equations.

3.3.3 Bacteria detection and concentration Estimation \textit{in vitro}

Our challenge is to decode the measured affinities of some unknown bacterial species reacting to UMD molecular beacons. With the predicted hybridization affinities of \( N=3 \) bacteria (\textit{E. coli}, \textit{S. aureus} and, \textit{F. tularensis}) to \( M=3 \) random probes stored in the computationally obtained \( \Phi_{3 \times 3} \), we used a variant of the Orthogonal Matching Pursuit (OMP) algorithm, as described in [28], and successfully detected all three species. To further provide the physician with a metric quantifying how close the measured affinity vector is to the other bacteria in the database, we calculated (Figure refmainresult) the inner product of the normalized measured affinity vectors from the three experiments to the columns of the centered and sum-normalized matrix \( \Phi_{3 \times 3} \) (also used in [28] to define one-sided coherence, see the background chapter for
more info). In all experiments, the hybridization affinity vector had the highest inner product with the bacterial species that was present, as OMP suggested. We also estimated the bacterial concentrations using OMP with errors of 7.74, 23.5, and 2.94%, for *E. coli*, *S. aureus* and, *F. tularensis*, respectively (Figure 3.3.3).

<table>
<thead>
<tr>
<th>Inner product</th>
<th>Exp. 1</th>
<th>Exp. 2</th>
<th>Exp. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>0.83</td>
<td>-0.98</td>
<td>-0.49</td>
</tr>
<tr>
<td>S. aureus</td>
<td>-0.78</td>
<td>0.96</td>
<td>0.42</td>
</tr>
<tr>
<td>F. tularensis</td>
<td>-0.99</td>
<td>0.93</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Figure 3.11: The inner product of the experimentally obtained hybridization affinity (Figure 3.3.2) and center sum normalized predicted (by thermodynamic modeling) hybridization affinity for three bacterial DNA samples (*E. coli*, *S. aureus*, and *F. tularensis*) as a measure of the similarity of the probe measurements to the bacteria in the database. Orthogonal Matching Pursuit (OMP) correctly recovers the highlighted bacterium that was present.

Figure 3.12: Comparison of the predicted concentrations of three bacterial DNA (*E. coli*, *S. aureus*, and *F. tularensis*) (gray bars) with the experimentally measured values (black bars).

To estimate the experimental and modeling noise level and to assess the consis-
tency of the measured affinities ($\mathbf{y}$) to the predicted affinity vectors ($\mathbf{\phi}^p$), we evaluated the ratio of the predicted binding vector energy, $||\mathbf{\phi}^p||$, to the binding error $||\mathbf{\phi}^p - \mathbf{y}||$ in dB for *E. coli*, *S. aureus*, and *F. tularensis*. The predicted binding signal-to-noise-ratio ranged from 15.85 to 34.95dB, suggesting high accuracy of the hybridization affinity prediction model (Figure 3.3.3).

![Figure 3.13](image-url): The ratio of the hybridization pattern signal to the hybridization pattern noise ranges from 23 to 35 dB.

### 3.3.4 UMD on most common human pathogens

To prove the UMD concept when $M \ll N$, we expanded the reference genome database to contain 38 genera listed by the Centers for Disease Control and Prevention as most commonly pathogenic to humans [37]. One signature species from each genus was picked to construct a full genus level hybridization affinity matrix $\mathbf{\Phi}$, with $N=38$ columns and $M=3$ rows, representing the predicted binding between
each of the 38 genus-level bacterial DNAs with the 3 random probes used in the experiments above. The OMP algorithm correctly identified the *F. tularensis* genus from the 37 other genera. *S. aureus* was correctly identified as being closest (having highest inner product) to *Staphylococcus* and one other genus (*Rickettsia*). Finally, *E. coli* was identified as closest to *Escherichia* and 6 other genera (*Clostridium, Coxiella, Helicobacter, Shigella, Vibrio, Yersinia*) with similar binding patterns, suggesting more random probes are required for more sensitive and specific genus level detection within this group of 38 genera.

### 3.4 Extension by simulation

#### 3.4.1 Genus level single bacteria detection using UMD

To numerically evaluate the minimum number of random probes $M$ required to perform complete genus-level single bacterium detection using UMD *in silico*, we generated 300 random probes with three different GC contents (50, 56, and 60%) in equal proportions. We then randomly selected $M = 1, 2, \ldots, 8$ probes and predicted the hybridization affinity matrix $\Phi_{M \times 38}$ using an alternative fast binding prediction method.

