USING COMPUTATIONAL INTELLIGENCE FOR KNOWLEDGE DISCOVERY FROM THE HUMAN MICROBIOME

BENJAMIN WINGFIELD BSc (Hons) MSc

FACULTY OF COMPUTING, ENGINEERING AND THE BUILT ENVIRONMENT OF THE UNIVERSITY OF ULSTER

A thesis submitted for the degree of
Doctor of Philosophy

March 2019

"I confirm that the word count of this thesis is less than 100,000 words"

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professor Sonya Coleman, Professor Martin McGinnity, and Professor Tony Bjourson for their outstanding support, guidance, and advice throughout my PhD, which has prepared me well for academic life and working in research.

Undertaking my PhD would not have been possible without financial support from the Department for Employment and Learning whose scholarship supported this project and enabled me to attend several conferences.

Major parts of the research presented in this thesis required the collection, processing, and sequencing of biological samples. I would like to thank everybody involved in the process, including Professor Tony Bjourson, Dr. Coral Lapsley and Dr. Elaine Murray from the Northern Ireland Centre for Stratified Medicine for providing the data which were analysed in this thesis.

I would also like to thank my friends and family for their support and encouragement throughout my PhD. Finally, I would like to thank my fiancée Natalie for her patience and support. I am fortunate that we both undertook a PhD at the same time, and I wish her the best of luck for her thesis submission and defence.

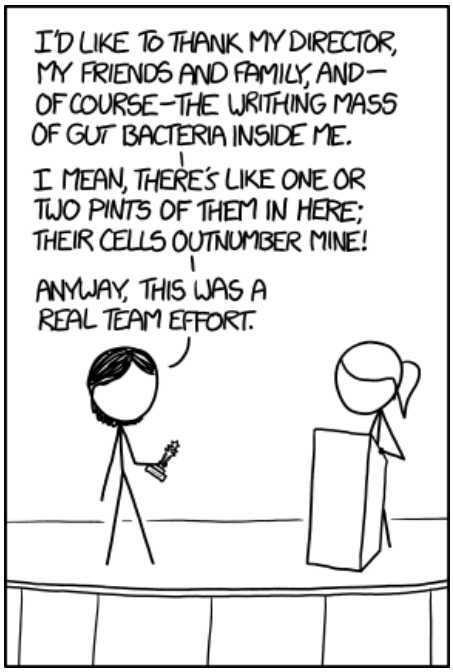


Figure 1: "Team Effort" by Randall Munroe.

ABSTRACT

Subtle changes in microbial populations that inhabit different areas of the human body — known as microbiomes or microbiota — can contribute to disease development, and restoring these imbalances may provide a cure. Localised and systemic diseases such as Inflammatory Bowel Disease (IBD) and depression have been linked with alterations to microbiota across the human body. Our understanding of how both diseases develop contains significant gaps, and the microbiome — described by some as our "second genome" — offers a compelling new area for knowledge discovery. This thesis aimed to advance the field of microbiome research and is an account of the work conducted whilst investigating the human gut and oral microbiome for links with IBD and depression. In this thesis, a hybrid model and aggregating ensemble feature selection (EFS) approach are applied to microbiome census data gathered from subjects with IBD. Microbial ecology techniques are applied to identify alterations to the oral microbiome in depressed subjects, and a multimodal Computational Intelligence (CI) classification paradigm known as a Super Self-Organising Map (sSOM) is applied to predict depression from a saliva sample. Finally, a rough set characterisation approach was developed and applied to gut and oral microbiome census data in depressed subjects to avoid destructive data normalisation and to enable knowledge discovery. The outcomes from the development of the hybrid model and aggregating EFS approach include the accurate non-invasive prediction of IBD, and the identification of novel and robust alterations to the gut microbiome in an adult cohort of IBD patients. The result provides a potential alternative to invasive colonoscopy, improve the time to diagnosis and treatment of IBD, and delivers new insights into the aetiology of IBD. The investigation of the oral microbiome identified novel alterations in depressed subjects for the first time. The changes to the structure and composition of the oral microbiome were significant enough to enable the accurate prediction of depression from a saliva sample. The results contribute to the microbiome-gut-brain axis theory by associating alterations to the oral microbiome with depression for the first time, and offer an alternative to subjective criteria for diagnosing depression, which currently relies on patient self-report and clinical judgement. The rough set microbiome characterisation approach replicated existing results and identified previously undescribed alterations to the gut microbiome in depressed subjects. The results provide an alternative approach to destructive normalisation techniques that are often applied to microbiome census data (identifying an optimal approach is an open research question), and contribute to our understanding of the microbiome-gut-brain axis, which could lead to psychobiotic treatments of depression in the future.

GLOSSARY

α-diversity Within-sample diversity. 112, 113, 125, 127

β-diversity Across-sample diversity. 127

k-mer Also known as words; all possible subsequences of length k that are contained in a nucleotide sequence. 78

de novo "Of new" — in the context of OTU picking, cluster sequences relative to each other (without use of reference databases). 76, 79

in silico performed on a computer or via computer simulation. 103

in vitro performed in a laboratory outside of a living organism. 121

`DESeq2` `R` software package: Differential gene expression analysis based on the negative binomial distribution. Incorporates shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. 122, 124

`DESeq` `R` software package: Differential gene expression analysis based on the negative binomial distribution. 114, 115

`R` A programming language for statistical computing. 121–124

`dada2` `R` software package: denoises sequenced amplicon data. 111

`picante` `R` software package: provides a suite of tools for analyzing the phylogenetic and trait diversity of ecological communities. 123

`vegan` `R` software package: enables multivariate analysis of ecological communities. 121, 122

16S ribosomal ribonucleic acid (16S rRNA) A component of the prokaryotic ribosome (the ribosome is responsible for biological protein synthesis). The gene encoding 16S rRNA is widely used as a universal marker gene to identify bacterial species. 52–54, 58, 73, 75, 77, 82, 84, 108, 109

adrenocorticotropin hormone (ACTH) A hormone involved in stress response, part of the HPA axis. Stimulates the production of cortisol. 43

amplicon sequence variant (ASV) 16S rRNA gene sequences that have been denoised to a single nucleotide resolution. 62, 63, 75, 78–80, 94, 96, 97, 100, 101, 104, 121–125, 133, 135, 154, 160, 173, 174

ancestral state reconstruction (ASR) Extrapolation of individuals to common ancestors. 77

Area Under the Receiver Operating Characteristic (AUROC) A statistical method for measuring the diagnostic capability of a binary classifier. 32, 84, 88, 89, 101

Artificial Intelligence (AI) The theory and development of computer systems able to perform tasks that typically require human intelligence. 1, 4, 170

Artificial Neural Network (ANN) A biologically inspired statistical pattern recognition technique. 19–21, 34, 44, 69, 85, 118, 124

base call A process that converts raw sequence data (e.g. a time-series of chromatogram peaks) to text-based nucleotide sequences. 49, 54, 55, 71, 81

biological marker (biomarker) A measurable indicator of a biological state (e.g. disease) 4

bipolar disorder (BD) Also known as manic depression. Distinct from MDD, a mental disorder that causes periods of depression and periods of elevated mood (mania). 41

black box A model that is viewed only in terms of its inputs and outputs, without any knowledge of its internal processes. Its implementation is opaque. 19, 137, 174

clade A group of organisms that include a common ancestor and all of its extant and extinct descendants 38

Computational Intelligence (CI) "The ultimate goal of researchers in this field was mimicking Nature with artificial technologies to replicate the basic mechanisms of Nature in engineering systems for the benefit of humanity... CI technologies are living approaches to tackle real-world problems... created as answers to the needs of applications" (Fulcher, 2008) ix, 1–5, 7–9, 23, 25, 31, 33–35, 69, 72, 104, 109, 137, 140, 165, 169, 170, 175, 178

corticotropin-releasing hormone (CRH) A hormone involved in stress response, part of the HPA axis. Stimulates the production of adrenocorticotropin hormone. 43

crohn's disease (CD) a chronic condition where the entire gastrointestinal tract becomes inflamed. 39, 74, 76, 77, 83, 88, 91, 94, 96, 100, 171

cross-sectional study An epidemiological study that describes the characteristics of a population. Data are collected at a fixed point in time and the relationships between characteristics (e.g. the composition of the microbiome and disease status) are considered. Cross-sectional studies cannot establish a causative effect. 39

deoxynucleoside triphosphate (dNTP) The building blocks of nucleic acid molecules. Required for PCR to amplify target sequences. 54

Diagnostic and Statistical Manual of Mental Disorders (DSM) A manual published by the American Psychiatric Association used by clinicians and psychiatrists that describes symptoms, statistics, and treatment options for mental health disorders. 41

divisive amplicon denoising algorithm (DADA) A bioinformatics algorithm that infers the true DNA sequence present in amplicon data (denoising) to a single-nucleotide resolution. 63, 72, 80

ensemble feature selection (EFS) A feature selection methodology that improves the robustness of feature selector output. ix, 4, 5, 30, 34, 35, 74, 91–94, 100, 103, 104, 137, 162, 171, 173

functional Magnetic Resonance Imaging (fMRI) A method of studying brain activity by measuring blood flow associated with changes (e.g. a particular task). 139

guanine One of the four main nucleobases found in DNA. Pairs with cytosine. 80

halophile Organisms that live in highly saline environments. They require salinity to survive. 104

halotolerant Organisms that can tolerate highly saline environments, but do not require salinity to survive. 104

hypothalamic-pituitary-adrenocortical (HPA) The set of organs that constitute the HPA axis: a major neuroendocrine system responsible for a variety of mechanisms including stress and immune response and digestion. 43, 45, 71

Indeterminate Colitis (IC) IBD cases that are impossible to diagnose as ulcerative colitis (UC) or crohn's disease (CD). 77

Inflammatory Bowel Disease (IBD) An umbrella term for a group of inflammatory diseases of the gastrointestinal tract, including Crohn's Disease and ulcerative colitis. ix, 2–5, 33–35, 38, 39, 43, 44, 71, 73–75, 81–92, 94, 96, 100–104, 107, 136, 162, 169–171, 173, 174, 178

KEGG Ortholog (KO) A database of molecular-level functions. Part of the KEGG collection. 123

Knowledge Discovery in Databases (KDD) The process of discovering useful knowledge from a collection of data, also known as discovery science or discovery-based science. 8

Kyoto Encyclopedia of Genes and Genomes (KEGG) A collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. 123

linear discriminant analysis (LDA) Statistical pattern recognition technique. 10

Linear discriminate analysis effect size (LEfSe) A software package that aims identify biological markers from relative abundance microbiome census data. 123

major depressive disorder (MDD) Also known as depression. A mental disorder that causes a persistent low mood, low self esteem, and chronic anhedonia. 40

minimum entropy decomposition (MED) A sequence clustering algorithm that partitions high-throughput sequencing data into ecologically meaningful and phylogenetically homogeneous units. 63, 72

multidimensional scaling (MDS) A set of related ordination techniques for visualising complex data. 10, 11

multilayer perceptron (MLP) A type of feedforward artificial neural network. 83, 88–90

Negative Predictive Value (NPV) Proportion of true negative results. 134

omics A field of biological research that ends with -omics. For example, genomics involves the study of the genome. In molecular biology the -ome suffix refers to "all constituents considered collectively". 1

operational taxonomic unit (OTU) Literally "the thing(s) being studied". Typically used to describe marker gene sequence reads clustered at a similarity threshold to approximate a bacterial taxonomy (e.g. 97% similarity is considered equivalent to a bacterial species). 58–63, 65, 66, 68, 70–72, 75, 76, 78, 79, 82, 84, 91, 96, 111, 113, 160

paediatric crohn's disease activity index (PCDAI) A subjective criteria that stratifies the severity of Crohn's disease in paediatric patients. 82

Permutational multivariate analysis of variance (PERMANOVA) A non-parametric multivariate statistical test. 121

Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) Bioinformatics software package that predicts metagenome functional content from marker gene survey data. 77, 78, 80, 82, 86, 87, 123, 133, 176

polymerase chain reaction (PCR) A technique that can create multiple copies of a chosen DNA sequence. This is used to amplify the marker genes to enable sequencing and bioinformatic analysis. 53, 56, 57

Positive Predictive Value (PPV) Proportion of true positive results. 134

principal component analysis (PCA) A statistical method that transforms a set of features into a subset of linearly uncorrelated features (principal components), often used as a dimensionality reduction technique. 10, 33

probiotics A substance that stimulates the growth of microorganisms (particularly those with beneficial properties). 108

psychobiotic Live bacteria that when ingested confer mental health benefits through interactions with commensal gut bacteria. ix, 4, 43, 108, 175, 177

Quantitative Insights Into Microbial Ecology (`QIIME`) A bioinformatics pipeline for performing microbiome analysis from sequenced amplicon data. 60, 61, 76, 78, 79, 84

quantitative polymerase chain reaction (qPCR) A molecular biology laboratory technique that monitors the amplification of targeted DNA sequences in real time. 91

Random Forest Ensemble statistical pattern recognition technique. 73, 82, 88, 90, 92–94, 96, 101

Recursive Feature Elimination (RFE) A feature selection method. 86, 92, 93

robust The consistency of feature selector output when small changes are made to input data (e.g. by adding or removing a sample). In this work, robust biomarkers were identified for inflammatory bowel disease in Chapter 4 using aggregating ensemble feature selection. Additionally, the robustness of the rough set characterisation process was investigated in Chapter 6. 74

robustness *See* robust 73

Robustness-Performance Trade-off (RPT) A variant of the F1-score (F-measure). 94, 96

Rough Set Theory (RST) First described by Pawlak, a rough set is a pair of "crisp" (conventional) sets: a lower approximation set and an upper approximation set. 5, 6, 31, 137, 138, 165

Self-Organising Map (SOM) A type of artificial neural network; a pattern recognition technique. 11, 118, 119, 137, 174

Sequence Read Archive (SRA) Bioinformatics database that provides a public repository for high-throughput DNA sequence data. 75

Short Chain Fatty Acid (SCFA) Fatty acids with between 2 and 6 carbon atoms. 172, 178

Sparse Correlations for Compositional data (`SparCC`) A similarity-based network inference tool that can tolerate sparse and compositional data. 116, 122, 134

Super Self-Organising Map (sSOM) A type of multimodal artificial neural network; a pattern recognition technique. ix, 109, 118, 119, 124, 134, 135

support vector machine (SVM) A statistical pattern recognition technique based on a separating hyperplane. 18, 21, 73, 83, 85, 86, 88–90, 93, 94, 96, 101

Synthetic Minority Over-sampling Technique (SMOTE) An oversampling algorithm that generates synthetic data to mitigate class imbalance. 92

thymine One of the four main nucleobases found in DNA. Pairs with adenine. 81

ulcerative colitis (UC) A chronic condition where the large intestine becomes inflamed. 39, 74, 76, 77, 83, 88, 91, 94, 96, 100, 171

zero-radius operational taxonomic unit (zOTU) An operational taxonomic unit that contains identical elements, a proposed name for amplicon sequence variants. 63, 79

DECLARATION

I hereby declare that with effect from the date on which the thesis is deposited in Research Student Administration of Ulster University, I permit:

1. the Librarian of the University to allow the thesis to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or for inclusion within the stock of another library.

2. the thesis to be made available through the Ulster Institutional Repository and/or EThOS under the terms of the Ulster eTheses Deposit Agreement which I have signed.

Benjamin Wingfield

INTRODUCTION

> Where shall I begin? Which of all my important nothings shall I tell you first?
>
> Jane Austen

Bioinformatics and computational biology are challenged by the growth of extremely complex, highly dimensional, and noisy data gathered from a range of sources (Holzinger et al., 2014). Huge quantities of omics data from a vast array of fields (including but not limited to genomics, metagenomics, proteomics, transcriptomics, and metabolomics) are generated each day, and it is estimated that by 2025 omics data will present some of the most demanding computational challenges for data acquisition, storage, distribution, and analysis (Stephens et al., 2015). The generated data provides opportunities to enable the generation of data driven hypotheses and aid the development of stratified medicine. However, extracting useful knowledge from the mountains of generated data (in a process known as knowledge discovery) is not an easy task, and a variety of approaches are required to do so. Knowledge discovery can be enabled by applying techniques such as statistical models, Artificial Intelligence (AI), machine learning methods, and Computational Intelligence (CI) methods. The knowledge discovery process consists of a series of steps with the ultimate goal to transform data into knowledge (Fayyad et al., 1996):

Data selection The process of identifying a dataset for analysis and selecting a data subset for data mining if appropriate;

Data pre-preocessing Organising and tidying (Wickham et al., 2014) information to remove outliers, perform data normalisation, and mitigate missing data;

Data transformation Data are made appropriate for data mining via transformations (e.g. applying summary statistics or dimensionality reduction techniques)

Data mining "The process of discovering useful patterns and trends in large data sets" (Larose and Larose, 2014);

Evaluation Interpreting the output of data mining (e.g. extracted data visualisation).

