4-25-2018

Genomic Methods for Bacterial Infection Identification

Duy Tran Pham

GENOMIC METHODS FOR BACTERIAL INFECTION IDENTIFICATION

by

Duy Tran Pham

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Bioinformatics

The University of Memphis
May 2018

Acknowledgements

I would like to thank my family for supporting me and giving me strength to complete my graduate studies. Without them, I would not have the motivation to complete this thesis. I would also like to thank Dr. Garzon for his time and patience. He is always around to assist with any trouble and has helped guide me towards the right direction for completing this thesis. Furthermore, he has given me life-long experiences and encouragement that I will keep as I continue to pursue my future goals. I would also like to thank Dr. Homayouni and Dr. Venugopal for serving as a committee member and providing me with useful feedback for improving this thesis. Thanks to my labmate, Sambriddhi Mainali, for her comments and suggestions as well. Finally, I want to thank the University of Memphis Bioinformatics Program for awarding me a Graduate Research Assistantship which has made all of my graduate studies possible.

Abstract

Hospital-acquired infections (HAIs) have high mortality rates around the world and are a challenge to medical science due to rapid mutation rates in their pathogens. A new methodology is proposed to identify bacterial species causing HAIs based on sets of *universal* biomarkers for next-generation microarray designs (i.e., nxh chips), rather than *a priori* selections of biomarkers. This method allows arbitrary organisms to be classified based on readouts of their DNA sequences, including whole genomes. The underlying models are based on the biochemistry of DNA, unlike traditional edit-distance based alignments. Furthermore, the methodology is fairly robust to genetic mutations, which are unlikely to reduce accuracy. Standard machine learning methods (neural networks, self-organizing maps, and random forests) produce results to identify HAIs on nxh chips that are very competitive, if not superior, to current standards in the field. The potential feasibility of translating these techniques to a clinical test is also discussed.

TABLE OF CONTENTS

List of Tables

Table Page

List of Figures

Figure Page

List of Figures

Chapter 1: Introduction

Identifying the species of an unknown bacterium is important in the clinical field, especially for pathogens in hospitals since they have high morbidity and mortality rates around the world. This problem dates back to 1847, when Ignaz Semmelweis showed that childbed fever was transmissible through unsanitary hands of health-care workers. Since then, surgeons and health care affiliated workers have adopted new aseptic and antiseptic techniques to reduce patient's exposure to HAI pathogens. Still, preventative programs established to reduce patient's susceptibility remain difficult and challenging, as bacteria become antibiotic resistant and methods that provide timely and accurate identification of the infectious agents are lacking for appropriate patient diagnosis for treatments. Therefore the ability to accurately identify the infectious agents in an appropriate time and manner is a critical step for epidemiological surveillance and public health decisions. This thesis develops such a method to classify bacterial pathogens, with emphasis on hospital-acquired infections (HAIs.)

1.1. Statement of the Research Problem

An example of the classification problem is shown in Figure 1 in its most general form. In biology, this problem requires grouping together biological organisms based on similar characteristics in order to determine whether an unknown organism is a member of an established group.   Perhaps the best examples are the attempts by biologists to establish a standard taxonomy and phylogeny of all life on earth. There are groups within different levels of the biological taxonomy (i.e., domain, kingdom, phylum, class, order, family, genus or species) where an organism can be classified. Difficulty arises, however, as the classification of an organism to a taxon shifts from coarser levels (like domains) to finer levels (like genera and species.) Thus, research over time has tried to improved methods and technologies that can provide enough resolution to classify an organism at the narrowest level, e.g. species. This particular problem is known as the species identification problem, see Figure 2. I will propose a new method that will provide accurate identification of species of HAI pathogens, along with robustness to mutations.

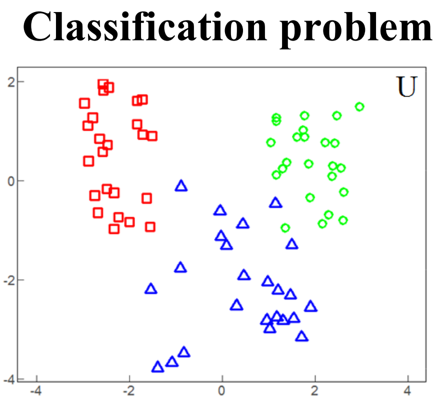

Classification problem

Figure 1: A *classification problem* for a collection of objects U calls for placing an object in one of a pre-determined mutually exclusive finite set of categories (e.g., three shapes/colors above) that partition the full collection *U*. In the simplest case of binary classification, the partition is given by a certain set of objects *L* and its complement *U* \ *L* in *U*. This is a well-known and difficult problem in computer science, usually unsolvable or NP-complete in full generality.

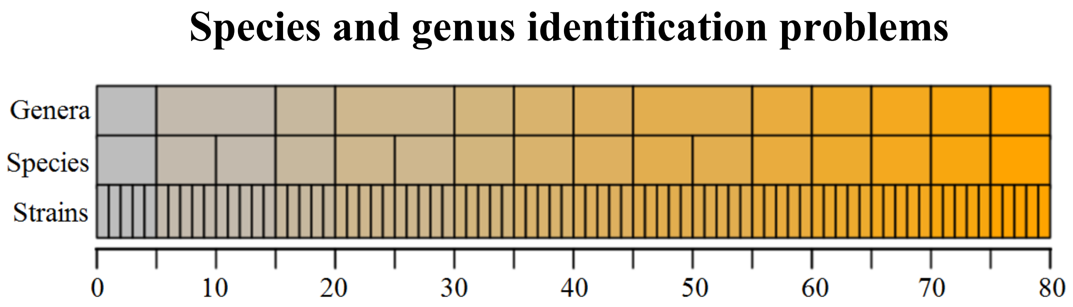

Species and genus identification problems

Figure 2: The *species (genus) identification* problem is a classification problem where each category is a specific species (16, middle row) or genus (13, top row), respectively. An object is an organism (80 of them in the sample, bottom row), as represented by some biomarker such as a gene (like 16S sRNA or COI) or its whole genome sequence (WGS), for example.

1.2. Significance of the Research Problem

Research and development for the species identification problem is important, especially for the identification of microbial communities, since scientists in different industries (e.g., agriculture, clinical microbiology, and food production) use these tools to improve the quality of life.

In particular, a hospital-acquired infection (HAI) is defined as an infection that develops within 48 hours after the patient is admitted to a hospital. Of all acquired infections, they account for 7% in developed and 10% in developing countries (Khan et al., 2017.) In the U.S., an estimated 1 of 25 patients admitted to hospitals gets infected with an HAI (Magill et al., 2014.) A summary of the top pathogens known to cause HAI is shown in Table 1 below.

Patients with an HAI can expect to pay thousands of dollars in healthcare cost depending on the HAI site as shown in Table 2 (Zimlichman et al., 2013.) These costs are mostly due to increased length of hospital stay and lack of efficient tools for diagnostic and treatments. In Table 1, *Clostridium difficile* is the top ranked HAI pathogen and infections from *C.difficile* alone cost roughly $1.51 billion in healthcare cost in the U.S annually (Table 2.) Despite a number of preventative programs established to reduce the rate of infection and healthcare cost, the need for a fast and accurate method to identify HAI is still needed due to the rapid rate of mutation of bacterial genomes. The development of such a method will assist with patient diagnosis and treatment, as well as controlling future outbreaks.

Table 1: Top 10 hospital acquired bacterial infections (HAIs) in the USA for 2011 (Magill et al., 2014)

Pathogen	**Rank**
Clostridium difficile	1
Staphylococcus aureus	2
Klebsiella pneumonia	3
Escherichia coli	4
Enterococcus species	5
Pseudomonas aeruginosa	6
Candida species	7
Streptococcus species	8
Coagulase-negative *staphylococcus* species	9
Enterobactor species	10

Table 2: **Financial burden of the top 5 types of HAI** (Zimlichman et al., 2013)

Types of HAIs	**Average Cost (in USDs)**	**Annual Cost (Billion)**
Central line-associated bloodstream infections (CLABIs)	45,814	1.85 (18.9%)
Ventilator-associated pneumonia (VAPs)	40,144	3.09 (31.6%)
Surgical site infections (SSIs)	20,785	3.30 (33.7%)
Clostridium difficile infections	11,285	1.51 (15.4%)
Catheter-associated urinary tract infections (CAUIs)	896	0.03 (00.4%)

Chapter 2: Concepts and Methods

2.1. Molecular Biology Background

Before the development of methods that rely on genomic information for characterizing unknown organisms, the best procedures for this problem were based on phenotypic characteristics. Micro-organisms could be classified based on their structure, cellular metabolism, growing conditions, or differences in their cellular components. In 1884, Hans Christian Gram developed a well-known procedure that characterized bacteria based on the detection of peptidoglycan in their cell walls. He distinguished bacteria into two groups, Gram-stain positive (stained violet) and Gram-stain negative (no stain.) This method has been used in clinical microbiology laboratories and is sometimes the preliminary procedure for bacterial identification (Srinivasan et al., 2012.) However, like other phenotypic methods, this procedure does not afford enough resolution to identify bacteria at the species level, since many closely-related species share similar phenotypes. It was not until 1953, when Watson and Crick solved the three-dimensional structure of DNA, that phenotypic methods were replaced by molecular techniques.

Deoxyribonucleic acid (DNA) is a macromolecule that is abundantly present in nearly all living organisms on planet earth. It is considered to be the blue print of life since it contains the genetic instructions necessary for an organism to be born and live. A sequence of nucleotides represented by A, C, G, and T (described in Figure 3) form the genetic code (genome) that constitutes the unique characteristics of each individual. A well-known functionality of DNA sequence is to help produce proteins. Part of the DNA that codes for genes are transformed into a single stranded molecule known as RNA (ribonucleic acid.) This molecule contains the same nucleotides as DNA except thymine (T), which is converted to uracil (U) during the transcription process. The resulting messenger RNA (mRNA) strand is then transported from the nucleus and into the cytoplasm for protein synthesis. This step is known as the translation process, where molecular machines known as *ribosomes* and adaptor molecules (tRNA), link together chains of amino acids to produce proteins. Proteins regulate the functions within cells and are basic structures required of most cellular components.

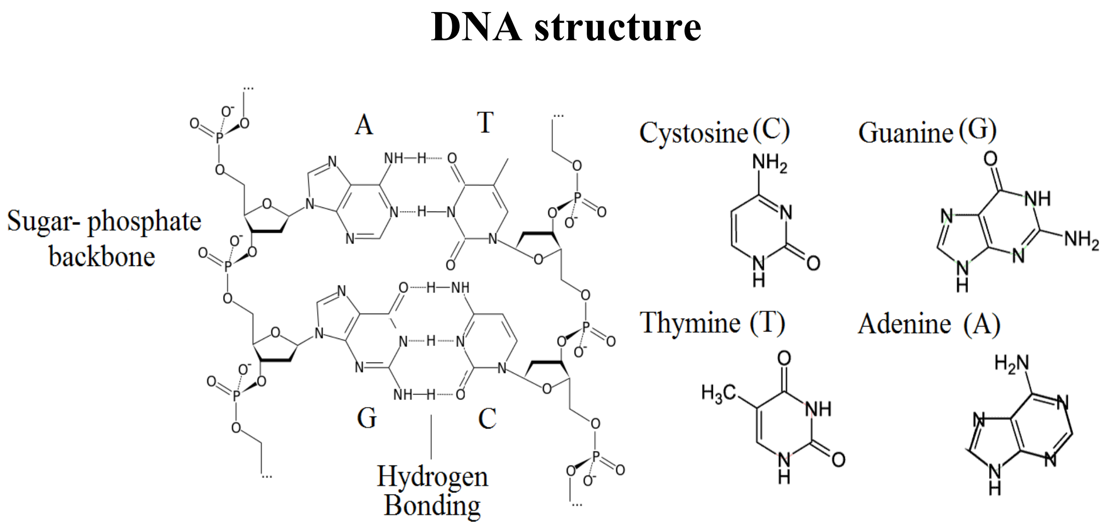


Figure 3: A DNA contains the genetic instruction for an organism to develop, live, and reproduce. Nearly every living organism carries a unique DNA genome that essentially defines what they are. A single-stranded DNA (ssDNA) contains nucleotides A (adenine), C (cytosine), G (guanine), and T (thymine) and a sugar-phosphate backbone. Two single strands form a double-helix through hydrogen bonding between A-T and C-G.

Over time and during reproduction, DNA can be just copied or can change due to random mutations such as insertion, deletion, or substitution. A *conserved region* in a genome is a subsequence of the genome that codes for an important functionality and undergoes very low evolutionary changes by natural selection. This type of sequence occurs across species and is useful for differentiating species from one another. The ribosomal RNA (rRNA) for example, is essential for protein synthesis and is present in nearly all living organisms. For bacteria, the 16S rRNA gene (~1500 bp) is widely used to identify species of unknown organism since the similarity is at least 97% between strains within the same species. Still, single conserved regions, like the 16S rRNA, cannot provide enough resolution for most bacteria at the species level (Janda & Abbot, 2007.) For example, *B.globisporus* and *B. psychrophilus* have 99% similar 16S rRNA sequences (Nguyen et al., 2016.) Additionally, the strain, *Escherichia coli* K12, has two copies of the 16S rRNA gene that differs by about 5% (Nguyen et al., 2016.) Alternative methods such as multi-locus sequence analysis (MLSA) take 6-7 housekeeping genes (~ 500 bp) and align each of the sequenced genes to a database to determine the species of an unknown bacterium. This idea may seem to improve the resolution; however, only a limited number of species can be compared within a universal database (e.g., https://pubmlst.org/databases/.) Limitations arise from the fact that conserved regions of the genome need to be known *a priori* to discriminate species from one another. In this method, no universal sets of housekeeping genes can be used for all bacteria (Glaeser & Kämpfer, 2013.) This makes the species identification task laborious as different protocols may be required for different species. Therefore, effective classification methods should be universal and must prove capable of working with large scale genomic, or even proteomic, data to ensure accurate species identification for all bacteria since the classification may, in principle, depend on every nucleotide in the sequence.

2.2. Current methods for HAIs

Currently, there are various techniques used to identify bacterial pathogens (Fournier et al., 2014; Emerson et al., 2008.) Table 3 provides some of the commonly used methods in clinical laboratory settings. In this section, popular methods based on large-scale genomic or proteomic data will be summarized including matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, pulse-field gel electrophoresis, whole-genome sequencing, and DNA microarray.