To simulate the hybridization affinity vector *y in silico*, we modeled the measurement noise and modeling error as i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$ (with standard deviation $\sigma = 0.025$ for measurement noise and $\sigma = 0.123$ for modeling error), added to the predicted hybridization affinities. The detection simulation was conducted as follows. A realization of a $K = 1$-sparse signal $x$ was generated (for example, $[0, 0, 0, 1, 0, \ldots, 0]$
in transpose), where only one bacterium is present with unit concentration. A set of $M$ probes was drawn, and the hybridization affinity matrix $\Phi_{M \times 38}$ for that probe set was calculated using the fast hybridization affinity prediction method described earlier in this chapter. Then, the hybridization affinity vector $\mathbf{y}$ was calculated as $\mathbf{y} = D \Phi_{M \times 38} + \mathbf{n}$, where $D$ denotes the column sum normalized version of $\Phi_{M \times 38}$. The i.i.d. Gaussian noise $\mathbf{n}$ was added to simulate real experimental and modeling noise with the variances observed in the experiments. The simulated hybridization affinity vector $\mathbf{y}$ and matrix $\Phi_{M \times 38}$ were fed into the OMP detection algorithm described in [28] as sketched earlier in the background. This procedure was repeated for 100 random probe selections for each of the 38 realizations of sparse vector $\mathbf{x}$, and then averaged over to obtain an overall metric for detection performance.

For any given $M$, the average sensitivity (true positive rate) of the recovery algorithm over 100 iterations and 38 realizations of sparse vector $\mathbf{x}$ is calculated as,

$$\text{Sensitivity} = \frac{\text{Total number of true positives}}{\text{Total number of positives}} = \frac{\text{Total number of true positives}}{(100 \times 38)}.$$

Specificity (true negative rate) is calculated as

$$\text{Specificity} = \frac{\text{Total number of true negatives}}{\text{Total number of negatives}} = \frac{\text{Total number of true negatives}}{(100 \times 38 \times 37)}.$$

The denominator indicates that there were 37 negatives per realization of $\mathbf{x}$ per random probe trial. Figure 3.4.1 shows the average sensitivity and average specificity in recovering the simulated vector $\mathbf{y}$ of all 38 bacteria corrupted by two noise profiles (measurement and modeling noise) using the OMP algorithm from $M$ random DNA probes. The error bars show the standard
deviation over 100 trials with different random probes. Only three random probes are sufficient to recover all 38 bacteria at the measurement noise level, while five random probes are sufficient after incorporating the modeling error.

![Genus-level single bacterium detection](image)

**Figure 3.14**: Genus-level single bacterium detection. Blue lines show the average sensitivity (true positive rate) of the OMP algorithm in recovering 38 species from the most common human infectious genera, simulated by adding i.i.d. Gaussian noise with standard deviation $\sigma = 0.025$ (dashed line) to account for the measurement noise and $\sigma = 0.123$ to account for the modeling error. The probes were randomly selected from a dictionary of 300 probes (100 probes each of GC content 50, 56, 60%). The red lines show the average specificity for the two different noise levels. The error bars show the standard deviation of 100 simulation trails. There exist sets of $M = 3$ and $M = 5$ random probes (among the 300 probes) that enable detection with 100% sensitivity and specificity at these measurement and modeling noise levels.

### 3.4.2 Species level single bacteria detection using UMD

To evaluate UMD performance for species-level bacterial detection, we focused on differentiating Methicillin-resistant *Staphylococcus aureus* (MRSA) strain USA300 TCH1516 from eight other species of the same genus (*Staphylococcus*) *in silico* with
the method described above. We followed the same procedure but using bacterial species from a single genus. The hybridization affinity matrix $\Phi_{M \times 9}$ was calculated using the 9 Staphylococcus species listed in the Appendix A. Only five random probes (Figure 3.4.2) were sufficient for detecting MRSA from other Staphylococcus species despite the presence of high genome similarity. This further underscores UMDs potential to differentiate pathogens at high taxonomic resolution. Random probes bind to an organisms whole genome, including the regions causing antibiotic resistance; their differential binding in various pathogenic serotypes enables sub-species level detection.