Common tasks for data mining include (Larose and Larose, 2014):

- Describing patterns and trends in data;
- Approximating a categorical target variable from a larger data set (classification);
- Approximating a numeric target variable from a larger data set (regression);
- Predicting future events (e.g. the share price of a company in 3 months);
- Clustering observations into similar groups;
- Identifying association rules (finding features that co-occur).

CI and machine learning provide powerful tools for extracting knowledge from data and have been successfully applied to many domains for knowledge discovery such as sociodemographic analysis and financial market analysis (e.g. identifying if a payment is fraudulent; Larose and Larose, 2014). However, to date microbiome census data (described below) have rarely been modelled using approaches other than standard machine learning algorithms.

Microbiota across the human body have been implicated in a vast number of localised and systemic diseases over the past decade, including colon cancer, rheumatoid arthritis, and Inflammatory Bowel Disease (IBD). Microbiota are defined as "the assemblage of microorganisms present in a defined environment" (Marchesi and Ravel, 2015). The term microbiome refers to "the entire habitat, including microorganisms, their genomes, and the surrounding environmental conditions" (Marchesi and Ravel, 2015). Evidence first presented in animal studies suggested that altering the microbiome could influence host behaviour. Further work developed this empirical evidence into the microbiome-gut-brain axis theory, which describes the complex signalling events that occur between the central nervous system, endrocrine and immune systems, the enteric nervous system, and the gastrointestinal microbiome.

Research has linked the microbiome-gut-brain axis to depression and other psychiatric disorders (Foster et al., 2017). Despite decades of research it is unclear how depression originates: no single biological mechanism or environmental factor has been shown to fully explain the aetiology of depression, and to date no empirical diagnostic tests are currently in clinical use. The microbiome has been referred to as the second genome of the human body: the number of bacterial cells associated with the human body greatly outnumbers the amount of human cells present. Thus, the microbiome is a promising area to identify new markers for disease.

Attempts have been made to identify microbiome dysregulation of the gut-brain-axis in association with depression in preclinical studies. One of the core

pathways that is thought to link the microbiome-gut-brain axis and depression is the "leaky gut" phenomenon. Stress is thought to cause the epithelial barrier of the gastrointestinal tract to become compromised, causing an increased translocation of bacterial cells across the mucosal lining. The translocated bacterial cells then interact with the enteric nervous system and immune cells, activating an immune response that increases the production of inflammatory mediators. The resultant inflammatory response contributes to depression, which has close links with chronic inflammation. Despite a large amount of animal work supporting the microbiome-gut-brain axis theory, limited and conflicting preclinical evidence is apparent in humans. To study the gut microbiome, faecal samples must be collected, which can be a challenging process for large epidemiological studies. Challenges include sample collection, processing, transportation, and subject recruitment barriers. The oral microbiome can be investigated via the collection of saliva. Saliva is a cost effective non-invasive biomarker source. Despite the oral microbiome being part of the gastrointestinal microbiome, the oral microbiome has not been investigated to date for associations with depression.

The oral microbiome is one of the most diverse microbiomes in the human body, and influences the microbiota found in the rest of the gastrointestinal tract. Alterations to the oral microbiome have been linked to both oral and systemic diseases with an inflammatory aetiology such as IBD and neurological diseases such as Alzheimer's disease. Three salivary glands are the source of nearly 90% of saliva fluid, which have the potential to absorb blood based biomarkers of disease. This suggests that saliva can contain important disease information. Therefore charting the oral microbiome for links to depression is a promising area for further research. Such work could potentially provide new insights into depression aetiology, and help to identify novel diagnostic and therapeutic response biological markers. CI provides a range of tools that can identify such biological markers. In addition, CI approaches can be applied to gain a greater understanding of complex microbial communities and the role they play in disease.

1.1 Objectives of the thesis

The aim of this research is to develop computational models of microbiomes across the gastrointestinal tract in order to investigate the mechanisms that are involved in the aetiology of disease, with a focus on depression. The first models are applied to IBD data, as many large publicly available datasets were available while analytical models were developed and applied. During this time, in collaboration with the Northern Ireland Centre for Stratified Medicine, an oral microbiome dataset was collected, containing control and depressed subjects' data for analysis and further study. Diseases linked to the microbiome are often systemic and have

an unclear aetiology. Both diseases are currently difficult to diagnose: there are no empirical tests for depression in clinical use, and IBD requires invasive colonoscopy. Gaining a better understanding of these diseases will advance treatment by assisting clinicians with disease identification. In addition to prediction, models can inform the underlying aetiology of disease. Such insights can be used to deliver novel treatments from this understanding, including psychobiotics. To achieve the aim of this thesis the following objectives have been determined:

1. Review computational approaches including CI and machine learning that have been applied for knowledge discovery from biological data;
2. Review microbiome literature to identify how the microbiome is thought to be linked with diseases (with a focus on depression), how microbiome census data are created, and CI applications to microbiome census data;
3. Identify methodologies that overcome current limitations in the application of computational models to microbiome census data;
4. Develop computational models that predict IBD from microbiome census data and enable knowledge discovery;
5. Using AI and CI techniques, identify associations between the oral microbiome and depression in a cohort of young adults;
6. Develop an approach that could characterise microbial environments while preserving data semantics that are destroyed by standard normalisation procedures.

1.2 Thesis contributions

The research outlined in this thesis provides novel contributions to the area of microbiome research. The work has been peer reviewed in a published conference paper (Wingfield et al., 2016) and journal paper (Wingfield et al., 2018c). Furthermore, the work has contributed to the development of two other papers that are under preparation for submission to peer reviewed journals (Wingfield et al., 2018b; Wingfield et al., 2018a). The primary contributions of this thesis are:

1. The extension of existing predictive models for IBD by the development of a hybrid model;
2. Generating robust microbial biological markers (biomarkers) for IBD with ensemble feature selection (EFS);

3. Identifying a range of alterations in the oral microbiome of a depressed cohort;
4. Accurately predicting depression from a saliva sample;
5. Developing and applying Rough Set Theory (RST) characterisation to microbiome census data gathered from depressed cohorts which reproduces empirical findings and delivers novel insights regarding the microbiome-gut-brain axis.

1.3 Outline of the thesis

Chapters 2 through to 6 outline the contributions of this work in detail. Chapter 7 provides a conclusion and proposes future work. A brief summary of Chapters 2 through to 7 is provided below:

Chapter 2 provides a critical review of CI approaches for knowledge discovery from biological data and identifies their limitations. Additionally, applications of CI to stratified medicine are reviewed and assessed for their ability to overcome deficiencies in the microbiome research knowledge base.

Chapter 3 reviews human microbiome research, the role the microbiome plays in disease, and the applications of CI to microbiome research. An outline of the different stages of a microbiome experiment is provided due to the interdisciplinary nature of the topic. The outline begins with extracting DNA from environmental samples and covers each stage through to the processing of sequenced genomic data to produce microbiome census data. A huge variety of bioinformatics algorithms exists for the generation of microbiome census data, and a critical review of microbiome census data paradigms and algorithms is provided. Furthermore, the links between depression and human microbiomes are thoroughly explored via a review of the microbiome-gut-brain axis.

Chapter 4 outlines the development of a hybrid model for IBD prediction and the generation of robust microbial markers for IBD prediction. The hybrid model evaluates the predictive power of a wide range of microbiome data, including taxonomic, functional, and clinical data. The application of EFS identified a subset of bacterial taxa that could accurately predict IBD from stool in a large paediatric cohort.

Chapter 5 identifies the first documented alterations to the oral microbiome of a depressed cohort of young adults using a variety of

microbial ecology analysis techniques. Furthermore, multimodal CI algorithms are applied to enable the prediction of depression from a saliva sample. The results provide new insights regarding the microbiome-gut-brain axis theory and have the potential to have great impact on the microbiome knowledge base.

Chapter 6 describes a RST approach to characterise the oral and gut microbiomes of depressed cohorts. The work throughout Chapters 4 and 5 focused on prediction, but analysing predictive power is only a small part of microbiome census data analysis; RST provides a transparent way to transform data into knowledge. Furthermore, the application of RST provides a solution to an open research question regarding identifying an optimal normalisation technique for microbiome census data. The results of the characterisation are compared with previous empirical findings.

Chapter 7 concludes this thesis with a discussion of the research findings. The research outlined in this thesis is compared with the related literature to demonstrate how the contributions of this work have advanced the understanding of the microbiome-gut-brain axis and the development of microbiome census data analysis strategies. The chapter closes with an outline for proposals of future work which could extend the research presented in this thesis.

COMPUTATIONAL INTELLIGENCE FOR KNOWLEDGE DISCOVERY

> Any sufficiently advanced technology is indistinguishable from magic.
>
> Arthur C. Clarke

2.1 Introduction

This chapter provides an evaluation of Computational Intelligence (CI) approaches for knowledge discovery from biological data, decision and data fusion, and discusses applications of CI to stratified medicine. Stratified medicine aims to improve health care efficiency and efficacy by delivering the right treatment to the right patient at the right time. Advances in molecular biology have caused an exponential growth of biological information. New technology has democratised science by making experiments that generate vast quantities of data widely available, and most funding agencies require experimental data to be made available in publicly accessible archives. At the end of 2015 the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) stored nearly 80 petabytes (8×10^7 gigabytes) of biological data (Cook et al., 2015). Knowledge discovery is an important process for extracting meaningful information from vast quantities of data — the volume of biological data is projected to exceed astronomical data by 2025 (Stephens et al., 2015). CI is often used for the purpose of knowledge discovery in databases, as many CI algorithms are capable of dealing with uncertainty, vagueness, and incomplete data - all of which are common in complex biological data. A generalised pipeline for knowledge discovery in databases is presented in Section 2.2, and the background behind various CI techniques is discussed. Section 2.3 describes the concept of decision and data fusion. Decision fusion covers the concept of ensembles for learning, feature selection, and hybrid systems, and relies on combining the outputs of multiple weak models to outperform a single strong model. Data fusion covers supervised multi-modal classification, which can incorporate multiple types of data to create a more holistic model of a system. Section 2.4 provides a brief overview of CI techniques that have been applied to stratified medicine. In Section 2.5 the chapter concludes with a summary of findings from the literature.

2.2 The knowledge discovery process

Knowledge Discovery in Databases (KDD) is the process of extracting valuable intelligence from information. Chapters 4 and 6 rely on data gathered from databases; Chapter 5 uses data gathered from experiments carried out in collaboration with the Northern Ireland Centre for Stratified Medicine. The KDD process is identical for both approaches from the data cleaning stage forward (see Figure 2.1). The vast quantities of data made available on publicly accessible archives such as EMBL-EBI, described earlier, contain much undiscovered biologcal knowledge. The CI techniques described in this chapter are discussed in the context of a KDD pipeline, as the ultimate aim of this thesis is to extract new knowledge from complex biological data. The background of applicable CI techniques (e.g. rough set theory) will be also be covered. The data cleaning and processing stage of KDD is thoroughly discussed in Chapter 3 (specifically the process of calculating the abundance of bacteria from short DNA sequences). Data that represent the microbial communities that inhabit the human body are often highly dimensional. These types of data are common across many disciplines, and typically large amounts of the data are not useful for modelling the problem at hand. The terms redundancy and irrelevance are useful to describe such data (John et al., 1994). Redundant data are highly correlated with other data, and relevant data have predictive power. If relevant data are removed from a dataset the predictive ability of a model decreases. Weakly relevant data may contribute to the predictive power of a learning model in combination with other data. This section begins with a discussion about reducing the complexity of gathered data via a process known as dimensionality reduction.

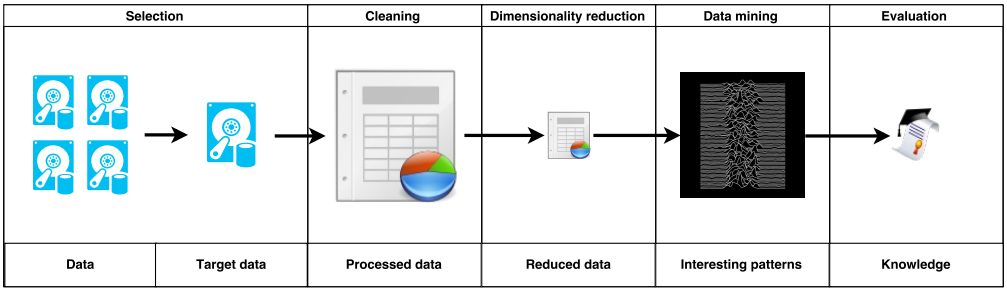


Figure 2.1: Knowledge discovery process (adapted from Fayyad et al., 1996).

2.2.1 Dimensionality reduction

It is common for many features in a high dimensional dataset to be irrelevant to a classification problem (Saeys et al., 2007). Superfluous features affect many aspects of the model fitting process. Mitigating the effects of the irrelevant features is a

common step in CI workflows for a variety of reasons, all of which are extremely useful for knowledge discovery from complex biological data:

- To improve the ability of researchers to interpret a fitted model;
- To reduce the time it takes to fit a model;
- To avoid the curse of dimensionality (Keogh and Mueen, 2011);
- To reduce overfitting (see Figure 2.2), which improves the ability of fitted models to generalise to unseen data.

The curse of dimensionality describes a range of problematic side-effects that occur as a result of analysing highly dimensional data (Keogh and Mueen, 2011). Broadly speaking traditional dimensionality reduction falls into three broad strategies (James et al., 2013):

Shrinkage Shrinkage is also known as regularisation. Shrinkage penalises complex models to reduce variance and overfitting. It does this by fitting a linear model to all f features. The size of the estimated coefficients is penalised according to a complexity parameter a. Shrinkage can cause estimated coefficients to be exactly zero. It is common to perform subset selection, explained below, by fitting a new model using only non-zero coefficients. Popular models that implement shrinkage include Ridge regression (Zou and Hastie, 2005), the Lasso (Meinshausen and Bühlmann, 2006), and the Elastic Net (Zou and Hastie, 2005).

Feature extraction This approach involves transforming the f features into an N-dimensional subspace, where $N < f$. An example of this approach is Principal Component Analysis. The transformed features are then used to fit a model.

Feature subset selection Feature subset selection is also known as feature selection. It aims to identify a subset of f features that are related to a response. The feature subset is used to fit a model.

Feature extraction, feature subset selection, and CI approaches such as rough set theory are described below. Shrinkage is implemented across many models. Although shrinkage improves the predictive performance and computational complexity of a model (see Figure 2.2), unless feature subset selection is used in tandem with shrinkage every f feature is still present in the final model. As this chapter is focussed on knowledge discovery from biological data, shrinkage is discussed in tandem with feature subset selection in this section.

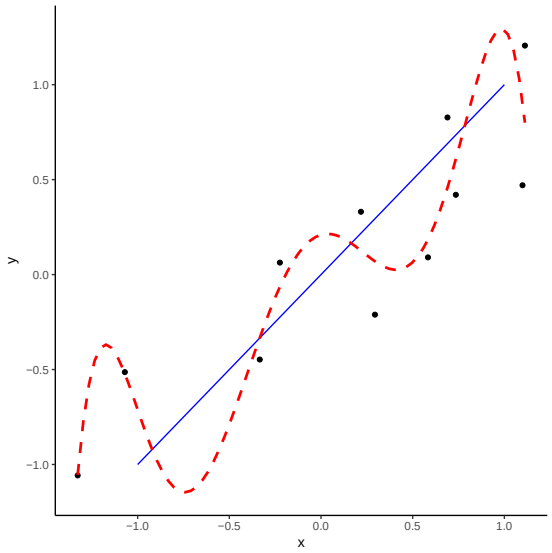


Figure 2.2: An example of overfitting. Overfitting occurs when a trained model performs very well on training data, but performs poorly on unseen data. The red dashed line is overfitted to the training data. The blue line is an ideal fit. Shrinkage penalises complex models (such as the red dashed line) to reduce overfitting.

