University of Alberta

Development of Statistical Methods for Analysis of High-Dimensional Biological Data

by

Qiaozhi Li

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Epidemiology

Department of Public Health Sciences

Fall 2013
Edmonton, Alberta

Dedication

I would like to dedicate this thesis to my husband, Feng Dai, for offering me constant encouragement, understanding and support, to my daughter, Zihan Dai, for her lovely spirit bringing endless joy to my life, and to my parents, Miaolin Li and Guizhen Liu, for always being positive and supportive about my goals and ambitions.

Abstract

High-dimensional biological data have been increasingly made available for tackling complex health problems. As with any Big Data opportunities, this has led to methodological challenges for extracting relevant information from such data, particularly in settings where biologically-sensible and statistically-appropriate methodologies that are practical and effective in public health practice or healthcare delivery have not been established.

This thesis aims at developing statistical methods specifically for two heath problems with high-dimensional biological data: I) A logic-regression-based genetic biomarker discovery method for environmental health, identifying the source/host of *Escherichia coli* using its genomic data; and II) An image analysis method for automatic tuberculosis (TB) detection in resource-limited settings, where the modern TB detection methods are not employable, using high-throughput sputum-culture images.

My research has developed these methods that are aimed to be implemented in the respective fields to advance effectiveness of the public health practice.

Key words: high-dimensional biological data, automatic TB detection, image analysis, E. coli host specificity, logic regression

Acknowledgement

I would like to sincerely express my deepest gratitude to my thesis examining committee, Profs. Norman Neumann, Gian Jhangri, and Yutaka Yasui, for their expertise and support for this academic endeavor.

I would like to hereby thank my supervisor, Prof. Yutaka Yasui, for his patient guidance, invaluable advice and inspiring encouragement throughout my time in the program. I have been extremely lucky to have him as my supervisor and a mentor, who makes the best effort to pass his rigorous academic knowledge to students, who cares not only about our research work but also our academic development and future career.

The work in Chapter 2 of this thesis, the biomarker discovery study on *E. coli* host-specificity, was in collaboration with Prof. Neumann and his PhD. student Zhi Shuai. I would like to thank them for kindly providing the data and for their valuable expert insights on the methodology development. I really enjoyed my time working with Dr. Neumann and Shuai.

The work in Chapter 3 of this thesis, an image-analysis methodology development for automatic tuberculosis detection in resource-limited settings, was in collaboration with TB/HIV Research Fundation, Chiang Rai, Thailand. I would like to thank my colleagues in Thailand, Mr. Boonchai Chaiyasirinroje and Miss Oranuch Nampaisan, for their great effort in preparing the data and for their unselfish help throughout the study. This work was funded by Alberta Ingenuity Center for Machine Learning. I would like to thank them for their kind support.

I would also like to acknowledge my colleagues in Dr. Yasui's team for their suggestions and comments on my work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CUP - carbon utilization 4
E. coli - *Escherichia coli* 3
FM - Fitch-Margoliash 7
LTB - latent TB 28
LTIIa - liable toxin IIA 6
M.tuberculosis - *Mycobacterium tuberculosis* 28
MEA - minimum evolution algorithms 7
MGIT - Mycobacteria Growth Indicator Tube 29
ML - maximum likelihood 7
MLST - multilocus sequence typing 4
MODS - microscopic-observation drug-susceptibility 29
MP - maximum parsimony 7
MSST - multi-spacer sequence typing 4
MST - microbial source tracking 4
NAAT - nucleic acid amplification test 28
NJ - neighbour joining 7
PCR - polymerase chain reaction 13
PFGE - pulsed field gel electrophoresis 4
PNB - p-nitrobenzoic acid 32
SNP - single-nucleotide polymorphism 11
STII - stable toxin gene II 6
TB - tuberculosis 28
THRF - TB/HIV Research Foundation 30
TSB - Tryptic Soy Broth 12
UPGMA - unweighted pair group method using arithmetic average 7

1 Introduction

1.1 Rationale

Data collection and data analysis have experienced tremendous evolution in the 21st century as a result of accelerating development and involvement of advanced technology [1]. As a result, high-dimensional data, wherein each observation is accompanied by thousands of measurements, have been becoming rapidly prevalent and easier/cheaper to obtain. Specifically, the application of advanced data-acquisition technologies in biology such as high-throughput imaging and high-throughput genomic and proteomic technologies have made high-dimensional biological data increasingly available for health research [1,2. 3].

The introduction of high-dimensional biological data in health research has significant impact on tackling complex health problems providing rich information of study observations. However, as with any Big Data opportunities, this has also introduced the challenges of working with high-dimensional data in health research. For example, it has led to methodological challenges for statistically-appropriately extracting relevant information from such data [1,2, 3]. High-dimensional data analysis requires new concepts and proper techniques. This is particularly crucial in settings where biologically-sensible and statistically-appropriate methodologies that are practical and effective in public health practice or health care delivery have not been established.

1.2 Purpose

The purpose of this thesis is to develop methodologies specifically for two health problems with high-dimensional data in settings where biologically-sensible and statistically-appropriate methodologies have not been established.

The first problem is on a genetic biomarker discovery for contamination source tracking in environmental health. The major question is to identify the source/host of *E. coli* collected from the contamination site using the microbe's high-dimensional genomic data. In addition to the high-dimensional issue, this problem is difficult to tackle due to complex hypotheses of *E. coli* host selection. This thesis will develop a logic-regression-based statistical methodology for analyzing the high-dimensional genomic data of *E. coli* in order to study *E. coli*'s evolution toward host selection, which involves two hypotheses of *E. coli*'s host-specificity and host-generality.

For the second problem, the thesis aims to develop an image-analysis-based statistical methodology for automatic TB detection in resource-limited settings using high-throughput microscopic sputum-culture images. The difficulty of this problem lies mainly in the quantitative characterization of culture-growth from the noisy high-dimensional culture images. In this thesis, a fast and robust methodology will be developed to process the high-dimensional image data and to form an automatic computer-based TB detection in resource-limited settings.

1.3 Thesis organization

In this chapter, the rational and purpose of this thesis were introduced. The two problems related to high-dimensional biological data that are mentioned in Section 1.2 will be presented in the following two chapters. In Chapter 2, the logic-regression-based analysis of *E. coli* genome assessing its host-specificity will be presented. An image analysis automatic TB detection for the Automated MODS in resource-limited settings will be presented in Chapter 3, followed by a discussion section in Chapter 4.

Bibliography

1. D. L. Donoho (2000). High-dimensional data analysis: the curses and blessings of dimensionality. Lecture on August 8, 2000, to the Americal Mathematical Society "Math Challenges of the 21st Century". Available form http://www-stat.stanford.edu~donoho/.
2. R. Clarke, H. W. Ressom, A. Wang, et al. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer, 8(1): 37-49.
3. J. Quackenbush (2007). Extracting biology from high-dimensional biological data. The Journal of Experimental Biology, 210: 1507-1517.

2 Logic-regression-based Analysis of *E. coli* Genome Assessing its Host-specificity

2.1 Background

This section is organized as follows. First, a brief review of *E. coli* and evidence for host specificity is given, followed by a review of tools for investigating host specificity of *E. coli*. The focus of this thesis is on statistical methods, in particular, methods for evaluating DNA sequence-based information. Two types of statistical methods are compared: supervised and unsupervised learning methods, and the reason for using *supervised learning*, as opposed to unsupervised learning, for this thesis is explained. Finally, the challenges associated with considering two hypotheses on *E. coli* host selection, namely, host-specificity and host-generality, are discussed, which motivate the use of a particular supervised learning method, logic regression, in this thesis.

2.1.1 Review of *E. coli* and evidence for host specificity of *E. coli*

Escherichia coli (*E. coli*), a Gram negative and facultative anaerobe, is widely distributed in the intestine of human and warm-blooded animals and is one of the best-studied model microorganisms since its discovery in 1885. Most *E. coli* living in the gastrointestinal tract of humans and animals are commensal strains [1]. Only a small proportion of *E. coli* strains are pathogenic and which can cause intestinal or extra-intestinal disease. Six well-characterized pathovars of *E. coli* have been described: enteroinvasive, enteropathogenic, enterohaemorrhagic, enterotoxigenic, enteroaggregative, and uropathogenic [1].

Host preference and/or specificity are not uncommon in microbial world [2-5]. For example, different mycorrhizal fungal species were found to be preferably associated with different genotypes of the orchid *Corallorhiza maculate* [3]. Some *Mycoplasma* species isolated from different bird species were found to be serologically distinct [6]. Evidence for some degree of host adaptation and preference of *E. coli* has also been observed. First, phylogenetic analysis of the *E. coli* strains using several typing methods have demonstrated that this species consists of four major phylogenetic groups (A, B1, B2, and D) [7]. The four phylogenetic groups differ in their ecological niches. It was observed that group A (40.5%) and B2 (25.5%) are more frequently isolated from humans while group B1 (41%) was most prevalent in animals [8]. The distribution of *E. coli* phylogenetic groups in different hosts indicates that *E. coli* may display a certain level of host preference. Second, it was found that avian septicemia *E. coli* strains were more virulent to chicks than *E. coli* strains isolated from newborn human meningitis cases [9]. A human-specific *E. coli* clone of the B2 lineage was recently reported by Clermont [10]. In a study by Kim *et al.* [11], it was found that, based on pulsed field gel electrophoresis (PFGE) patterns, most human isolates (81.8%) can be typed into one group, and most bovine strains also clustered together (78%). Therefore, *E. coli* populations may be hypothesized to adapt and are selected for in the gastrointestinal tract of specific hosts.

2.1.2 Review of tools for investigating host specificity of *E. coli*

Several tools have been used to identify host specific patterns in *E. coli*, with the goal of identifying the host sources of fecal contamination in order to track and control human and animal fecal inputs in the environment. Collectively, these tools are often referred to as microbial source tracking (MST) tools. Among these tools are multi-spacer sequence typing (MSST), PFGE, ribotyping, multilocus sequence typing (MLST), carbon utilization profiling (CUP), and host-specific genetic marker identification.

PFGE is a DNA fingerprinting method that uses rare cutting restriction enzymes to cleave bacterial genomic DNA (10 to 800kb in length), which is then electrophoresed under alternating electric currents to produce DNA fingerprints for each isolate. It was demonstrated that using PFGE, 89% of *E. coli* strains can be correctly assigned to its host source, though the specificity was only 50% [12]. Conversely, another study demonstrated little or no association between PFGE profiles and host sources [13]. PFGE is renowned for its high sensitivity of detecting small genetic differences due to the specific and rare enzymatic cutting sites of the restriction enzyme on the bacterial genome. However, this high sensitivity to small genetic differences may be a disadvantage in discriminating host sources of the bacteria from diverse origins ecologically and geographically. This defect is further exacerbated by the great plasticity of *E. coli* genomes [14].

Ribotyping is also a DNA fingerprint method that uses restriction endonucleases to cut bacterial genomic DNA, followed by electrophoresis of the fragments on an agarose gel, and hybridization of the rRNA gene fragments with oligonucleotide probes. Ribotyping has been proven to be effective for microbial source tracking in multiple studies[15,16]. In a study by Carson *et al.* [15], ribotyping was used to classify *E. coli* isolates from eight known animal sources, the average rate of correct classification was 73.6%. However, in other studies ribotyping performed less effectively[12,17]. In one study, only 27% of the indicator strains were correctly assigned to their sources [17]. The limited diversity on ribosomal DNA due to its conservativeness in evolution may decrease its discrimination power. Moreover, as a restriction based DNA fingerprint method, only a portion of DNA sequence information can be utilized, which further challenge ribotying's ability in subtyping the diverse *E. coli* population.

MLST is a DNA-sequence-based molecular typing method, in which the sequences from several genes (usually housekeeping genes) are compared for genetic variations to classify strains, identify clonal groups and determine phylogenetic relationship. Several studies found that MLST had high levels of discriminatory power [18,19], while in others, MLST performed poorly. For instance, MLST showed the least discriminatory power in an evaluation study which used PFGE, rep-PCR and MLST to type *E. coli* O157:H7 isolates from cattle, food, and infected humans [20]. In another study, Adiri *et al* [21] used MLST to study *E. coli* O78 strains from human, avian and cattle and no host specificity distribution was observed. DNA sequences in MLST are derived from housekeeping genes, which are highly conserved: therefore, it may offer limited sequence diversity in comparison to other genetic regions such as intergenic DNA sequence (discussed later in this section).