Table 3: Clinical methods used for bacterial identification (Fournier et al., 2014; Emerson et al., 2008.)

Method	Technology	Type
Fingerprinting	Pulse-field gel electrophoresis (PFGE)	Genomic
	Riboprinting	Genomic
	Restriction Fragment Length Polymorphism (RFLP)	Genomic
DNA Sequencing	Small-subunit ribosomal gene	Genomic
	Multi-locus sequence (MLS)	Genomic
	Whole-genome sequencing (WGS)	Genomic
Hybridization	DNA Microarray	Genomic
Mass Spectrometry	Matrix Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF-MS)	Proteomic
	Electrospray Ionization Mass Spectrometry (ESI-MS)	Proteomic
	Surface-Enhanced Laser Desorption/Ionization (SELDI)	Proteomic

2.2.1. Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry

The use of mass spectrometry (MS) for bacterial identification was first proposed by Anhalt and Fenselau in 1975. At the time, MS could only analyze small molecules and was limited due to differences in growth conditions and media. Soft ionization MS such as electron spray ionization (ESI) and matrix assisted laser desorption ionization time-of-flight (MALDI TOF) were later developed in the 1980s to extend the application to larger biomolecules like ribosomal proteins (Singhal et al., 2015.) Both techniques measure protein mass by converting the proteins to ions with the addition or removal of one or more protons. This method is known as *peptide mass fingerprinting*, and has been useful for characterizing bacteria at the species and genus level. In recent years, however, MALDI TOF MS has come to dominate ESI MS for bacterial identification (Singhal et al., 2015.) The advantage of MALDI TOF MS over ESI MS is that the data is easier to interpret since MALDI TOF MS produce single charged ions. Additionally, the chromatography step in ESI MS is not required with MALDI TOF MS, and so the simplicity and speed to result is highly desirable. Therefore, for evaluation of the novel methods proposed in this thesis, MALDI TOF MS will be discussed since it is the preferred MS method used in clinical laboratories for microbial identification.

MALDI TOF MS was developed by Michael Karas, Franz Hillenkamp, and their colleagues in 1985, although emergence of its application for clinical use did not begin until 1996. There are several reasons for the decade long delay prior to common use (Nomura, 2015.) Microbiologists speculated whether spectra patterns could consistently reproduce established hierarchical order of microbial taxonomy, since protein patterns are expected to change depending on growing conditions. Additionally, MALDI TOF MS is almost fully automated, so the simplicity and speed seemed unreliable in comparison to other methods that required an expert to perform the task. Moreover, databases for clinically relevant pathogens were still under development since MALDI TOF MS requires a comprehensive database for accurate identification. It was not until 1996, when MALDI TOF MS was shown to capably produce spectral fingerprints from whole bacterial cells without pretreatment, that microbiologists began to publish reports of its use in identification of bacteria, fungi, yeast, and mycobacteria (Nomura, 2015.)

There are two types of MALDI TOF MS approved by the US Food and Drug Administration (FDA) for routine bacterial identification; MALDI Biotype CA System by Bruker Daltonics Inc. (https:// bruker.com/) and VITEK MS by bioMérieux Inc. (http:// biomerieux-usa.com.) Both differ in the identification step described below (Clark et al., 2013.) Figure 4 (below) shows a typical workflow of MALDI TOF MS for bacterial identification. First, a small portion of cultured bacteria colony are smeared onto a spot of a stainless steel plate and then mixed with a formic acid solution called a matrix. Typically, the matrix comprises of alpha-cyano-4-hydroxycinnamic acid (CHCA) that is mixed with organic solvents and water, although other type of matrices have been explored to achieve better performance (Nomura, 2015.) The plate is set aside to allow time for the sample and matrix solution to crystallize. Finally, the plate is deposited into a mass spectrometer for scanning and automatic measurements. Within the mass spectrometer, each spot on the stainless-steel plate is hit with a laser beam that converts the sample into a gas. The matrix material must be strong enough to absorb the laser's energy to create ions from molecules with minimal fragmentation. In the gas phase, charge is transferred from the matrix solution to the microbial molecules through random collisions. The charged ions are then accelerated through a TOF analyzer and measurements are determined by the length of time it takes for the ions to travel the length of the tube. In the end, a spectrum of the organism's protein mass is generated. This spectrum is compared to other spectra in a database by pattern recognition algorithms for peak detection to finally determine a species or genus match. The Biotype produces a score between 0 and 3.000. A score ≥ 2.000 is a species match and a score between 1.700 and 1.999, inclusively, is a genus match. A score < 1.700 indicates that there is no reliable match in the database. The VITEK system has two approaches for identification. The first approach uses a pattern matching algorithm, while the second approach groups together peaks shared by a minimum number of strains to create a reference 'super-spectra' weighted according to the species or genera. A confidence value between 0 to 100% is also outputted for comparison between the unknown spectrum and super-spectra.

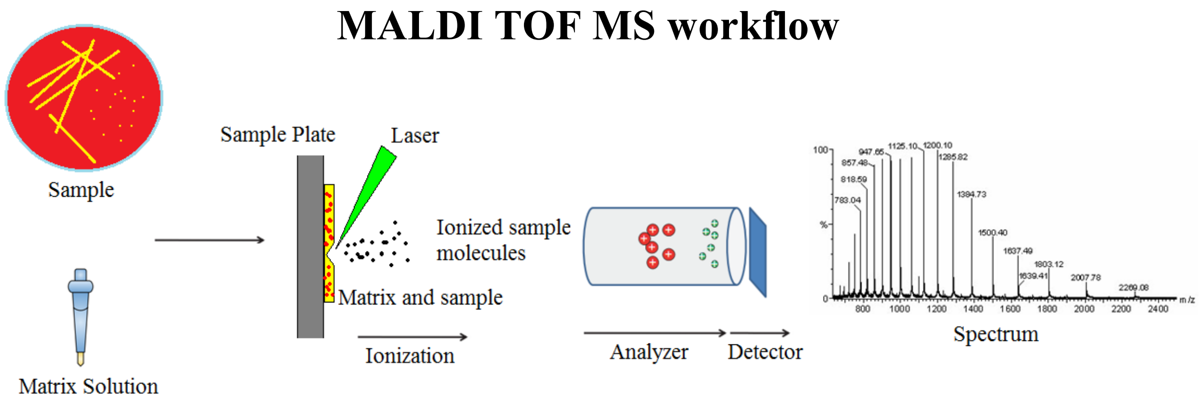


Figure 4: For species/genus identification using MALDI TOF MS, a small sample of a cultured bacterial colony is smeared onto a stainless steel plate. A matrix solution, typically alpha-cyano-4-hydroxycinnamic acid (CHCA) mixed with organic solvents and water, is applied to the spot containing the sample. After being air-dried and placed into the mass-spectrometer, a beam of laser is fired at the sample and the matrix mixture, causing the molecules in the sample to ionize. The ions then travel up the analyzer by size and are then detected by a sensor to obtain a spectrum of the mass-to-charge ratio. The spectrum is then compared to other spectra in a database by a pattern recognition algorithm to classify the species/genus of the sample.

The initial cost of these instruments currently exceeds $150,000, with an additional annual maintenance fee that can be troublesome for many clinical laboratories. However, those that can overcome the cost can expect to save more money compared to other conventional identification methods. A study by (Tran et al., 2015) explored the annual cost between traditional methods and MALDI TOF MS (Vitek System) between April 1, 2013, and March 31, 2014. The total annual cost for the traditional methods was $142,533 versus $68,887 with MALDI TOF MS. This resulted in a laboratory savings of about 52%. Moreover, the average hands-on time per specimen is about 5 minutes (Dhiman et al., 2011.) A meta-analysis study by (Zhou et al., 2017) showed that MALDI TOF MS can identify bacterial species with about 84% accuracy with a confidence interval [81.2%- 88.9%] and bacterial genus with about 91% accuracy with a confidence interval [88.3%-93.3%], both at the 95% confidence level. The accuracy will vary depending on the species and comprehensiveness of the database. The sensitivity of MALDI TOF MS will also depend on the concentration of bacterial cells. It is suggested that a minimum of approximately 10^4 cells per sample is required for reliable identification. Nevertheless, the cost, speed, accuracy, and automated features of MALDI TOF MS have made it the gold standard technology for bacterial identification, although microbiologists would prefer a tool that is smaller and cheaper in capital costs (van Belkum et al., 2017.)

2.2.2. Pulse-Field Gel Electrophoresis

DNA fingerprinting is a method used to uniquely characterize an organism based on patterns of DNA fragments. Before 1984, conventional electrophoresis methods were used to separate DNA fragments according to size; however, only a single electric field was applied. This effectively separated fragments of size up to ~20 kb, but larger fragments would cluster at the top of the gel and appeared as a large band when photographed for analysis. In 1984, Schwartzand and Cantorand remedied the issue by inventing pulse-field gel electrophoresis (PFGE.) This method is different from conventional electrophoresis in that an alternating electric field is applied to modify the direction and speed in which the fragments migrate. Up to ~10 Mbps (Mega base pairs) can be separated using this technique.

Various types of PFGE (e.g., CHEF, FIG, AFIGE, OFAGE, PHOGE and PACE) have been developed for separating and typing DNA molecules with large fragments present (Parizad et al. 2016.) Each method's output goal is the same, but the separation process and maximum fragment size that can be separated differ. Among the various types of PFGE, the Contour Clamped Homogeneous Electric Field (CHEF) is widely used (Parizad et al. 2016.) Figure 5 shows the typical workflow for PFGE based on CHEF's separation technique.

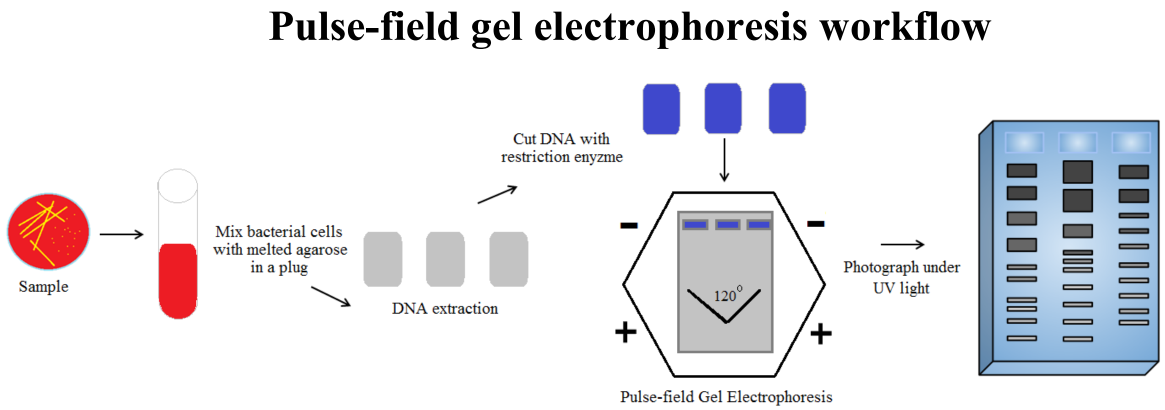

Pulse-field gel electrophoresis workflow
Sample
Mix bacterial cells with melted agarose in a plug
DNA extraction
Cut DNA with restriction enyzme
–
–
+
+
120°
Pulse-field Gel Electrophoresis
Photograph under UV light

Figure 5: Pulsed-field gel electrophoresis (PFGE) is a highly discriminative molecular typing technique that is based on DNA banding patterns. Cells from bacteria are embedded in a melted agarose gel plug and lysed for DNA extraction. Next, the DNA is cleaved into different size fragments using restriction enzymes. The DNA fragments are then separated according to size during electrophoresis using an electric field of alternating polarity. Finally, the DNA banding patterns can be seen under UV light and photographed to be stored in a database for later identification from unknown samples.

Bacterial cells are first cultivated within 24 hours. Next, a cell suspension is prepared using an appropriate buffer. The bacterial cells are then mixed with agarose gel in a plug and a biochemical is added to release the DNA from the cell membrane. The plug, which now contains DNA, is washed several times to remove proteases and cell debris. Next, a restriction enzyme is added to the plug. The choice of restriction enzyme for DNA shearing is important in the PFGE process since they are able to locate a specific nucleotide sequence and cleave the DNA from that exact place. The length of the fragments is important since it will be later used to distinguish between bacterial strains. Once the DNA in the plug has been cleaved to several fragments, the plug is placed into a cavity at the top of the electric field. The CHEF's system contains 24 electrodes aligned in a hexagonal arrangement. The top portion of the hexagon contains negative electrodes while the bottom portion contains positive electrodes. Since DNA is negatively charged, the DNA fragments will migrate towards the positive field. The flow of voltage from the positive to negative electrodes are altered at an angle of 120 degrees, which allows more time for the larger DNA fragment to migrate. The timing between electrodes switch is precisely controlled to ensure that all fragments are represented. Finally, the gel is stained so that the banding patterns can be seen under ultraviolet (UV) light. A photograph is taken and the image is stored into a database for comparison.

PFGE is still widely used today in epidemiology and microbiology studies. It is commonly used for bacterial typing since it provides high discrimination power at the strain level. Although widely used, PFGE is not ideal for routine use in clinical settings. The drawback of this method is that it is time consuming and costly since it can take up to four days or more to obtain results (Aguilera-Arreola et al., 2015.) Moreover, this method lacks discrimination power of bands that are nearly identical in size, which ultimately makes it difficult to interpret the banding patterns.

2.2.3. Whole-Genome Sequencing (WGS)

In 1995, the first bacterium to have its entire genome sequenced was *Haemophilus influenzae*. Since then, over 564 million whole genomes have been sequenced and deposited into Genbank as of February 2018 (https://www.ncbi.nlm.nih.gov/genbank/statistics/.) This was made possible due to rapid advances in sequencing technologies and in particular, whole-genome sequencing (WGS), over the last decade. Previous sequencing methods known as first-generation sequencing were developed based on the Sanger method; however, these technologies have been superseded by next-generation sequencing (NGS) techniques (Deurenberg et al., 2017.)

Second-generation (NGS2) sequencing technologies are currently the most commonly used methods for sequencing genomes. The procedure uses a “shot-gun” based approach, which requires library preparation, amplification, and an assembling step described in Figure 6. Third-generation sequencing (NGS3) technologies aim to sequence the nucleotides directly at the molecular level in order to increase the sequencing reads while reducing biases and achieving higher throughput (Deurenberg et al., 2017.) Fourth generation technologies (NGS4) sequence the nucleic acid *in situ*, directly in fixed cells and tissues. Both third-generation and fourth-generation sequencing are still under development and are not yet implemented for broad use due to current the lack of robustness of the methods.