Feature extraction via transformations

Dimensionality reduction can be achieved by processes that irreversibly transform a dataset from a high-dimensional space to a feature subspace. This data transformation process is also known as feature extraction. Linear transformation techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and multidimensional scaling (MDS) are popular for visualising the variance present in a microbial community (Dinsdale et al., 2013) or analysing gene expression data (Lee et al., 2007). The transformations try to reveal the Euclidean structure of the data. PCA transforms a set of features to a smaller set of uncorrelated features (principal components) that represent the largest amount of variance present in the data (Abdi and Williams, 2010). The underlying assumption of PCA is that high variance represents useful information, and transformed features with low variance can be removed. PCA is a straightforward process: one first calculates the eigenvectors of a covariance matrix of data instances X (e.g. samples). A subset of k eigenvectors that corresponds to the k largest eigenvalues (where k is less than the number of original dimensions d) is used to build a transformation matrix M. $X \cdot M$ yields the transformed feature subspace Y. LDA is similar to PCA but aims to find a feature subspace that maximises the separation between different classes (Fisher, 1936); PCA does not take into account class labels (e.g. disease

status, Balakrishnama and Ganapathiraju, 1998). MDS uses similarity measures to measure the proximities of data instances to visualise the structure of data (Borg and Groenen, 2005). The transformation procedure aims to preserve the proximities between data instances. A variant of MDS known as nonmetric MDS is widely used to visualise the structure of microbial communities (Kuczynski et al., 2010). Metric (classical) MDS is incompatible with incomplete, asymmetric, or ordinal data. One important limitation of the algorithms described above is that they cannot capture non-linear relationships that are common in data, particularly in complex biological systems (Lee et al., 2007).

Methods capable of capturing non-linear relationships include manifold-based extensions of linear techniques such as Isomap (Tenenbaum et al., 2000) and neural approaches such as the Self-Organising Map (SOM) (Kohonen, 1998) or auto-encoders (Masci et al., 2011). Manifold approaches assume that data are present in an embedded non-linear manifold with fewer dimensions than the original feature space (Hira and Gillies, 2015). Isomap builds a manifold by connecting points between K clustered data instances. The pairwise distance between each point is calculated as the geodesic distance. Isomap can identify non-linear patterns in data because the geodesic distances can represent the lower-dimensional manifold. If K is too small the geodesic distance cannot be accurately calculated in a sparse graph, and if K is too large "short circuit" edges can be introduced into the graph. Short circuit edges occur when points that are not geodesically close are joined; this fails to represent a manifold's topology. Taken together these problems mean that the Isomap approach can struggle to deal with sparse or noisy data (Balasubramanian and Schwartz, 2002), which makes the approach poorly suited for many complex biological datasets. Neural approaches to dimensionality reduction perform well but are vulnerable to poor generalisation (overfitting) and noisy data (Hira and Gillies, 2015). The background and applications of neural approaches are discussed thoroughly in the context of supervised learning in Section 2.2.2.

The key disadvantage of feature extraction processes is a loss of data interpretability. The irreversible data transformation process destroys the semantics of the original data. This is particularly important when knowledge discovery is the goal of an experiment: for example a domain expert will be more interested in the action of a subset of predictive genes than a subset of principal components (Abeel et al., 2010). Linear transformations are incapable of truly representing many of the phenomena present in complex biological systems (Lee et al., 2007), and the non-linear transformations described above struggle with noisy and sparse data (Balasubramanian and Schwartz, 2002), which are common properties of biological data. Therefore the rest of this subsection focuses on other types of dimensionality reduction, including feature selection and rough set theory approaches.

Feature subset selection

Feature subset selection algorithms are diverse, but fall into three main categories: filter methods, wrapper methods, and embedded methods (see Table 2.1). Semi-supervised and unsupervised feature selection algorithms are available for data that are fully or partially missing labels or for experiments that aim to investigate the structure of the data (Ang et al., 2016). In this section only supervised feature selection algorithms are discussed as the data used throughout this thesis are labelled.

Filter feature selection algorithms select features without building a model, and aim to reduce dimensionality by directly operating on the dataset with criteria such as correlation, redundancy, or information gain (Guyon and Elisseeff, 2003). Filter methods are quick and relatively simple to implement at the expense of model performance. Wrapper feature selection algorithms use a multi-objective optimisation approach to maximise model performance and minimise feature subset size (Guyon and Elisseeff, 2003). Wrappers search through the space of possible feature subsets using the constructed model as a performance measure (e.g. classification accuracy). The search method can range from simple (combinatorial) to complex (computational intelligence approaches such as genetic algorithms). Although wrapper methods provide better results than filter methods they have a high computational cost and tend to overfit (Guyon and Elisseeff, 2003). Embedded feature selection algorithms use internal data from the classification model to enable feature selection (e.g. feature rankings of Random Forests). Embedded methods provide a balance between computational complexity and performance (Guyon and Elisseeff, 2003), and often appear at the top of feature selection algorithm benchmarks. A comprehensive review of multiclass classification and feature selection algorithms found that embedded feature selection algorithms performed best across 8 metagenomic datasets (Statnikov et al., 2013). A ranked list of features generated by an embedded feature selector is often combined with a feature elimination procedure such as recursive feature elimination to generate a feature subset (Guyon et al., 2002). Regularisation can be thought of as a type of embedded feature selection. Models such as the Elastic net penalise (regularise) the coefficient of some features towards zero (Zou and Hastie, 2005). Features with a coefficient of zero can be considered irrelevant and removed from the model.

Table 2.1: Overview of feature selection techniques

Type	Advantages	Disadvantages	Examples
Univariate filter	Scales well to large data sets	Ignores interactions between features	Information gain (Hall and Smith, 1998)
	Independent of model	Independent of model	Euclidean distance
Multivariate filter	Faster than wrapper methods	Slower than univariate filter methods	Fast correlation based feature selection (Yu and Liu, 2004)
	Independent of model	Independent of model	
	Considers interactions between features		INTERACT (Zhao and Liu, 2007)
Deterministic wrapper	Simple to implement	Selection depends on model	Sequential forward selection (Aha and Bankert, 1996)
	Interacts with model	More likely to overfit	Sequential backward elimination (Aha and Bankert, 1996)
	Considers interactions between features	Can get stuck in local optima	
Randomised wrapper	Less likely to get stuck in local optima	Computationally expensive	Genetic algorithms (Yang and Honavar, 1998)
	Considers interactions between features	Selection depends on model	Particle swarm optimisation (Wang et al., 2007b)
	Interacts with model	More likely to overfit	
Embedded	Considers interactions between features	Selection depends on model	SVM weight vectors (Guyon et al., 2002)
	Faster than wrapper methods		Random Forests (Díaz-Uriarte and De Andres, 2006)
	Interacts with model		

Table 2.2: Decision system example

Features			Decision
Headache	Thirst	Regret	Hungover
Yes	High	Medium	Yes
No	Low	Low	No
Yes	Medium	Very High	Yes

Rough set theory

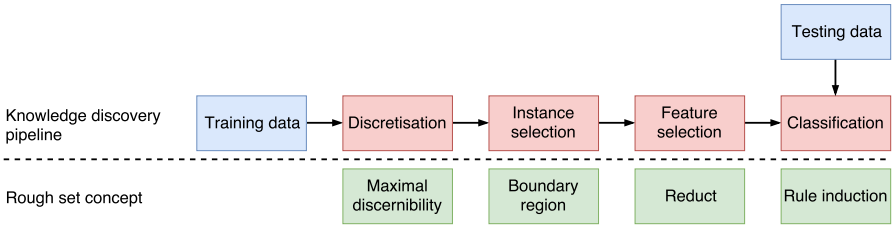


Figure 2.3: Overview of rough set theory concepts and their application to the knowledge discovery process.

Rough set theory (Pawlak, 2012) enables the modelling of imprecise or imperfect knowledge, and provides a range of concepts that are useful for knowledge discovery (see Figure 2.3). Rough sets have been applied to gene expression data for the purposes of dimensionality reduction and classification rule discovery (Dai and Xu, 2013). Using rough set theory it is possible to identify a subset of features (called a reduct) that are most informative, and non-informative features can be removed without any information loss (e.g. classification accuracy is not reduced). The reduct can be identified without any additional kind of data while simultaneously preserving the semantics of the data. The background to rough set theory is described and applications to dimensionality reduction are discussed in this section.

One of the most important aspects of rough set theory is the concept of indiscernability (Pawlak, 1998). Let $IS = (\mathbb{U}, \mathbb{A})$ be an information system - a data table where rows are objects and columns are features, where $\mathbb{U}$ is a nonempty finite set of objects (the universe of discourse), and $\mathbb{A}$ is a nonempty finite set of features such that $a : \mathbb{U} \mapsto V_a$ for every $a \in \mathbb{A}$ where V_a is the value set of feature a. Information systems can be extended into decision systems with the addition of decision features (e.g. class labels, see Table 2.2; Pawlak, 1998). A decision system can be defined as $DS : T = (\mathbb{U}, \mathbb{A} \cup \{d\})$ where $d \notin A$ is the decision feature. The elements of A are known as condition features. Let P be an arbitrary subset of

$\mathbb{A}$. With any $P \subseteq \mathbb{A}$ there is an associated indiscernability relationship IND(P) (Jensen and Shen, 2008):

$$\text{IND}(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P, a(x) = a(y)\} \tag{2.1}$$

where x and y are arbitrary objects of $\mathbb{U}$. The relation correponds to the equivalence relation if and only if the objects have the same vectors for the features in B. The indiscernability relation induces a partition in the universe $\mathbb{U}$, which is the set of equivalence classes generated by IND(P), and is denoted as $\mathbb{U}/\text{IND}(P)$ (Jensen and Shen, 2008):

$$\mathbb{U}/\text{IND}(P) = \otimes\{\mathbb{U}/\text{IND}(\{a\}) | a \in P\} \tag{2.2}$$

where:

$$A \otimes B = \{X \cap Y | X \in A, Y \in B, X \cap Y \neq \emptyset\} \tag{2.3}$$

Instances x and y are said to be indiscernible by features if and only if $(x, y) \in$ IND(P). Let X be a subset of $\mathbb{U}$, X can be approximated with the information contained within P by defining the P-upper and P-lower approximations, denoted as $\underline{P}X$ and $\overline{P}X$ respectively (Jensen and Shen, 2008):

$$\underline{P}X = \{x | [x]_p \subseteq X\} \tag{2.4}$$
$$\overline{P}X = \{x | [x]_p \cap X \neq \emptyset\} \tag{2.5}$$

the order pair $\langle \underline{P}X, \overline{P}X \rangle$ is a rough set of X. The boundary region can be defined via the lower approximation and upper approximation (Jensen and Shen, 2008):

$$\text{BN}_p(X) = \overline{B}(X) - \underline{B}(X) \tag{2.6}$$

The boundary region consists of objects that cannot be certainly classified into X in B (see Figure 2.4). A rough set is crisp if the boundary region is empty. Let P and Q be feature sets that induce equivalence relations over $\mathbb{U}$. The positive and negative regions can be defined as (Jensen and Shen, 2008):

$$\text{POS}_p(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{2.7}$$
$$\text{NEG}_p(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{2.8}$$

The negative region consists of objects that certainly do not belong to X, and the positive region consists of objects that are certain to belong to X.

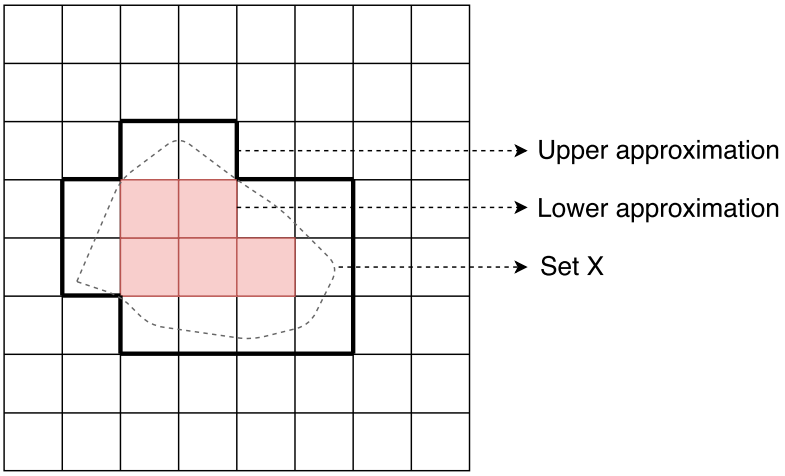


Figure 2.4: Lower and upper approximations of rough set X.

It is common for many of the features present in a decision system to be superfluous. Rough sets can be used to identify the minimal (most concise) representation of a decision system, called a reduct. Reducts aim to only keep features that preserve the indiscernability relation; there are often multiple subsets that do this but the most minimal are called reducts (Pawlak, 1998). First the dependency between features must be defined. From this the significance of individual features can be measured. A set of Q features depends on a set of P features if all feature values in Q are determined by the feature values in P. The degree k $(0 \leq k \leq 1)$ that Q depends on P $(P \rightarrow_k Q)$ is defined by (Jensen and Shen, 2008):

$$k = \gamma_p(Q) = \frac{|\text{POS}_p(Q)|}{|\mathbb{U}|} \tag{2.9}$$

if $k = 1$ then Q totally depends on P, if $0 < k < 1$ then Q depends partially on P, and if $k = 0$ then Q does not depend on P. The significance of a feature can be calculated by estimating the change in dependency when a feature is removed from the set of all features. A highly significant feature will cause a large change in dependency if removed. The significance of feature $x \in P$ on Q can be calculated by (Jensen and Shen, 2008):

$$\sigma_p(Q, a) = \gamma_p(Q) - \gamma_{p-\{a\}}(Q) \tag{2.10}$$

A reduct is the minimal subset R from the original feature set $\mathbb{C}$ so that for a given feature set D, $\gamma_R(\mathbb{D} = \gamma_{\mathbb{C}}(\mathbb{D}))$. In a minimal subset no features can be removed without affecting the dependency degree (Pawlak, 1996). Data sets can have many

reducts, and the collection of all possible reducts is defined as (Jensen and Shen, 2008):

$$R = \{X | X \subseteq \mathbb{C}, \gamma_X(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D}); \gamma_{X-\{a\}}(\mathbb{D}) \neq \gamma_X(\mathbb{D}), \forall a \in X\} \tag{2.11}$$

the intersection of all reducts is called the core. In the context of feature selection it is common to search reducts to identify a reduct of minimal cardinality (Jensen and Shen, 2008):

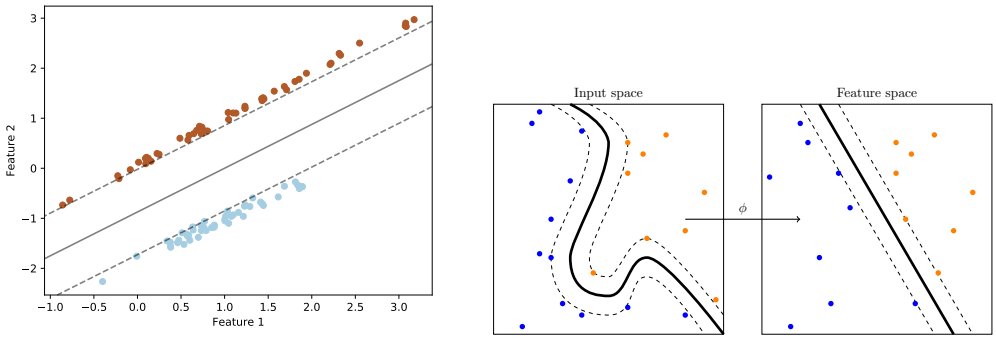
$$R_{\text{min}} = \{X | X \in R_{\text{all}}, \forall Y \in R_{\text{all}}, |X| \leq |Y|\} \tag{2.12}$$

The rough set procedures outlined above have been applied extensively to microbiome census data described in Chapter 3) throughout Chapter 6. The key motivation for applying rough set theory is that it no additional information about input data is required, such as statistical probability distributions and degree of membership in fuzzy set theory (Pawlak, 1996).

2.2.2 Data mining with supervised learning

Data mining techniques enable the search for valuable information from large volumes of data (Liao et al., 2012). Supervised learning is a data mining technique that includes classification, regression, and structured output learning. The aim of supervised learning is to identify a function $f : x \mapsto y$ that can generalise well to unseen data, where x is the feature space of predictors (the independent variables) used to make a prediction, and y is the response. Classification is a type of supervised learning problem where y is a discrete value (qualitative output), and regression is a type of supervised learning problem where y is a continuous value (quantitative output; Hastie et al., 2009). In structured output learning y is a structured object, such as the automated annotation of biological macromolecules (Jiang et al., 2014). Most models are capable of a combination of output types (e.g. classification and regression), but in this chapter only classification is described in detail, as the supervised learning problems approached in this thesis are all classification tasks.