Figure 3.15 : **Species-level single bacterium detection.** The average sensitivity (blue lines) and specificity (red lines) in species level detection of MRSA among 8 Staphylococcus species using OMP. There exist many sets of $M = 3$ and $M = 5$ probes (among the 300 probes) that enable detection with 100% sensitivity and specificity at the same measurement and modeling noise levels as above.
3.4.3 Genus level double bacteria detection using UMD

UMD has the unique advantage that it can recover more than a single ($K > 1$) organism in the sample. To evaluate the minimum number of probes $M$ required for this task we used the Basis Pursuit De-Noising (BPDN) algorithm (as described in the background) for $K = 2$ equi-concentration bacterial species detection, given a $\Phi_{M \times 38}$ from 38 human pathogenic bacteria (Figure 3.4.3). Given 38 possible bacterial species, we simulated all $(38 \times 37)/2$ different $K = 2$ sparse realizations of $x$ (for example, $[0,0,0,\ldots,0,1,0,\ldots,0,1,0,\ldots,0]$ in transpose). Both species were constrained to have equal concentrations, a more difficult recovery constraint for CS decoding algorithms than if they were different [38]. For every possible realization of $x$, we selected $M$ random probes and simulated noisy measurement and recovery as described in the previous section. This was iterated 100 times for each value of $M$ in order to average over recovery results. Average sensitivity (true positive rate) in recovering $K = 2$ species is calculated as: Sensitivity = Total number of true positives / $(100 \times 38 \times 37/2)$. Specificity (true negative rate) is calculated as Specificity = Total number of true negatives / $(100 \times 38 \times 37 \times 36/2)$. The total number of negatives per simulation was 36; the denominator sums this over all $(100 \times 38 \times 37)/2$ simulations.

We found that $M = 9$ and $M = 17$ probes achieve sensitivity and specificity of 100%, respectively, suggesting that we can recover multiple bacterial species in the same sample by using additional probes.
Figure 3.16: **Genus-level two bacteria detection.** Basis Pursuit De-Noising (BPDN) was utilized to recover the noise-corrupted sum of all the bacteria pairs ($K = 2$) present in the database. At the above measurement and modeling noise levels, $M = 9$ and $M = 17$ probes achieve sensitivity and specificity of 100%, confirming our intuition that additional probes are required to recover $K = 2$ bacteria.

### 3.4.4 Greedy probe selection

While just a small number of random probes are sufficient for detecting a single bacterium ($M = 5$ probes) or two bacteria ($M = 17$ probes) with 100% sensitivity and specificity, there is a chance that a set of randomly selected probes could have a weaker performance (as quantified by the variance error bars in Figures 3.4.1, 3.4.2, and 3.4.3). The set of $M$ probes with the best detection performance can be found by searching the entire probe space. This combinatorial search becomes computationally intractable when the number of probes grows large. We thus developed a rapid probe selection method that we dub Greedy Probe Selection (GPS) (see the Supplementary Material). Figure 3.4.4 compares the detection performance and computational time.
for GPS probes (solid line) with naïve combinatorial search (dashed line). With a small sacrifice in sensitivity, GPS found the best performance probe in a few seconds, exponentially faster than the naïve search method. Figure 3.4.4 show the simulated UMD confusion matrix in detecting pathogenic bacteria using $M = 3$ and $M = 7$ probes selected by GPS respectively. As CS theory guarantees, when the number of probes grows, detection performance improves for all the species in the database universally.

Figure 3.17: **Performance and computational complexity comparison of probe selection algorithms.** Figure shows the average sensitivity in detecting 38 human infectious species of different genera using the proposed Greedy Probe Selection (GPS) algorithm (blue line). GPS-selected probes achieve 100% sensitivity with $M = 7$ probes, while on average an $M = 5$ random probe set (see A) achieves the same performance. However, the computational time for selecting probes using GPS grows only linearly with the number of probes (purple solid line), while the time for searching the complete random probe space grows exponentially (purple dashed line).
Figure 3.18: A confusion matrix summarizing UMD pathogenic bacterial detection in simulation, using $M = 3$ and $M = 7$ probes selected by GPS.
Chapter 4

Discussion and Conclusion

In this thesis we have described the application and implementation of the Universal Microbial Diagnostics (UMD) platform to the problem of bacterial detection. We see that exploiting the sparsity inherent to a detection problem can bring several different advantages to microbial detection schemes.

The UMD approach to sensing advocates a paradigm shift in biological detection methods: universal sensing. UMD probes are universal in the sense that a fixed set of probes can capture the salient information required to distinguish between members of a growing database of species, up to a certain point of growth [39]. Instead of creating probes that are specific to what we are interested in sensing, we create a probe set that is drawn completely independent of any targets. This approach opens up the possibilities for detection of undiscovered, mutated or newly engineered bacteria, which would otherwise go unnoticed if specific, tailored probes are used. Given the millions of bacterial species estimated to exist naturally both with harmful and vital purposes, and the potential dangers of ill-purposed synthetic bacterial species being used as bio-weapons, it is important for humankind to look to new methods of nonspecific sensing in this domain.