The role of whole genome sequencing (WGS) technologies is becoming increasingly important in public health and hospital infection control-affiliated laboratories. This approach provides comprehensive genomic information for understanding infectious diseases and better resolution in characterizing strains. Due to the reduction of cost and speed of current technologies, WGS in the future may replace other sequence-based methods that rely on conserved regions (16s rRNA and Multi-Locus Sequencing (MLS)) for bacterial identification. However, future work on improving NGS workflow, such as automatic pipelines for data analysis, external quality controls for proficiency testing, and shorter runs of NGS platforms, are needed in order for this method to become widely accepted for patient guidance and infection control management (Deurenberg et al., 2017.)

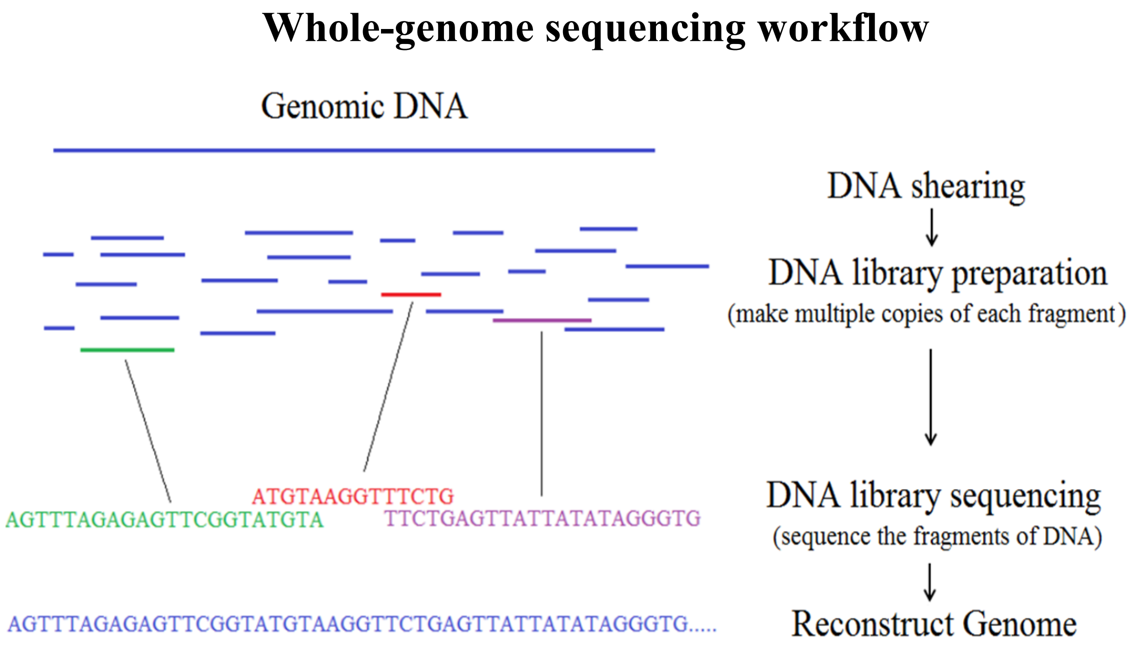

Whole-genome sequencing workflow
Genomic DNA
DNA shearing
DNA library preparation
(make multiple copies of each fragment)
ATGTAAGGTTTCTG
AGTTTAGAGAGTTCGGTATGTA
TTCTGAGTTATTATATAGGGTG
DNA library sequencing
(sequence the fragments of DNA)
AGTTTAGAGAGTTCGGTATGTAAGGTTCTGAGTTATTATATAGGGTG.....
Reconstruct Genome

Figure 6: For whole-genome sequencing (WGS), DNA is extracted from bacterial cells and sliced into fragments using enzymes or mechanical disruption. Many copies of each DNA fragment are produced using polymerase chain reaction (PCR) to create a DNA library, which are loaded into a sequencer to obtain DNA reads. Assembling software is used to put together the millions of DNA reads in the correct order as one re-constructed sequence for further analysis.

Generally, alignment (e.g. BLAST and Average Nucleotide Identity (ANI)) or alignment-free based methods (such as *k*-mers) are used to classify bacteria using their entire genome sequence. For alignment based methods, an algorithm will find local regions of a reference genome where the queried sequence has high identical base pair matches. Usually, a percentage that indicates how identical the two sequence is given in the end. The drawback to this method is that there is no universal cut-off percentage threshold to determine the identity of an unidentified genome (Zielezinski et al., 2017.) Cut-off values may vary between 95-97%. Additionally, alignment-based methods are time consuming and computationally extensive since the number of possible alignments increases as the length of the sequence gets larger, especially for multiple-sequence alignments (Zielezinski et. al., 2017.) Dynamic programming algorithms can resolve this issue; however, the time complexity remains in the order of the product of the length of the sequences (Zielezinski et. al., 2017.) Alignment-free based methods such as *k*-mers are popular tools due to their speed to result. These alignment-free based methods use fast algorithms to count the frequency of all possible *k*-mers (short DNA oligonucleotides of length *k*) and use some distance metric (e.g., Euclidean distance) to determine sequence similarities. Other methods such as CLARK will find unique *k*-mers that are specific to a species (Ounit et al., 2015.) Still, a major disadvantage to alignment-free methods is that memory consumption can be relatively large if a large *k* is selected (there are 4^k possible k-mers) (Zielezinski et. al., 2017.) Moreover, other *k*-mer based algorithms may be more memory efficient and computationally inexpensive; however, the trade-off is usually failure to identify bacteria at the narrowest taxonomic levels (e.g., species.)

2.3. Next Generation Microarrays (nxh chips)

DNA microarrays are powerful tools used to capture large-scale genomic information by means of hybridization between a nucleotide sequence to its (nearly) complementary strand. A typical microarray contains a collection of DNA spots that is fixed to a solid glass surface arranged in rows and columns. These spots are referred to as probes in this thesis (sometimes contrary to the reverse convention used by biologists.) They consist of a number of copies of oligonucleotide (oligo) sequences as shown on the left in Figure 7. Fragments of DNA or RNA molecules, consequently called targets herein, hybridize to these probes as explained on the right in Figure 7.

In 1975, the first DNA microarray was created by Grunstein and Hogness to identify DNA of interest cloned within *E.coli* plasmids by quantifying the hybridization to radiolabeled probes. Since then, microarrays have gone through dramatic improvements, which include automated steps and fluorescent detections of hybridization. They are mostly used for gene expression analysis, which, in a clinical setting, is useful for identifying bacteria, antibiotic resistancy, and virulence factors based on a specific set of genes.

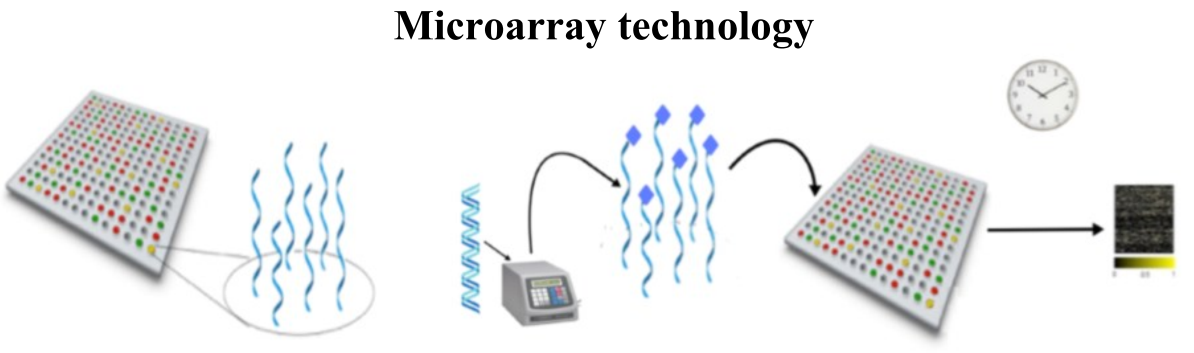


Figure 7: **Left**: a DNA microarray is a collection of spots on a solid surface where DNA oligonucleotides (oligos) are attached. Each oligo, referred to herein as a "probe", may hybridize with a homologous fluorescently-labeled free fragment from a target to generate a signal when exposed to ultraviolet light. **Right**: targets are poured on the microarray and time is allowed for hybridization before a readout (or signature) is obtained with a quantitative aggregated measurement of the signals or each probe. Permission for the use of this image was given by (Garzon & Mainali, 2017.)

Although DNA microarrays are useful for a wide variability of applications, there are several drawbacks and limitations for this tool. First, the reliability and reproducibility of the data varies between laboratories since hybridization failure of targets to probes occur (Garzon & Mainali, 2017.) In this situation, target strands that provide useful information are missed and thus, this results in low reliability and accuracy of the probe intensities value. This is in part, due to the fact that probes are arranged on the chip without consideration that the probes do no fully capture all possible targets. Moreover, no constraints are provided to ensure minimal crosshybridization between probes (Garzon & Mainali, 2017.) Without such restrictions, target strands will not have a chance to hybridize to probes that are designed to capture a specific strand. Secondly, for microbial analysis, microarrays are designed from a previously known reference strain. This is not useful since the genomes of bacteria tend to be highly variable and even fairly similar bacteria may require different microarrays. For example, the gene content of

A. actinomycetemcomitans, differs by as much as 20% between any two strains (Bumgarner, 2013.) Therefore, a lot of information is missed that could be useful for identifying two strains that belong to the same species.

In recent years, researchers have tried to address the limitations of microarrays. In particular, Garzon et al. (Garzon & Bobba, 2012) have proposed ways to address these problems in a next generation of microarrays, where the probes are designed in such a way that oligos will not hybridize with each other or to themselves. The Codeword Design problem described by (Garzon & Bobba, 2012) seeks to find sets of DNA oligonucleotides with such noncrosshybridizing (nxh) properties. The problem of finding these designs has been proven to be hard, even NP complete to solve in full generality (Garzon & Bobba, 2012; Phan et al, 2009.) The difficulty arises mostly due to lack of knowledge of the structure of the Gibbs Energy of strands of a fixed size, which governs hybridization between oligonucleotides. Such sets will afford a next generation microarray, called *noncrosshybridizing (nxh) chips*.

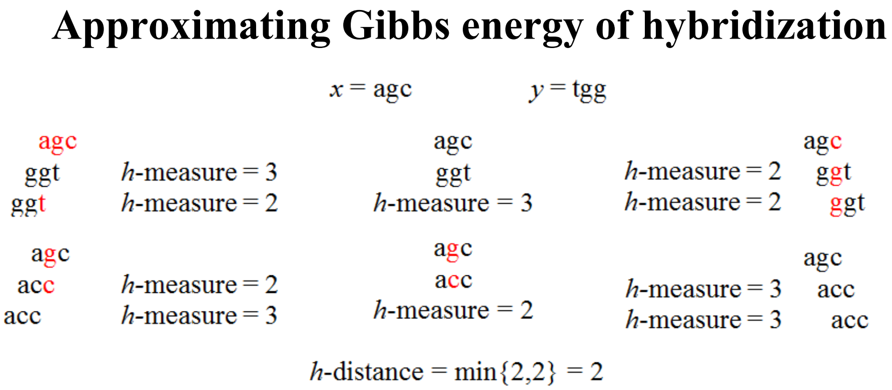
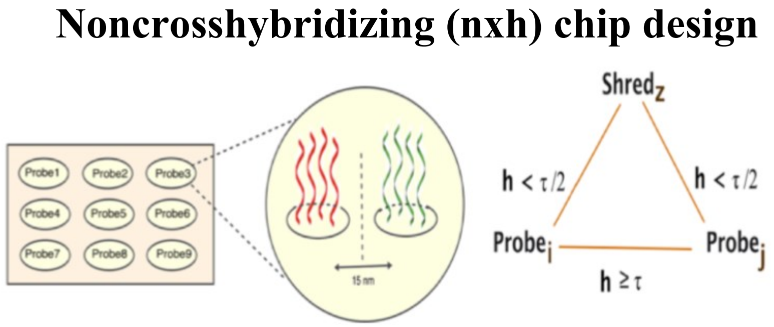


Figure 8: Computation of the *h*-distance *h(x, y)* between strands *x* and *y* of common length *n*. The strands x and the reverse y^R of *y* are aligned, and the minimum difference from *n* of the number of WC-complementary matches (in red) across all possible frameshifts is taken as the value for the *h*-measure *hm(x, y)*. This procedure is then repeated for the Watson-Crick complement (WC) *y'* of *y*. The *h*-distance between *x* and *y* is the minimum of the two *h*-measures *hm(x, y)* and *hm(x, y')* (Garzon & Bobba, 2012.)

The Codeword Design problem (CWD) can be depicted by arranging the oligos onto a geometric-pack sphere, where the oligos are coordinated in such a way that neighboring high hybridization affinity are mapped into a geometric frame similar to that of a Euclidean space (Garzon & Bobba, 2012.) A new model was proposed for this representation called the hybridization distance (*h*-distance.) This model measures the likelihood that two oligos, *x* and *y,* will hybridize with each other (Garzon & Bobba, 2012.) An example of the computation of the *h*-distance between two oligos *x* and *y*, where *x* = agc and *y* = tgg, is shown in Figure 8. The *h*-distance model is shown to be a feasible but reasonable approximation of the Gibbs Energy of duplex formation for DNA hybridization in that hybridization decisions based on the *h*-distance agrees with decision based on the Gibb's Energy Nearest Neighbor model over 80% of the time (Garzon & Bobba, 2012.) The *h*-distance does not distinguish between an oligo and its Watson-Crick complement, and so the term *p*-mer (for complementary *poligomer pairs*) will be used to refer to such pairs. Now, the *h*-distance can be treated like the ordinary Euclidean distance (because it satisfies the same key properties, including the triangle inequality) and used to quantify the amount of noise inherent in a microarray design (Garzon & Bobba, 2012; Garzon & Mainali, 2017.)