Support Vector Machines



(a) (b)

Figure 2.5: (a) Linear Support Vector Machine classification of a two-class synthetic dataset showing the maximum margin separating hyperplane; (b) Kernel functions enable linear SVMs to solve non-linear classification problems by remapping features.

The core concept behind support vector machines (SVMs) is to construct a maximal margin hyperplane or set of hyperplanes in high dimensional space that can be used to separate two classes of data (see Figure 2.5a; Vapnik, 1998). A larger separation between two classes indicates improved generalisation capability in the model. SVMs are capable of learning classification problems that are not linearly separable with the application of kernel functions (Scholkopf and Smola, 2001). Once an SVM has been trained new objects are classified according to which side of the hyperplane they fall on. Kernel functions can be used to map data to higher dimensional space, where a hyperplane can be found by the SVMs (see Figure 2.5b; Scholkopf and Smola, 2001). Examples of popular kernel functions include polynomial, radial basis function, or sigmoid functions. SVMs are highly effective in high dimensional spaces, are computationally efficient, and work well in $p \gg n$ classification problems (where the number of features is much larger than the number of data objects), which are common in biological datasets (Statnikov et al., 2008). SVMs can also employ regularisation to improve the generalisation ability of the model by reducing overfitting. For these reasons SVMs are often considered to be "best of class" for DNA microarray classification tasks and benchmarks have confirmed that SVMs are superior to Random Forests for problems such as cancer diagnosis and clinical prognosis from gene expression datasets (Statnikov et al., 2008). SVMs are typically binary classifiers but can be extended to multi-class classification. A common approach is to use a series of one-versus-all classifiers,

where a single class is compared to all other classes binned into an agglomerated "all" class. This process is repeated for all unique classes (Noble, 2006). A different approach is one-versus-one multiclass classification, where a separate classifier is trained for each set of paired labels (Noble, 2006). Given N classes, one-versus-one classification trains $\frac{N(N-1)}{2}$ classifiers, while one-versus-all classification trains N classifiers. The one-versus-all paradigm is adopted where applicable throughout this thesis because of the computational expense of one-versus-one classification (Rifkin and Klautau, 2004).

Artificial Neural Networks and Deep Learning

Artificial Neural Networks (ANNs) are a computing paradigm that are capable of learning to perform a task by iteratively considering examples (Patterson, 1998). The design of ANNs is inspired by the connections between neurons in biological nervous systems (Basheer and Hajmeer, 2000). Biological nervous systems are capable of performing extremely complex tasks (e.g. pattern recognition) much faster than an electronic equivalent. Brains are often thought of as non-linear highly parallel computers, and the ultimate goal of ANNs is to mimic the processing capability of a brain (Jain et al., 1996). There are approximately 10^{11} neurons in the human brain (Herculano-Houzel, 2009), and each neuron can have thousands of connections to other neurons called synapses (estimates for the total number of synapses range from 10^{14} to 5×10^{14} in human adults; Drachman, 2005).

An ANN typically consists of a series of inputs, an input layer, a hidden layer, and an output layer (see Figure 2.6a; Jain et al., 1996). Each neuron is often fully connected to forward neurons; and each connection is weighted. Weight values are summed and passed to an activation function which defines an output that is passed to the output units (see Figure 2.6b; Jain et al., 1996). A learning algorithm controls how the weight of each connection is changed in response to newly presented data. ANN network architecture is varied, but a typical example is a backpropagated fully connected feed-forward network (see Figure 2.6a), also known as a multilayer perceptron (Noriega, 2005). Multilayer perceptrons are capable of approximating any function (Hornik et al., 1989) which makes them well suited for learning complex biological systems. ANN variants are widely applied for the purpose of data mining (Liao et al., 2012), including the reconstruction of regulatory gene networks from time series DNA microarray gene expression data (Ma and Chan, 2007) and predicting the structure of microbial communities (Larsen et al., 2012). It is important to note that studying the architecture of a trained ANN will not provide any insight into the structure of the function being approximated (they are black box algorithms), which limits their suitability for knowledge discovery. "White box" algorithms, such as decision trees or rule-based expert systems, can provide insights into how the model makes a decision

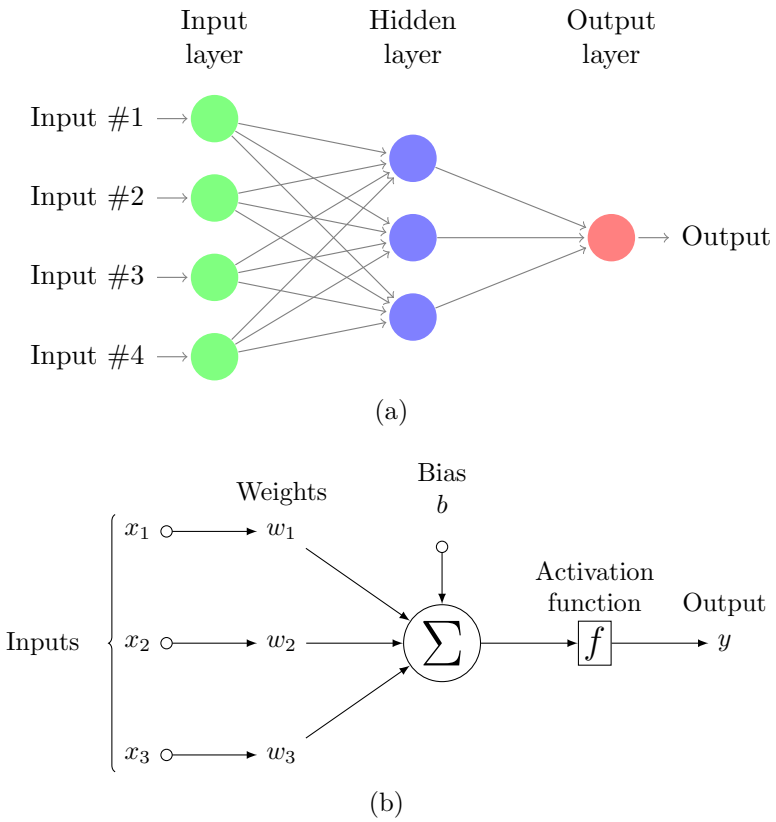


Figure 2.6: (a) A fully connected feed-forward ANN with a single hidden layer (a multilayer perceptron). (b) Overview of an artificial neuron model.

(Kononenko, 2001).

Deep learning is a recently developed paradigm that allows computational models composed of several processing layers to automatically learn the representation of input data (LeCun et al., 2015). Deep learning has caused breakthrough improvements to the performance of speech to text transcription (Hinton et al., 2012) and image recognition applications (Krizhevsky et al., 2012). The traditional machine learning paradigm required manual feature extraction from complex data such as images (represented as an array of pixel values; Guyon and Elisseeff, 2006). In deep learning feature extraction can take place over multiple layers, and the final output of a deep learning algorithm uses a combination of the layers to match objects. The core concept of deep learning is that the multi-layer feature extraction approach occurs automatically, with no human feedback, and the feature extractions are learned from the data (LeCun et al., 2015). A deep learning architecture simply requires a set of non-linear mappings: most deep learning applications use multi-

layer perceptrons with many hidden layers. This approach was not computationally feasible until it was implemented using graphics processing units, which reduced training times by up to 20 times (Schmidhuber, 2015).

Although deep learning has been widely applied to biomedicine (Mamoshina et al., 2016) it has not been implemented in this thesis due to the limitations of the paradigm. Deep learning models are black boxes. It is impossible to understand complex relationships that are present in biological data using a deep learning model. In addition, large datasets are a prerequisite for deep learning. Fields where deep learning excel typically have tens of thousands of examples (e.g. CIFAR-10 has 60,000 examples; Krizhevsky and Hinton, 2009). Training a deep learning model on a small dataset (microbiome datasets larger than a thousand samples are rare) could easily lead to overfitting. Other limitations include high computational costs and the requirement of extremely complex analysis pipelines.

Decision trees

Tree-based methods segment the feature space of predictors into a group of simpler regions (see Figure 2.7; Safavian and Landgrebe, 1991). Predictions are made by taking the average (mean or mode) of the training samples that belong to the predicted region. Trees are simple to implement and it is easy to understand why a model has assigned a particular output to new data, but generally speaking they do not perform as well as other models such as ANNs or SVMs (James et al., 2014). The uncompetitive performance has lead to the development of ensembles of de-correlated decision trees, called Random Forests, which are described further in Section 2.3.1 (Breiman, 2001). Decision trees are capable of learning non-linear classification problems (see Figure 2.8). Building a classification tree consists of two steps (James et al., 2014):

1. Divide the feature space of predictors (the set of possible values for $X_1, \ldots, X_n$) into J disjoint regions ($R_1, \ldots, R_J$)
2. For each observation that falls into region R_J take the average of response values of training observations in R_J

The process used in the first step to divide the feature space of predictors is known as recursive binary splitting (James et al., 2014). Recursive binary splitting is a top-down greedy approach that begins at the top of the tree (where all observations belong to one region) and greedily splits the feature space with two new branches. Greedy splitting only considers the current node of the tree, even if it is theoretically possible to improve the overall tree by changing the splits at a later stage of the algorithm. The first stage of recursive binary splitting is to select predictor X_j and cutoff s so that splitting the feature space of predictors

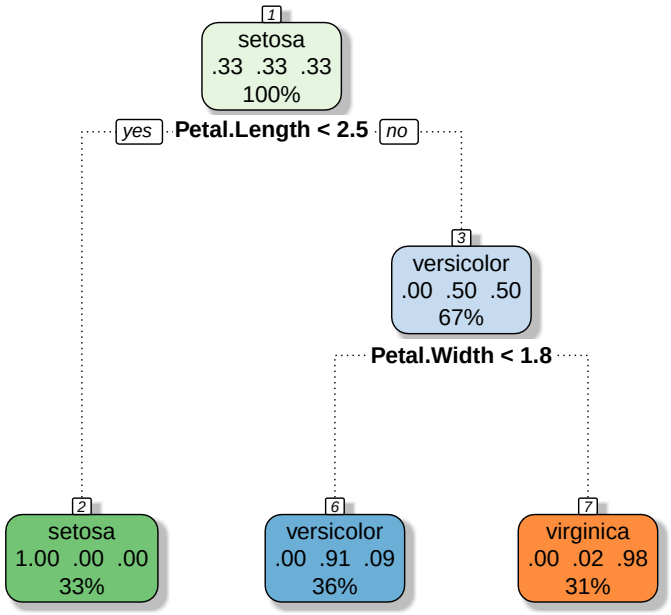


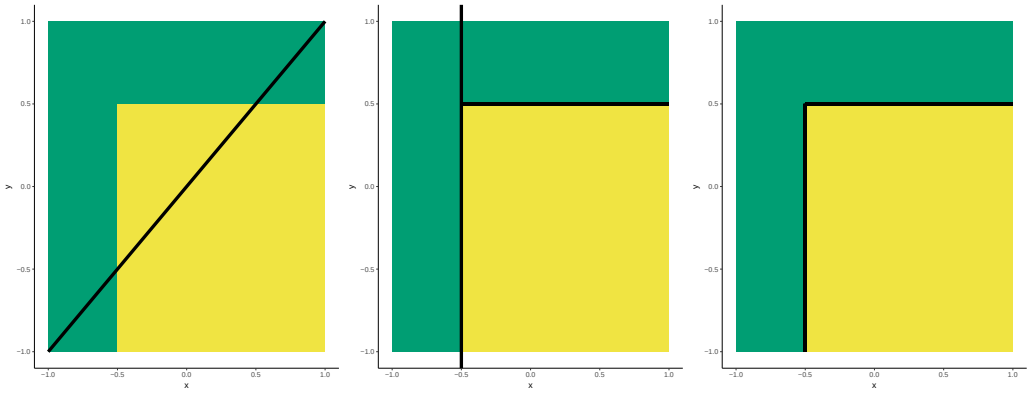
Figure 2.7: Decision tree fitted to the iris data set (Anderson, 1935) for predicting the species of a flower. Nodes show the classification, the probability of each class at that node, and the percentage of observations used at the node.

into the region $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ reduces the error rate. $\{X|X_j < s\}$ describes a feature space region where the value of X_J is less than s. During the tree-growing process the Gini index is typically used to measure the variance across K classes (James et al., 2014):

$$\text{Gini} = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \tag{2.13}$$

where $\hat{p}_{mk}$ is the proportion of observations in the m-th region from the k-th class. The Gini index measures the purity of a node because a small Gini index shows that the majority of samples are assigned to a single class. The process is repeated to identify the best predictor and best cutoff for splitting the data to minimise the classification error rate for each putative region, but only one of the regions is split instead of the entire feature space. The process continues until a breakpoint is reached, typically a minimum number of observations that must be present in a node for a split to be attempted (Apté and Weiss, 1997).

As the number of features increases, the tendency for decision trees to overfit also increases. This is because the tree rapidly becomes very complex. A common strategy to overcome this limitation is to prune a complex tree to obtain a simpler subtree (Apté and Weiss, 1997). The subtree will have lower variance at the cost



(a) Linear SVM (b) Decision tree (c) Multilayer perceptron

Figure 2.8: A two-dimensional two-class classification example. Classes are indicated by the shaded regions. Linear models fail to model the non-linear decision boundary.

of bias. A common metric for pruning the tree is the classification error rate E, which is the fraction of training observations of a region that do not belong to the most common class, and is given by (James et al., 2014):

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{2.14}$$

Rule-based expert systems

Rule-based systems are one of the simplest forms of CI (Grosan and Abraham, 2011). Rule-based systems use `IF-THEN` rules to represent and encode knowledge into a computer system. The rule definitions depend entirely on the task the expert system is built to do (see Table 2.3). Rule-based systems are capable of encoding a domain expert's knowledge and experience of a niche topic into an automated computer system. Rule-based systems consist of the following elements (Grosan and Abraham, 2011):

Fact set A collection of data and conditions (i.e. features) that is relevant to the starting state of the expert system. In the fact headache = yes, headache is the data and the condition is yes.

Rule set The rule set contains all possible actions that should be taken for a particular problem. `IF` relates rules to the fact set, and `THEN` relates to actions.

Stopping criterion Once a solution has been found (if one can be found) the expert system should terminate to avoid infinite loops.

Table 2.3: Example rule-based system.

Facts		
Data	Conditions	Rules
Influenza diagnosis		**Premises**
headache	true, false	`IF` headache true `AND`
temperature	$< 38, \geq 38$ celsius	`IF` temperature ≥ 38 `AND`
muscle pain	low, medium, high	`IF` muscle pain medium `OR`
		`IF` muscle pain high
		Conclusion
		`THEN` influenza is true

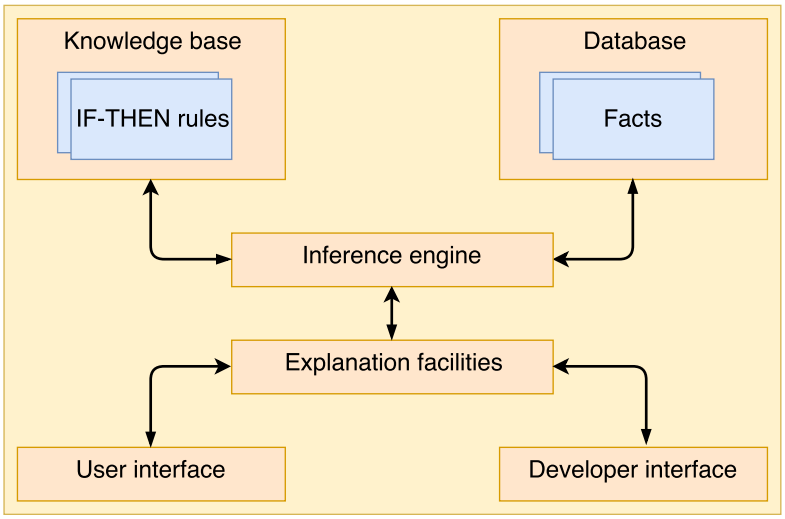


Figure 2.9: Structure of a rule-based expert system.

`IF-THEN` rules consist of the premise (antecedent) and the conclusion (consequent), and are stored in the knowledge base (see Figure 2.9; Liao, 2005). The facts which correspond to the `IF` rules are stored in the database. The inference engine represents all of the protocols that process the knowledge base to identify information requested by a user. The explanation facilities analyses the structure of an answer provided by the expert system to provide an explanation of why the inference engine has provided particular information. The user interface provides bidirectional communication between the expert system and a user, which typically consists of queries and answers provided via a graphical user interface. The developer interface provides bidirectional communication between the expert system and a knowledge engineer (a computer scientist that works with a domain expert

to represent knowledge in the expert system; Grosan and Abraham, 2011).