UMD also offers a platform that contend with this data deluge from new microbial
species being discovered and sequenced every day [40]. We speculate that the recovery condition required by CS generally hold in detecting any newly sequenced DNA strand. The hybridization affinity of random probes can, therefore, be computed for that DNA strand universally and utilized in UMD for detection.

The UMD platform also holds promise to eliminate PCR. Compared to the conventional unique 16S or 23S ribosomal DNA identifier approach, here not only does each probe oversample the target set by binding to multiple targets, each probe over samples each genome it binds to by hybridizing in multiple places. It is this multiple probe-binding mechanism throughout a genome that suggests the approach of whole genome sensing instead of single gene-based sensing. Random probes bind with genomes on the principle that there are multiple binding spots on each genome which are both unique and shared by various bacteria, and by exploiting them we are able to be flexible in our probe generation method, leading to advantages like nonspecific sensing. The hypothesis that PCR may not be necessary in the case of whole genome sensing is that with enough binding sites on a genome for any single probe type, there is enough cumulative fluorescent intensity that amplification is unnecessary.

Moreover, while most other biological sensing methods (such as PCR based methods, group testing, etc.) reduce to selecting among a finite set of hypotheses (absence or presence of an organism), UMD is unique in that it can estimate a linear function of hybridization affinity data from the random probe measurements [39]. This gives UMD a potentially important future proof property: a fixed set of measurement
probes can be used to detect and estimate the concentration of even novel organisms not present in the library. To detect a new organism, the software can be adjusted to take into account how the new organism will react to the existing probe set; however, no new capture probes are required.

Several other pathogen detection schemes are currently under investigation [4, 41, 42, 43]. To the best of our knowledge, the UMD platform is the only approach that enables a unified representation of all known and unknown bacterial organisms in a low dimensional geometric space. The sloppy beacon approach in [4] uses only six probes to non-specifically detect a group of bacteria, however, by using the melting temperature as the signature profile of each bacteria this methods is limited in detecting more than a single bacteria at a time or even having an estimation of the concentration of the bacteria sample at hand. In [41] authors leverage the CS theory to decode the hybridization affinity of a set of known bacteria to DNA microarray probes, however, because of fixing the design matrix $\Phi$ they can provide no further claims on detecting a novel organism. Also the probe design procedure in this methods is burdensome. Mohtashemi, et. al. [42] investigate the feasibility of a target blind sensing approach using randomly chosen DNA probes from pathogenic sequence implemented on a DNA microarray but the requirement for more than 10,000 probes limits further applicability of this approach.

The theory of CS provides rigorous recovery guarantees and suggests algorithms to leverage this geometry to both detect bacteria and estimate their concentrations. Our
successful implementation of UMD confirms that random DNA probes empirically satisfy the incoherency requirements imposed by CS theory and can be used for universal microbial sensing. The UMD platform has the potential to rapidly direct the physician to use appropriate antibiotics or treatment and thus minimize the risk of antibiotic resistance. It can be utilized in bio-defense applications to detect multiple unknown and mutant agents. We envision that, with further optimization of probe design and detection schemes, UMD can sense an even wider range of organisms (e.g. viruses, fungi) and varied biomolecules of interest (genes, proteins).

Here we discuss some of the limitation of our current implementation of UMD to be utilized in real clinical setting and we will expend on different methods to tackle the problem as our future work.

4.1 Future work

The first and probably the biggest challenge in the current UMD detection platform is the presence of background contaminant DNA. For example in the case of clinical diagnostics we expect the infectious sample from human body to have a certain concentration of human DNA even after purification steps. The traditional molecular based bacterial detection suffer from this to lower extents because the PCR process provides the required SNR after multiple steps of amplification of the certain maker genes (e.g., 16S, etc.). However, the 3 billion base pair long human genome is apt to cause some confusion in sensing bacterial patterns using UMD. Here we propose
some practical ways to tackle the problem.

If the concentration of contaminant DNA in not overwhelmingly high to make all the probes hybridized to itself, we envision to solve this problem either algorithmically and/or by adding extra probes specific to human DNA to estimate its presence. Once the concentration of human DNA is estimated we can treat the whole human DNA molecule as a new organism in our target bacterial library and therefore add a column to the hybridization affinity matrix $\Phi$ to absorb the structured noise profile of the contaminant.