CUP, another phenotypic method, is based on differences among bacteria in their ability to use a wide range of carbon and nitrogen sources for energy and growth from the diet of their host animals. Uzoigwe *et al.* [22] demonstrated that on average 89.5% of *E. coli* isolates can be correctly classified into host sources by CUP. However, in another CUP based study, although the average rate of correct classification was 73%; the false positive rate was as high as 66% [23]. When *E. coli* isolates from more diverse hosts were analyzed, no host specific pattern was found [24]. Phenotypic traits of bacteria can vary as the environment changes. The environmental differences between culturing conditions and the bacteria's original niches, as well as complexity of the genetic diversity of *E. coli*, challenge CUP's discriminatory power on host specificity.

Host-specific toxin genes in *E. coli* have also been identified and have the potential to be used as MST markers [25-27]. In a study by Khatib *et al.* [26], heat liable toxin IIA (LTIIa) gene from enterotoxigenic *E. coli* was used in a cattle-specific PCR assay and 87% of environmental samples from cattle waste and lagoons were LTIIa positive. In another study, a pig-specific heat stable toxin gene II (STII) was identified by Khatib *et al.*[28], and was useful in distinguishing *E. coli* isolates between swine waste and other animal sources. Unfortunately, the fact that some organisms do not have these toxin markers hinders the application of these methods for identifying host sources of fecal contamination. Moreover, horizontal gene transfer enables toxin genes to be transferred among the microbes from different host sources making it difficult to attribute host sources of pollution to these genetic markers.

Although evidence supports the emerging concepts of host-specificity within *E. coli* population, it is still generally believed that no perfect technique or method has been found to confirm or refute the concept. To study the relationships between *E. coli* and their host source, a new approach was used in this thesis, and based on MSST in combination with novel statistical tools applied during the analysis of DNA sequence data. MSST is a DNA sequencing based method targeting intergenic spacer regions in the genome. Intergenic regions, often containing promoter and enhancer elements, regulate the expression of genes and therefore relate to changes in cell phenotypes, functions, and sensation [29-32]. Intergenic regions are under less stringent selection pressure; they carry more genetic variations that can be used for characterization when compared to gene coding regions. Additionally, intergenic sequences recruit factors coded on other loci of the genome to initiate and regulate transcription, so they may be more informative of general metabolic sensitivity to environmental conditions (i..e, the varying physiological conditions of different animal gastrointestinal systems).

Therefore, we believe intergenic regions represent a unique target for assessing DNA sequence polymorphisms associated with host specificity.

Unlike host specific genetic marker methods, which can infer host source directly, MSST is a DNA-sequence-alignment-based method, which needs statistical classification to infer host sources. Therefore, selection of statistical methods is also critical step in applying MSST.

2.1.3 Review of statistical methods of DNA phylogenetic sequence snalysis

The commonly used methods for clustering fingerprints or DNA sequences alignments and constructing phylogenetic trees include unweighted pair group method using arithmetic average (UPGMA), neighbour joining (NJ), Fitch-Margoliash (FM), minimum evolution algorithms (MEA), maximum parsimony (MP), and maximum likelihood (ML) [33]. UPGMA [34] and NJ [35] are clustering-type methods based on the pairwise similarities of samples computed on the basis of sequence alignment. This type of methods construct a tree to reflect the structure presented in the pairwise similarities of samples, starting from the most similar sequence pairs followed by step-wise adding one sample to the tree. FM [36] and MEA [37] are also based on the pairwise similarities of samples comparing many alternative tree topologies and selecting one that has the best fit between estimated distances in the tree and the actual evolutionary distances. The major drawback of these methods is that the actual sequence information is lost when all the sequence variation is reduced to pairwise similarities [33]. Another type of phylogenetic-tree-construction methods is based directly on the sequence characters rather than on the pairwise similarities of samples, called character-based methods. Maximum parsimony [38] and ML [39] methods are the two most popular character-based methods. They count mutational events accumulated on the sequences and study evolutionary dynamics of each character. Maximum parsimony chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths. It works by searching for all possible tree topologies and reconstructing ancestral sequences that require the minimum number of changes to evolve to the current sequences. Maximum likelihood uses probabilistic models to choose the best tree that has the highest probability or likelihood of reproducing the observed data.

This method searches every possible tree topology and considers every position in an alignment to find a tree that most likely reflects the actual evolutionary process. It works by calculating the probability of a given evolutionary path for a particular extant sequence, i.e., calculates the total probability of ancestral sequences evolving to internal nodes and eventually to existing sequences, where the probability values are determined by a substitution model.

2.1.4 Two types of classification methods

The clustering methods such as UPGMA, NJ, FM, MEA, MP and ML listed above, all belong to a type of classification methods called *unsupervised learning*. Unsupervised learning first discovers groups of patterns in data and then classifies samples into these groups: the group labels are not observed as data [40]. Unsupervised learning is useful for finding hidden structure in unlabeled data. The other type of classification methods, often contrasted to unsupervised learning, is *supervised learning*. It is used to infer patterns in observed data associated with observed group labels [40]. In supervised learning, each sample has an input set of data (typically a vector) and a label of the grouping. For data with known group labels, supervised learning methods are more powerful for finding the patterns/structures that is related to the group labels. In other words, the two types of classification methods, supervised learning and unsupervised learning, have some clear distinctions, these being: 1) the former aims to infer a classification function that can be used to classify new samples into the groups that are observed and known; while the latter aims to classify the samples into groups where the groups and their labels are not observed; and 2) the former requires labeled data, while the latter uses unlabeled data. In the microbial-source-tracking context, UPGMA, NJ, FM, MEA, MP and ML work similarly and they are all unsupervised learning methods.

2.1.5 Use of supervised learning

In this thesis, it is hypothesize that host-specificity and host-generality are present simultaneously in *E. coli* populations within any given host and consequently the goal of supervised learning methods are to try to find specific patterns of *E. coli* genes that are consistent with the host-specificity. We attempt to achieve this using a supervised learning method since the *E. coli* can be labeled with respect to their host sources (i.e., group labels are based on the isolation of *E. coli* from fecal samples collected from specific hosts). This is a distinguishing feature from the methods described above that used unsupervised learning.

There are multiple reasons for using a supervised learning method for this problem. First, the supervised learning will identify the most informative sites and patterns in *E. coli* DNA sequence pertaining to their hosts, and patterns formed by these sites can be used to infer the hosts of new samples; while the unsupervised learning will only cluster similar samples (or distinguish dissimilar samples) and the results will not necessarily inform which animal hosts are found within a set of new samples. Second, the group label of each sample, i.e., its host source, is provided as part of data. Unsupervised learning methods leave this useful information (i.e., host group labels) unutilized; Third, unsupervised learning uses the information of *all* single nucleotides in *E. coli* gene sequence alignments as a distinguishing characteristic between different hosts, but information in many single nucleotides are irrelevant to host-specificity. Using irrelevant information in *all* single nucleotides to infer the differences between *E. coli*'s hosts may be result in misleading, poor-performing classification. Conceptually, the genetic information on only a small number of genes, pieces of genes, or SNPs may be decisive for host-specificity. A supervised learning method can select the most relevant information on host specificity from the labeled data.

There are many supervised learning methods available for classification problems. Some of these are analytical methods that give simple, yet explicit analytical functions for classification: these have good interpretability. These analytical methods include logistic regression, log-binomial regression, and logic regression. This thesis uses logic regression the motivation of which will be discussed in 2.1.7 below. Other supervised learning methods are non-analytical and that give complex classification functions and which are usually not interpretable. These methods include Boosting [41], Artificial Neural Network [42], and Support Vector Machines [43]. An interpretable classification function is generally preferable for assessing a scientific hypothesis.

2.1.6 Challenges in assessing *E. coli* host specificity

As discussed in 2.1.1, host specificity in *E. coli* implies that some *E. coli* may specialize towards colonization of one host, while host generality states that some *E. coli* may specialize in the colonization of multiple hosts. The two hypotheses have to be tested quantitatively. Towards this end, robust methodologies for testing these hypotheses are critically needed. The methods described in 2.1.3 cannot be used to handle the two hypotheses simultaneously. Even if these methods are useful for evaluating host-specificity, under the two hypotheses, it is not surprising to see the *E. coli* samples from the same host may not always cluster together. This is because the host-generalist *E. coli* will be present across different hosts or spread throughout the constructed phylogenetic trees. In this thesis, we try to fill this gap by proposing a statistical methodology for testing and exploring the two hypotheses of host selection in *E. coli* simultaneously

A question relevant to *E. coli* host specificity might be: "Is this *E. coli* isolate from human or from other animal sources?" Under the hypothesis of host-specificity with MSST, this question can be rephrased as: "does the genetics of *E. coli* indicate whether its host is human or another animal?" Supervised learning is useful for addressing similar questions in a traditional classification problem setting, where *E. coli* genetics classifies the sample to be specific to either human or other animals.

In our study, however, if host-specificity and host-generality operate simultaneously in *E. coli* host selection, the question stated above becomes more complex. For an *E. coli* isolate collected from a host such as a bovine, we precisely know that the *E. coli* sample can colonize in that host. However, we are not sure whether the *E. coli* sample can colonize only in that host or whether it can colonize in a group of hosts. This presents a major analytic challenge. We address this challenge in separating the analysis into two stages. In the first stage, we will identify genetic patterns of host-specific *E. coli* by setting the classification specificity as close to 100% as possible: this will eliminate the likelihood that the genetic patterns discovered in the first stage represent host generalists. In the second stage, we will eliminate the samples that were host-specific and search host-generalist patterns in the remaining samples.

Specifically, we consider host-generality as the potential to colonize more than 1 host (e.g., human and bovine), and repeat the high-specificity supervised learning/classification for the specific host generality group.

Another challenge is the choice of the form of statistical association between *E. coli*'s genetic patterns and its host(s). *E coli*'s host selection is unlikely to be determined by a single single-nucleotide polymorphism (SNP): the host selection may involve multiple SNPs and genes. Thus, we also need to look into multiple SNPs/genes to find an *E. coli* genetic pattern associated with its host.

Moreover, interactions of multiple SNPs/genes, instead of a simple sum of individual SNP/gene effects, may be critical in determining the host-selection of *E. coli*. In a statistical model, SNP interaction is often expressed quantitatively as a multiplication of two or more SNP-genotype indicators: this is one form of epistasis [44]. Choosing biologically-plausible forms of SNP interaction or epistasis, which may be different from the standard form of interaction in statistical modeling, is important.

2.1.7 Logic regression

Logic regression is a supervised learning method that is used to model an outcome with Boolean combinations of potential predictors that are binary, such as the indicators of SNP genotypes [45]. It has been used in analyzing human genetics data for identifying genetic biomarkers that modify the risk of a phenotype. For example, it has been applied to the genome-wide association studies data to identify genetic markers for the risk of developing Crohn's Disease [46]. The rationale for method selection was motivated based on using a biologically sensible method, because the model uses specific forms of SNP-SNP interactions, i.e., SNP intersections and unions, which have biologically plausible interpretations [46]. An SNP intersection requires that all of the SNPs in a specified group of SNPs must take their respective high-risk genotypes in order to increase the disease risk. This form is similar to a group of (sequential) mutations that must accumulate before a cell transforms into cancerous in the multistage carcinogenesis theory. SNP union, on the other hand, allows disease risk to be elevated through multiple independent ways: this form captures genetic heterogeneity.