The key advantage of expert systems is that any decisions made by the system can be queried and the rationale behind the decision can be provided (Efraim et al., 2001). In the context of knowledge discovery, the key disadvantage of rule-based systems is that domain knowledge is not always easily encoded into rules, and that highly dimensional data causes exponential growth in the number of rules (Liu et al., 2000). A large number of rules will hinder the computational performance of an expert system, and make it more difficult to interpret the decision of an expert system. The combination of rule-based systems with rough sets can simultaneously solve both problems, via the application of reducts and a process known as rule induction to automatically extract rules from data.

2.3 Decision and data fusion

It is common in CI to create a strong model from a combination, or ensemble, of weaker models (Qi, 2012). Ensemble methods reduce some of the main causes of error for learning algorithms: noise, bias and variance. This approach is most commonly applied to the output of models, and a wide variety of procedures are available to tune the process depending on the goals of the model, which are described in Sections 2.3.1 and 2.3.2. Often a committee is formed from the combined models, and a vote of the output is tallied. Depending on the process votes can be weighted or unweighted. There are many different weighting systems: sensible strategies including linking vote weight to the confidence of a fuzzy or probabilistic model or to the performance of a model. The tallied votes can be averaged (for regression) or a simple majority determined (for classification). In Section 2.3.3 the decision fusion concept is expanded to combine the weak output of multiple feature selectors in order to find a subset of strong and stable features that can be used to fit a model.

A number of reasons have been proposed for an ensemble of weaker models outperforming single models (Dietterich et al., 2000). The classification problem might be able to be solved by different but equally optimal hypotheses: an ensemble reduces the risk that the model makes a decision that uses a non-optimal hypothesis. The hypothesis space is expanded by using multiple models, and a single model is unable to represent the true function. Ensembles reduce the risk that a model will get stuck in local optima, which can give a better approximation of the function being learned.

The same idea can be applied to combine the input of models. Consider the following classification problem: a model is trying to classify the genre of a song from a music video. The video data have three different modalities: the sequence of video frames, the audio file, and the lyrics encoded as subtitles (for karaoke

fans). The modalities are visual, audio, and textual, respectively. It is likely that the predictive performance of the model will be improved by combining the three different types of data. This approach is useful for many different types of experiment: it is unlikely that a single type of data can be used to model complex natural phenomena. A naïve approach would be to concatenate the three types of data into a single input matrix. A more elegant approach is to leverage the knowledge of the domain experts by combining and weighting the input of each data type independently known as multimodal data fusion, discussed in Section 2.3.4.

2.3.1 Decision fusion with ensemble multiclassifiers

Bagging and boosting

Bagging, or bootstrap aggregation, is a technique that has been widely applied to improve many learning models, including decision trees. Given a set of observations $\{z_1, \ldots, z_n\}$ with variance σ^2, the mean variance is $\frac{\sigma^2}{n}$. Hence averaging a set of observations reduces the variance of a learning model. Having many different independent training sets is unrealistic for many supervised learning experiments. Different training sets can instead be built by repeatedly sampling a single data set with replacement (Bühlmann, 2012; creating B bootstrapped training sets):

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x). \tag{2.15}$$

where $\hat{f}^1(x), \ldots, \hat{f}^B(x)$ is a set of learning models. Although bagging improves the performance of decision trees it is harder to interpret a bagged model (because many different models must be simultaneously interpreted). Increasing the number of bags does not increase the risk of overfitting; it is important to balance bagging test error and computational complexity when deciding how many bags to use.

The AdaBoost M1 algorithm (Freund and Schapire, 1997) is a popular implementation of the boosting paradigm, which can reduce bias and variance (see Figure 2.10). Given a two-class classification problem the error rate produced by classifier $G(X)$ (let X be a vector of input features) on a training sample is (Hastie et al., 2009):

$$\text{error} = \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq G(x_i)) \tag{2.16}$$

Boosting produces a set of weak classifiers $G_1(x), \ldots, G_M(x)$. The output predictions for the set of classifiers are combined through a weighted majority voting

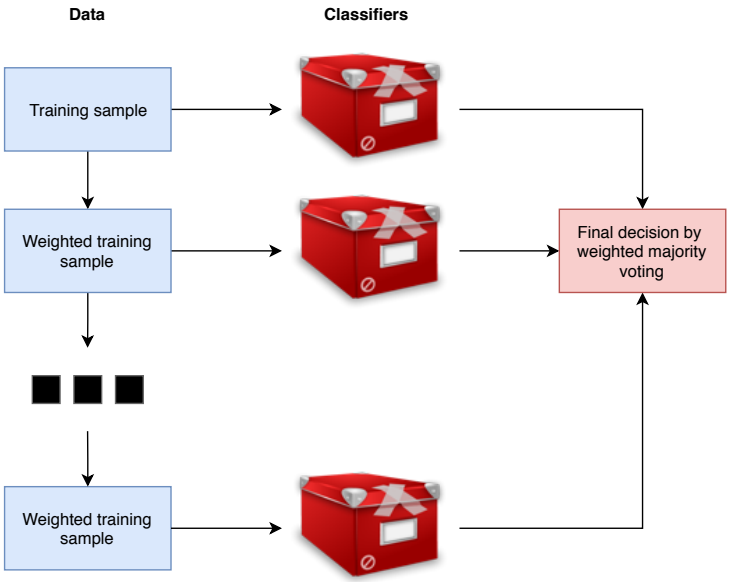


Figure 2.10: Boosting. Classifiers are trained on weighted samples. The output of each classifier is combined via weighted majority voting to come to a final prediction.

scheme (Hastie et al., 2009):

$$G(x) = \text{sign}\Big(\sum_{m=1}^{M} \alpha_m G_m(x)\Big) \tag{2.17}$$

where $\alpha_1, \ldots, \alpha_m$ and classifier weights $(G_m(x))$ are calculated by the boosting algorithm. The aim of the process is to give more accurate classifiers higher weights. At each stage of the boosting process weights $w_1, \ldots, w_N$ are applied to each training sample. Weights are initialised to $w_i = 1/N$ (unweighted). For each stage of the boosting process $m = 1, \ldots, M$ the weights are modified and the classifier is reapplied to the weighted samples. Samples that were missclassified by classifier $G_{m-1}(x)$ have their weights increased, and correctly classified samples have their weights decreased. Thus samples that are difficult to classify become more important and each classifier in the sequence is focused on these examples.

Random Forests

Random Forests are an ensemble of decorrelated decision trees (Qi, 2012). The decorrelation procedure involves a small adjustment to recursive binary splitting. At each split point a random sample of n features is taken, and the splitting

procedure is only permitted to consider features present in the n feature subset (typically $n \approx \sqrt{f}$, where f is the total number of features). By preventing the model from considering the majority of available features the decorrelation process prevents a single dominant feature or group of strong features from creating very similar decision trees across the bags. The benefits of bagging are inhibited for strongly correlated decision trees (variance reduction). It is important to note that if $n = f$ then there is no difference between a random forest model and standard bagged decision trees. Random Forests do not require cross-validation to estimate their generalisation ability (Qi, 2012). When each tree in a Random Forest is generated from different bootstrap samples, approximately a third of samples are left out of and not used to construct a tree. By predicting the class of each unused sample for each tree and calculating the proportion of misclassification errors the "out of bag" error can be estimated.

Random Forests have the ability to measure the importance of features, which is particularly useful for embedded feature selection (Díaz-Uriarte and De Andres, 2006). For each tree in a forest each "out of bag" sample can be predicted and the number of votes for the correct class summed. After the value for each feature m in each out of bag sample is randomly permuted and fresh predictions are made, an error rate can be estimated by subtracting the number of correct votes in the permuted data from the number of correct votes in the original data. The average of the error rate across all trees is called the raw importance score for feature m. By measuring the importance score of all features a ranked list of important features can be assembled.

2.3.2 Decision fusion with ensemble hybrid methods

In ensemble hybrid methods different types of models are combined to form a new system, which will be referred to as a meta-model. This is in contrast to ensemble multiclassifier method, in which the same model is repeatedly applied to different data resamples. There is a variety of strategies for implementing a hybrid meta-model, but they fall into two broad categories (Woźniak et al., 2014). The most popular is parallel hybrid fusion, in which a meta-model combines multiple models working on the same classification problem. Serial hybrid fusion implements different types of models on different classification problems. This attempts to decompose complex problems into a series of solvable modules (Woźniak et al., 2014). Consider the following classification problem: determining the presence and subtype of a complex disease. Traditionally multi-class classification would be conducted (e.g. absent, subtype_1, ..., subtype_n). An alternative approach would be to be decompose the classification problem into two steps: determining disease presence (two-class classification) and determining the subtype of the disease (multi-class classification). This is particularly useful for complex diseases, where

different models might perform better on different problems due to the inherent properties of the model. For example, a linear SVM might perform well on highly dimensional data on one classification problem. Multilayer perceptrons can approximate any function, including non-linear classification problems. A neural network could perform better on less highly dimensional data for a different classification problem. According to the no free lunch theorem, there is no universally best classification model (Wolpert, 2002). Serial decision fusion permits decomposing complex problems into a series of easier steps.

2.3.3 Aggregating Ensemble feature selection

Feature selection is particularly useful for knowledge discovery from high-dimensional datasets. Domain experts will be interested in investigating a ranked list of features in a top-down iterative fashion to gain new insights into the problem that the model is attempting to learn. When knowledge discovery is a priority, an important aspect of feature selection to consider is the robustness of a feature selection algorithm. The stability of a feature selection algorithm is defined as the variation in feature subset output caused by small changes to input data (Saeys et al., 2008). Small changes can occur at the instance level (e.g. adding or removing a sample) or the feature level (e.g. by adding noise). For example, an unstable feature selection algorithm can return a completely different subset of features if an instance is removed from a dataset. Domain experts will have more confidence in stable feature subsets, as further analysis is usually costly (particularly for biological data).

The robustness of a feature selection algorithm can be estimated via a similarity based approach. By resampling a dataset and repeatedly performing a pairwise comparison of the feature selection algorithms output, a global similarity measure can be calculated. The outputs of a robust feature selection algorithm to resampled data will be more similar compared to the outputs of a weak feature selection algorithm, and the similarity measure will be higher. A global similarity measure can be defined as (Saeys et al., 2008):

$$S_{\text{global}} = \frac{\sum_{i=1}^{k} \sum_{j=i+1}^{k} S\left(F_i, F_j\right)}{k\left(k-1\right)} \tag{2.18}$$

where f_i is the output of a feature selection algorithm applied to resample i (k resamples total) and $S(f_i, f_j)$ is the similarity measure between f_i and f_j. Once a robust feature selector has been selected a consensus based approach is used to combine the output of an ensemble of chosen selectors into a final list which is is useful for validation purposes. The resampling, comparison, and consensus paradigm is known as ensemble feature selection (Saeys et al., 2008) - the rationale of which stems from ensemble learning (where multiple models can be combined to

perform better than a single model). It is important to note that a single Random Forest model qualifies as a type of non-aggregating ensemble feature selection (EFS), because the final feature ranking is derived from analysing an ensemble of decision trees.

Robustness must always be considered in tandem with classification performance, so the output of feature selection algorithms must be paired with a classification model in order to evaluate the EFS strategy. This requirement means that embedded feature selectors have an advantage over filter and wrapper methods, as they combine feature selection and classification during training (they are less computationally expensive). A method to automatically balance robustness and classification performance is required when evaluating a consensus based EFS strategy. A variation of the F-measure - the harmonic mean between specificity and sensitivity (Vickery, 1979) - called the robustness-performance trade off (RPT) has been proposed (Saeys et al., 2008):

$$RPT_\beta = \frac{(\beta^2 + 1)\, RP}{\beta^2 R + P} \tag{2.19}$$

where R is the robustness (measured by a chosen similarity measure (Spearman rank correlation coefficient for ranked features in this thesis), P is the performance of the classifier (accuracy), and β is a parameter that balances the importance of robustness versus performance (typically 1 by default to give equal importance to robustness versus performance).

2.3.4 Multimodal classification

Multimodal data fusion is defined as the analysis of several data sets such that different data sets can interact and inform each other (Lahat et al., 2015). In many scientific fields information about a phenomenon can be recorded from different types of sensors, across multiple experiments, and in different conditions. Each of these different recording methods is referred to as a modality. Multimodality is particularly important for biological data because a single modality will rarely provide complete knowledge about a complex system. The motivation for multimodal data fusion lies in its ability to deliver a holistic model of a complex system and its ability to improve decision making (Lahat et al., 2015). Multimodal data fusion has been applied to the task of developing non-invasive diagnosis techniques. Multimodal prediction was used to identify patients that would progress from mild cognitive impairment to Alzheimer's disease with an accuracy of 73% (Ritter et al., 2015). The data modalities were diverse, including medical history (e.g. examinations, demographic, and neuro-physical tests), imaging data (e.g. MRI or PET scans), and laboratory data (e.g. cerebrospinal fluid examinations). Multimodal

classification is applied to enable holistic modelling of the oral microbiota and host to predict depression from the oral microbiome in Chapter 5.

2.4 Applications to stratified medicine

The rate of biological information gathering has increased massively over the past several decades. This has been driven by a wide variety of fields, including but not limited to, genome sequencing, protein expression, gene expression data, and metagenomics. CI algorithms are regularly combined with biological data for both bioinformatics and computational biology applications (Hassanien et al., 2008). Biological data are often imprecise and incomplete, which can violate the assumptions of standard statistical models; CI models have minimal prior assumptions (Lahat et al., 2015). A CI framework such as Rough Set Theory (RST) can handle uncertainty, vagueness, and missing data (Petit et al., 2014).

The goal of bioinformatics is to build software tools and methods to understand biological data (i.e. an engineering approach), while the goal of computational biology is to understand biological systems via the application of computational methods (often using bioinformatics tools). This thesis is concerned with using computational biology and CI techniques for the application of stratified medicine specifically. Stratified medicine has a variety of definitions, and can be split into two categories (Schleidgen et al., 2013):

- a holistic approach centred around individual patients;
- targeting treatment at specific population subgroups (e.g. based on the presence or absence of a particular gene).

The goal of stratified medicine is to identify the best treatment for each individual patient to maximise the benefit of treatment and to minimise any harmful side effects. Current non-stratified practice in medicine means that polypharmacy, the use of multiple treatments which can lead to the administration of more medications than are clinically required, is extremely common in the elderly population (Hajjar et al., 2007). Polypharmacy causes negative health outcomes, and stratified medicine is an approach that can mitigate this problem.

This section will review applications of CI to a wide variety of different types of data for the purpose of dimensionality reduction and data mining in stratified medicine. Using different types of information to reach a decision is a widely used approach in stratified medicine applications, as information gathered by a single method will rarely be able to completely describe complex biological systems or phenomena. A multimodal information fusion paradigm can be implemented with many CI techniques. Multimodal data are widely present in many fields such as

medical imaging, remote sensing, speech recognition, and omics (e.g. genomics, metabolomics, metagenomics etc.). Both multimodal and single mode approaches are discussed below.

Rough set theory has been widely applied for feature selection and knowledge discovery from complex biological data for stratified medicine purposes. Knowledge discovery with rough sets typically involves rule induction from discretised data, which enables easily interpretable descriptions of complex biological data. Domain experts are often more interested in understanding how the features are used by the model to predict a condition rather than solely optimising the predictive power of a model. Applications include describing artery damage after cannulation, analysing drug-induced changes to gene expression data, and the prediction of various cancers from gene expression data, described below. Rough set theory has been combined with an inductive learning approach to automatically acquire knowledge for an expert system (Azar et al., 2015). The expert system aimed to model artery damage that arose after cannulation of the radial and dorsalis pedis arteries from a set of clinical attributes in 46 patients (Azar et al., 2015). Rough set theory has been applied to gene expression data for the purpose of analysing drug-induced changes to gene expression profiles (Petit et al., 2014). The expression profiles of 17 genes were recorded for a variety of different drugs which were thought to cause phospholipidosis (the accumulation of phospholipids in tissue). The process was used to generate descriptive rules (not predictive: not enough data were present to independently test the predictive power of the rules).