Probes for a next generation nxh chip must be carefully selected in order to prevent crosshybridiziation. To do so, a stringency parameter τ must first be selected as a threshold to decide hybridization using the *h*-distance between a probe and a target fragment. The probes are judiciously selected such that each probe is at the minimal distance τ from all others. An nxh chip design, also called a *basis*, will have a sufficient number of probes to ensure that each target will hybridize to at least 1 probe and, ideally to at most one probe. As shown in Figure 9 (right), hybridization to at most 1 probe is possible with a threshold τ/2, since a shred *z* with *h*-distance less than τ/2 for two of the probes *i* and *j*, cannot hybridize to both probes due to the triangle inequality and the minimal separation of τ in the *h*-distance between any pair of probes (a property that the Gibbs energy does not possess.) Additionally, each spot on an nxh basis contains the same finite number of an oligo and its Watson-Crick complement, separated by an appropriate physical distance to prevent crosshybridization, as illustrated in Figure 9 (left.) Once the probes that satisfy the above properties have been selected, genomic information can be captured from an organism using any DNA biomarker to obtain a digital signature. The process for obtaining a digital signature will be similar to standard microarray technology, though probes will be fluorescently labeled here instead of targets to avoid pseudo-signals due to hybridization among targets. An illustration of using an nxh chip is shown in Figure 10.



Noncrosshybridizing (nxh) chip design

Figure 9: **Left**: A noncrosshybridizing (nxh) chip design as described in (Garzon & Mainali, 2017.) All the probes on the chip are at *h*-distance of at least τ apart. Each spot consists of a fixed number of copies of oligos and the same number of copies of their Watson-Crick complements, spaced at a distance to prevent crosshybridization. **Right**: A target shred *z* is assumed to be able to hybridize with a probe if and only if its *h*-distance to the probe is less than *τ*/2. Copies of a random *z* cannot hybridize to two probes otherwise contradicting the triangle inequality, which the *h*-distance property follows. In this way, the amount of noise is reduced, thus addressing a major problem for standard microarray technology. Permission for the use of this image was given by (Garzon & Mainali, 2017.)

Computing digital signatures on an nxh chip

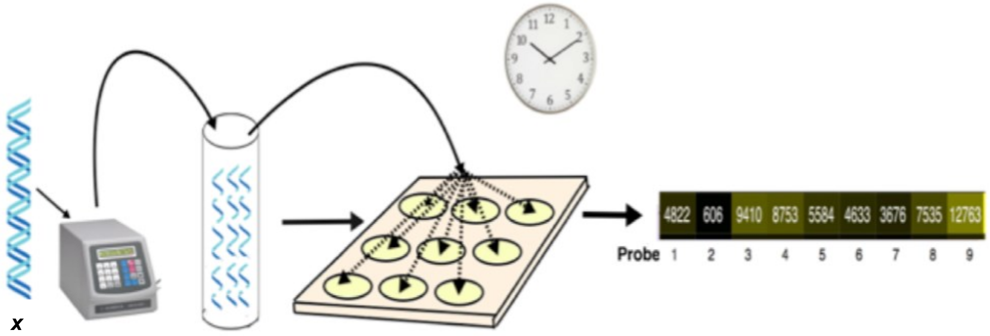


Figure 10: A sequence *x* from an organism is first shredded by sonication. Each shred is then fluorescently labeled before being poured to the nxh chip to acquire a digital by standard microarray procedures. Permission for the use of this image was given by (Garzon & Mainali, 2017.)

2.4. Computer Science Background

In this thesis, machine learning techniques were used on digital signatures obtained from nxh chips to identify the species of bacterial strains. A brief review of the machine learning methods used will be given in this section.

2.4.1. Machine Learning Methods

Machine Learning methods are becoming commonly used in clinical settings. Progress over recent years has enabled these tools to perform accurate evaluation of complex patterns observed in most clinical data. In particular, random forest, artificial neural networks, and deep learning are extensively used in the field of bioinformatics for classification tasks. Moreover, they have been combined with microarray expression datasets to accurate classify different types of cancer (Chu et al., 2014.) Therefore, the following sections will provide a brief overview of neural networks, self-organizing maps, and random forests which will be used to identify HAI species based on digital signatures obtained from nxh chips.

2.4.1.1. Neural Networks (NNs)

An artificial neural network consists of a finite number of neuronal units connected through directed synaptic links that resemble the synapses in the mammalian brain (Garzon & Pham, 2018.) They can solve classification problems in a way similar to how the human brain works based on given set of input features (Hassoun, 1995.) A neuron can be in one of several states of activation characterized by a real number at any given time, but can change its activation in the next time step by applying its characteristic transfer function to the net input of its neighboring neurons (obtained by a weighted sum of the states of other neuron with a synaptic link into it.) This process is iterated until all neurons have been updated for any single data point as described in Figure 11. A major advantage of using this method is that a prior deep analysis of the data is not required. Instead, the model can be trained by a learning algorithm (e.g., back propagation (Hassoun, 1995)) to classify data by passing a number of data points labeled with the expected correct answer from a training set for an appropriate number of times (or epochs), until the answers are mostly right (Figure 12.) The quality of the model is measured by how accurate the model predicts the actual labels in a testing set of data that the NN has not seen prior to the training phase.

Definition of neural networks (NNs)

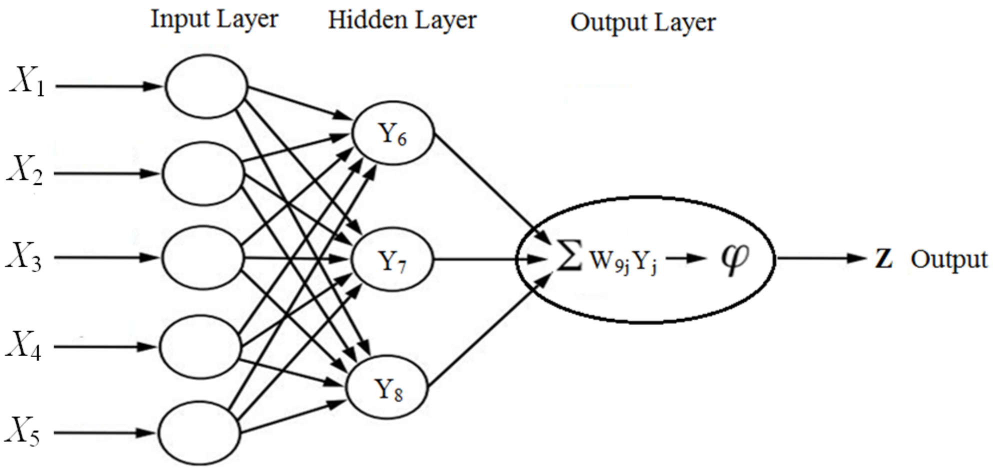


Figure 11: An artificial neural network consists of a finite number of neuronal units connected through directed synaptic links. They can solve classification problems in a way similar to how the human brain does it based on input features. The particular kind of NN above is a feed-forward neural net (FNN), where the neurons are arranged in layers, each receiving signals from neurons in the previous layer. The first is an input layer of neurons (in 1-1 correspondence) reading the features in a data point, the last layer producing a prediction as to which category the datum in the input belongs, and several neurons arranged in (so-called hidden) layers that try to distinguish characteristic features in the input feature vector. Each of the inputs is multiplied by an originally established weight, is continuously changed by an initial created threshold value and sent to an activation function (φ) to map its output.

How to train neural networks

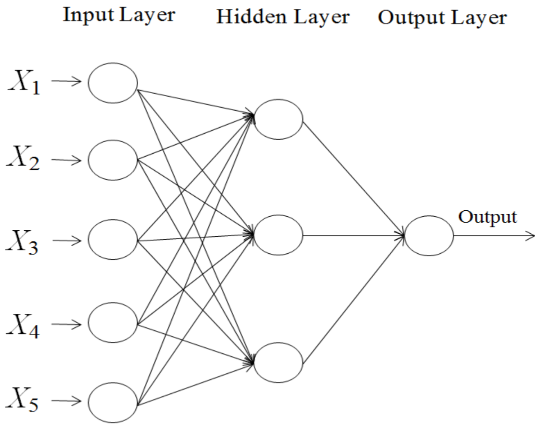


Figure 12: The neural network model can be trained by a learning algorithm (e.g. *backpropagation*) to classify data by passing a number of data points labeled with the expected correct answer (category where they belong) from an input space for an appropriate number of times (or *epochs*), until the answers produced are mostly right.

The following procedure to identify HAI species using NNs on digital signatures obtained from nxh chips is described in (Garzon & Pham, 2018.) Various feed forward neural networks (FNN) were trained for 4000 epochs by using the ‘h2o’ library package in R (Arora et al., 2006) with a hyperbolic tangent function as a smooth transfer function (nonlinearity.) The single output unit produced normalized decimal values between 0 and 1. For species identification, each *k* of the 16 species was assigned a range of outputs in the interval of radius 0.03125 centered at (*k*/16) – 0.03125, for *k* = 1… 16. The data corpus was partitioned into a learning set (80% of them, randomly assigned) and the remaining (20%) for the testing set. As is customary in machine learning, various combinations of nxh bases, neuron types, and hidden layers were tried in an attempt to optimize performance. As an example, 3mE4-[4-3-2-1] describes a FNN with 4 input features, two hidden layers with 3 and 2 neurons, providing input to a single 1 neuron in the output layer using the four feature signature vectors on the nxh basis 3mE4-2-at1.1, as described in Table 4. The *h2o.predict* function was then used to test the accuracy of the model. The accuracy was based on whether the predicted value obtained from the *h2o.prediction* function after training of the network fell within the correct interval coding for the corresponding species of a data point (strain.) This process was repeated for 32 different models, and the average of the 32 accuracies was taken to determine the performance of a given neural network architecture for the training and testing set. Additional performance measures including sensitivity, specificity, and precision, were computed for the testing set using the ‘caret’ package (Kuhn, 2008.)

2.4.1.2. Self-Organizing Maps (SOMs)

Self-organizing maps are a type of artificial neural network developed by Tuevo Kohonen in the 1980s as a type of solution to classification problems (Kohonen, 1982.) This method is different from NNs in that the training process uses a competitive unsupervised learning technique, which is useful for clustering together groups that contain similar feature patterns without additional labels in the data. A unique property of SOM is that it is able to map high-dimensional data onto a two-dimensional map while preserving the topology of the input space, as shown in Figure 13. Like other artificial neural networks, though, SOM contains neurons (also called units or categories) that read weighted values of the input space and are located onto a low-dimensional plane, usually in the form of a rectangular or hexagonal grid. Classification using SOMs is achieved by a ‘winner-takes-all’ approach described in Figure 14.

SOMs have been used to explore the relationship of genes based on microarray gene expression data. In particular, high classification of cancer for different type of tissue samples can be achieved using oligonucleotide microarray with ~80% accuracy (Covell et al., 2003.) In this study, we take a similar approach using SOMs, except that the classification of bacterial species is determined by the features in a digital signature on the probes of an nxh chip from DNA biomarkers.

Definition of self-organizing maps (SOMs)

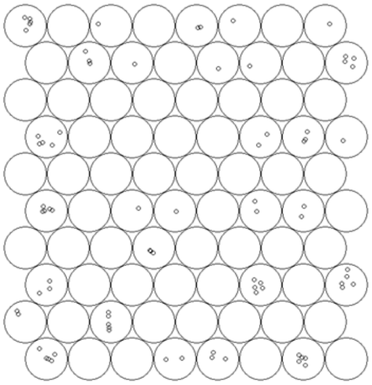


Figure 13: Self-organizing maps (SOMs) are a special kind of artificial neural network topology that is able to learn from unlabeled input data (i.e., with no knowledge of their corresponding categories.) This form of unsupervised learning is useful for clustering together data with similar feature patterns. A unique aspect of SOMs is that the cluster representation is embedded in a tessellation of a 2D plane, usually a rectangular grid or, as shown above, a hexagonal grid.

How to train self-organizing maps

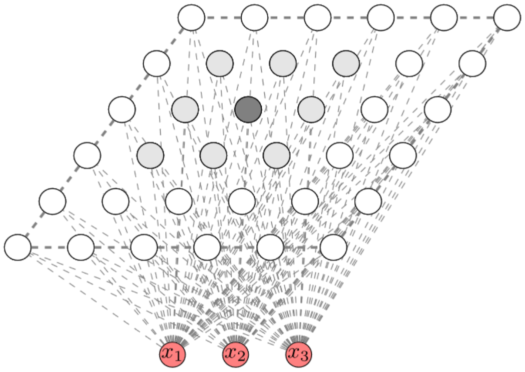


Figure 14: SOMs are trained by randomly initializing weights from features in the inputs to every neuron on the map. Starting with random weights, the Euclidean distances between the training data points and all nodes in the map are computed. Weights of the winner-takes-all (the closest) node are adjusted so that next time it will have a better chance of winning, while the weights into other (particularly neighbor) nodes are adjusted so that next time they will have less chance of winning (recognizing the data point.) Thus, after training, the weights into every node have been adjusted to recognize common features that are similar (but different) from their neighbors' and form a single cluster if recognized by the same winner. After training, a labeling phase labels each neuron sites to identify which category they recognize uniquely. Image is free to share and use at (https://commons.wikimedia.org/wiki/File:Self-organizing-map.svg.)

The following procedure to identify HAI species using SOMs on digital signatures obtained from nxh chips is described in (Garzon & Pham, 2018.) In order to perform classification with SOMs, the *supersom* function from the ‘kohonen’ library package in R and its default settings were used to obtain a SOM classifier since it supports supervised learning and prediction (Wehrens & Buydens, 2007.) The strains were classified into 16 species labeled 1-16. The data was partitioned into sets of 80%/20% for training and testing data. Features of the nxh basis were scaled to *z*-scores to mimic the example in the package documents. Prediction was performed using the *predict* function without species label and accuracy was based on whether the SOM correctly identified the unlabeled species with given input features of the nxh basis. This process was repeated for 32 models to obtain accuracy measurements. The mean of the 32 accuracy is reported. Additional performance measures including sensitivity, specificity, and precision were also reported for the testing set using the ‘caret’ package (Kuhn, 2008.)

2.4.1.3. Random Forests (RFs)

In computer science, a binary tree is a data structure consisting of nodes connected by edges. Each node is adjacent (nodes that share the same edge) to at most two others nodes of the parent-child type. The tree begins by initializing a root node that contains no incoming edges. All other nodes in the tree will have exactly one incoming edge. A node that contains an outgoing edge is known as an internal node, while a node that does not is called a leaf node. Decision trees are similar to binary trees and have been useful for tackling the classification problem as shown in Figure 15 (Song & Lu, 2015.) From a root node, the next level is determined by a threshold for the best attribute values that partitions the data set into roughly two halves; the procedure is repeated with each subtree using a common additional feature by a *recursive binary splitting* algorithm until a leaf node that contains a label is reached. The tree is then pruned in order to reduce the complexity (Figure 16.) Classification for a datum begins at the root node and continues along an edge that best describes the characteristic of a particular feature until a leaf node is reached. The datum is classified as the label of the leaf node it reaches.