In this thesis, logic regression is used as a method of classification for distinguishing between *E. coli* samples from a certain host (or a certain group of hosts) and those from other hosts. The motivations for using logic regression include: it is a supervised learning method; it is an analytical supervised learning method which generates clear interpretable functions for classifications; and it incorporates biologically-plausible forms of SNP-SNP interactions in the classification function. In our context, SNP intersection captures a situation where two or more *E. coli* SNPs jointly influence biological functions related to host-selection, while SNP union captures a situation where two or more *E. coli* SNPs are redundant in their biological effects (i.e., genetics heterogeneity). In an SNP intersection, all of the relevant SNPs in the intersection set must take their respective specific genotypes for *E. coli* to live in a certain host, where one, or a subset, of the set is insufficient. On the other hand, in an SNP-SNP union, any SNP in the union set taking a specific genotype is sufficient for *E. coli* to live in a certain host.

2.2 Data and materials

The study uses 780 E. coli samples in total, which were samples with known sources, which includes human and 13 other animals, provided by Dr. Neumann (School of Public Health, University of Alberta). Human E. coli was obtained through the Provincial Laboratory for Public Health, Alberta, from stool samples submitted for clinical analysis. The animal E. coli samples were obtained from Dr. Ed Topp (Agriculture and Agri-Food Canada) and Dr. Tom Edge (Environment Canada). The host of each E. coli sample was labeled with its original source.

All samples have been verified as *E. coli* through biochemical analysis using a Vitek Bacterial Identification System (BioMerieux). In order to avoid evolutionary selection associated with culture-based growth condition, all *E. coli* samples were grown in Tryptic Soy Broth (TSB) only once and three pellets were collected for each isolate to maintain genetic stability. One pellet was used directly for DNA extraction and analysis and the other two will be archived at -80°C for future use.

Genomic DNA was extracted from *E. coli* TSB cultures using DNeasy Blood & Tissue kits (QIAGEN) according to the manufacturer's instructions. Three intergenic regions (agfD, ompF and flhDC) were amplified separately by polymerase chain reaction (PCR). All PCR products were sequenced bidirectionally by Macrogen Inc. (Korea) and The University of Calgary Genetic Analysis laboratory. All sequences were aligned using ClustalX 2. Each SNP's genotype is coded with binary indicators for A, G, C, T, and – (notation for mutation).

Three *E. coli* genes (agfD, ompF, and flhDC) were selected in the study based on previous research. The intergenic regions of the three genes were used.

2.3 Method

2.3.1 A logic-regression-based analytical approach

In this thesis, logic regression is used as a supervised learning method of classification for distinguishing between E. coli samples from a certain host (or a certain group of hosts) and those from other hosts. SNP intersections and unions are expressed mathematically as Boolean logics such as $(X1 \wedge X2) \vee X3^c$, where X's are indicators of SNP genotypes. "$\wedge$", "$\vee$" and "^{c}" represents intersection (AND), union (OR), and complement (NOT), respectively. The systematic part of the logic regression model applied here takes the form

$$\text{Logit } (E[Y]) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots + \beta_p L_p$$

where Y is a binary variable, an indicator for being from one E. coli host versus being from the other hosts, β_0, β_1,… β_p are the parameters, and L_1, L_2, …, L_p are Boolean combinations (called "trees") of indicators of SNP genotypes in the E. coli genes (called "leaves"). A massive number of potential models can be built with varying sizes, i.e. the number of trees and leaves. Thus, the model building requires great computing resources. A Simulated Annealing algorithm is used to select the trees and leaves adaptively based on deviance as the model fit measure for finding the model with the best. For further limiting computational burdens, the maximum size of the model is limited to two trees and 10 leaves.

Since a complex hypothesis of simultaneous existence of host specificity and host generality, a logic-regression-based two-step analytical approach is proposed for studying host specificity and host generality of *E. coli*.

In the data, each *E. coli* sample is labeled with the name of the host from which it was collected. Although each sample was collected from the labeled host, it may be able to live in other host(s) under the mechanism hypothesized by the host generality. Both of the hypothesized host-selection patterns, i.e. host specificity and host generality, need to be considered and tested simultaneously in the analysis.

Considering potential existence of host specificity and host generality in *E. coli* (Table 2.1), the proposed analytical approach consists of two steps. In the first step, the focus will be on the hypothesis of host specificity. In the second step, the focus will be on the hypothesis of host generality. In each step, logic regression model is used as a biomarker discovery tool to distinguish between *E. coli* samples from one host (or a group of hosts) and those from other hosts. The biomarkers to be identified are SNP interactions that are potentially important for the host-specific selection in *E. coli*.

Table 2. 1 Two groups of samples under the host specificity and host generality

	Host-specific Samples	**Host-general Samples**
Human	H1: Samples that can only live with Human	H2: Samples that can live with Human and some other hosts
Bovine	B1: Samples that can only live with Bovine	B2: Samples that can live with Bovine and some other hosts
Cat	C1: Samples that can only live with Cat	C2: Samples that can live with Cat and some other hosts
Dog	D1: Samples that can only	D2: Samples that can live with Dog

	live with Dog	and some other hosts
Deer	E1: Samples that can only live with Deer	E2: Samples that can live with Deer and some other hosts
Goose	G1: Samples that can only live with Goose	G2: Samples that can live with Goose and some other hosts
Chicken	K1: Samples that can only live with Chicken	K2: Samples that can live with Chicken and some other hosts
Moose	L1: Samples that can only live with Moose	L2: Samples that can live with Moose and some other hosts
Muskrat	M1: Samples that can only live with Muskrat	M2: Samples that can live with Muskrat and some other hosts
Horse	O1: Samples that can only live with Horse	O2: Samples that can live with Horse and some other hosts
Pig	P1: Samples that can only live with Pig	P2: Samples that can live with Pig and some other hosts
Coyotes	Q1: Samples that can only live with Coyotes	Q2: Samples that can live with Coyotes and some other hosts
Gull	S1: Samples that can only live with Gull	S2: Samples that can live with Gull and some other hosts
Beaver	V1: Samples that can only live with Beaver	V2: Samples that can live with Beaver and some other hosts
Sheep	Y1: Samples that can only live with Sheep	Y2: Samples that can live with Sheep and some other hosts

a. Step One

In the first step, a logic regression model is built for each host distinguishing between the samples from the host under analysis and those from other hosts. In each of the analyses, 0/1 binary outcome is set up for each sample according to its host label and the host under analysis. For host X, the outcome would be 1="sample that is collected from the host X" and 0="sample that is collected from other hosts". In the analysis for host “Human”, specifically, *E. coli* samples in the set of H1 or H2 in Table 2.1 have outcome 1, and all other samples have outcome 0.

In the following, it will be demonstrated that the SNP patterns identified in Step One with high specificity are host-specific. Assuming that host specificity and host generality exist simultaneously in *E. coli* host selection, it is reasonable to suppose that the host-specific *E. coli* samples (H1, B1, etc.) have their host-specific SNP pattern (Human-specific, Bovine-specific, etc.), and among the host-general *E. coli* samples (H2, B2 etc.), those can colonize in a particular group of multiple hosts have a host-group-specific SNP pattern. For example, samples in H1 share a Human-specific SNP pattern and samples in B1 share a Bovine-specific SNP pattern. Samples in H2 and B2 that can colonize in, for example, both human and Bovine but not in other hosts share a Human-Bovine-specific SNP pattern. Samples in H2, B2 and C2 that can colonize in Human, Bovine and Cat but not in other hosts share a Human-Bovine-Cat-specific SNP pattern, and so on.

The SNP pattern identified by comparing samples from one host (e.g., H1 and H2) with the rest in Step One could be a SNP pattern that is either host-specific (corresponding to the host under analysis) or host-group-specific (corresponding to a particular group of hosts involving the host under analysis). However, a SNP pattern from Step One with very high specificity (close to 100%) could only be a host-specific pattern rather than a host-group-specific pattern. This is because if a host-group-specific pattern is identified, *E. coli* samples collected from other hosts in the group rather than the host under analysis can also have the identified host-group-specific pattern. This would make its specificity lower than 100%.

After identifying the SNP patterns for host-specific *E. coli* samples, the possibility of host generality involving human *E. coli* will be explored in the next step.

b. Step Two

Under host generality, host-general *E. coli* samples can transmit and colonize in multiple hosts. Host generality involving human *E. coli* is of particular interest for contamination source-tracking. In the Step-Two analysis, the focus will be on exploring possible groups of multiple hosts including human within which host-general *E. coli* sharing specific SNP patterns can transmit and colonize. Unlike the analyses in Step One, It is not known exactly which hosts could form such a host group and which *E. coli* samples belong to a particular host group. In this step, a set of possible host groups that contain human and only one animal host, such as Human-Bovine, Human-Cat, will be explored. This is an attempt to address host-generality with a limited scope, i.e., host-generality involving only human and one animal.

Specifically, the *E. coli* samples collected from each host are partitioned into two sets, a host-specific set and a host-general set, based on the host-specific SNP patterns identified in Step One (Table 2.2). For each host, the host-specific set includes the *E. coli* samples that are classified as specific to that host by the Step-One model (H1', B1', etc.); the host-general set includes the *E. coli* samples that are classified as not specific to that host (H2', B2', etc.). Based on this partition, a potential host group for host generality is defined as the union of two host-general sets from human and one animal host, such as Human-Bovine group including all samples in H2' and B2'.

Table 2. 2 Partition of *E. coli* samples by the Step One classification into host-specific and host-general sets

	Classified as Host-specific in Step One	**Classified as Host-general in Step One**
Human	H1': Samples that are classified as specific to Human	H2': Samples that are classified as not specific to Human
Bovine	B1': Samples that are classified as specific to Bovine	B2': Samples that are classified as not specific to Bovine
Cat	C1': Samples that are classified as specific to Cat	C2': Samples that are classified as not specific to Cat
Dog	D1': Samples that are classified as specific to Dog	D2': Samples that are classified as not specific to Cat
Deer	E1': Samples that are classified as specific to Dog	E2': Samples that are classified as not specific to Cat

Goose	G1’: Samples that are classified as specific to Goose	G2’: Samples that are classified as not specific to Goose
Chicken	K1’: Samples that are classified as specific to Chicken	K2’: Samples that are classified as not specific to Chicken
Moose	L1’: Samples that are classified as specific to Moose	L2’: Samples that are classified as not specific to Moose
Muskrat	M1’: Samples that are classified as specific to Muskrat	M2’: Samples that are classified as not specific to Muskrat
Horse	O1’: Samples that are classified as specific to Horse	O2’: Samples that are classified as not specific to Horse
Pig	P1’: Samples that are classified as specific to Pig	P2’: Samples that are classified as not specific to Pig
Coyotes	Q1’: Samples that are classified as specific to Coyotes	Q2’: Samples that are classified as not specific to Coyotes
Gull	S1’: Samples that are classified as specific to Gull	S2’: Samples that are classified as not specific to Gull
Beaver	V1’: Samples that are classified as specific to Beaver	V2’: Samples that are classified as not specific to Beaver
Sheep	Y1’: Samples that are classified as specific to Sheep	Y2’: Samples that are classified as not specific to Sheep

Then a logic regression model is build for each of the potential host groups that include Human. In each of these analyses, the 0/1 binary outcomes are 1=“*E. coli* sample that is in the potential host group” and 0=“*E. coli* sample that is not in the potential host group”. For example, for the analysis of Human-Bovine group, *E. coli* samples in set H2’ and B2’ have outcome 1, and all other *E. coli* samples have outcome 0.

Similar to the logic in Step One, it can be demonstrated that a SNP pattern identified by logic regression in the Step-Two analysis with very high specificity (close to 100%) is host-group specific, i.e. it is specific to *E. coli* samples that can transmit and colonize in the group of hosts but not in other hosts. The rational is as follows. Under the diagram of Table 2.1, a SNP pattern identified by the logic regression in Step-Two for a host group (such as Human-Bovine) could be specific to 1) “the host group under analysis” (such as Human-Bovine), 2) “one of the host in the group and some other hosts out of the group” (such as Human-Cat, or Bovine-Dog), and 3) “all hosts in the group plus some other hosts out of the group” (such as Human-Bovine-Cat). If the identified SNP pattern is not host-group-specific, i.e. it is specific to 2) or 3), then the *E. coli* samples that are collected from other hosts out of the group (Cat or Dog in 2); and Cat in 3) will make the specificity of the identified SNP pattern lower than 100%.