The aforementioned contamination removal method can be improved (even in the presence of high concentration of DNA contaminants) with an intelligent way of selecting the DNA probes. Instead of propagating DNA probes completely randomly we can randomly choose them from a set of probes whose hybridization affinity patterns to contaminant DNA is sufficiently orthogonal to the hybridization affinity of the probes to the target microbial organisms. More rigorously, we promote using probes whose hybridization affinity to the contaminant molecule is in the null space of the designed matrix $\Phi$. This is biologically a feasible design procedure since the various contaminant organisms (e.g., human, fruit, and vegetable DNA) genome posses regions with sufficient sequence variations (from bacteria) to be exploited.

It is worth to note here a computational challenge and a potential remedy in modeling the hybridization affinity of DNA probes to large DNA sequences (such as human DNA). As we discussed earlier in this chapter the accurate thermodynamical
alignment of probes to bacterial target is a time consuming task (~12 hour for a single probe vs. a bacterial genome). We have also proposed a fast method of approximating the hybridization affinity that ignores bacteria-bacteria interactions. In the case of dealing with human DNA the accuracy of this approximation procedure has to be further investigated. One potential remedy that can alleviates this completely is to obtain the hybridization affinity in the experiment, i.e., do a separate experiment with only the contaminate DNA.

Apart from removing contaminations, there are further strategies to improve the CS recovery performance in UMD. One strategy is to leverage the prior informations of the present bacteria in a microbial sample and the prior information of the taxonomy of current sequenced organisms (tree of life structure). We know there exist some bacteria that are only present together in certain environments. This prior information can be utilized (e.g., using a group sparsity penalty) to improve the recovery performance. The taxonomy of the organisms in the library can be used to define a metric of their genomic distance and this helps to cluster them in groups that can be leveraged in the recovery procedure; e.g., apply the ideas from model-based CS to further improve the recovery bounds.
Appendix A: Human pathogenic bacterial species dataset

To evaluate the UMD platform for genus level bacterial detection we selected 38 species from 38 different genera listed as most commonly pathogenic to humans by the Center for Disease Control and Prevention (CDC) [37]. The following genome sequences were obtained from the NCBI website:

*Acinetobacter baumannii* ATCC 17978,

*Aeromonas salmonicida* subsp. salmonicida A449,

*Bacillus cereus* ATCC 14579,

*Bartonella henselae* str. Houston-1,

*Bordetella pertussis* Tohama,

*Borrelia burgdorferi* B31,

*Brucella abortus* S19,

*Campylobacter jejuni* subsp. jejuni 81116,

*Chlamydia trachomatis* B/Jali20/OT,

*Clostridium botulinum* B1 str. Okra,

*Corynebacterium jeikeium* K411,

*Coxiella burnetii* RSA 331,

*Enterococcus faecalis* V583,

*Escherichia coli* str. K-12 substr. MG1655,

*Francisella tularensis* subsp. holarctica LVS,

*Fusobacterium nucleatum* subsp. nucleatum ATCC 25586,
Haemophilus influenzae F3047,
Helicobacter pylori B38,
Klebsiella pneumoniae 342,
Legionella pneumophila str. Corby,
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130,
Listeria monocytogenes 08-5578,
Mycobacterium leprae TN,
Mycoplasma pneumoniae M129,
Neisseria meningitidis MC58,
Prevotella melaninogenica ATCC 25845,
Propionibacterium acnes KPA171202,
Proteus mirabilis HI4320,
Pseudomonas aeruginosa LEB58,
Rickettsia rickettsii str. Iowa,
Salmonella enterica subsp. enterica serova,
Paratyphi A str. ATCC 9150,
Serratia proteamaculans 568,
Shigella sonnei Ss046,
Staphylococcus aureus subsp. aureus USA300-FPR3757,
Streptococcus pneumoniae ATCC 700669,
Treponema pallidum subsp. pallidum SS14,
**Vibrio cholerae** MJ-1236, *Yersinia pestis* CO9.

For genus level bacterial detection 9 different species from *Staphylococcus* genus were selected:

*Staphylococcus aureus* subsp. aureus USA300-FPR3757,

*Staphylococcus carnosus* subsp. carnosus TM300,

*Staphylococcus epidermidis* ATCC 12228,

*Staphylococcus haemolyticus* JCSC1435,

*Staphylococcus hominis* SK119 contig00059,

*Staphylococcus lugdunensis* HKU09-01,

*Staphylococcus pseudintermedius* HKU10-03,

*Staphylococcus saprophyticus* subsp. saprophyticus ATCC 15305,

*Staphylococcus warneri* L37603.
Bibliography