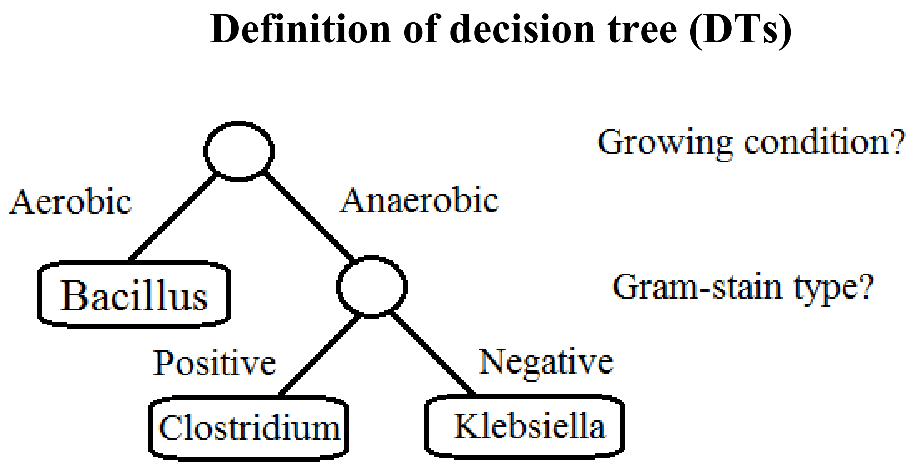


Figure 15: A tree is a directed acyclic graph with a finite number of nodes and adjacencies of the type parent-child. A decision tree solves a classification problem for certain data (e.g. bacteria) based on certain features associated with the levels of the tree. The most determinant feature determines the top adjacency, for example the growing condition, is shown at the top level. A datum is classified into two subsets, ideally roughly partitioning the data set into two halves. The process continues in each subtree at the second level based on a second common feature, and culminates in a childless node (leaf) with the category label There are as many leaves as there are categories in the classification, for example three (*Clostridium, Klebsiella,* and *Bacillus*) in this tree. Thus, a bacterium in this (unrealistic) example would be classified as *clostridium* genus if and only if it grows in anaerobic conditions and Gram-stain positive.

How to train decision trees

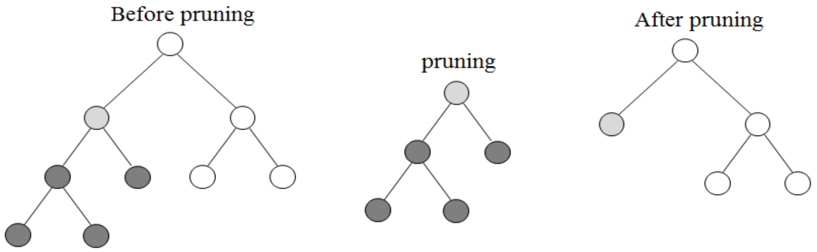


Figure 16: Decision trees are grown by a greedy approach called *recursive binary splitting*. This divides the input space into different subtrees where each node (split point) is evaluated and chosen to minimize a cost function (e.g., the Gini Index for classification trees.) Simpler trees are preferred and are less likely to overfit the data, so a tree is pruned by measuring how the decision trees perform without some of the subtrees. There are different types of pruning methods. The above figure shows a postpruning method where the trees are grown fully and then each subtree are pruned until a simpler and more generalized decision tree is achieved with minimal classification error.

A random forest (RF) is a supervised classification method that is built from a finite collection of decision trees used to make a prediction on unknown data as shown in Figure 17 (Breiman, 2001.) Each tree is grown from a bootstrap sampling of *k* features from the total of *m* features of the original dataset and the best split points are calculated using the subset of *k* features. Unlike decision trees, RF trees are unpruned to obtain low biases. Additionally, the randomized selection of features ensures low correlation between the trees. This process is repeated until the random forest consists of *r* number of decision trees as shown in Figure 18. A prediction is made on an unknown datum based on a majority vote across all trees. This training technique is responsible for making RFs as one of the most accurate classification machine learning techniques. Some of the advantages of using RF with microarray dataset are its running efficiency on large datasets, reduction of overfitting, and its predictive performance even when most variables are noise (Díaz-Uriarte & Alvarez de Andrés, 2006.) This eliminates the need for a pre-selection step of features. Instead, RFs is able to perform variable importance measurements to achieve the best results.

Definition of Random Forests (RFs)

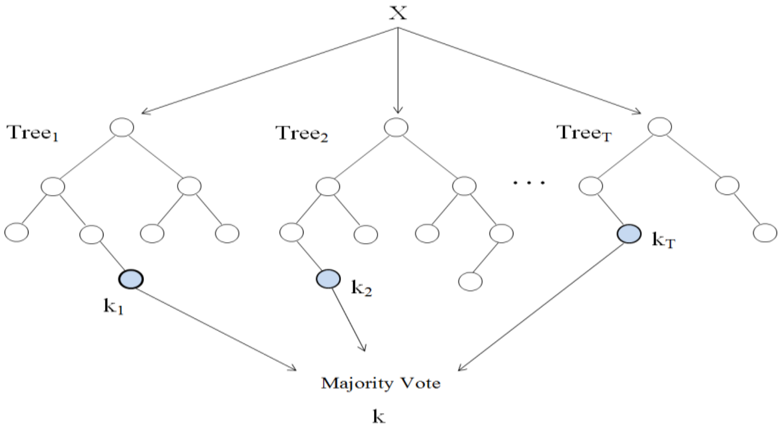


Figure 17: A random forest (RF) is another solution to a classification problem where the target data are heterogeneous and belong to several fairly distinct types, or there are several competing criteria to consider in the classification. It consists of a finite collection of decision trees, one for each type or criterion as the feature at the top of the subtree. The figure shows a typical architecture for a random forest. There are *T* decision trees grown from random bootstrap sampling; a simple majority voting can be used to determine a category to classify a target datum.

How to train random forests

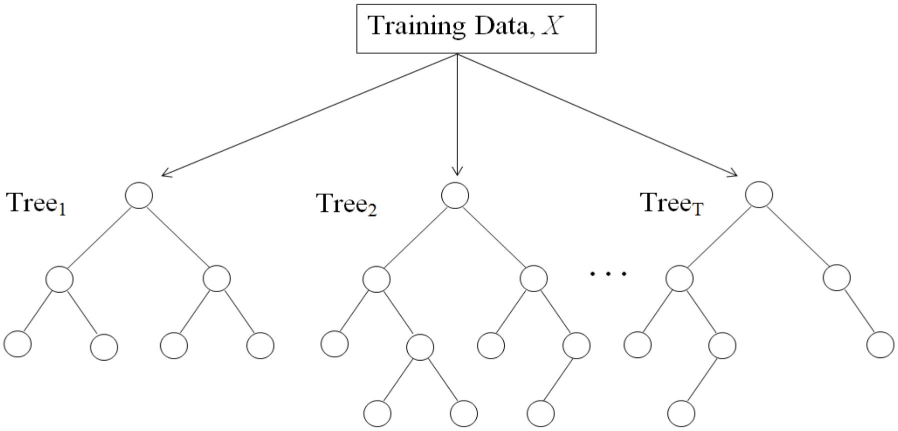


Figure 18: Training a random forest requires taking N random bootstrap samples from the data and constructing a decision tree for each bootstrap sample. Each node in a tree is trained on the corresponding random subset of features, as described in Figure 15. This will ensure that trees are not correlated and the random forest ensemble will have low variance. The process is repeated until *T* number of trees are grown that make up the random forest.

The following procedure to identify HAI species using RFs on digital signatures from nxh chips is described in (Garzon & Pham, 2018.) In order to classify the species of HAI, the *randomForest* function from the ‘randomForest’ package in R was used to train a data set and make predictions (Liaw & Wiener, 2002.) The strains were classified into 16 species labeled 1-16. As before, the data was partitioned into sets of 80% for training and 20% for testing. A RF model was fitted by formulating the species label as the response variable, to the feature values in the digital signatures, for each of several combinations of nxh bases. The default settings of the *randomForest* were retained (ntree = 500.) Prediction was done using the *predict* function without the species label for each strain and with only their features depending on the nxh basis. Accuracy performance was determined by whether the RF correctly identified the strain to its own species. Both training and testing accuracy were measured. The training was run 32 times for 32 different models, where each repetition selected different samples of training and testing data to obtain an accuracy percentage. The mean and standard deviation of the 32 prediction accuracy percentages for each of the training and testing are reported. Like NNs and SOMs, performance measures of sensitivity, specificity, and precision were also reported for the testing set using the ‘caret’ package (Kuhn, 2008.)

2.5. Data Collection and Processing

The designs of each nxh basis used in this thesis are shown in Table 4 below. Each nxh basis was given by (Garzon et al., 2017.) An example of one of the designs in Table 4 such as A (3mE4-2-at1.1) contains 3 probes of length 4, separated at an *h*-distance of 2 between any pair of probes, and two strands will only hybridize with each other if and only if their *h*-distance is less than 1.1. Whole-genome sequences of 80 HAI pathogens were then downloaded from GenBank (Benson et al., 2013.) The distribution of species, along with the overall species genome size is shown in Table 5. There were a total of 5 strains per species, and each species was selected based on some of the top HAIs listed by the Centers for Disease Control and Prevention (CDC) in Table 1 and from the World Health Organization (WHO) list of top pathogens harmful to human health (Dontinga, 2017.)

In order to obtain a 'digital signature' for a strain, the whole genome sequence was shredded to fragments of length *n* equal to the length of the probe of an nxh basis. The frequency of all *p*-mers was collected using a Python script. The Python script selects one of the two sequences of a *p*-mer based on lexicographical order. Next, a Perl script was used to compute the *h*-distance between the *p*-mers and the probes. If the *h*-distance between a *p*-mer and a probe was less than the stringency condition, τ, then the frequency of the *p*-mer was accumulated to that probe. Finally, the probes were normalized by dividing by the total number of *p*-mer counts. Digital signatures from a sample of two strains per species are shown for each nxh basis in Figure 19, 20, and 21. All visualization of the digital signatures was completed using *pheatmap* (Kolde, 2012.)

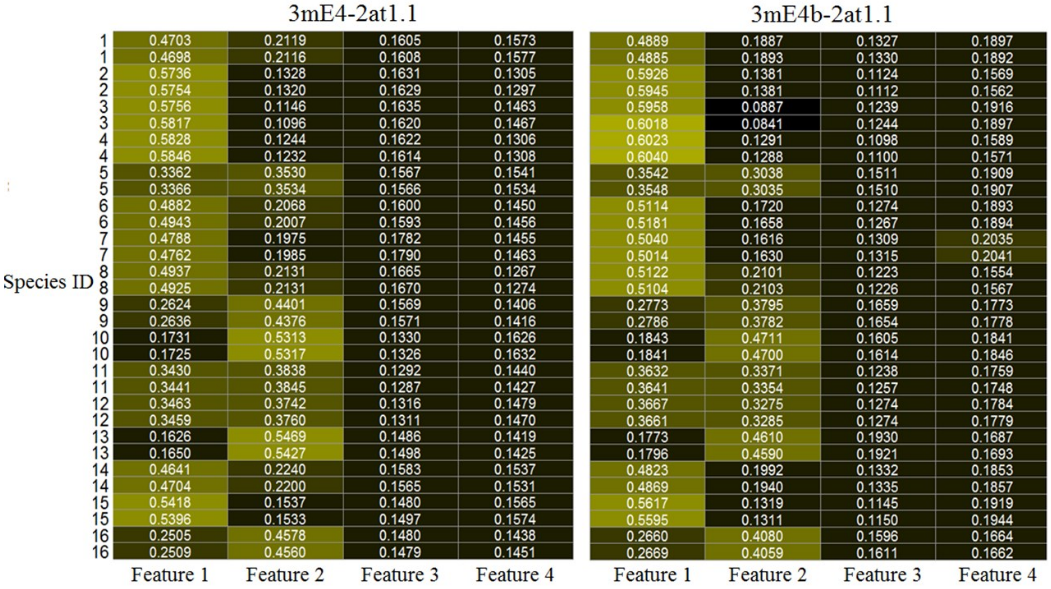
Table 4: **Nxh chip designs used to obtain digital signatures** (Garzon, 2017)

Basis	Length of Probes	Number of Probes	T
A: 3mE4-2-at1.1	3	4	1.1
B: 3mE4b-2at1.1	3	4	1.1
C: 4mP3-3at2.1	4	3	2.1
D: 8mP10-4at4.1	8	10	4.1

Table 5: **Species in the data sample and their relative genome size**

ID	Species	Genome Size (Mbps)
1	*Acinetobacter Baumannii*	20.108
2	*Campylobacter Coli*	8.610
3	*Campylobacter Jejuni*	8.309
4	*Clostridium Difficile*	20.808
5	*Escherichia Coli*	25.929
6	*Klebsiella Pneumoniae*	26.389
7	*Proteus Mirabilis*	20.491
8	*Serratia Marcescens*	26.161
9	*Enterococcus Faecalis*	13.784
10	*Enterococcus Faecium*	14.182
11	*Helicobacter Pylori*	8.082
12	*Mycobacterium Tuberculosis*	22.008
13	*Neisseria Gonorrhoeae*	10.967
14	*Neisseria Meningitidis*	10.755
15	*Pseudomonas Aeruginosa*	33.723
16	*Staphylococcus Aureus*	14.713