The SNP patterns for all possible host groups involving Human and one other animal host are compared and discussed to find which host groups are likely to exist.

2.3.2 Five-fold cross validation for evaluating the performance of the identified SNP patterns

The logic-regression-based model building discussed above is a supervised-learning approach, which fits a model using a set of observations and their known outcomes. This approach may raise a common issue, for all supervised-learning methods, called model overfitting. Cross-validation is a commonly used approach to (1) prevent model overfitting and/or (2) evaluate the performance of the model unbiasedly. It involves randomly partitioning a set of data into mutually exclusive subsets and fitting and testing a model with different subsets.

In this thesis, a 5-fold cross-validation approach was applied for evaluating the performance of the identified SNP patterns that is not inflated by overfitting. The training and testing of a model are performed in 5 rounds after the *E. coli* samples are randomly partitioned into five mutually-exclusive groups S1, S2, …, S5. For the *i*-th round (i=1, …, 5), all other groups except the *i*-th group are used as training data to fit a logic regression model. Then the *i*-th group is used as a test data to assess the performance of the model in this round. The final validation results of sensitivity and specificity are averaged over all the 5 rounds.

2.4 Results

This analysis includes *E. coli* samples collected from 15 different sources (Table 2.3). Of the total 780 isolates, 120 (15.4%) were bovine *E. coli* isolates and 105 (13.5%) were human isolates, and the other 555 (71.1%) were collected from 14 other animal species (Table 2.3).

Table 2. 3 Cross-validated results of Step One

E coli Source	Number of Samples	Sensitivity	Specificity
Bovine	120	0.29	0.97
Cat	21	0.04	0.99
Dog	61	0.21	0.98
Deer	48	0.82	0.98
Goose	54	0.05	0.99
Human	105	0.53	0.98
Chicken	59	0.54	0.99
Moose	14	0.67	0.99
Muskrat	56	0.77	0.99
Horse	44	0.36	1.00
Pig	49	0.44	0.99
Coyotes	44	0.61	0.99
Gull	18	0.05	0.99
Beaver	40	0.35	1.00
Sheep	47	0.46	0.99

2.4.1 Step one

In the first step of the analyses, a very strong association was found between a certain *E. coli* SNP pattern, represented by a logic regression model, and the human *E. coli* samples (Table 2.3). Distinguishing between human *E. coli* samples and other *E. coli* samples, logic regression obtained 54% sensitivity and 99% specificity evaluated by a 5-fold cross validation taking the overfitting problem into account. The SNP pattern identified by this logic regression model is Human-specific having a very high specificity. Host-specific patterns were also found for dog, deer, chicken, moose, muskrat, coyotes, and sheep with a moderately high sensitivity and a high specificity that is close to 100%.

2.4.2 Step two

Associations were found in Step-Two analyses between some specific SNP patterns, identified by logic regression, and the *E. coli* samples from some potential host groups involving Human (Table 2.4), such as the union of H2’ and B2’, and the union of H2’ and E2’ (Table 2.2). The identified SNP patterns were able to distinguish between *E. coli* samples from the potential host groups and other *E. coli* samples with high specificity. The host-group-specific SNP patterns that showed good performance were those for the host groups of Human-Bovine and Human-Deer. Specifically, for groups of Human-Bovine and Human-Deer, the sensitivities were 41% and 42%, respectively, with specificities 94% and 100%, respectively.

Table 2. 4 Analysis results for exploring host generality of *E. coli*

Host Group (Involving Human)	**Cases and Controls * (Based on Step 1)**		**Sensitivity and Specificity (Based on Step 2)**	
	# of Cases	**# of Controls**	**Sensitivity**	**Specificity**
Bovine	110	670	0.41	0.94
Cat	36	744	0.25	0.99
Dog	41	739	0	1.00
Deer	19	761	0.42	1.00
Goose	47	733	0.13	1.00
Chicken	30	750	0	1.00
Moose	14	766	0.07	1.00
Muskrat	23	757	0	1.00
Horse	35	745	0	1.00
Pig	38	742	0.37	1.00
Coyotes	25	755	0.28	1.00
Gull	19	761	0.16	1.00
Beaver	34	746	0	1.00
Sheep	25	755	0.16	1.00

* Cases are samples that are in the host group; controls are the samples that are not in the host group

2.5 Discussion

Subject to the limitations we describe below, the two-step approach proposed in this thesis nonetheless provides a possibility of being applied to the analysis of other microbial genetic data involving host-specificity and host-generality simultaneously. It opens a door for exploring the complex patterns of microbial host-selection that involves multiple hypotheses.

The logic-regression-based two-step data-analytical method proposed in this thesis can be useful for identifying important SNP patterns of *E. coli*'s host-selection. The question tackled in Section 2.1.6, namely, "does the genetics of *E. coli* indicate whether its host is human or other animals?" could be answered with the study findings in the Step-One analyses under simultaneous existence of host-specificity and host-generality. The identified host-specific SNP patterns may have important implications in contamination source tracking using *E. coli* samples collected from an environmental sample (i.e., food or water).

The host-group-specific SNP patterns identified in the Step-Two analysis for host groups involving human could be applied in studying human risk related to zoonotic *E. coli*. These identified SNP patterns can be used as biomarkers to identify *E. coli* samples that could transmit between human and other animals. The analytical methods mat be potentially useful for identifying the source of the pathogenic *E. coli*, which is especially significant in an *E. coli* outbreak.

On the other hand, there are important limitations in our study. First, the identified host-specific SNP patterns may not be sufficiently robust in practice. Due to the high adaptability and short-term genomic evolution of *E. coli*, the *E. coli* samples collected from a contamination site may have significantly different genetic patterns compared to its original host-specific ancestors. This raises a concern that the identified SNP patterns may be specific to the environment of the contamination site from which *E. coli* samples were collected as well as their host. Thus, analyzing a small set of samples will fail to identify the patterns specific to an *E. coli*'s original host that is robust against the variety of environments of the contaminations sites and other factors unrelated to the host. In future research, the robustness of the findings needs to be assessed and it is critical to incorporate a large variety of samples from the same host in this regard.

The second limitation is that the identified SNP patterns, including those for host-specificity in Step One and those for host-generality for Step Two, cannot classify all the samples we had into certain host-specific or host-generality groups. While our findings had appreciable proportions classified into these groups, it leaves a question as to whether a more broader host-generality analysis could classify the remaining samples or not.

The logic regression model used in this thesis has a fixed tree size, which may not be appropriate for all the analyses. The logic regression uses a Simulated Annealing Algorithm to conduct a massive search in the high-dimensional space to fit a model. This may not always reach a globally optimal result, and the result may depend on the random seed used for originating the stochastic search.

Bibliography

1. T. V. Nguyen, P. Le Van, C. Le Huy, et al. (2005). Detection and characterization of diarrheagenic escherichia coli from young children in hanoi, vietnam. Journal of Clinical Microbiology, 43(2):755-760.
2. L. Xiao, R. Fayer, U. Ryan, et al. (2004). Cryptosporidium taxonomy: Recent advances and implications for public health. Clinical Microbiology Reviews, 17(1):72-97.
3. D. L. Taylor, T. D. Bruns, S. A. Hodges (2004). Evidence for mycorrhizal races in a cheating orchid. Proceedings of the Royal Society B: Biological Sciences, 271(1534):35-43.
4. M. J. Mandel, M. S. Wollenberg, E. V. Stabb, et al. (2009). A single regulatory gene is sufficient to alter bacterial host range. Nature, 458(7235):215-U7.
5. M. Fauvart, J. Michiels (2008). Rhizobial secreted proteins as determinants of host specificity in the rhizobium–legume symbiosis. FEMS Microbiology Letters, 285(1):1-9.
6. O. Tenaillon, D. Skurnik, B. Picard, E. Denamur (2010). The population genetics of commensal escherichia coli. Nature Reviews Mcrobiology, 8(3):207-217.
7. J. B. Poveda, J. Giebel, J. Flossdorf, et al. (1994). Mycoplasma buteonis sp. nov., mycoplasma falconis sp. nov., and mycoplasma gypis sp. nov., three species from

birds of prey. International Journal of Systematic and Evolutionary Microbiology, 44(1):94.

8. R. R. Chaudhuri, I. R. Henderson (2012). The evolution of the escherichia coli phylogeny. Infection, Genetics and Evolution, 12(2):214-226.
9. Z. R. Eliora (2006). Host specificity of septicemic escherichia coli: human and avian pathogens. Current Opinion in Microbiology, 9(1):28-32.
10. O. Clermont, M. Lescat, C. L. O'Brien, et al. (2008). Evidence for a human-specific escherichia coli clone. Environmental Microbiology,10(4):1000-1006.
11. J. Kim, J. Nietfeldt, A. K. Benson (1999). Octamer-based genome scanning distinguishes a unique subpopulation of escherichia coli O157:H7 strains in cattle. Proceedings of the National Academy of Sciences of the United States of America, 96(23):13288-13293.
12. S. P. Myoda, C. A. Carson, J. J. Fuhrmann, et al. (2003). Comparison of genotypic-based microbial source tracking methods requiring a host origin database. Journal of Water and Health, 1(4):167-180.
13. S. Parveen, N. C. Hodge, R. E. Stall, et al. (2001). Phenotypic and genotypic characterization of human and nonhuman escherichia coli. Water Research, 35(2):379-386.
14. O. Lukjancenko, T. M. Wassenaar, D. W. Ussery. Comparison of 61 sequenced escherichia coli genomes. Microbial Ecology, 60(4):708-720.
15. C. A. Carson, B. L. Shear, M. R. Ellersieck, A. Asfaw (2001). Identification of fecal escherichia coli from humans and animals by ribotyping. Applied and Environmental Microbiology, 67(4):1503-1507.
16. T. M. Scott, S. Parveen, K. M. Portier, et al. (2003). Geographical variation in ribotype profiles of escherichia coli isolates from humans, swine, poultry, beef, and dairy cattle in florida. Applied and Environmental Microbiology, 69(2):1089-1092.
17. D. F. Moore, V. J. Harwood, D. M. Ferguson, et al. (2005). Evaluation of antibiotic resistance analysis and ribotyping for identification of faecal pollution sources in an urban watershed. Journal of Applied Microbiology, 99(3):618-628.
18. S. J. Peacock, G. D. de Silva, A. Justice, et al. (2002). Comparison of multilocus sequence typing and pulsed-field gel electrophoresis as tools for typing staphylococcus aureus isolates in a microepidemiological setting. Journal of Clinical Microbiology, 40(10):3764-3770.

19. S. R. Nallapareddy, R. W. Duh, K. V. Singh, B. E. Murray (2002). Molecular typing of selected enterococcus faecalis isolates: Pilot study using multilocus sequence typing and pulsed-field gel electrophoresis. Journal of Clinical Microbiology, 40(3):868-876.

20. S. L. Foley, P. F. McDermott, S. Zhao, et al. (2004). Evaluation of molecular typing methods for escherichia coli O157:H7 isolates from cattle, food, and humans. Journal of Food Protection, 67(4):651-657.

21. R. S. Adiri, U. Gophna, E. Z. Ron (2003). Multilocus sequence typing (MLST) of escherichia coli O78 strains. FEMS Microbiology Letters, 222(2):199-203.

22. J. C. Uzoigwe, E. H. O'Brien, E. J. Brown (2007). Using nutrient utilization patterns to determine the source of escherichia coli found in surface water. African Journal of Environmental Science and Technology, 1(1):007-013.

23. V. J. Harwood, B. Wiggins, C. Hagedorn, et al. (2003). Phenotypic library-based microbial source tracking methods: Efficacy in the california collaborative study. Journal of Water and Health, 1(4):153-166.

24. A. P. White, K. A. Sibley, C. D. Sibley, et al. (2011). Intergenic sequence comparison of escherichia coli isolates reveals lifestyle adaptations but not host specificity. Applied and Environmental Microbiology, 77(21):7620-7632.

25. T. M. Scott, T. M. Jenkins, J. Lukasik, J. B. Rose (2005). Potential use of a host associated molecular marker in enterococcus faecium as an index of human fecal pollution. Environmental Science & Technology, 39(1):283-287.