Rough set classifiers have been used to identify a subset of biomarkers from gene expression data that can classify gastric carcinomas (Nørsett et al., 2004). The expression of 2504 genes was measured in tumour biopsies from 17 patients; a bootstrap t-test was used to identify a subset of differentially abundant genes (between 10 – 40). Rules were induced from the discretised (e.g. low, medium, or high expression) subset of differentially abundant genes for a variety of different classification tasks (e.g. growth patterns, remote metastastis, etc.). A cross-validated Area Under the Receiver Operating Characteristic (AUROC) of between 0.66 – 1.00 was reported for the six different classification problems. Due to the small size of the dataset the classifiers were validated by comparing the performance of the normal classifier against 2000 classifiers with randomly permuted class labels with a bootstrap t-test. Three of the classifiers were statistically significant ($p < 0.05$) and are likely to generalise well. A combination of rough sets and decision trees have been used to predict the location of the primary tumour in metastatic adenocarcinoma (Dennis et al., 2005) from the expression profiles of 27 genes. Identifying the site of the primary tumour is important to guide care and to improve the patient's prognosis. On unseen data the decision trees performed extremely well, with an accuracy of 88% for the prediction of seven different

primary tumour sites. The common rationale for applying rough sets to biological data lies in their ability to identify minimal feature sets (reducts) and generate human-interpretable rules from linguistic variables (discretised data). Additionally, the concept of approximation is useful for dealing with noisy data (biological data are often noisy). However, rough sets can only deal with discretised feature values - continuous feature values are common in the real world and in some circumstances it is preferable to use continuous values instead of linguistic variables. The limitations of rough sets are often overcome by combining rough set theory with other CI methodologies, such as evolutionary computing or fuzzy logic, described below.

Evolutionary rough sets have been used for feature selection (Banerjee et al., 2007). As mentioned previously, multiple reducts exist for any rough set. Identifying reducts is a nondeterministic polynomial time hard problem (NP-hard, brute forcing would take an unreasonably long time; Skowron and Rauszer, 1992), and some approaches have applied heuristics to find an optimal reduct for feature selection purposes because of this (Zhong et al., 2001). Genetic algorithms provide an alternative efficient search technique that works well in large solution spaces, based on the theory of evolution (Kumar et al., 2010). A multiobjective genetic algorithm has been applied to identify a reduct of minimal genes that can be used to predict cancer from DNA microarray data (Banerjee et al., 2007). The fitness functions evaluated the size of a reduct and the number of object combinations the reduct could discern. The classification performance of the reducts was found to outperform the classification performance features derived from PCA with a k-NN classifier. The authors proposed that the core of the reducts (the intersection of reduct features) could be useful for future experimental work by biologists.

Fuzzy rough sets have been used for feature selection for tumour classification (Dai and Xu, 2013). Crisp rough sets cannot represent continuous data. Gene expression data are usually continuous, and must be discretised before they can be input to a rough set. Fuzzy rough sets combine vagueness (fuzzy set theory) and indiscernability (rough set theory) into a single framework. Dai and Xu introduced the gain ratio, a metric popular for growing decision trees, into fuzzy rough theory and developed a feature selection algorithm utilising the gain ratio. The classification accuracy of a colon cancer dataset processed with the fuzzy rough feature selection protocol was higher compared with standard crisp rough set alternatives (Dai and Xu, 2013).

Feature selection algorithms are a standard preprocessing step in a knowledge discovery pipeline, and so they are widely applied to complex biological data. Feature selectors applied to Inflammatory Bowel Disease (IBD) (the topic of Chapter 4) are described below. Filter methods have been applied for the purpose of knowledge discovery in IBD on gene expression (Wei et al., 2013) and proteomic (Chen et al., 2009) data. Wrapper methods have been applied to imaging (Schüffler

et al., 2013) and spectroscopy (Bezabeh et al., 2009) data. Embedded feature selection is widely applied to biological data because it offers a good balance between computational complexity and performance, including for IBD classification from metagenomic data (Tong et al., 2013; Papa et al., 2012). Benchmarks on metagenomic data sets have shown that embedded feature selectors offer the best overall performance (Statnikov et al., 2013). The feature selection tasks included identifying body habitats (e.g. skin or faeces), psoriasis, and gastrointestinal disorders such as reflux esophagitis from the microbial community present in samples. Aggregating EFS has been applied for the purpose of knowledge discovery from DNA microarray and mass spectrometry datasets (Saeys et al., 2008) for the purpose of cancer classification. Aggregating EFS combines the output of multiple feature selection algorithms into a single consensus list of a feature subset to improve knowledge discovery.

2.5 Summary

CI is a critical tool for the purpose of knowledge discovery from complex biological data. It enables comprehensive searches through large amounts of imprecise and noisy data for hidden information, and has been widely applied to biological data sets. However, the application of CI to microbiome census data — 16S marker gene survey data gathered from environmental samples that can describe the structure and composition of microbial communities, described further in the next chapter — has been limited to standard supervised learning algorithms and feature selectors to date. Although these tools are invaluable for analysing such complex data, other approaches such as ensemble hybrid methods, rough set theory, fuzzy set theory, and ANNs are yet to be applied.

Aggregating EFS is a process that improves knowledge discovery by increasing the confidence domain experts can have in the output of a feature selector. Aggregating EFS results in a final consensus feature ranking by merging the decisions of a group of feature selectors. Minor changes to the input of a feature selector can cause large changes to the output of typical feature selectors. This is for a variety of reasons: the stability of the feature selector output is not a key target of standard feature selectors, and multiple different feature subsets can be equally optimal for a given classification problem. The community of microorganisms that live on humans — the human microbiome, described in Chapter 3 — is highly variable across individuals. Data that describe the microbiome are often noisy and imprecise. Feature selection in metagenomics has neglected the concept of feature stability or robustness, which aggregating feature selection has been shown to improve (Saeys et al., 2008).

Multimodal classification is applied to many domains such as robotics or image

processing. While multimodal classification has been applied to medical and biological data, its application to metagenomic data is sparse. The systems captured in biological data are immensely complex and dynamic: a single measuring paradigm will rarely be able to capture all of the information about a phenomenon. The microbiome does not exist in isolation: constant interactions are present between the host (e.g. the human) and microbiota. The use of multimodal classification could introduce the ability to holistically model the microbiome for the first time.

Rough set theory provides a suite of useful concepts that are invaluable for knowledge discovery from complex data but has not to date been applied to metagenomic data. Reducts can be used to remove redundant or irrelevant features from biological data. In bioinformatics and computational biology feature selection is almost always a prerequisite for model building due to the dimensionality of the data being studied (Saeys et al., 2007). The concept of boundary regions is useful as Aristotelian logic is not capable of representing health meaningfully: health is not the absence of disease (Torres and Nieto, 2006). The use of `IF-THEN` rules in rough decision systems enables transparency, which is paramount for knowledge discovery from complex biological data.

The following three chapters will apply CI algorithms on metagenomic data for knowledge discovery about human diseases. Chapter 4 will focus on IBD, as many public data sets are available and it is a key research topic regarding the effect of the human microbiome on health. Aggregating EFS will be applied to metagenomic data for the purpose of knowledge discovery and to aid the development of a non-invasive diagnostic test — IBD must currently be diagnosed via invasive colonoscopy. Chapter 5 explores the human oral microbiome for links to depression. A range of microbial ecology techniques are applied to the data that describe the first documented changes to the oral microbiome in a depressed cohort. A multimodal classification algorithm that enables the holistic modelling of the oral microbiome to predict depression. Finally, Chapter 6 applies rough set theory to metagenomic datasets collected from a depressed cohort to remove irrelevant features and transparently describes the microbial community dynamics present in the gut and mouth. These three chapters will outline the novel contributions this thesis brings to the knowledge base.

> Tell me what you eat, and I will tell you what you are.
>
> Jean Anthelme Brillat-Savarin

3.1 Introduction

This chapter provides a review of research regarding the human microbiome, the role it plays in disease, and the applications of computational intelligence in microbiome research. Microbiome research has shown that complex interactions between microbes and various host processes (e.g. the immune system) can drive disease in the host, even if the microbes present are not pathogenic per se (an overview of this phenomenon is provided in Section 3.2). An outline of the general stages of a microbiome experiment is provided in Section 3.3 (see Figure 3.2), and methods for generating microbiome count data are reviewed in depth. The problems introduced by clustering sequences with a global similarity threshold are noted, and alternative denoising strategies discussed. Section 3.4 gives a brief overview of the application of computational intelligence to microbiome research, an underexplored area to date. In Section 3.5 the chapter concludes with a summary of findings from the literature and an overview of the deficiencies in microbiome knowledge and protocols that are addressed in this thesis.

3.2 The role of the microbiome in disease

Many different microbial communities exist throughout the human body. The complete collection of taxa (a taxon is a group of bacteria considered to be a single unit) present in a microbial community is known as the microbiota; the microbiome includes the collective genomes of the microbiota (Human Microbiome Project Consortium, 2012). The rapid increase in publications reporting on the analysis of microbial communities that inhabit the human body has led to some confusion in terminology. The terms microbiota, microbiome, and metagenome are often used interchangeably. For consistency, the term microbiome will be used instead of the term microbiota. Below are some definitions of terms commonly used in microbiome research:

Metagenomics: The functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample (Riesenfeld et al., 2004)

Metagenome: The collection of host genetic content and all microbiomes present (Brüls and Weissenbach, 2011)

Human microbiome: The bacteria, archaea, viruses, and eukaryotic microbes (and their collective genomes) that exist throughout the human body (Shreiner et al., 2015)

Gut flora: Synonymous with the gastrointestinal microbiome. Flora can be used to describe other microbiomes (e.g. oral flora).

3.2.1 How does the microbiome influence disease?

There is some disagreement about when bacteria first colonise human foetuses, but the womb is traditionally thought to be sterile (Morgan and Huttenhower, 2012). During and after birth every body surface is colonised by microbes via a variety of processes. This includes bacteria, archaea, fungi, and viruses. The microbiome provides key functions for the host, such as nutrient metabolism (Kamada et al., 2013) and helping to educate and develop the host immune system (Hooper et al., 2012). Many microbes in the microbiome provide no benefit directly, but their presence can prevent the development of pathogenic microbes. The members of a microbial community in a healthy host exist in a state of constant competition (Coyte et al., 2015). Although disease can be associated with low diversity and the dominance of specific bacterial clades it is important to note that high diversity is not always healthy; highly diverse communities are not inherently superior to simpler communities (Shade, 2017).

Some of the earliest work in microbiome research found different patterns in the gut microbiome in conditions such as obesity (Turnbaugh et al., 2006) and Inflammatory Bowel Disease (IBD) (Elson et al., 2005). The Human Microbiome Project (Turnbaugh et al., 2007) was launched in 2008 to identify and characterise microbes that live in healthy and diseased humans. The MetaHIT project was also launched in 2008, with a particular focus on obesity and IBD (MetaHIT Consortium, 2011). The following subsections focus on IBD and depression, which are explored throughout this thesis. IBD data was used as a starting point to develop bioinformatics software pipelines, as most public human microbiomic data was gathered from IBD and control (healthy) subjects. Depression was investigated in collaboration with the Northern Ireland Centre for Stratified Medicine. Possible links between depression and the microbiome are numerous and well-documented (described below) but the area remains underexplored.

3.2.2 The gastrointestinal microbiome and Inflammatory Bowel Disease

The microbiota that reside in the gastrointestinal tract represent a complex and diverse microbial community (Bäckhed et al., 2005). Although the microbiota provide key beneficial functions for the host, the presence of a large microbial community in close proximity to the host provides a constant challenge for the immune system. A healthy gut immune system can tolerate the normal microbiome (Brown et al., 2013), and maintains homeostasis of the microbial communities by containing the microbiota to the the lumen and outer mucus layers of the gut (Johansson et al., 2008). A dense sterile inner mucus layer that contains antimicrobial peptides is responsible for segregating the microbiota from intestinal epithelial cells (Vaishnava et al., 2011).

IBD is a group of disorders that cause persistent inflammation of the gut. IBD caused 53,000 deaths worldwide in 2013 and its prevalence is increasing (Molodecky et al., 2012). Urbanisation has been linked with autoimmune diseases, including IBD (Zuo et al., 2018), and IBD is a growing problem in many parts of Asia, the Middle East, and South America (Zhao et al., 2013; Ng et al., 2013). IBD is an umbrella term that covers both crohn's disease (CD) and ulcerative colitis (UC). For many years the aetiology of IBD was poorly understood, as the disease is characterised by unpredictable periods of active inflammation and remission. Responses to treatment are also unpredictable, with some patients not responding to steroid treatments, and a proportion requiring surgical removal of badly affected sections of the gut (Vester-Andersen et al., 2014). This caused speculation that a decrease in the diversity and changes to the composition of the intestinal microbiome, known as dysbiosis, could contribute to the development of the disease (Tamboli et al., 2004). A number of cross-sectional studies have confirmed that dysbiosis of the intestinal microbiome is present in patients with IBD during disease onset (before treatment; Gevers et al., 2014), or after IBD has been diagnosed in a clinical setting (Sokol et al., 2008; Willing et al., 2008; Papa et al., 2012; Tong et al., 2013). It has been recently shown that dysbiosis of the gut microbiota precedes the onset of colitis-induced inflammation in mice (Glymenaki et al., 2017). The aetiology of IBD is thought to involve complex interactions between the gut microbiome, the environment, the host immune system, and the host genome (Wallace et al., 2014). It is thought that a genetic susceptibility for a dysregulated mucosal immune system causes a larger than normal immunological response to the gut microbiome in some patients. This response can cause shifts in the composition of the bacterial community, further increasing the immunological response from the mucosal immune system (Prosberg et al., 2016).

3.2.3 The microbiome and depression

Depressive disorders are a broad collection of mental health disorders associated with a range of emotional, physical, cognitive, and behavioural symptoms. Depression is typically characterised by a loss of enjoyment in ordinary life and low mood (National Collaborating Centre for Mental Health, 2010). Alongside severity, persistence must also be taken into account when characterising depression; typically changes must last at least two weeks. Changes can be episodic in nature. Stratifying depressed subjects into different categories of depression can be particularly challenging. It is important to note that the subtypes proposed below are only based on symptomatic differences (see Table 3.1); there is limited evidence to date that suggests different underlying diseases are the cause.

Table 3.1: Major depressive disorder specifiers (i.e. subtypes) (American Psychiatric Association et al., 2013).

Depression features	Main features
Anxious	Excessive restlessness or fear
Mixed	Changes that occur in a person's behaviour that appear to be exaggerated or boastful (e.g. inflated self-esteem, very talkative)
Chronic	MDD diagnosed for at least two years
Melancholic	Anhedonia, early morning awakening, loss of appetite, excessive guilt
Catatonia	Unusual movements or behaviours such as persistent immobility or mutism
Atypical	Weight gain, hypersomnia, mood reactivity
Peripartum onset	Develops before or close to childbirth, extreme mood fluctuations and excessive concern over their child's wellbeing
Seasonal pattern	Depressive patterns coincidence with specific seasons (e.g. onset of winter)

In this thesis the term depression refers to the specific mental disorder major depressive disorder (MDD), also known as clinical depression (depression can be used to refer to a simple state of low mood). Subjects with depression often have a comorbid physical or psychiatric diagnosis (Brown et al., 2001). For example, postnatal depression is associated with childbirth, and vascular depression is associated with the elderly (Sneed and Culang-Reinlieb, 2011). In England 1 in 6 adults reported experiencing common mental health problems, including anxiety and depression (McManus et al., 2016). In England 3.3% of adults reported

having a depressive episode in the last week in the 2014 adult psychiatric morbidity survey (McManus et al., 2016). The prevalence of depression has been found to be consistently higher in women (Waraich et al., 2004). Depressive disorders were ranked as the fourth leading cause of burden in 1990 across the world. In 2000 this increased to the third leading cause of burden. By 2010 depressive disorders were the second leading cause of disease burden globally (Ferrari et al., 2013), with 8.5% of total years lived with disability being attributed to depression.

Depression is diagnosed by general practitioners using measures in line with Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria (American Psychiatric Association et al., 2013). Diagnosis relies on self-reported symptoms and clinical judgement. Subjective criteria (e.g. the Hamilton depression rating scale; Hamilton, 1960) are used to empirically identify depression and measure the severity of depression during treatment. Subjective criteria are the only methods in clinical practice for diagnosing depression, and no diagnostic tests are currently used (e.g. a blood test). The criteria consists of a series of questions about the symptoms of the interviewee in line with DSM diagnostic criteria. These questions can include emotional state (e.g. suicidal tendencies), insomnia, psychomotor retardation, and weight loss. Each question has responses ranked according to severity (e.g. 0 = no symptoms, 3 = severe symptoms). The responses are summed; increased score correlates with increased severity. The Hamilton depression scale was originally developed to test the efficacy of first-generation antidepressants in the 1950s; it has been proposed that the Hamilton depression scale is unfit for purpose and should be rejected entirely, to be replaced with a new diagnostic paradigm (Bagby et al., 2004).