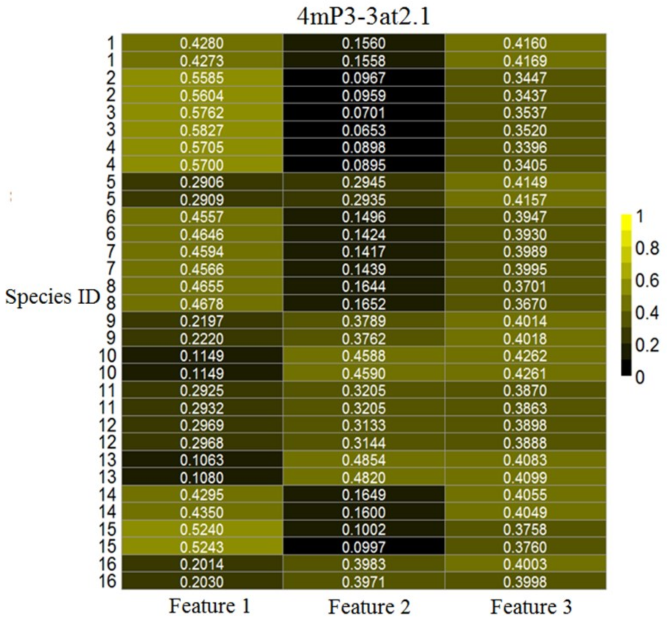
Representative digital signatures for 3mE4-2at1.1 and 3mE4b-2at1.1



Species ID	3mE4-2at1.1				3mE4b-2at1.1			
	Feature 1	Feature 2	Feature 3	Feature 4	Feature 1	Feature 2	Feature 3	Feature 4
1	0.4703	0.2119	0.1605	0.1573	0.4889	0.1887	0.1327	0.1897
1	0.4698	0.2116	0.1608	0.1577	0.4885	0.1893	0.1330	0.1892
2	0.5736	0.1328	0.1631	0.1305	0.5926	0.1381	0.1124	0.1569
2	0.5754	0.1320	0.1629	0.1297	0.5945	0.1381	0.1112	0.1562
3	0.5756	0.1146	0.1635	0.1463	0.5958	0.0887	0.1239	0.1916
3	0.5817	0.1096	0.1620	0.1467	0.6018	0.0841	0.1244	0.1897
4	0.5828	0.1244	0.1622	0.1306	0.6023	0.1291	0.1098	0.1589
4	0.5846	0.1232	0.1614	0.1308	0.6040	0.1288	0.1100	0.1571
5	0.3362	0.3530	0.1567	0.1541	0.3542	0.3038	0.1511	0.1909
5	0.3366	0.3534	0.1566	0.1534	0.3548	0.3035	0.1510	0.1907
6	0.4882	0.2068	0.1600	0.1450	0.5114	0.1720	0.1274	0.1893
6	0.4943	0.2007	0.1593	0.1456	0.5181	0.1658	0.1267	0.1894
7	0.4788	0.1975	0.1782	0.1455	0.5040	0.1616	0.1309	0.2035
7	0.4762	0.1985	0.1790	0.1463	0.5014	0.1630	0.1315	0.2041
8	0.4937	0.2131	0.1665	0.1267	0.5122	0.2101	0.1223	0.1554
8	0.4925	0.2131	0.1670	0.1274	0.5104	0.2103	0.1226	0.1567
9	0.2624	0.4401	0.1569	0.1406	0.2773	0.3795	0.1659	0.1773
9	0.2636	0.4376	0.1571	0.1416	0.2786	0.3782	0.1654	0.1778
10	0.1731	0.5313	0.1330	0.1626	0.1843	0.4711	0.1605	0.1841
10	0.1725	0.5317	0.1326	0.1632	0.1841	0.4700	0.1614	0.1846
11	0.3430	0.3838	0.1292	0.1440	0.3632	0.3371	0.1238	0.1759
11	0.3441	0.3845	0.1287	0.1427	0.3641	0.3354	0.1257	0.1748
12	0.3463	0.3742	0.1316	0.1479	0.3667	0.3275	0.1274	0.1784
12	0.3459	0.3760	0.1311	0.1470	0.3661	0.3285	0.1274	0.1779
13	0.1626	0.5469	0.1486	0.1419	0.1773	0.4610	0.1930	0.1687
13	0.1650	0.5427	0.1498	0.1425	0.1796	0.4590	0.1921	0.1693
14	0.4641	0.2240	0.1583	0.1537	0.4823	0.1992	0.1332	0.1853
14	0.4704	0.2200	0.1565	0.1531	0.4869	0.1940	0.1335	0.1857
15	0.5418	0.1537	0.1480	0.1565	0.5617	0.1319	0.1145	0.1919
15	0.5396	0.1533	0.1497	0.1574	0.5595	0.1311	0.1150	0.1944
16	0.2505	0.4578	0.1480	0.1438	0.2660	0.4080	0.1596	0.1664
16	0.2509	0.4560	0.1479	0.1451	0.2669	0.4059	0.1611	0.1662

Figure 19: A selection of normalized digital signatures of two strains is shown for each of the HAIs species (Table 5) on the first two nxh bases (Table 4) used.

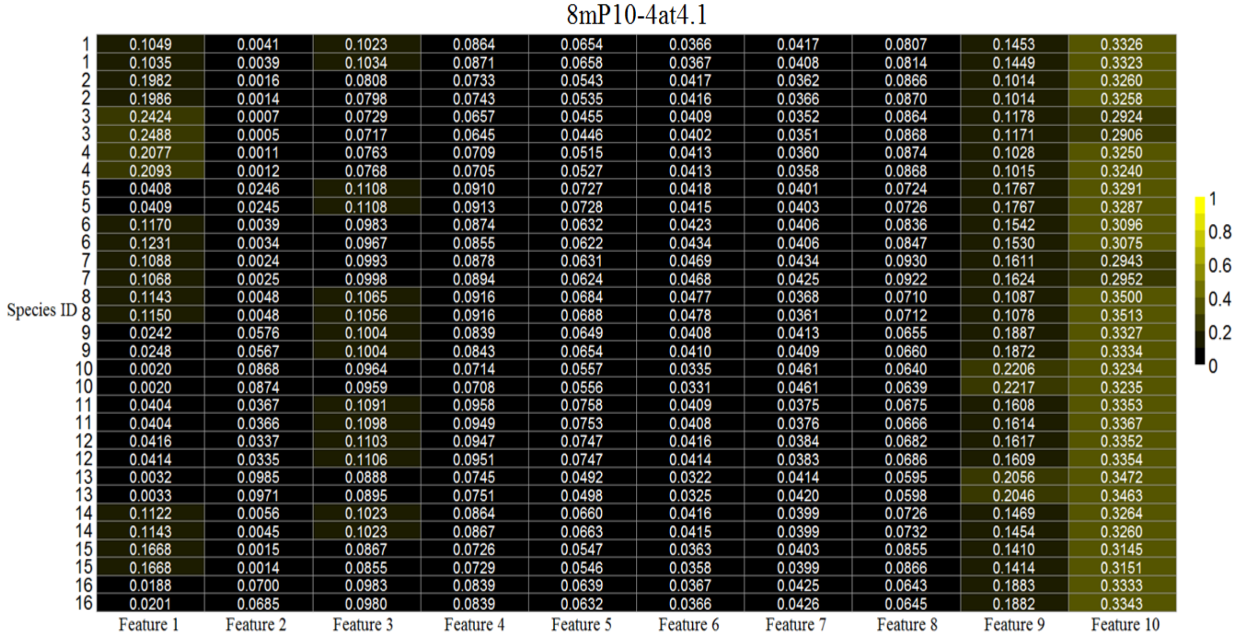
Representative digital signatures for 4mP3-3at2.1



Species ID	Feature 1	Feature 2	Feature 3
1	0.4280	0.1560	0.4160
1	0.4273	0.1558	0.4169
2	0.5585	0.0967	0.3447
2	0.5604	0.0959	0.3437
3	0.5762	0.0701	0.3537
3	0.5827	0.0653	0.3520
4	0.5705	0.0898	0.3396
4	0.5700	0.0895	0.3405
5	0.2906	0.2945	0.4149
5	0.2909	0.2935	0.4157
6	0.4557	0.1496	0.3947
6	0.4646	0.1424	0.3930
7	0.4594	0.1417	0.3989
7	0.4566	0.1439	0.3995
8	0.4655	0.1644	0.3701
8	0.4678	0.1652	0.3670
9	0.2197	0.3789	0.4014
9	0.2220	0.3762	0.4018
10	0.1149	0.4588	0.4262
10	0.1149	0.4590	0.4261
11	0.2925	0.3205	0.3870
11	0.2932	0.3205	0.3863
12	0.2969	0.3133	0.3898
12	0.2968	0.3144	0.3888
13	0.1063	0.4854	0.4083
13	0.1080	0.4820	0.4099
14	0.4295	0.1649	0.4055
14	0.4350	0.1600	0.4049
15	0.5240	0.1002	0.3758
15	0.5243	0.0997	0.3760
16	0.2014	0.3983	0.4003
16	0.2030	0.3971	0.3998

Figure 20: A selection of normalized digital signatures of two strains is shown for each of the HAIs species (Table 5) on 4mP3-2at2.1.

Representative digital signatures for 8mP10-4at4.1



Species ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
1	0.1049	0.0041	0.1023	0.0864	0.0654	0.0366	0.0417	0.0807	0.1453	0.3326
1	0.1035	0.0039	0.1034	0.0871	0.0658	0.0367	0.0408	0.0814	0.1449	0.3323
2	0.1982	0.0016	0.0808	0.0733	0.0543	0.0417	0.0362	0.0866	0.1014	0.3260
2	0.1986	0.0014	0.0798	0.0743	0.0535	0.0416	0.0366	0.0870	0.1014	0.3258
3	0.2424	0.0007	0.0729	0.0657	0.0455	0.0409	0.0352	0.0864	0.1178	0.2924
3	0.2488	0.0005	0.0717	0.0645	0.0446	0.0402	0.0351	0.0868	0.1171	0.2906
4	0.2077	0.0011	0.0763	0.0709	0.0515	0.0413	0.0360	0.0874	0.1028	0.3250
4	0.2093	0.0012	0.0768	0.0705	0.0527	0.0413	0.0358	0.0868	0.1015	0.3240
5	0.0408	0.0246	0.1108	0.0910	0.0727	0.0418	0.0401	0.0724	0.1767	0.3291
5	0.0409	0.0245	0.1108	0.0913	0.0728	0.0415	0.0403	0.0726	0.1767	0.3287
6	0.1170	0.0039	0.0983	0.0874	0.0632	0.0423	0.0406	0.0836	0.1542	0.3096
6	0.1231	0.0034	0.0967	0.0855	0.0622	0.0434	0.0406	0.0847	0.1530	0.3075
7	0.1088	0.0024	0.0993	0.0878	0.0631	0.0469	0.0434	0.0930	0.1611	0.2943
7	0.1068	0.0025	0.0998	0.0894	0.0624	0.0468	0.0425	0.0922	0.1624	0.2952
8	0.1143	0.0048	0.1065	0.0916	0.0684	0.0477	0.0368	0.0710	0.1087	0.3500
8	0.1150	0.0048	0.1056	0.0916	0.0688	0.0478	0.0361	0.0712	0.1078	0.3513
9	0.0242	0.0576	0.1004	0.0839	0.0649	0.0408	0.0413	0.0655	0.1887	0.3327
9	0.0248	0.0567	0.1004	0.0843	0.0654	0.0410	0.0409	0.0660	0.1872	0.3334
10	0.0020	0.0868	0.0964	0.0714	0.0557	0.0335	0.0461	0.0640	0.2206	0.3234
10	0.0020	0.0874	0.0959	0.0708	0.0556	0.0331	0.0461	0.0639	0.2217	0.3235
11	0.0404	0.0367	0.1091	0.0958	0.0758	0.0409	0.0375	0.0675	0.1608	0.3353
11	0.0404	0.0366	0.1098	0.0949	0.0753	0.0408	0.0376	0.0666	0.1614	0.3367
12	0.0416	0.0337	0.1103	0.0947	0.0747	0.0416	0.0384	0.0682	0.1617	0.3352
12	0.0414	0.0335	0.1106	0.0951	0.0747	0.0414	0.0383	0.0686	0.1609	0.3354
13	0.0032	0.0985	0.0888	0.0745	0.0492	0.0322	0.0414	0.0595	0.2056	0.3472
13	0.0033	0.0971	0.0895	0.0751	0.0498	0.0325	0.0420	0.0598	0.2046	0.3463
14	0.1122	0.0056	0.1023	0.0864	0.0660	0.0416	0.0399	0.0726	0.1469	0.3264
14	0.1143	0.0045	0.1023	0.0867	0.0663	0.0415	0.0399	0.0732	0.1454	0.3260
15	0.1668	0.0015	0.0867	0.0726	0.0547	0.0363	0.0403	0.0855	0.1410	0.3145
15	0.1668	0.0014	0.0855	0.0729	0.0546	0.0358	0.0399	0.0866	0.1414	0.3151
16	0.0188	0.0700	0.0983	0.0839	0.0639	0.0367	0.0425	0.0643	0.1883	0.3333
16	0.0201	0.0685	0.0980	0.0839	0.0632	0.0366	0.0426	0.0645	0.1882	0.3343

Figure 21: A selection of normalized digital signatures of two strains is shown for each of the HAIs species (Table 5) on 8mP10-4at4.1.

Chapter 3: The Classifiers

With the preliminaries in place, I can now present the classifiers for HAIs and an assessment of their performance with respect to the current state-of-the-art methods.

3.1. Statistical Solutions

Traditionally, statistical methods that test for significant differences (e.g., students' t-test) have been the choice for solutions to problems where uncertainty plays a large role, such as HAI infections. Naturally, this kind of method had to be tried first, beginning with the simpler problem of statistically significant discrimination of bacterial strains and/or species, particularly using decision trees. Two strains were considered significantly different if the z-score of any feature was at least $\sigma = 1$ unit of standard error (SE) difference apart. Figure 22 shows how well each nxh basis could discriminate each strain at different taxa level. This statistical method was able to significantly discriminate strains at the genus level with 99.18% accurate discrimination using all features in nxh bases A, B, C, and D combined.

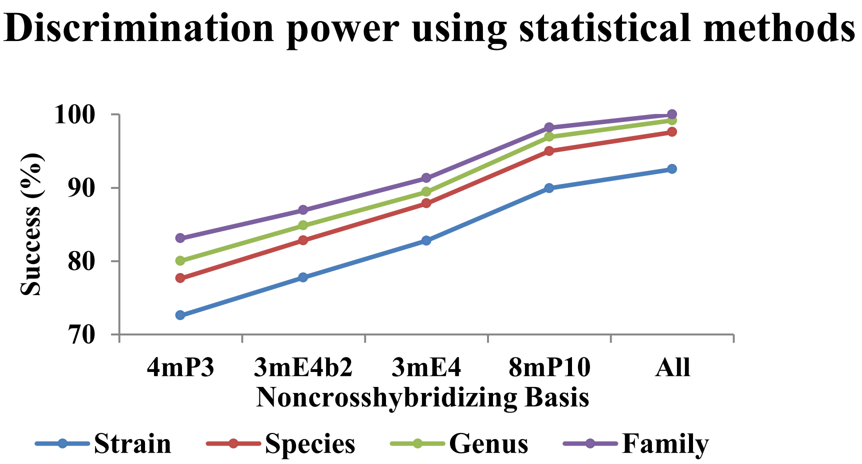


Figure 22: The success rate of discrimination within different levels of taxa by statistical methods using each nxh basis in Table 4 is shown. As the level of taxa decreases, so does the discrimination power. It can be observed that as more features are included, the discrimination power increases (3/10/21 features in 4mP3/8mP10/All bases, respectively.)