26. L. A. Khatib, Y. L. Tsai, B. H. Olson (2002). A biomarker for the identification of cattle fecal pollution in water using the LTIIa toxin gene from enterotoxigenic escherichia coli. Applied Microbiology and Biotechnology, 59(1):97-104.

27. K.L. FAU, T. Y. FAU, B. H. Olson. A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic escherichia coli. Applied Microbiology and Biotechnology *JID - 8406612*. 0309.

28. L. A. Khatib, Y. L. Tsai, B. H. Olson (2003). A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic escherichia coli. Applied Microbiology and Biotechnology, 63(2):231-238.

29. C. He, H. Saedler (2005). Heterotopic expression of MPF2 is the key to the evolution of the chinese lantern of physalis, a morphological novelty in solanaceae.

Proceedings of the National Academy of Sciences of the United States of America, 102(16):5779-5784.

30. G. A. Wray, M. W. Hahn, E. Abouheif, et al. (2003). The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution, 20(9):1377-1419.

31. P. R. Schofield, J. M. Watson (1986). DNA sequence of rhizobium trifolii nodulation genes reveals a reiterated and potentially regulatory sequence preceding nodABC and nodFE. Nucleic Acids Research, 14(7):2891-2903.

32. W. K. Smits, O. P. Kuipers, J. W. Veening (2006). Phenotypic variation in bacteria: the role of feedback regulation. Nature Reviews Microbiology, 4(4):259-271.

33. J. Xiong (2006). Essential bioinformatics (chapter 11). Cambridge University Press, New York. 42-169

34. R. R. Sokal, C. D. Michener (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38: 1409–1438.

35. N. Saitou, M. Nei (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4):406-425.

36. W. M. Fitch, E. Margolias (1967). Construction of phylogenetic trees. Science, 155:279-284

37. A. Rzhetsky and M. Nei (1992). A simple method for estimating and testing minimum-evolution trees. Molecular Biology and Evolution, 12:163-167.

38. W. M. Fitch. Toward defining the course of evolution:minimum change for a specified tree topology. Systematic Zoology, 20(4): 406-416.

39. J. Felsenstein (2004). Inferring phylogenies. Sinauer Associates, Sunderland, MA.

40. M. Mohri, A. Rostamizadeh, A. Talwalkar (2012). Foundations of Machine Learning, The MIT Press, ISBN 9780262018258.

41. Y. Freund and R. E. Schapire (1999). A short introduction to boosting, Journal of Japanese society for artificial intelligence, 14(5): 771-780.

42. J. A. Anderson (1995). An Introduction to Neural Networks, MIT Press.

43. C. Cortes, V. N. Vapnik (1995). Support-Vector Networks, Machine Learning, 20.

44. D. G. Clayton (2009). Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes. PLoS Genetics 5(7): e1000540. Doi:10.1371/journal.pgen.1000540.

45 I. Ruczinski, C. Kooperberg, M. LeBlanc (2004). Exploring interactions in high dimensional genomic data: an overview of logic regression, with applications. Journal of Multivariate Analysis, 90, 178–195.

46. I. Dinu, S. Mahasirimongkol, Q. Liu, et al. (2012) SNP-SNP interactions discovered by logic regression explain Crohn's disease geneics. PLoS ONE, 7(10): e43035. Doi:10.1371/journal.pone.0043035.

3 Image Analysis Method for Tuberculosis Detection for Automated MODS in Resource-Limited Settings

3.1 Background

3.1.1 Review of Tuberculosis (TB) and TB diagnosis

TB is an airborne infectious disease. It is the second leading cause of death from an infectious disease worldwide [1] being prevalent mostly in developing countries [2]. According to the World Health Organization, about 2 billion people around the world were infected with TB, with an estimated global incidence rate of 139 cases per 100,000 population and a mortality rate of 20 deaths per 100,000 population in 2008 [3]. More than 80% of estimated new cases and deaths occurred in developing countries [2, 4].

Most TB cases are caused by *Mycobacterium tuberculosis* (*M.tuberculosis*). Latent TB (LTB) infection occurs when people carry small amount of *M. tuberculosis* bacilli in their body but the manifestation of disease is essentially controlled by an infected person's immune system. TB disease, or active TB, occurs when the bacterial load overcomes the person's immune defences and causes TB symptoms such as bad cough, chest pain and/or weight loss. People with LTB infection are not infectious and would be negative to most TB tests, due to the low levels of actively growing bacteria [4]. Patients with active TB may be infectious, with infections most commonly spread by inhalation of infected droplets coughed up by patients [4]. Although there are more LTB infections than active TB cases, identifying LTB is very difficult because LTB cases do not exhibit classical TB symptoms and the only test, the tuberculin skin test, is not accurate [6]. But since LTB is generally not considered an infectious state, it is more important for public health to identify active TB cases in order to treat and curb the spread of TB bacteria. TB diagnostic tests include culture-based methods, smear microscopy, nucleic acid amplification tests (NAAT), or through chest x-rays.

To curb the transmission of TB bacteria, the key is to detecting active TB accurately and quickly. Accurate and rapid TB diagnostic methods have been developed and are available mostly in the developed countries. Based on a systematic review by the

National Health Service's Health Technology Assessment Programme, NAAT testing, has far superior accuracy when applied to respiratory samples and is also more rapid than most current culture-based methods [7]. The Mycobacteria Growth Indicator Tube (MGIT) is a fully automated liquid-culture-based instrument intended for the fast detection and recovery of mycobacteria from clinical specimens with high accuracy [8]. However, these types of methods (NAAT and MGIT) are expensive and sometimes require highly trained technologists, which make them not widely accessible to resource-limited settings where TB is most prevalent but where financial resources and manpower are limited [9]. On the other hand, the low-cost TB diagnosis currently used in those settings, such as smear microscopy, have poor accuracy, and consequently new low-cost and accurate diagnostic mehtods for TB are urgently needed to replace those diagnostics used in those settings [9].

The microscopic-observation drug-susceptibility (MODS) assay is a low-cost culture-based TB diagnostic platform developed in 2006 [10]. Comparative analysis with other automated mycobacterial culture systems such as the MBBacT system (bioMerieuX) and culture on Löwenstein-Jensen medium with the proportion method, two reference methods commonly used in developing and industrialized countries, respectively, has demonstrated the utility of MODS for detection of *M. tuberculosis* infections. The estimated sensitivity and specificity of detection for MODS was 97.8% and 99.6%, for the automated mycobacterial culture was 89.0% and 99.9%, and for Löwenstein-Jensen was 84.0% and 100.0%. [10]. Thus, the MODS system has the potential for being used in resource-limited settings [11]. One principle for the MODS assay for detection of tuberculosis directly from sputum relies on the following principles: first, that *M. tuberculosis* grows faster in liquid medium than in solid medium; second, that characteristic TB cord formation can be visualized microscopically in liquid medium at an early stage [10]. In MODS, 24-well tissue-culture plates are used for preparing MODS assay. The cultures are examined under an inverted light microscope at a magnification of 40× every day (except Saturday and Sunday) from day 4 to day 15, on alternate days from day 16 to day 25, and twice weekly from day 26 to day 40. Since the method relies on microscopic observation of TB characteristic formations in liquid culture, manually delivery of liquid sputum-culture causes two critical limitations for use in practice: (1) biosafety; and (2) efficiency in handling a large number of samples.

The Automated MODS, developed recently by the TB/HIV Research Foundation (THRF), Chiang Rai, Thailand, is an automated MODS system designed to address the needs for an inexpensive, rapid and efficient TB diagnosis in resource-limited settings, overcoming the limitations of the original MODS. Specifically, the Automated MODS minimizes the risk of biohazard and cross-contamination by isolating sputum cultures using individual culture tubes sealed with individual caps, and using an auto-image-capture/reading system in an auto-running machine, which records culture images over time without manual handling of the culture. Of vital importance in the Automated MODS system is an automated image analysis program installed in a local computer or a network through which the recorded images can be reviewed and analyzed. Such an automated program can be used to detect TB growth without manual efforts and therefore reduces the laboratory workload and training appreciably.

3.1.2 Review of Image analysis methods for TB detection

Image analysis is the extraction of meaningful information from images, especially from digital images, using digital image processing and pattern recognition techniques [12]. There are many different techniques used in automatically analyzing images, each of which may be useful for certain tasks. However, there still aren't any known image analysis methods that are generic enough for wide ranges of tasks, compared to image analyzing capabilities of humans themeselves. The applications of digital image analysis are continuously expanding through all areas of science and industry including astronomy, materials science, machine vision, security, robotics, etc. as well as in microscopy.

Microscope image processing is the use of digital image processing techniques to process, analyze and present images obtained from a microscope. Analysis of images will vary considerably according to application. Typical analysis includes determining where the edges of an object are, counting similar objects, calculating the area, perimeter length and other useful measurements of each object. A common approach is creating an image mask that contains only those pixels that match certain criteria, and then performing simpler scanning operations on the resulting mask..

The use of image analysis techniques in automatic TB detection through culture images is quite new. A simplified automated image analysis method was used for detection and phenotyping of *Mycobacterium tuberculosis* on porous supports by monitoring growing microcolonies [13]. A key success of the image analysis depends on the use of a porous support for microcolony to grow in a liquid culture. The image analysis itself then involves the identification of microcolonies (distinct black spots) on a relatively simple and light background using existing publically available algorithms, the details of which was not shown. Another application of image processing techniques is for identifying *M. tuberculosis* in Ziehl-Neelsen stains, which is a solid culture method [13]. It involves recognition of ZN-stained AFB in digital images and using prior knowledge about the distinctive ZN stain color (red green blue values that were significantly different from non-TB objects) to label a given image pixel as a 'TB object' or a 'non-TB object', followed by shape/size analysis to refine detection. Other work also identifies bacteria particles based on image segmentation [15□17], which separate the bacteria pixels and non-bacteria pixels based on color information as well as morphological information. They often involves multiple steps of image processing and pattern recognition techniques such as image segmentation (edge detection) to detect the borders of the suspected objects, morphological operation to fix broken structures belonging to the objects, classification (based on complex classifiers, statistical techniques) for identify the bacteria particles based on shape or other heuristic descriptors [18,19].

The image analysis methods previously used in this area aim to automatically identify the distinctive morphology of microcolonies, requiring advanced techniques from segmentation and shape recognition. Some of them have shown great performance for detecting TB bacilli for solid culture or AFB stains. However, there is no duplicated mature image analysis method for detecting TB bacilli in liquid culture, which is more accurate in TB detection. The reason is that the liquid culture has very complex background so that the complex image segmentation and shape recognition tasks are difficult to automate for identification of floating and growing TB microcolonies with a high degree of sensitivity and specificity [20]. In addition to that, due to the special low-cost design of the Automated MODS, TB-specific microcolony formation was hard to identified in the Automated MODS images with human visual observation alone. The method for identifying bacterial particles and then classification as TB or non-TB is not a current application for the Automated MODS. New concept and image analysis methods for automatic TB detection for the Automated MODS need to be developed.

3.2 Data and materials

3.2.1 Sputum samples

The study was conducted in collaboration with THRF, using 81 sputum samples from suspicious pulmonary TB patients who were seen at the Chiang Rai Provincial Hospital in October 2010. These sputum samples were cultured with the Automated MODS at the TB Laboratory in Chiang Rai Provincial Hospital for the study. The TB status of each sample is verified by AFB test using MODS culture as test sample, which is used as a reference standard of TB diagnosis in the study. As verified by the reference standard, of the 81 sputum samples collected, 50 (61.7%) were positive, 29 (35.8%) were negative for *M.tuberculosis*, and two were "indeterminate": because the two indeterminate had both of the drug-free cultures contaminated by bacterial overgrowth, they were removed from the analysis.