A non-subjective diagnostic test for depression would be invaluable, improving speed of diagnosis and first round response to treatment. Relying on subjective criteria for diagnosis contributes to the heterogenity that characterises depression. Misdiagnosis or slow diagnosis has serious consequences for patients. For example, antidepressant drug treatments are not effective for bipolar disorder (BD) subjects. Treatment outcomes are worse for patients that are misdiagnosed with depression and subsequently correctly diagnosed after several episodes of illness (Swann et al., 1999). The response rate for depression patients to first rounds of pharmacalogical interventions is around 30% (Trivedi et al., 2006). Non-responders are cycled through different types and classes of drugs until a response is apparent. Each cycle can last up to 12 weeks (Papakostas et al., 2008), prolonging impairment or even exacerbating the condition. Pharmacalogical interventions remain the fastest and most effective way of treating the most severe forms of depression (Kirsch et al., 2008).

Depression has a complex aetiology. The first antidepressants were found by chance, approximately 65 years ago. Iproniazid, an irreversible monoamine

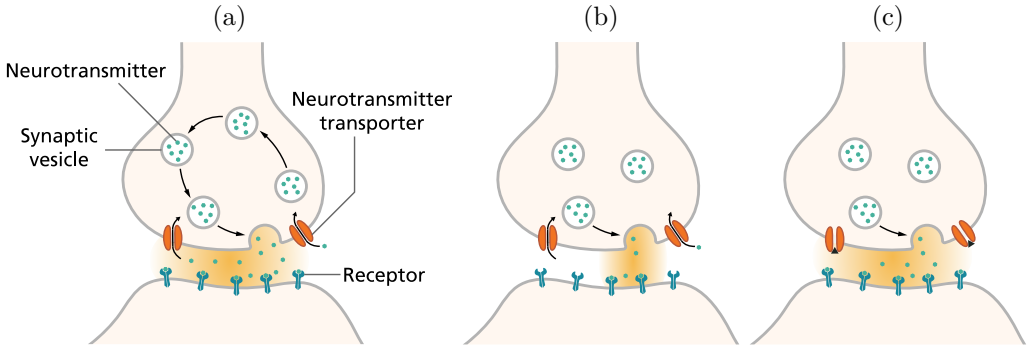


Figure 3.1: Monoamine hypothesis of depression. **(a):** A healthy synapse. **(b):** A depressed synapse. **(c):** Treated depressed synapse. Derived from "Schematic of a synapse" by Thomas Splettstoesser, distributed under a CC BY-SA 4.0 license.

oxidase inhibitor, was originally used to treat tuberculosis but was found to cause patients to become "inappropriately happy" (López-Muñoz and Alamo, 2009). It was found that the first antidepressant drugs had a similar mechanism of action: they increased the concentration of monoamine neurotransmitters serotonin and noradrenaline in the brain (first observed in extracellular brain fluid, see Figure 3.1). The increased concentration was caused by decreased catabolism (breakdown: the opposite of anabolism) and decreased re-uptake in the postsynaptic neuron (Castrén, 2005). This is now known as the monoamine hypothesis of mood disorders, and the idea of a chemical imbalance in the brain causing depression has been generally accepted across the field and general public (Deacon and Baird, 2009). According to the monoamine hypothesis of depression in a healthy brain monoamine neurotransmitters are secreted and bind to receptors on the postsynaptic neuron; reuptake of neurotransmitters by transporters ends the transmission (see Figure 3.1a; Hirschfeld, 2000). In a depressed brain a mood disorder is produced by the low concentration of monoamine neurotransmitters (see Figure 3.1b). Blocking of the re-uptake transporters by antidepressant drugs increases the concentration of monoamine neurotransmitters to treat depression (see Figure 3.1c). However, the monoamine hypothesis has many inconsistencies: drugs such as reserpine which deplete monoamine neurotransmitters cannot induce depression in healthy subjects, despite many claims to the contrary (Baumeister et al., 2003). It is important to note that brains treated with antidepressant drugs do not return to normal. A cohort of healthy and formerly depressed (in remission after successful antidepressant treatment) subjects were given 200mg dopamine D2/D3 receptor antagonist sulpiride. In healthy subjects, there was no change in mood, while in formerly depressed subjects profound depression quickly returned (Willner et al., 2005).

It is common sense that chronic stress can manifest as disease; it is less easy to explain how. Stress is persistently accompanied by poor health: it can quadruple the chance of an adverse medical outcome (Sandberg et al., 2004), and accelerate the progression of chronic conditions such as coronary heart disease (Smith et al., 2005). A potential mechanism that explains the biological manifestation of stress is the hypothalamic-pituitary-adrenocortical (HPA) axis. The axis is activated by the secretion of corticotropin-releasing hormone (CRH) by the hypothalamus. The pituitary gland will secrete adrenocorticotropin hormone (ACTH) in response to an CRH signal. In turn, the adrenal glands will secrete cortisol in response to a ACTH signal. Cortisol is an extremely important hormone: it can affect the central nervous system (including learning and emotion), the metabolic system (glycogen regulation), and the immune system via inflammatory regulation (Sapolsky et al., 2000). Dysfunction of the HPA axis is one of the most consistent markers of depression (Molcrani et al., 1997). Antidepressants have been found to normalise HPA axis function, via mechanisms of action that are independent of monoamine re-uptake inhibition (Willner et al., 2013). A thorough review of current theories about the pathophysiology of depression is outside the scope of this thesis, but many are available in the literature (Castrén, 2005; Shyn and Hamilton, 2010; Hodes et al., 2015).

A comprehensive review found that no biological markers for depression are available for inclusion in diagnostic criteria (Mössner et al., 2007). Neurotrophic factors (Shimizu et al., 2003), biochemical markers (Heuser et al., 1994), neuroimaging markers (Kempton et al., 2011), immunological markers (Maes et al., 1995), and neurophysiological tests (Gangadhar et al., 1993) were considered for inclusion. While the proposed markers significantly differ between healthy controls and depression patients, the tests lack sensitivity, specificity, or reproducibility. Recently discovered biomarkers show promise for specific subgroups of patients with severe depression. In severely depressed adolescent males, elevated morning cortisol acts as a biomarker (Owens et al., 2014). Brain glucose metabolism can be used as a predictive biomarker (McGrath et al., 2013), successfully guiding intervention strategies (response to psychological intervention vs pharmacological intervention). Recent work in mice has shown that the administration of bacterial probiotics can act as an antidepressant (*Lactobacillus rhamnosus JB-1*) or induce anxiety (*Campylobacter jejuni*; Foster and Neufeld, 2013). This raises several questions: if the microbiome can influence the mind and behaviour, what are the mechanisms behind the phenomenon? Could beneficial bacteria be used clinically as a psychobiotic treatment for mental health disorders? Could differences in the microbiome be used to diagnose depression? It is likely that a combination of factors will determine predisposition to depression, both genetic (including epigenetic and microbiomic) and environmental (life stresses and exposures).

IBD and depression are both diseases with heterogenous diagnosis, treatment, and outcomes. IBD symptoms are non-specific - primarily abdominal pain is reported - and a colonoscopy is required to confirmation. Diagnosis is often delayed in paediatric patients because of the invasive nature of the colonoscopy procedure, which can impact the child's development (e.g. stunted growth). In the first contribution of this thesis subsets of microbial markers are identified with ensemble feature selection that can accurately identify IBD with up to 97% accuracy in a large paediatric cohort. In future work this approach could be adapted to develop a highly accurate non-invasive test for IBD that would significantly decrease the time to diagnosis, improving patient outcomes. Diagnosing depression relies on subjective criteria rather than diagnostic tests. Misdiagnosis of depression is widespread, lowering response rates to antidepressant drugs. In the second contribution of this thesis an Artificial Neural Network (ANN) is used to diagnose depression with high accuracy from microbiome count data. Clinical validation of this approach could improve treatment outcomes by increasing the response rate to antidepressant drugs. Potential mechanisms that may explain how the microbiome could influence depression are explained below.

3.2.4 The gut-brain axis communication

Bidirectional communication between the gut and the brain was identified by many scientists in the 19th century, including Charles Darwin (Darwin, 1872). A well known example of this communication is the frequent co-occurrence of hunger and anger: low blood glucose has been associated with aggression in married couples (Bushman et al., 2014). This concept was dubbed the "gut-brain axis". The concept has been extended to include the microbiome after new evidence that bidirectional communication occurs at all levels - with several distinct mechanisms - between the microbiome, gut, and brain. The methods of communication include neural, metabolic, and immune pathways (El Aidy et al., 2015). The microbiome-gut-brain axis has been implicated in the aetiology of depression via altered neurotransmitter signalling, hypothalamic-pituitary-adrenal (HPA) axis modulation, and inflammation (Cryan and Dinan, 2012; Foster et al., 2017).

The intestinal microbiome produces many important neurotransmitters present in the human brain (Lyte, 2013; Lyte, 2014). Neurotransmitters are chemical messengers, distinct from hormones, that enable communication across the central nervous system. The abundance of serotonin precursors in the blood has been shown to be increased by Bifidobacteria. Serotonin is directly synthesised by many genera, including *Streptococcus* and *Escheridia*. Dopamine and acetylcholine are produced by *Bacillus* and *Lactobacillus*, respectively. Gamma-aminobutyric acid is produced by *Lactobacilli*. *Lactobacilli* have also been shown to alter the expression

of gamma-aminobutyric acid receptors present in the brain (Bravo et al., 2011). However, although the neurotransmitters are capable of crossing from the intestine into the blood stream (it is not certain if they do), most are thought to act locally, modulating the nervous system present in the gut (the enteric nervous system). The blood brain barrier, a membrane that separates circulating blood from brain extracellular fluid (Tran, 2011), is highly selective and thought to prevent any neurotransmitters crossing from the blood to the brain. Instead, the secreted neurotransmitters are thought to indirectly affect the brain by interacting with the enteric nervous system.

Metabolic communication is thought to occur between the intestinal microbiome and the brain primarily via the secretion of short chain fatty acids such as butyrate and propionate (Stilling et al., 2014). Short chain fatty acids are byproducts of microbial metabolism, but are epigenetic modulators via the action of histone deacetylases. Epigenetic modulation relates to the modification of gene expression via processes that are unrelated to DNA sequence modifications. Variation in the methylation of DNA alters gene expression, and the changes in methylation states are heritable (epigenetic traits). An epigenetic trait is defined as "stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" (Berger et al., 2009). It has been proposed that the synthesis of genes in brain cells could be altered via the epigenetic effects of short chain fatty acids, altering host behaviour, as short chain fatty acids can freely pass through the blood brain barrier (Stilling et al., 2014). The modified gene synthesis could then cause changes in behaviour.

Cytokine molecules have been shown to enable communication between the intestinal microbiome and the brain via immune signalling (El Aidy et al., 2014). Although it is unlikely that cytokines are capable of crossing the blood brain barrier, the barrier is not perfect. Certain areas of the brain (circumventricular organs; Fry and Ferguson, 2007) are not protected in order to allow chemical signals to be sent from the brain to the rest of the body, such as the median eminence of the hypothalamus. Cytokines such as interleukin-1 and interleukin-6 are thought to activate the HPA axis, which releases cortisol. Increased concentrations of cortisol occur after a psychological stressor is introduced to depressed subjects (Burke et al., 2005).

Inflammation concomitant with the onset of depression may lead to a dysfunctional intestinal epithelium barrier ("leaky gut") due to the opening of intercellular tight junctions (Kelly et al., 2015). The translocation of bacterial cells, bacterial byproducts, and inflammatory mediators across the leaky gut is thought to drive a chronic pro-inflammatory state and subsequently activates the HPA axis (Kelly et al., 2016). The bacterial products and inflammatory components can cross the blood brain barrier and initiate a central inflammatory response via the activation

of microglia cells (the primary immune cells of the central nervous system; Yirmiya et al., 2015). Repeated modification of microglial cells causes chronic brain inflammation, which is thought to play a role in the structural and functional brain alterations associated with mental health disorders (Stein et al., 2017).

3.3 Counting the uncountable

Bacteria are omnipresent across the surface of the earth. By extension, bacteria are ubiquitous on the surface of organisms which live on earth, having been exposed to the surface of the earth during their life. It should be noted that the surface epithelium that forms the gastrointestinal and respiratory tracts is, despite being inside the body, actually an exterior surface (i.e. equivalent to skin). Bacteria are very small - *Escherichia coli* cells are on average 2 µm long - and very numerous. An estimated 3.8×10^{13} bacterial cells reside in the human body (which weigh around 0.2 kilograms; Sender et al., 2016). This is larger than the total number of human cells present (3.0×10^{13}). How can we count and catalogue such a vast number of bacteria? This section will discuss this process. Identifying the specific sequence of nucleotides present in a DNA molecule is a complex procedure and only briefly discussed. Detailed discussion regarding sample collection and sequencing strategies is available in the literature, and references are provided below. The focus of this section is on identifying ecologically unmixed units of bacteria from DNA sequence fragments and the processes that ensure that generated sequence counts are accurate.

3.3.1 What is a bacterial species?

Schoolchildren are taught that organisms of the same species can interbreed to produce fertile offspring. Unfortunately bacteria do not have sex, which complicates matters considerably (Cohan, 2002). When Carl Linnaeus began assigning plants into groups with binomial nomenclature in the *Species Plantarum*, he used simple physical characteristics such as the structure of stamen to do so. Viewed under a microscope, the majority of microbes resemble colourless blobs. Stains and dyes can be used to differentiate microbes, but the most widely used classification systems use carbohydrate utilisation tests and other biochemical methods (Goodfellow et al., 1997). Molecular systematics uses differences in the structural composition of DNA to group and determine the evolutionary relationships of bacteria (Blaxter, 2003). The majority of bacterial species in a complex community cannot be identified or grouped from their physical and chemical characteristics alone (Blaxter, 2003). The magnitude of microbial diversity has only been realised since the application of molecular systematics to environmental samples (Lynch and Neufeld, 2015).

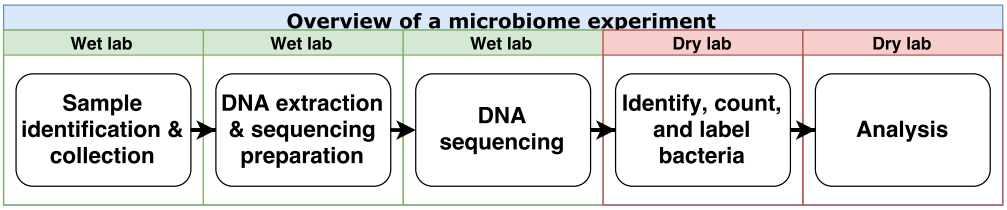


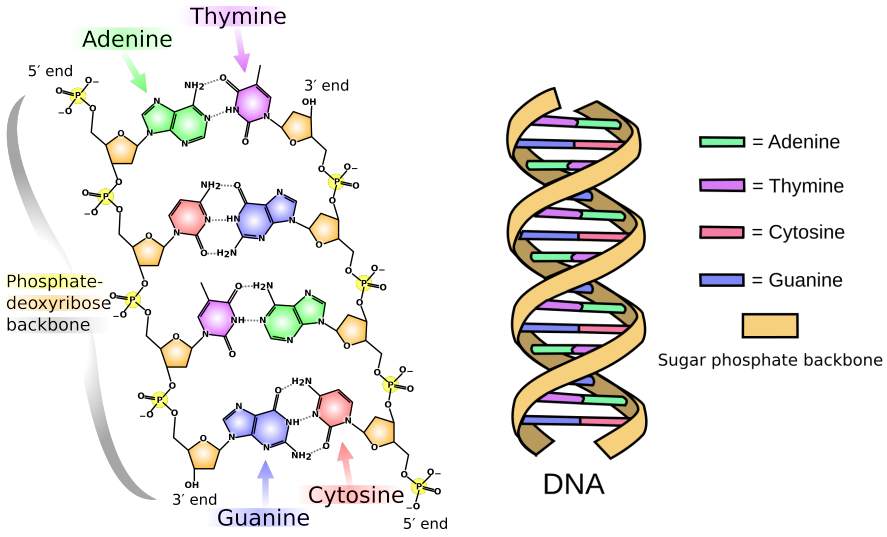
Figure 3.2: Broad overview of stages of a microbiome experiment

Molecular systematics help to measure differences between organisms, but it does not explain how to group them. Standard cutoff points of similarity have been adopted by the scientific community. When two organisms are found to be at least between 97% and 99% similar (Huse et al., 2010), they are considered to be the same species. If this approach was adopted for macroscopic organisms humans could be grouped with chimpanzees as a single species (via original estimates of 98.5% DNA similarity; Hoyer et al., 1972). This blunt approach has clear limitations, but is a significant improvement over phenotypic methods. However, there is no unified species concept for bacteria (Doolittle and Zhaxybayeva, 2009), and taxonomic units defined by similarity thresholds are theoretical constructs.