There is a trend in successful discrimination (e.g., including more features.) 4mP3-3at2.1 contains 3 features, whereas 'All' is the combination of all features from bases in Table 5 achieves the best discrimination power. This method, however, fails to discriminate strains within the same genera and thus provides poor resolution for species discrimination, which is an easier problem than the full identification problem. Therefore, this method is not ideal for species identification since closely related species could not fully be discriminated. Therefore, more powerful machine learning techniques have to be used.

3.2. Neural Networks Solutions

For NNs, single and paired combination of nxh basis performed poorly, with average accuracies below 85%. Results improved when they were trained on triplet or quadruplet combinations of nxh features. As shown in Figure 23, the triplet and quadruplet combination of nxh features could consistently achieve an average accuracy above 90% with low variability using various NN architectures. However, features from 3mE4-2at1.1, 3mE4b2-2at1.1, and

4mP3-3-at2.1 nor 3mE4-2at1.1, 3mE4b2-2at1.1, and 8mP10-4at4.1 combined could actually achieve accuracy above 90% on the average. The best triplet combinations of features are from 3mE4-2at1.1, 4mP3-3at1.1, and 8mP10-4at4.1 using 3 hidden layers (14, 10, and 8.) This combination achieves an average accuracy of 94.14% on the testing set. The combination of features from all four bases with a NN architecture of 3-hidden layers (18, 14, and 6), however, achieves an average accuracy of 94.34% on the testing set, as well as higher sensitivity, specificity, and precision, as shown in Table 6. The average accuracy, sensitivity, specificity, and precision between both combinations were not significantly different (*p*-value > 0.05, a student's t-test between the measurements of the two combinations of nxh was conducted to test for significant with α = 0.05) and so triplet combinations of nxh features will be sufficient to achieve high accuracy for species identification. In addition to higher accuracy, the combination of three or four bases achieves at least 91% in sensitivity, specificity, and precision on the testing set, as can be seen in Table 6.

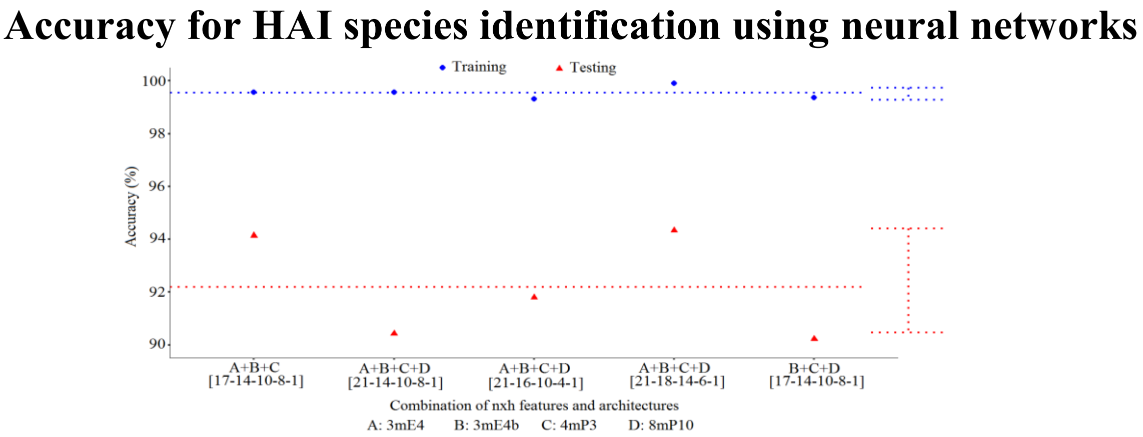


Figure 23: Neural Networks can consistently achieve nearly over 90% accuracy on HAI identification of bacterial strains. The top five performing feedforward neural networks (FNNs) are shown above, with shown standard deviation (blue/red error bars) from the mean (blue/red dotted line.) Neural networks achieved high accuracy with the combination of the features in all bases. The combination of bases and neural net architectures are described along the *x*-axis.

Table 6: **Performance of neural networks (NNs) on HAI identification**
The combination of all features in the bases in Table 4 achieves the best performance in terms of accuracy for training and testing, along with low variance and higher sensitivity, specificity, and precision.

Basis and Architecture	Training Accuracy	Training Accuracy Std	Testing Accuracy	Testing Accuracy Std	Testing Sensitivity	Testing Specificity	Testing Precision
A+C+D [17-14-10-8-1]	99.56	0.99	94.14	7.09	94.87	99.42	95.16
B+C+D [17-14-10-8-1]	99.37	1.48	90.23	7.09	91.65	99.08	92.05
A+B+C+D [21-14-10-8-1]	99.56	1.14	90.43	7.61	91.82	99.14	91.98
A+B+C+D [21-16-10-4-1]	99.32	1.48	91.80	7.18	92.50	99.23	93.44
A+B+C+D [21-18-14-6-1]	**99.90**	**0.38**	**94.34**	**5.30**	**95.17**	**99.46**	**95.64**

These results show that classification with NNs is capable of discriminating power at the species level with only the combination of features from 3mE4-2at1.1, 4mP3-3at2.1, and 8mP10-4at4.1, whereas the statistical method requires all features combined to achieve a high discrimination power. Moreover, NNs could discriminate between species within the same genus as indicated by the high accuracy, unlike the statistical method. More significantly, NNs with combined nxh features also outperform MALDI TOF MS (84%) in terms of accuracy, as described in section 2.2.1.

3.3. Self-Organizing Maps Solutions

SOMs turned out to be one of the two best performers, consistently achieving an average accuracy of 97.36% on testing with low variability (std of 2.03) using most combinations of features from an nxh basis. As shown in Figure 24, 4mP3-3at2.1 achieves the lowest accuracy on the testing set with an average accuracy of 92.77%.

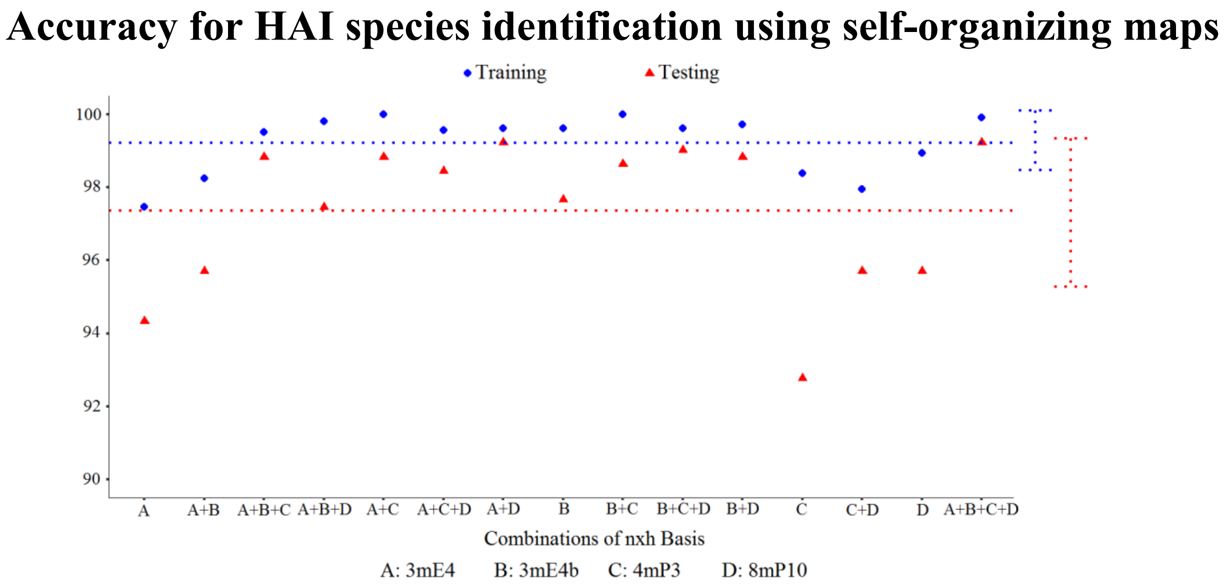


Figure 24: Self-organizing maps can achieve nearly 100% accuracy on HAI identification of bacterial strains by a genomic method on next generation microarrays (nxh chips.). Virtually any combinations of nxh bases can consistently identify the species in each strain with probability at least 92% accuracy on the testing set and low standard deviation (blue/red error bars) from the mean (blue/red dotted lines.)

The best performance on testing in terms of accuracy is the combination of all features from the four basis and features from 3mE4-2at1.1 and 8mP10-4at4.1 combined. Both achieve 99.22% accuracy, as shown in Table 7. However, the standard deviation of the accuracies from 32 runs is lower using 3mE4-2at1.1 and 8mP10-4at4.1 (3.07) than combining features from all four bases (3.46.) Both combinations shared similar sensitivity and specificity and were not significantly different (*p*-value > 0.05 in a similar test as described above.) Thus, the combination of features from 3mE4-2at1.1 and 8mP10-4at4.1 would reduce the need for trials of combinations of features from three or more nxh bases for high accuracy, which will speed the process of species identification.

Table 7: **Performance of self-organizing maps (SOMs) on HAI identification**
The combination of features from all nxh bases (A+B+C+D) from Table 4 and 3mE4-2at1.1 (A) and 8mP10-4at2.1 (D) combined achieves the highest testing accuracy (99.22%.) 'A+B+C+D' achieves the highest specificity while the combination of 'A+D' achieves the best sensitivity on the testing set. 'A+D' has lower variation of testing accuracy (3.07) than any other combination of features from nxh bases.

Basis	Train Accuracy	Train Accuracy Std	Test Accuracy	Test Accuracy Std	Test Sensitivity	Test Specificity	Test Precision
A	97.46	4.69	94.34	10.45	95.03	99.51	97.02
A+B	98.24	4.04	95.70	9.04	96.35	99.65	97.18
A+B+C	99.51	2.08	98.83	4.03	98.81	99.90	99.08
A+B+D	99.80	0.77	97.46	6.52	98.63	99.77	98.44
A+C	**100.00**	**0.00**	98.83	3.70	98.57	99.90	98.85
A+C+D	99.56	1.99	98.44	5.50	98.55	99.86	98.70
A+D	99.61	2.21	**99.22**	**3.07**	**99.46**	99.92	99.37
B	99.61	2.21	97.66	5.20	97.87	99.80	97.74
B+C	**100.00**	**0.00**	98.63	4.69	98.81	99.88	98.81
B+C+D	99.61	2.21	99.02	3.92	99.24	99.92	99.53
B+D	99.71	1.66	98.83	6.63	99.15	99.90	**99.65**
C	98.39	3.11	92.77	12.41	94.33	99.35	96.54
C+D	97.95	4.21	95.70	8.90	96.09	99.61	97.29
D	98.93	3.06	95.70	11.82	96.69	99.60	98.06
A+B+C+D	99.90	0.55	**99.22**	3.46	99.40	**99.93**	99.40

These results show that SOMs outperform NNs in that fewer nxh features were required to achieve accuracy near 100%. Moreover, unlike NN, the optimal architecture to produce quality results did not need to be optimized. Furthermore, SOMs also outperform MALDI TOF MS (84%) at the species level, on average accuracy.

3.4. Random Forests Solutions

RFs performs the best, achieving an average accuracy of 98.5% on the testing set with low variability (std of 0.4) using any combination of features from nxh basis as show in Figure 25. The lowest accuracy of the testing set was 97.85% using 3mE4b-2at1.1. The best performance across all performance metrics except specificity, are features from 3mE4-2at1.1 as shown in Table 8. A triplet combination of features from 3mE4-2at1.1, 3mE4b-2at1.1, and 8mP10-4at4.1 had the highest sensitivity (99.37%), although it was not significantly different (*p*-value > 0.05 in a similar test as described above) from the combination of features from 3mE4-2at1.1 (99.28%). Using RFs as a classifier with nxh features as input will further reduce the time and work required for bacterial identification since only a single basis is needed for high performance.

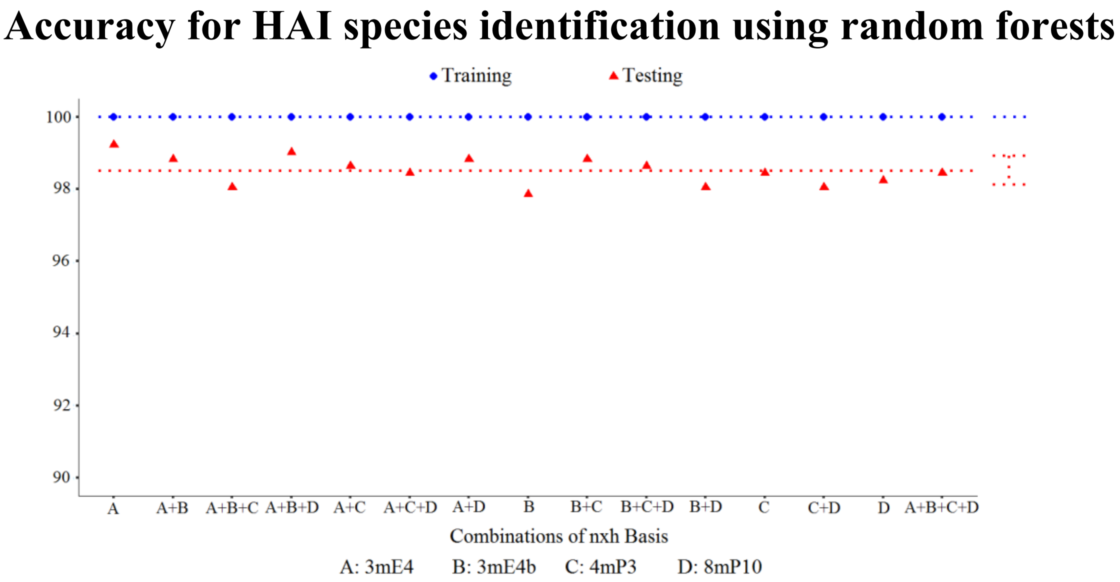


Figure 25: Random Forests with 500 trees achieve at least 97% accuracy on HAI identification of bacterial strains with combinations of nxh bases shown in Table 5, outperforming neural networks and comparable with or better than self-organzing maps. The accuracy rate is consistent across any combination of nxh bases with low standard deviation (blue/red error bars) from the mean (blue/red dotted lines.)