3.2.2 MODS assay

In the TB laboratory, MODS assays were performed for each sputum sample according to the previously published MODS protocol [11]. Briefly, sputum samples were digested and decontaminated using N-Acetyl-L-Cysteine-sodium hydroxide for 15 minutes, dissolved with Phosphate Buffered Saline pH 6.8, and the mixture was centrifuged for 15 minutes at 3,000 rpm at 4° Celsius. The supernatant was dispersed using 7H9-OADC-PANTA (mixture of 7H9 (liquid Middlebrook broth), OADC (oleic acid, albumin, dextrose, and catalase), and PANTA (polymyxin B, amphotericin B, nalidixic acid, trimethoprim, and azlocillin)). The 900ul of the final mixture was inoculated in 3 individual 2ml plastic tubes as follows: two drug-free tubes containing 100ul 7H9-OADC-PANTA solutions, one drug tube containing 100ul p-nitrobenzoic acid (PNB), a drug used to kill *M.tuberculosis*. TB bacteria would grow in the two drug-free tubes but not in the PNB tube since PNB kills TB bacteri*a*; while other bacteria/fungus would grow in all three tubes. Therefore, a TB case can be identified by comparing culture growth in the drug-free tubes and the PNB tube. The culture tubes were kept at 37° Celsius for 35 days in the closed transparent incubator of the Automated MODS.

3.2.3 Image data

Starting from the first day of incubation, cultures were imaged daily by an automated digital microscope of the Automated MODS with a total magnification of ×40 and a field size of 480×480 pixels. The digital microscope of the Automated MODS is installed at the end of a 3-axis moving head connected with a computer-controlled motor, which is programmed to automatically take the culture images from the center of each tube. A culture image covers the entire culture area (Figure 3.1) and saved as a JPG file. The digital images of all cultures were transferred from Chiang Rai to the research team at the University of Alberta to develop image analysis methods for automated TB detection.

For each of the 79 patients, three sequences of culture images can be observed: I_{CL1}, I_{CL2}, I_{PNB}, each containing 35 images taken daily from day 1 to day 35. I_{CL1} and I_{CL2} denote the sequences of images for cultures in the two drug-free tubes, and I_{PNB} denotes the sequence of images for the culture in the PNB tube.

3.3 The proposed image analysis method

The rational of proposing a new image analysis method for the Automated MODS is that the previously used image analysis methods (not necessarily for MODS) for TB detection are not suitable for MODS. Those methods typically focus on identifying distinctive morphology of microcolonies and TB-specific morphological characteristics in culture images. They can be implemented for some solid culture [5-8] that have very clean background and for microcolonies with very simple morphological characteristics. However, MODS, as a liquid-based culture test, presents complicated background and complex morphology of floating microcolonies in culture images. The previously used methods are not suitable for MODS since the noisy background of MODS affect the robustness of their performances in detecting any microcolonies, not to mention identifying TB-specific morphological characteristics. Actually, identifying distinctive morphology of microcolonies and TB-specific morphological characteristics with such complicated MODS-culture images is difficult even for a visual inspection by an experienced technologist.

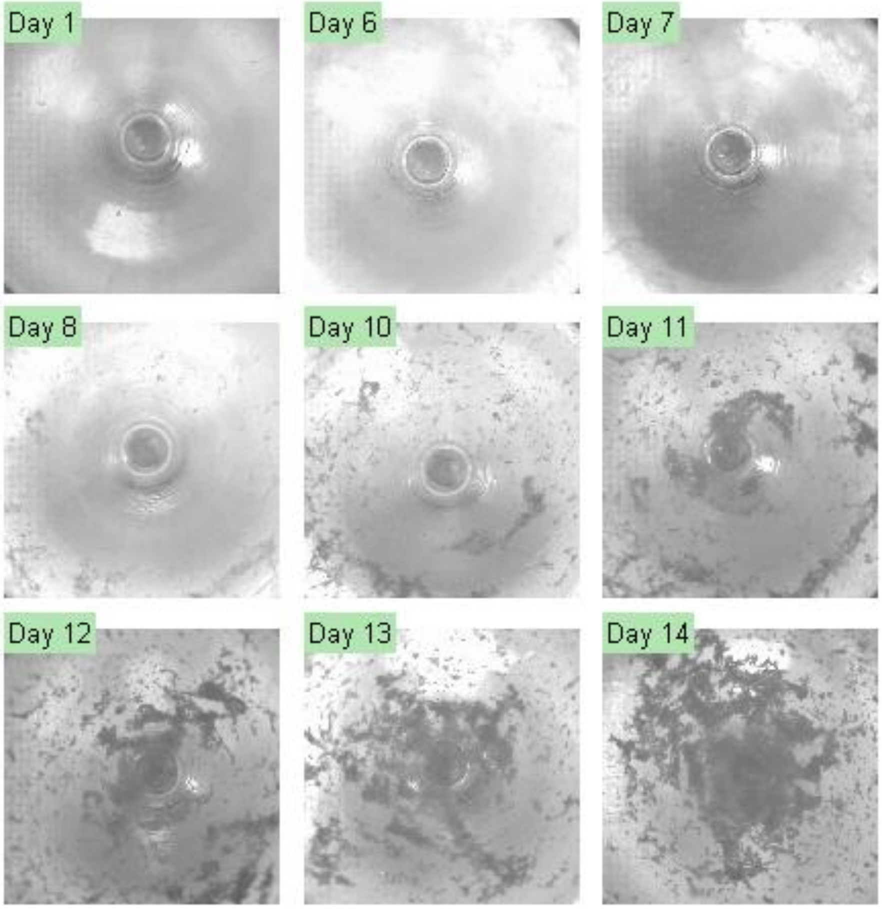


Figure 3. 1 Examples of positive MODS culture images taken at inoculation (Day 1) and between 6 to 14 days of inoculation

The proposed image analysis method aims to identify TB-specific culture growth patterns in the culture images instead of identifying TB-specific morphological characteristics from floating microcolonies. Of crucial importance in the method is a successful mathematical characterization of the culture growth pattern via a newly defined function of pixel intensities in images, called the λ-function, which largely reduces influences of image background, and overcomes the limitations of previously used image analysis methods. To describe the proposed image analysis method, the definition of the λ-function and the related λ-method will be given first in the following subsection.

3.3.1 The λ-function

To define the λ-function, two difference operators, ∇_{x} and ∇_{y} , which are used widely in image analysis to quantify the differences of intensities between adjacent pixels, will be introduced first. A culture image u is an N by N matrix of intensity levels with values in {0, 1, …, 255}, where N denotes the size of the image (in this case, N=480×480). Alternatively, an image u is a function defined on {1, 2, …, N} ×{1, 2, … , N} taking values in {0,1,… , 255}, and $u(\mathrm{i},\mathrm{j})$ denotes the function value (image intensity value) at pixel (i, j). Given an image u, the difference operators ∇_{x} and ∇_{y} in x- and y-directions are defined by

$$\nabla_{\mathrm{x}}\, u(\mathrm{i},\mathrm{j}) = \begin{cases} 0 & \text{if } \mathrm{i}=1 \\ u(\mathrm{i},\mathrm{j}) - u(\mathrm{i}-1,\mathrm{j}) & \text{if } 2 \le \mathrm{i} \le \mathrm{N} \end{cases}$$

$$\nabla_{\mathrm{y}}\, u(i,j) = \begin{cases} 0 & \text{if } \mathrm{j}=1 \\ u(i,j) - u(i,j-1) & \text{if } 2 \le \mathrm{j} \le N^{\cdot} \end{cases}$$

Next, define

$$s_u(i,j) = \frac{|\nabla_x u(i,j)|}{1+\min\{u(i,j),u(i-1,j)\}} + \frac{|\nabla_y u(i,j)|}{1+\min\{u(i,j),u(i,j-1)\}} \qquad (3.1)$$

where i, j=1,⋯, N, and set u(l, k)=0 if either l=0 or k=0. The definition of $s_u(i,j)$ is based on the following two characteristics of microcolonies in culture images. First, microcolonies in a culture image normally appear at pixels with low intensity values, where both terms' denominators in (3.1) are relatively small. Second, at least one of the two terms' numerators of (3.1) will be relatively large at the pixels near the edges of colonies. Therefore, large values of $s_u(i,j)$ tend to indicate the existence of a microcolony with the pixel (i, j) being near the edge of the colony. On the other hand, the image background areas are much brighter than microcolonies. Pixels in the background areas have large denominators in the two terms in (3.1), which makes their s_u tend to have small values.

Now let us return to the definition of the λ-function. The pixels where their s_u values are small, that is, they are unlikely to be the edge of microcolonies, will be ignored. This idea leads to the definition of the λ-function:

Given an image u, let $s_u^*(1) > \cdots > s_u^*(N^2)$ be a sequence with descending order of the N^2 numbers $\{s_u(i,j)$; i, j=1, …, N}. An overall smoothness measure of image u is defined as

$$SS_u = \sum_{\mathrm{j}=1}^{a\mathrm{N}^2} \mathrm{s}_\mathrm{u}^*(\mathrm{j})$$

where $0 < a < 1$ is a parameter to be determined: this allows us to ignore the smallest values of s_u's in a given image. How to optimize and determine this parameter will be discussed in Section 3.4.1. Then the λ-function is defined as follows:

$$\lambda(u,v) = \frac{SS_u}{SS_v} \tag{3.2}$$

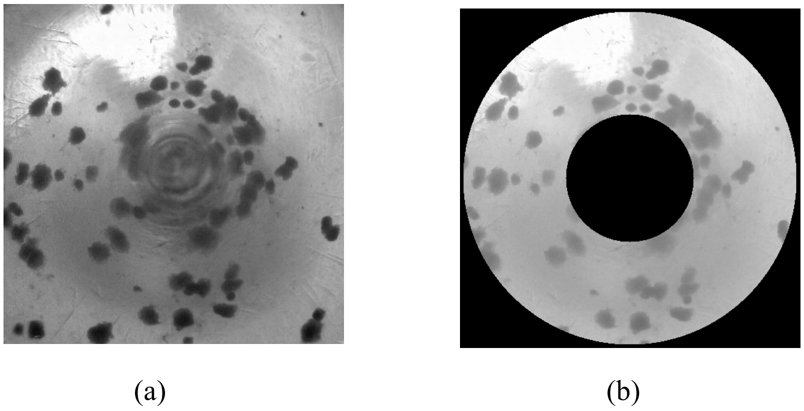
The meaning of λ-function is a ratio of overall smoothness measures comparing a culture images u to a reference culture image v.

To conclude this subsection, there are a few remarks on the λ-function. First, the definition (3.2) does not use smaller values of $s_u(i,j)$ and $s_v(i,j)$, which correspond to pixels near which microcolony is unlikely to exist. Thus, this definition allows us to ignore irrelevant information of the images. As a result, the influences of culture image background, which has relatively high intensity and is quite complex for culture images recorded automatically by digital microscopes, will be largely reduced in the ratio (3.2). Second, since smoothness measure SS_u reflects the amount of microcolonies in a culture, λ-function can be used to monitor culture growth in a culture by comparing the smoothness measure of an image to a reference image. In application a reference image is selected to be an image of a culture at its early inoculation time and compare the images of the same culture at later stages with it. Moreover, at the original stage, there are little inherent differences across different cultures. Therefore, λ-function can be used as a consistent measure of culture growth across cultures although different reference images need to be used.

3.3.2 Proposed image analysis method

The proposed method involves three steps, which will be described in details below.

In the first step, the aim is to eliminate the influences of inherent differences across cultures on the smoothness measure SS_u. Based on the data, it can be observed that the center and the edge areas of different culture tubes have quite different structures in images, and even for the same culture tube, the images of those areas will be quite different under different illumination conditions on different days. To eliminate these differences, an image mask is applied to each image to exclude its center and edge areas from further analysis. The image mask (the black area in Figure 3.2.b) is defined as follows: a pixel (i, j) will be masked if its distance from the center of the image d_c satisfy: $d_c > L_{max}$ or $d_c < L_{min}$, where L_{max}=235 and L_{min}=90 are chosen based on the actual center and edge areas in images.



(a) (b)

Figure 3. 2 (a) An original image recorded by the Automated MODS and (b) the output image processed by masking and Gaussian filtering

Since culture images of the the Automated MODS were automatically recorded by digital microscopes, and they normally contain image noise caused by the digital image capture device. The image noise is random variation of intensity in images and not present in the object imaged: however, this added spurious and extraneous information may distort the smoothness measure. In the first step, a Gaussian filter is also applied to smooth such image noise. It works by filtering an image with a Gaussian function:

$$g(x, y) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution. In this application, an 18x18 Gaussian filter (18 is the number of rows and columns of the Gaussian filter) with standard deviation 3 is chosen. The size of the Gaussian filter was chosen so that it is smaller than the size of any useful information, a visible microcolony in this context, but larger enough to blur the noisy information. The output image is a modified image of the original one with the same size. The value of each pixel in the output image is a weighted average of all its neighbors in the input image weighted by the Gaussian filter centered at the pixel. The neighbor pixels involved contains all pixels in the input image corresponding to the Gaussian filter. The standard deviation determines the degree of blur. It is chosen to make the averaged pixel value weighted enough towards the value of the central pixels considering the size of the Gaussian filter. Figure 3.2 (b) shows the image after being processed by masking and filtering in the first step.

In the second step, apply the λ-function to all sequences of processed culture images from the first step. Let I=$\{u_1, \ldots, u_{35}\}$ denote a sequence of 35 images of any incubated culture that are taken daily from day 1 to day 35, where u_i denotes the culture image taken on the i-th day. Since the bacteria-irrelevant culture contents, such as debris and air bubbles, will be stabilized in the culture on about day 5 after incubation while bacterial growth usually will not be visible before day 5, the reference image, v_0, is chosen to be the image taken on day 5; namely, $v_0 = u_5$. The image v_0

will be considered as a reference image in the λ-function applied to the sequence I. All the other images u_i, starting with day 5, with the image v_0 via the function $\lambda(u_i, v_0)$ are compared. A graph of the sequence of λ-function values

$$\lambda(I) = (\lambda(u_5, v_0), \lambda(u_6, v_0), \dots, \lambda(u_{35}, v_0)) \qquad (3.3)$$

can successfully characterize the culture growth pattern for the sequence I (see Figure 3.3 for an example).

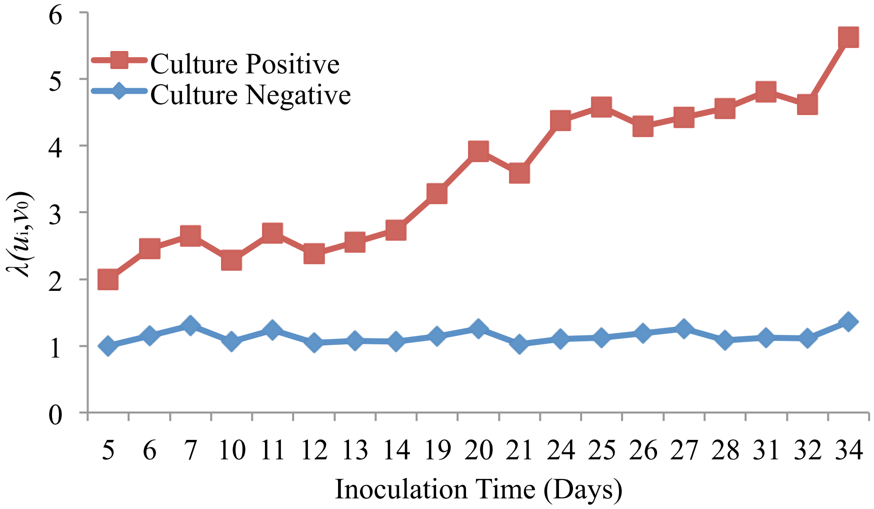


Figure 3. 3 Graphs of λ(I)

For each sequence I={u_1, …, u_{35}} of 35 culture images in the dataset, the formula (3.2) is used to compute the sequence **λ(I)** using u_5, the image on day 5 in each sequence, as the reference image v_0 for the sequence.

Finally, in the third step, a decision for culture positivity is made, i.e., a significant culture growth based on the data $\lambda(I)$ computed in the second step. A sequence I={u_1, …, u_{35}} of 35 culture images is said to be culture positive if *d* consecutive days $\lambda(u_{i+1}, v_0), \dots, \lambda(u_{i+d}, v_0)$ in the sequence $\lambda(I)$ given in (3.3) satisfy simultaneously the condition:

$$\lambda(u_j, v_0) > \lambda_t, \text{ j=i+1, i+2, …, i+d.}$$

The sequence I is said to be culture negative otherwise. The decision rule contains three crucial statistical parameters: the parameter a in the definition of the λ-function and two new parameters $\lambda_t > 1$ and $d \geq 2$.

Recall that, for each patient, three sequences I_{CL1}, I_{CL2}, and I_{PNB} of 35 culture images are obtained. The decision rule for TB positivity is defined based on the principle of the three-assay laboratory design as follows.

Decision Rule: On any given day, using image data up to that day, a patient is judged to be TB positive if the image sequences I_{CL1} and I_{CL2} for the cultures in the two drug-free tubes are *both* culture positive, but the image sequence I_{PNB} for the culture in the PNB tube is culture negative. The patient is judged to be TB negative if at least one of the two image sequences, I_{CL1} and I_{CL2}, is culture negative, or all three image sequences, I_{CL1}, I_{CL2} and I_{PNB}, are culture positive.

In the definition of TB positive image sequences, it is required that an interval {i+1, i+2, …, i+d} of d consecutive days, rather than a single day, to satisfy a condition. This helps to reduce the chance that some temporarily unusual conditions, such as an illumination condition which can yield outliers of the image data, to create a false positive.

3.4 Statistical analysis

3.4.1 Selection of the values for parameters a, λ_t and d

The decision rule in Section 3.3.2 relies on three parameters, a, λ_t, and d. The purpose of this section is to show the method for selecting the values of these parameters.

Note first that $0<a<1$, $\lambda_t \geq 1$ and d can be any integer in the set A_1={6, …, 35}. Based on the data, an image with λ_t =5 corresponds to an image with a large amount of culture growth. Thus, $1 \leq \lambda_t \leq 5$ is considered. The parameter intervals (0, 1) and [1, 5] of the parameters a and λ_t are partitioned as follows:

$$A_2 = \{0.05k\colon\ k = 1, \dots, 20\} = \{0.05, 0.1, 0.15, \dots, 0.95, 1\},$$
$$A_3 = \{1 + 0.05k\colon\ k = 0, 1, \dots, 80\} = \{1, 1.05, 1.1, \dots, 4.95, 5\}.$$

The main goal is to find the optimal values $\hat{a}, \widehat{\lambda_t}$, and $\hat{d}$, of the parameters a, λ_t, and d, respectively, subject to the conditions that d takes values in A_1, a in A_2 and λ_t in A_3. The optimal values can be determined as follows.

First, the training samples were selected by selecting randomly 50% of the samples: 25 of the total 50 TB positive patients, and 15 of the total 29 TB negative patients.

Next, for each given set of parameter values for d in A_1, a in A_2 and for λ_t in A_3, the decision rule of Section 3.3.2 is applied to classify the training samples. Let $N_1(a, \lambda_t, d)$ denote the number of the TB-positive patients in the selected samples who are also judged correctly to be TB-positive using the decision rule with the parameter values a, λ_t and d. Let $N_2(a, \lambda_t, d)$ denote the number of the TB-negative patients in the training sample who are also judged correctly to be TB-negative. Finally, the optimal parameters $\hat{a}$, $\widehat{\lambda_t}$, and $\hat{d}$ is computed as follows:

$$(\hat{a}, \widehat{\lambda_t}, \hat{d}) = \operatorname{argmax}_{a \in A_1, \lambda_t \in A_2, d \in A_3} \left(\frac{N_1(a,\ \lambda_t,\ d)}{25} + \frac{N_2(a,\ \lambda_t,\ d)}{15}\right)$$

That is, $\hat{a}$, $\widehat{\lambda_t}$, and $\hat{d}$ are the parameter values with which the sum of sensitivity and specificity, $\frac{N_1(a, \lambda_t, d)}{25} + \frac{N_2(a, \lambda_t, d)}{15}$, attains its maximal value. Since the sets A_1, A_2 and A_3 are all finite, this last step can be computed easily.

In this study, the following values of $\hat{a}$, $\widehat{\lambda_t}$, $\hat{d}$ were obtained using the above mentioned method:

$$\hat{a} = 0.2,\ \widehat{\lambda_t} = 1.7,\ \hat{d} = 3. \tag{3.4}$$

The sensitivity and specificity attained with these optimal parameter values will be studied in the next subsection.

3.4.2 Test-set validation

In this subsection, the method of test-set validation is used to evaluate the performance of the decision rule with parameter values selected in the last subsection.

Recall that the training sample consists of 25 randomly selected TB positive patients and 15 randomly selected TB negative patients. The remaining 39 patients, including N_1=25 TB positive and N_2=14 TB negative patients, are served as the testing samples. The decision rule with parameter values given in (3.4) is applied to these testing samples. Let N_3 (resp. N_4) denote the number of patients in the TB positive (resp. TB negative) testing samples that are ruled to be TB positive (resp. TB negative) as well according to the decision rule. Then the sensitivity and specificity of the test can be calculated as follows:

$$\text{Sensitivity} = \frac{N_3}{25}, \text{Specificity} = \frac{N_4}{14}.$$

In this context, sensitivity is the conditional probability that a sputum sample is identified as TB positive, given that the sputum sample is TB positive. Specificity is the conditional probability that a sputum sample is identified as TB negative, given that the sputum sample is TB negative. Their confidence intervals are calculated using the standard exact method for binomial proportion.

3.4.3 Comparison with manual image reading

The culture images of the 79 samples were reviewed to identify TB cases by a medical technologist in THRF, who is also responsible for performing the diagnosis using the standard method. Similar decision rules for TB positivity based on the three-assay laboratory design were used by the technologist, except that the culture growth is identified by subjective human judgement. The time to positivity is also recorded by the technologist for each sputum sample.

Sensitivity and specificity were calculated for the λ-method and the manual image reading using the testing samples. The differences in sensitivity and specificity between the two methods were tested statistically. Since the same testing samples were used, the results from λ-method and those from manual image reading are paired. The p-values of the differences in sensitivity and specificity were calculated using a McNemar's exact test, taking this pairing into account.

The two methods were also compared in terms of time to positivity. Time to positivity of case samples were plotted and compared for both methods.

3.5. Results

3.5.1 TB detection

The results of automated TB detection on the testing samples obtained via the λ-method, with parameter values $\hat{a} = 0.2$, $\hat{d} = 3$, and $\widehat{\lambda_t} = 1.7$, are listed in Table 3.1 below.

Table 3. 1 Results detected via λ-method

		Reference method	
		Positive (+)	Negative (-)
λ-method	Positive (+)	20	0
	Negative (-)	5	14

Table 3. 2 Results obtained via manual image reading

		Reference method	
		Positive (+)	Negative (-)
Manual image reading	Positive (+)	24	1
	Negative (-)	1	13

Based on table 3.1, the λ-method has an estimated sensitivity of 80% [95% CI: 58.7% - 92.4%], and an estimated specificity of 100% [95% CI: 73.2% – 100%]. In comparison, the manual image reading has an estimated sensitivity of 96% [95% CI: 77.7% - 99.8%] and an estimated specificity of 92.9% [95% CI: 64.2% - 99.6%]

(Table 3.2). P-values for the differences of sensitivities and specificities between two methods are 0.13 and 0.50, respectively, based on McNemar's exact test (Table 3.3).

Table 3. 3 Comparison of sensitivity and specificity between λ-method and manual image reading

(a) Results on true-positive testing samples

Positive testing samples		**Manual image reading**	
		Positive (+)	Conclusions different from those of AFB test
λ-method	Positive (+)	20	0
	Conclusions different from those of AFB test	4	1

(b) Results on true-negative testing samples

Negative testing samples		**Manual image reading**	
		Conclusions different from those of AFB test	Negative (-)
λ-method	Conclusions different from those of AFB test	0	0
	Negative (-)	2	12

The 25 TB positive testing samples verified by the standard method were included in the head-to-head analysis of time to TB positivity (Figure 3.4). For the λ-method, time to TB positivity for a true-positive subject was defined as the last day of the *d* consecutive days of TB positive by the λ-method. The median time to TB positivity of the λ-method was 20 days [interquartile range 10 to 29 days], which is longer than that of the manual image reading's 13 days [interquartile range 17 to 27] (Figure 3.4).