Table 8: **Performance of random forests (RFs) on HAI identification**
The features in basis 3mE4-2at1.1 achieve the best performance in the quality measurements of accuracy, specificity, and precision. 'A+B+D' achieved the highest sensitivity.

Basis	**Training Accuracy**	**Training Accuracy Std**	**Testing Accuracy**	**Testing Accuracy Std**	**Testing Sensivitiy**	**Testing Specificity**	**Testing Precision**
A	**100.00**	**0.00**	**99.22**	**2.10**	99.28	**99.93**	**99.31**
A+B	**100.00**	**0.00**	98.83	2.48	98.45	99.90	98.52
A+B+C	**100.00**	**0.00**	98.05	2.94	97.69	99.82	97.74
A+B+D	**100.00**	**0.00**	99.02	2.31	**99.37**	99.90	99.26
A+C	**100.00**	**0.00**	98.63	2.63	98.39	99.88	98.67
A+C+D	**100.00**	**0.00**	98.44	2.75	98.18	99.85	98.78
A+D	**100.00**	**0.00**	98.83	2.48	99.01	99.88	99.18
B	**100.00**	**0.00**	97.85	6.07	98.18	99.80	98.45
B+C	**100.00**	**0.00**	98.83	2.48	99.14	99.89	99.07
B+C+D	**100.00**	**0.00**	98.63	2.63	98.82	99.86	98.87
B+D	**100.00**	**0.00**	98.05	4.03	98.34	99.81	98.74
C	**100.00**	**0.00**	98.44	2.75	98.62	99.86	98.20
C+D	**100.00**	**0.00**	98.05	3.34	98.10	99.81	98.69
D	**100.00**	**0.00**	98.24	3.27	98.67	99.84	98.24
A+B+C+D	**100.00**	**0.00**	98.44	2.70	98.47	99.85	98.69

Classification with RFs outperformed NNs and produced accuracy comparable or better than SOMs for HAI species identification. Any combination of features could produce an accuracy of 97% on the average, whereas NNs required triplet or more combination of features to achieve high accuracies. Additionally, some combination of features from nxh bases were comparable using RFs and SOMs, although no single nxh basis could achieve an accuracy of at least 97% as shown with RFs. In Figure 24, we see that features from the nxh basis, 3mE4b-2at1.1 (C) achieve an accuracy of 92.77% but RFs achieve an accuracy of 98.44% using the same features. These results show that nxh features combined with a RF classifier will be an accurate and cost-effective alternative to other classifiers, and, at least in principle, better than MALDI TOF MS.

Chapter 4: Summary and Conclusions

In this thesis, a new method for the identification of HAIs (Hospital-Acquired bacterial Infection) species has been proposed based on genomic data about the species and next-generation microarray designs. Machine learning techniques yield classifiers (such as Neural Nets, Self-Organizing Maps, and Random Forests) for identification of HAI species that are competitive with state-of-the-art methods using MALDI and WGS. There are several additional advantages to this methodology. First, it is scalable to a wide range of biomarkers, including whole genomes. Second, it uses sets of microarray designs that are *universal, i.e.,* prior knowledge of DNA sequences that distinguish between species is not required, unlike with standard microarrays and other techniques based on conserved regions of the DNA. Moreover, this method does not rely on a cut-off threshold to determine a species match. All identifications are based on a prediction given by a machine learning algorithm that learns from input features (from an nxh base.) Furthermore, this method is robust in the sense that small variations in a genomic biomarker (e.g., low-rate mutations caused by development of antibiotic resistance) are very unlikely to alter the outcomes of the classification of a given strain, even if they were to cause changes in the microarrays readout for the strain. These properties stem from the optimization of microarray designs in nxh chips and properties of the *h*-distance model that allow for small mutations in the target strands up to the stringency condition τ, before causing the target strand to hybridize to a different probe. Third, the methodology is scalable also in the sense that longer probe sequences on the chip designs are likely to increase both the accuracy of the classifications and the robustness to mutations in the target.

However, the new methodology does not come without limitations. First, digital signatures were obtained *in silico*, where it is easy to enforce the assumptions of perfect shredding of the genomic sequences into fragments of equal length and noise-free hybridization of fragments to probes. This assumption is not likely to hold true if the methodology was to be implemented *in vitro* using current standard microarray technology. On the positive side, the fact that nxh chips are designed to be primarily error-free for hybridization makes it possible that further work may afford a feasible implementation of the technology *in vitro*, which will remove the bottleneck for these classifiers (acquiring digital signatures) to become very fast, even for full genomes. They would also substantially reduce the cost of the classification per sample. Thus the refinements of these methods to an actual clinical test remain an intriguing possibility, although that was not part of the original scope of this project.

References

1. Aguilera-Arreola, M. G. (2015). Identification and typing methods for the study of bacterial infections: a brief review and mycobacterial as case of study. *Archives of Clinical Microbiology*, vol. 7, no. 1:3. http://www.acmicrob.com/microbiology/identification-and-typing-methods-for-thestudy-of-bacterial-infections-a-brief-reviewand-mycobacterial-as-case-of-study.pdf.

2. Arora, A., Candel, A., Lanford, J., LeDell, E., & Parmar, V. (2006). Deep learning with H2O.

3. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research* 41, Database issue: D36–D42. https://doi.org/10.1093/nar/gks1195.

4. van Belkum, A., Welker, M., Pincus, D., Charrier, J-P., & Girard, V. (2017). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: what are the current issues? *Annals of Laboratory Medicine*, vol. 37(6), pp. 475–483. https://doi.org/10.3343/alm.2017.37.6.475.

5. Breiman, L. (2001). Random Forest. *Machine Learning*, vol. 45, pp. 5-32. https://doi.org/10.1023/A:1010933404324.

6. Bumgarner, R. (2013). DNA microarrays: types, applications and their future. *Current Protocols in Molecular Biology*, 0 22, Unit–22.1. https://doi.org/10.1002/0471142727.mb2201s10.

7. Carbonnelle, E., Mesquite, C., Bille, E., Day, N., Dauphin, B., Beretti, J., Ferroni, A., Gutmann, L., & Nassif, X. (2011). MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory. *Clinical Biochemisty*, vol. 44, pp. 104-109. https://doi.org:10.1016/j.clinbiochem.2010.06.017.

8. Chu, C.M., Yao, C.T., Chang, Y.T., Chou, H.L., Chou, Y.C., Chen, K.H., et al. (2014). Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees. *Disease Marker*s, vol. 2014, 634123. https://doi.org/10.1155/2014/634123.

9. Clark, A. E., Kaleta, E. J., Arora, A., & Wolk, D. M. (2013). Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry: a Fundamental Shift in the Routine Practice of Clinical Microbiology. *Clinical Microbiology Reviews*, vol. 26 (3), pp. 547–603. https://doi.org/10.1128/CMR.00072-12.

10. Covell, D.G., Wallqvist, A., Rabow, A.A., & Thanki, N. (2003). Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. *Molecular Cancer Therapeutics*, vol. 2 (3), pp. 317-332.

11. Dhiman, N., Hall, L., Wohlfiel, S.L., Buckwalter, S.P., & Wengenack, N.L. (2011). Performance and cost analysis of MALDI-TOF mass spectrometry for routine identification of Yeast. *Journal of Clinical Microbiology*, vol. 49 (4), pp. 1614-1616. https://doi.org/10.1128/JCM.02381-10.

12. Dontinga, R. (2017). Hospital infections top WHO's list of priority pathogens. *Frontline Medical Communications Inc*. www.the-hospitalist.org/hospitalist/article/136547/hospital-acquired-infections/hospital-infections-top-whos-list-priority.

13. Deurenberg, R.H., Bathoorn, E., Chlebowicz, M.A., Couto, N., Ferdous, M., Garcia-Cobos, S., et al. (2017). Application of next generation sequencing in clinical microbiology and infection prevention. *Journal of Biotechnology*, vol. 243, pp. 16-24.

14. Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, vol. 7, 3. https://doi.org/10.1186/1471-2105-7-3.

15. Emerson, D., Agulto, L., Liu, H., & Liu, L. (2008). Identifying and characterizing bacteria in an era of genomics and proteomics. *Bioscience*, vol. 58 (10), pp. 925-936. https://doi.org/10.1641/B581006.

16. Fournier, P.-E., Dubourg, G., & Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Medicine*, vol. 6, pp. 114. https://doi.org/10.1186/s13073-014-0114-2.

17. Garzon, M. (2017). Personal Communication.

18. Garzon, M. H. & Bobba, K.C. (2012). A geometric approach to Gibbs energy landscapes and optimal DNA codeword design. *Lecture Notes in Computer Science*, vol. 7433, pp. 73–85. https://doi.org/10.1007/978-3-642-32208-2_6.

19. Garzon, M. H. & Mainali, S. (2017). Towards a universal genomic positioning system: phylogenetics and species identification. *Lecture Notes in Computer Science*, vol. 10209, pp. 469–479. https://doi.org/10.1007/978-3-319-56154-7_42.

20. Garzon, M. H. & Mainali, S. (2017). Towards reliable microarray analysis and design. *International Society for Computer Applications,* pp. 6.

21. Garzon, M. H. & Pham, D. T. (2018). Genomic Solutions to Hospital-Acquired Bacterial Infection Identification. In: Rojas, I. & Ortuño, F. (eds), Bioinformatics and Biomedical Engineering, *Lecture Notes in Computer Science*, vol. 10813 (1), pp. 486-497, Springer, Cham. https://doi.org/10.1007/978-3-319-78723-7_42.

22. Glaeser, S. P. & Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology*, vol. 38 (4), pp. 237-245. https://doi.org/10.1016/j.syapm.2015.03.007.

23. Hassoun, M. H. (1995). Fundamentals of artificial neural networks. *MIT Press*, Cambridge.

24. Janda, J. M. & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, vol. 45 (9), pp. 2761–2764. https://doi.org/10.1128/JCM.01228-07.

25. Khan, H. A., Baig, F. K., & Mehboob, R. (2017). Nosocomial infections: epidemiology, prevention, control and surveillance. *Asian Pacific Journal of Tropical Biomedicine*, vol. 7 (5), pp. 478-482. https://doi.org/10.1016/j.apjtb.2017.01.019.

26. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, vol. 43 (1), pp. 59-69. https://doi.org/10.1007/BF00337288.

27. Kolde, R. (2012). Pheatmaps: pretty heatmaps. *R Package Version*, vol. 61, pp. 1-7.

28. Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, vol. 28 (5), pp. 1-26.

29. Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, vol. 2 (3), pp. 18–22.

30. Magill, S. S., Edwards, J. R., Bamberg, W., Beldavs, Z. G., Dumyati, G., Kainer, M. A., et al. (2014). Multistate Point-Prevalence Survey of Health Care–Associated Infections. *The New England Journal of Medicine*, vol. 370 (13), pp. 1198–1208. https://doi.org/10.1056/NEJMoa1306801.

31. Nguyen, N-P., Warnow, T., Pop, M., & White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms and Microbiomes*, vol. 2, 16004. https://doi.org/10.1038/npjbiofilms.2016.4.

32. Nomura, F. (2015). Proteome-based bacterial identification using matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS): a revolutionary shift in clinical diagnostic microbiology. *Biochimica et Biophysica Acta (BBA) – Proteins and Proteomics*, vol. 1854 (6), pp. 528–537. https://doi.org/10.1016/j.bbapap.2014.10.022.

33. Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16 (1), pp. 236. https://doi.org/10.1186/s12864-015-1419-2.

34. Parizad, E. G., Parizad, E. G., & Valizadeh, A. (2016). The application of pulsed field gel electrophoresis in clinical studies. *Journal of Clinical and Diagnostic Research*, vol. 10 (1), DE01–DE04. https://doi.org/10.7860/JCDR/2016/15718.7043.

35. Phan, V. & Garzon, M. H. (2009). On codeword design in metric DNA spaces. *Natural Computing*, vol. 8 (3), pp. 571-588. https://doi.org/10.1007/s11047-008-9088-6.

36. Singhal, N., Kumar, M., Kanaujia, P. K., & Virdi, J. S. (2015). MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology*, vol. 6, pp. 791. https://doi.org/10.3389/fmicb.2015.00791.

37. Song, Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, vol. 27 (2), pp. 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.

38. Srinivasan, U., Ponnaluri, S., Villareal, L., Gillespie, B., Wen, A., Miles, A., et al. (2012) Gram stains: a resource for retrospective analysis of bacterial pathogens in clinical studies. *PLoS ONE*, vol. 7 (10): e42898. https://doi.org/10.1371/journal.pone.0042898.

39. Tran, A., Alby, K., Kerr, A., Jones, M., & Gilligan, P. H. (2015). Cost savings realized by implementation of routine microbiological identification by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Journal of Clinical Microbiology*, vol. 53 (8), pp. 2473–2479. https://doi:10.1128/JCM.00833-15.

40. Wehrens, R. & Buydens, L. M.C. (2007). Self- and super-organizing maps in R: the kohonen package. *Journal of Statistical Software*, vol. 21 (5), pp. 1-19. https://doi.org/10.18637/jss.v021.i05.

41. Zhou, Y., Shen, N., Hou, H., Lu, Y., Yu, J., Mao, L., & Sun, Z. (2017). Identification accuracy for matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) for clinical pathogenic bacteria and fungi diagnosis; meta-analysis. *Internal Journal of Clinical and Experimental Medicine*, vol. 10 (2), pp. 4057–4076. www.ijcem.com/files/ijcem0035141.pdf.

42. Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, vol. 18, pp. 186. https://doi.org/10.1186/s13059-017-1319-7.

43. Zimlichman, E., Henderson, D., Tamir, O., Franz, C., Song, P., Yamin, C.K., Keohane, C., Denham, C.R., & Bates, D.W. (2013). Health care–associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA Internal Medicine*, vol. 173 (22), pp. 2039–2046. https://doi:10.1001/jamainternmed.2013.9763.