3.6 Conclusions and discussions

In this thesis, a novel image analysis method is proposed for the automated programming of the Automated MODS, i.e., for the objective detection of TB growth in the culture images recorded in the Automated MODS. The most important contribution of this thesis is the introduction of a new function of pixel intensities in images, called the λ-function, which can be used to successfully characterize the culture growth pattern in the culture images. Compared with many image analysis

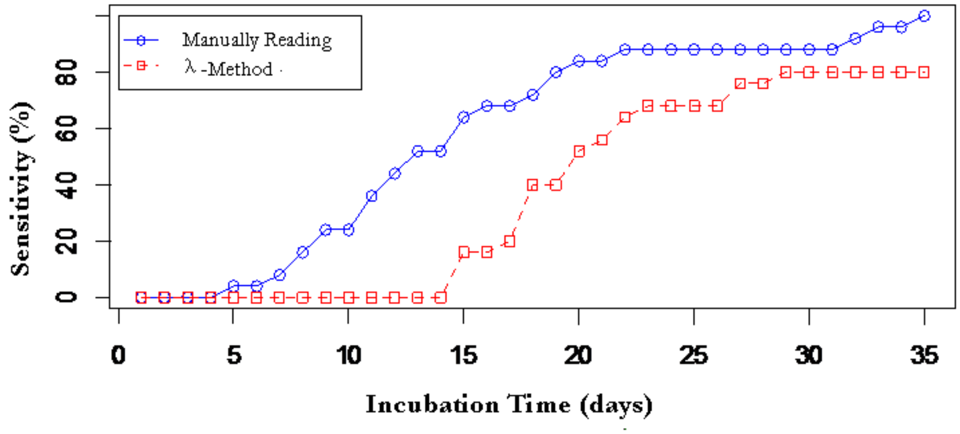


Figure 3. 4 Time to TB positivity of the λ-method and manual image reading

3.5.2 Time to TB positivity

methods previously used in this area [5-8], the proposed method has the following advantages. First, it is simpler, faster, and can be easily implemented using any programming language. Second, many of the Automated MODS images recorded by a digital microscope have complicated background for which even a highly skilled technologist can have difficulties for identifying TB culture growth. Due to the introduction of the λ-function, the method largely reduces influence of the image background, and therefore, is more robust and enjoys high sensitivity and specificity in TB detection.

The proposed automated image-analysis method for the Automated MODS has a high sensitivity and specificity evaluated against a standard TB diagnosis method. It is expected that this work will lead to an automated TB detection for the Automated MODS, being able to read and interpret culture images without medical technologists' time-consuming manual efforts. The new method significantly improves the efficiency of the Automated MODS, which uses a high-throughput imaging system. It makes the Automated MODS highly efficient and safe compared to the original MODS, with comparable sensitivity and specificity to manual image reading by an experienced medical technologist. Note that the manual reading used in this thesis was performed by a technologist who was aware of the sample's true positive/negative status, i.e., the manual reading was not done with blinding.

The proposed image-based automated TB detection method is an objective method that does not need human intervention once applied. The three critical parameters in the method can be determined by sound statistical methods based on training data from practical settings. In this study, the parameter determinations were based on the training samples and the estimates of the sensitivity and specificity were obtained using independent test samples and are, therefore, unbiased. In other words, the same performance would be expected for the method when it is used on similar patients and on sputum culture images obtained with the same the Automated MODS system.

The image analysis method shows a delay of 7 days in time to positivity, as compared to manual reading. Since TB is an infectious disease, taking longer time to perform the testing will leave longer time for TB bacteria to spread from TB cases to uninfected people. This remains a critical limitation of the Automated MODS compared to the advanced TB diagnostic such as NAAT.

Future research will be mainly focused on improving the λ-method in terms of the time to positivity, that is, to make the λ-method to detect sample positivity faster (i.e., in fewer days). Another direction for future research would be improving the sensitivity of the λ-method through the application of a more sensitive decision rule for bacterial growth.

Bibliography

1. G. L. Mandell, J. E. Bennett, R. Dolin (2010). Mandell, Douglas, and Bennett's principles and practice of infectious diseases (7th edition). Philadelphia, PA: Churchill Livingstone/Elsevier. Chapter 250. ISBN 978-0443068393.
2. Word Health Orgonization (2011). Global tuberculosis control, WHO report 2011. Available at: http://www.who.int/tb/publications/global_report/2011. Accessed June 12, 2012.
3. Dye C, S. Scheele, P. Dolin, et al. (1999). Global burden of tuberculosis: estimated incidence, prevalence and mortality by country. Journal of the American Medical Association, 282(7):677-686

4. K. Lönnroth, K. G. Castro, J. M. Chakaya, et al. (2010). Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet, 375(9728):1814-1829.
5. C. Dyer (2010). Tuberculosis (biographies of disease). Greenwood Press, Santa Barbara, California, USA.
6. M. Ishaq, I.M. Sameera, K.M. Miraj (2005). Tuberculin skin testing widely used as a diagnostic aid for tuberculosis, false negative outcome has questioned its specificity & sensitivity inspite of tuberculous infection. 5th Annual Meeting of the Federation-of-Immunology-Society, Boston, May 12-16.
7. P. Daley, S. Thomas, M. Pai (2007). Nucleic acid amplification tests for the diagnosis of tuberculous lymphadenitis: a systematic review. International Journal of Tuberculosis and Lung Disease, 11(11):1166-1176.
8. C. Scarparo, P. Piccoli, A. Rigon, et al. (2002) Evaluation of the BACTEC MGIT 960 in comparison with BACTEC 460 TB for detection and recovery of mycobacteria from clinical specimens. Diagnostic Microbiology and Infectious Disease, 44(2):157-161.
9. J. Foulds, R. O'Brien (1998). New tools for the diagnosis of tuberculosis: the perspective of developing countries. International Journal of Tuberculosis and Lung Disease, 2(10): 778-783.
10. D. A. J. Moore, C.A.W. Evans, R.H. Gilman, et al. (2006). Microscopic-observation drug susceptibility assay for the diagnosis of TB. The New England Journal of Medicine, 355:1539-50.
11. D. A. J. Moore, D. Mendoza, R. H. Gilman, et al. (2004). Microscopic observation drug susceptibility assay, a rapid, reliable diagnostic test for multidrug-resistant tuberculosis suitable for use in resource-poor settings. Journal of Clinical Microbiology, 42(10):4432-7.
12. C. Solomon, T. Breckon. (2011). Fundamentals of digital image processing: a practical approach with examples in Matlab. John Wiley & Sons. ISBN 0470844736.
13. A. L. den Hertog, D. W. Visser, C. J. Ingham, et al. (2010). Simplified automated image analysis for detection and phenotyping of Mycobacterium tuberculosis on porous supports by monitoring growing microcolonies. PLoS ONE, 5(6): e11008. doi:10.1371/journal.pone.0011008

14. P. Sadaphal, J. Rao, G. W. Comstock, M. F. Beg (2008). Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains. International Journal of Tuberfulosis Lung Disease, 12(5):579-582.

15. M. G. Forero, Cristobal, J. Alvarez-Borrego et al. (2003). Automatic identification techniques of tuberculosis bacteria, SPIE 2003 Nov 1; 5203 Appplications of Digital Image Processing XXVI: 71-81.

16. A. Divekar, C. Pangilinan, G. Coetzee, etc. Automated detection of tuberculosis on sputum smeared slides using stepwise classification, Proceeding SPIE Medical Imaging Conference (8315-123), Newport Beach, CA, Feb 4 – 9, 2012.

17. M. G. Forero-Vargas, E. L. Sierra-Ballen, J. Alvarez-Borrego, et al. (2001). Automatic sputum color image segmentation for tuberculosis diagnosis. Algorithms and Systems for Optical Information Processing V, Proceedings of SPIE, Vol. 4471.

18. K. Veropoulos, C. Campbell, G. Learmonth, et al. (1998). The automated identification of tubercle bacilli using image processing and neural computing techniques. Perspectives in Neural Computing, ICANN98.

19. J. O'Malley, R. Santiago-Mozos, M. G. Madden (2011). Image recognition of tuberculosis in sputum using K-NN with dynamic time warping. Proceedings of AICS-2011, the 22nd Irish Conference on Artificial Intelligence and Cognitive Science, Derry, Northern Ireland.

20. N. Pinto, D. D. Cox, J. J. DiCarlo (2008). Why is real-world visual object recognition hard? PLoS Computational Biology, 4(1): e27.

4 Conclusion and Discussion

In Chapters 2 and 3, two biologically-sensible methodologies for analyzing high-dimensional biological data, *E. coli* genomic data and sputum-culture image, were proposed for two complex public health problems. A common focus of the two methodologies was to extract information that is relevant for tackling the complex health problems utilizing high-dimensional biological data. Statistically-appropriate techniques were applied to this end.

In the analysis of *E. coli* genomic data (Chapter 2), model fitting in a high-dimensional data space was achieved by using an adaptive Simulated Annealing Algorithm, and performance of the model fitting was evaluated statistically by cross-validation. This method is successful in reducing computational cost, which is one of the major issues in high-dimensional data analysis, using an advanced adaptive method. In analyzing the sputum-culture images (Chapter 3), high-dimensionality of the image data is reduced dramatically by introducing a non-linear function of culture image without loosing the relevant information for measuring the level of culture growth. A key success of this method was in the handling of complex background and other noise issues through proper image-analysis methods along with an introduction of a novel mathematical function that captures the biological features of culture growth in tubes.

More importantly, the methodologies presented in this thesis considered the biological and clinical aspects of complex health problems, which are as important as the technical challenges caused by high-dimensionality of data. In Chapter 2, logic regression were carefully chosen as a basic model for analysis because of its special forms of SNP interactions, which are biologically more plausible than the commonly used multiplications of SNP indicators. Moreover, a two-step approach is adopted to incorporate the complex assumptions of *E. coli* host selection involving two distinct biological hypotheses. In Chapter 3, the high-dimensional sputum-culture images were transformed to measures of culture growth, which are clinically more meaningful for the problem than the high-dimensional intensity values in an image. The mathematical function constructed was based on the biological relationship between culture growth and image characteristics. Without considering these biological and clinical aspects of the complex health problems, the methodology development for the problems would have focused on technicality and would have produced methods that are hardly useful in practice.

As high-dimensional data (or Big Data) are increasingly available for research, high-dimensional data analysis is becoming more and more in demand: it is also becoming more difficult due to its increasing size. The advantages and disadvantages of high-dimensional data have been discussed widely in literatures [1-2]. Analytical approaches for dealing with high-dimensionality of the data have also been suggested [3]. When conducting high-dimensional data analysis, people should be aware of the so-called "curse of high dimensionality", in connection with the difficulty of optimization by exhaustive enumeration on high-dimensional spaces, and the effect of high dimensionality on statistical measures.

On the other hand, the successful experiences in the two complex health problems took great advantage of the high-dimensional data. This is the reason why high-dimensional data is increasingly applied in health problems as well as in medical imaging, marketing, finance, economics, and so on. It is expected that the demand for methodologies for analyzing these data will continue to grow. In particular, medical images and genomic data have been broadly used for innovations in solving different health problems, where similar methodological issues are encountered. I hope that the methodologies proposed in this thesis will be useful in a broader context in the future.

Bibliography

1. D. L. Donoho. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. Lecture on August 8, 2000, to the Americal Mathematical Society "Math Challenges of the 21st Century". Available form http://www-stat.stanford.edu~donoho/.
2. R. Clarke, H. W. Ressom, A. Wang, et al. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Review of Cancer, 8(1):37-49.

3. M. Verleysen (2003). Learning high-dimensional data. Limitations and Future Trends in Neural Computation. S. Ablameyko et al. (Eds.), IOS Press: 141-162.