3.3.2 A computer scientist's illustrated primer

This subsection provides a high-level overview of DNA sequencing, which is a crucial stage for any microbiome experiment (see Figure 3.2). DNA is a molecule that carries the blueprints for growth, development, and reproduction in all living organisms (Hunter, 1993). The majority of DNA is made of two polynucleotide strands joined in a double helix structure (see Figure 3.3(b)). Nucleotides are the monomer unit for nucleic acid polymers (i.e. nucleotides are building blocks: DNA strands are made from lots of nucleotides joined together; Sadava et al., 2009). Nucleotides are made up of a nitrogenous base - the term base is often used as a synonym for nucleotide - a five-carbon sugar (ribose or deoxyribose), and a phosphate group (see Figure 3.3(a)). There are four possible types of nitrogenous base in DNA: adenine, cytosine, guanine, and thymine (Sadava et al., 2009). Nucleotides are joined together by covalent bonds between the five-carbon sugar of one nucleotide and the phosphate group of the next nucleotide (this is called the "sugar phosphate backbone"). The nitrogenous bases of two complementary DNA strands can form hydrogen bonds according to base-pairing rules. When Crick and Watson discovered the structure of DNA in 1953 they identified base pairing rules: adenine bonds with thymine, and cytosine bonds with guanine (Watson, Crick et al., 1953). The two complementary DNA strands are antiparallel.

DNA sequencing is a process that measures the precise order of nucleotides in

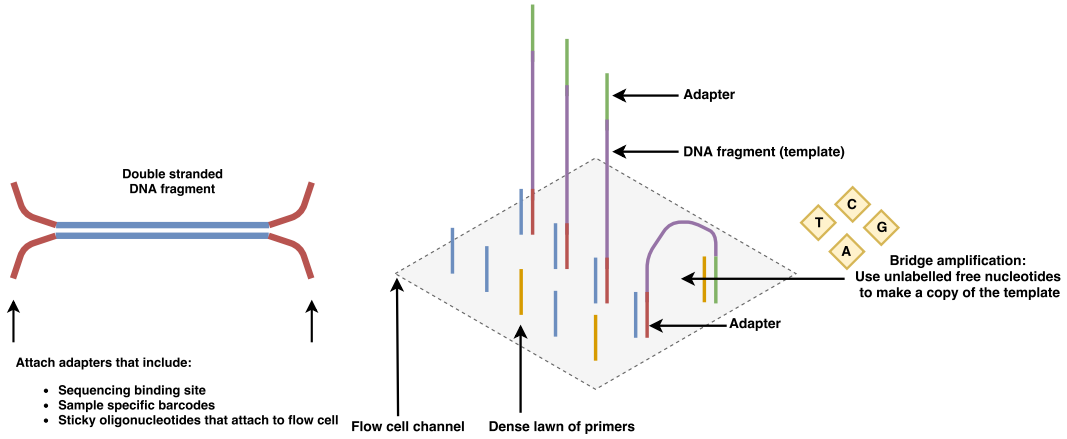


(a) (b)

Figure 3.3: (a) "Chemical Structure of DNA" by Madeleine Price Ball, licensed under CC-0. (b) DNA structure, derived from the diagram "DNA simple" in the public domain.

a DNA molecule (Sadava et al., 2009). The chemistry behind DNA sequencing can be very different depending on the paradigm used, but many rely on detecting millions of fluorescently labelled short DNA fragments in parallel (Glenn, 2011). Other approaches include ion semiconductor chip based sequencing (Merriman et al., 2012) and nanopore sequencing (Mikheyev and Tin, 2014). Throughout this thesis the sequencing data analysed are generated from "sequencing by synthesis" light-based methods such as pyrosequencing or paired-end Illumina sequencing as they are ideal for detecting microbes from environmental samples (Roesch et al., 2007; Fadrosh et al., 2014). Sequencing by synthesis will be described at a high level below. More detailed descriptions are available in the literature (Fadrosh et al., 2014).

The first step of sequencing by synthesis is sample preparation (also known as library preparation; Quail et al., 2008). DNA must first be broken into shorter fragments approximately 200-800 base pairs long by chemical (e.g. enzymatic degradation) or physical (e.g. sonication) means. Adapter sequences are then attached to the ends of the double stranded fragmented DNA, which are synthetic oligonucleotides (short DNA molecules) that enable the sequencing process to take place (see Figure 3.4(a)). After the adapters are attached, the double stranded DNA is denatured (separated) into single stranded DNA with heat (Meyer and Kircher, 2010). The single stranded DNA molecules are known as templates. The second

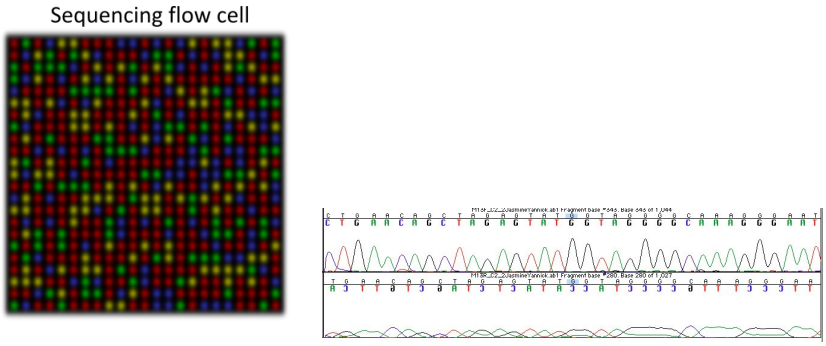


(a) (b)

Figure 3.4: (a) Sample preparation. (b) Cluster generation.

step is cluster generation, where the adapter sequence present on the templates causes the templates to stick to a flow cell, which is a glass plate with separate lanes. Many samples can be sequenced on the same lane because of the unique sample-specific barcodes present in the adapter regions (Quail et al., 2008). The templates that are tethered to the flow cell are clonally amplified via bridge polymerase chain reaction (many copies are simultaneously made; see Figure 3.4(b)).

Sequencing the flow cells that now contain millions of DNA fragments is simple: by synthesising a copy of the DNA fragments with fluorescently labelled nucleotides the precise order of nucleotides in the DNA fragments can be determined (Meyer and Kircher, 2010). The first base of the DNA fragments is determined by exciting the flow cell with a laser and recording a high resolution image. Each type of labelled nucleotide will fluoresce at a different wavelength and intensity (Mardis, 2008). By repeating this sequencing cycle the second nucleotide for all DNA fragments can be determined and so on until the entire DNA fragment has been recorded. If 200 nucleotide long DNA fragments are being sequenced then 200 images will be recorded. The images are processed to generate colour spectra for every pixel for every nucleotide position, called a chromatogram. Each pixel represents a different DNA fragment read, and each image represents a different nucleotide position (see Figure 3.5(a)). Sequencing is not a perfect process - as shown by the blurring of the flow cell in Figure 3.5(a) and overlapping signals in the chromatogram in Figure 3.5(b) - and it is necessary to interpret the chromatogram to determine which nucleotide is present at each position in the sequence. The process of converting information that measures the wavelength and intensity of fluorescing nucleotides to a text-based format that represents sequencing data



(a) (b)

Figure 3.5: (a) Example image of fluorescently labelled nucleotides on a flow cell, derived from "Microarray and sequencing flow cell" by Thomas Shafee, licensed under CC-BY-4.0. (b) Chromatogram derived from "Chromatogram" by Tom David, in the public domain.

is known as base calling (Ledergerber and Dessimoz, 2011). In many genomics applications the short DNA reads are often aligned to reference genomes and joined into a single contiguous sequence. When identifying microbes from environmental samples this is not usually done: instead the raw reads are used as input for a variety of clustering (Navas-Molina et al., 2013), or denoising algorithms (Rosen et al., 2012), discussed further in Sections 3.3.5 and 3.3.6.

3.3.3 From samples to sequences

Proper sampling is the foundation of any microbiome experiment. The most stunning analysis will be undermined by poor quality data. Samples must be identified, collected, and their DNA extracted (IHMS Consortium, 2015a). Good clinical practice is important while sampling human body sites to ensure that the samples accurately reflect the true bacterial composition present in an environment. It is also important to remember where the samples actually originated, as sampling strategies can introduce subtle differences into microbiome census data. For example, the classification accuracy of a Random Forest machine learning algorithm significantly improved with microbiome count data derived from intestinal biopsy versus data derived from a faecal sample, despite both representing the intestinal microbiome (Gevers et al., 2014). While planning a microbiome experiment a balance must be achieved between cohort size and the invasiveness of the sampling procedure. It may be infeasible to biopsy hundreds of subjects, and recruitment may be more difficult with invasive procedures such as colonoscopy. A thorough review of sampling protocols is outside the scope of this thesis, but many are

widely available in the literature (Aagaard et al., 2013; Sinha et al., 2016; IHMS Consortium, 2015a).

Once a sample has been collected the DNA present must be extracted for sequencing. The purity and quantity of DNA must be sufficient, or the sequencing process may fail or produce poor quality data. In brief, the membrane of the bacterial cells must be broken apart (lysed) while preserving the fragile DNA inside (IHMS Consortium, 2015b). Lysis methods can be physical, chemical, or mechanical (or a combination of the three). The DNA must then be separated from the cell remnants, which can contain enzymes which damage the fragile DNA, and stored in a stable environment. A thorough review of DNA extraction protocols is outside the scope of this thesis, but many are available in the literature (Bag et al., 2016; IHMS Consortium, 2015b). After the DNA has been extracted, it is sequenced (see section 3.3.2) for further analysis.

It is important to understand why, when attempting to determine the amount and number of microbes present in an environmental sample, sequencing microbial DNA is preferred over other methods. Historically, culture-dependent assays were used to determine the types of microbes present in a sample. For example: a patient presents to a doctor with a green and fuzzy wound. How would the doctor discover if pathogenic bacteria were present in the wound? An environmental sample would be taken from the wound (e.g. with a swab) and placed in an controlled nutrient-rich environment (a culture). This would promote rapid microbial growth. If no microbes were present then it can be said that the patient does not have a bacterial infection. If any microbes have grown on the media, identification can be attempted via stains, microscopy, and many other biochemical methods such as carbohydrate utilisation tests.

There are numerous problems with this approach which led to the development of culture-independent assays (Hugenholtz et al., 1998). The largest problem is that the majority of microbial life cannot be cultured (Breznak, 2002). This was first discovered when a difference was found between the number of bacteria observed via microscopy and the number of bacteria grown in a laboratory culture. The difference was several orders of magnitude in size, and was dubbed “The Great Plate Count Anomaly” (Staley and Konopka, 1985). Unculturable bacteria are unculturable because we lack the scientific understanding to create environments in which they can thrive (Stewart, 2012). Traditional growth media aim to provide a never-ending feast for bacteria. Only a small proportion of the microbial population are capable of taking advantage of the provided feast; the majority are limited by other factors (e.g. missing nutrients, dormancy, or competition). Thus, a culture-dependent assay will identify relatively few members of the original microbial community. Culture-dependent assays can only rarely be considered to be truly representative of the microbial population present in an environmental sample.

The other problem with culture-dependent assays is the process of actually distinguishing one bacterial species from another bacterial species. As described earlier in Section 3.3.1, the majority of bacteria resemble colourless blobs. In 2008 only 7,000 bacterial species were described (Achtman and Wagner, 2008), while in 2009 around one million valid species of insect were catalogued (Resh and Cardé, 2009). Estimates of the number of bacterial species range from hundreds of thousands to tens of millions. The discrepancy stems from a combination of the two problems described above. Bacteria must firstly be able to be distinguished from other bacterial species. This is usually tested via a battery of physical and chemical tests (e.g. high proteolytic activity, breaks down glucose, etc.). This process is difficult, and putative bacterial species must also be cultured in order to be recorded as a valid species. Culture-independent assays have begun to identify vast amounts of previously hidden bacterial diversity: the number of bacterial phyla has expanded from 11 in 1987 to at least 85 in 2012, the majority of which have no cultured representative species (Stewart, 2012). The number and variety of species present in a single phylum is vast - humans and sea squirts are in the same phylum (Chordata).

One example of a culture-independent assay is a marker gene survey, which requires a small section of microbial DNA to be sequenced (Tringe and Hugenholtz, 2008). Another example of a culture-independent assay is sequencing all microbial DNA present in a sample with metagenomic shotgun sequencing (Tringe and Rubin, 2005). This thesis uses data derived from marker gene surveys in all of the contributions, as metagenomic shotgun sequencing is infeasible for most laboratories due to its economic and computational expense. The gene encoding 16S ribosomal ribonucleic acid (16S rRNA) is often used as a universal marker gene (see Figure 3.6), because it has a number of interesting properties:

- The 16S rRNA gene is ubiquitous across all bacteria and archaea;
- The 16S rRNA gene is approximately 1500 nucleotides in length, which is short enough to be feasibly analysed;
- The gene sequence is highly conserved (similar across species) in some areas, allowing the comparison of distantly related species (Woese and Fox, 1977);
- In other areas the sequence is hypervariable, allowing the comparison of very closely related species (Clarridge, 2004);
- Some areas are completely conserved, which has aided the development of universal primers and protocols (Caporaso et al., 2011);
- Horizontal gene transfer - a process in which genetic material is shared between organisms and distinct from vertical gene transfer (i.e. parent to

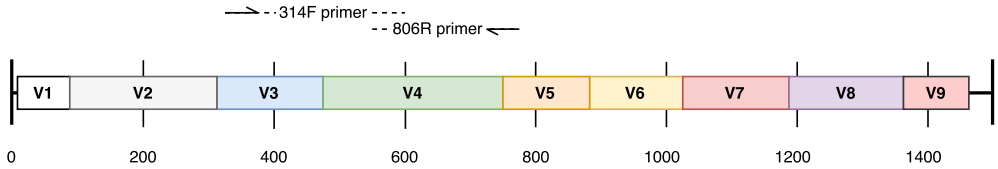


Figure 3.6: The hypervariable regions of 16S rRNA, which is around 1500 nucleotides long. The 314F and 806R primers are used for sequencing in Chapter 5.

child) - is thought not to occur in the 16S rRNA gene, ensuring that sequenced 16S rRNA genes originate from a specific bacterial cell (Jain et al., 1999).

A primer is a short strand of DNA (around 20 nucleotides long) that acts as a starting point for DNA synthesis (Sadava et al., 2009). Primers are designed to be complementary to specific target regions of DNA. Once bound to the targeted region DNA, DNA polymerase tethers itself to the primer and incorporates nucleotides complementary to the antisense DNA strand, generating a copy of the targeted DNA region (Sadava et al., 2009). Many popular sequencing technologies cannot sequence the entire length of the 16S rRNA gene. Instead, the various hypervariable regions are used to provide species-specific DNA signatures. These signatures can be used to create a bacterial census from a marker gene survey. Different hypervariable regions offer different levels of specificity and sensitivity for detecting distinct bacterial species. The V3 — V4, V4 only, or V4 — V6 regions are widely used (Yang et al., 2016).

In high-throughput sequencing a large number of samples are processed simultaneously via multiplex sequencing (Wong et al., 2013). Each sample is assigned an individual "barcode" sequence. The barcodes allow the reads to be separated and sorted after sequencing. Barcodes are attached to the sequences of interest with polymerase chain reaction (PCR), described below. By increasing the number of samples that can be processed simultaneously, multiplexing makes sequencing much more cost-effective than it otherwise would be by reducing time and reagent use (Wong et al., 2013). This has also helped to make the process of sequencing environmental samples, which have very large amounts of DNA present, more cost-effective.

Universal primers have been designed and made widely available (Caporaso et al., 2011) that bind to the conserved regions of the 16S rRNA gene that are near the hypervariable regions of interest. The primers are used to make many copies of 16S rRNA fragments present in environmental samples via a process known as the PCR. PCR was invented by Kary Mullis in 1983 (Mullis, 1990), for which he was awarded the 1993 Nobel Prize in Chemistry. From a single 16S rRNA gene millions of fragments can be created via a form of molecular photocopying that uses

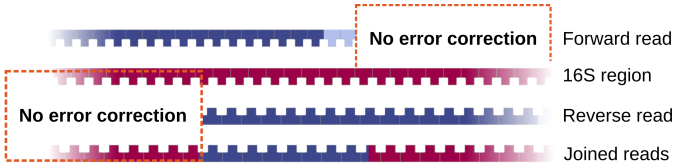


Figure 3.7: Paired end sequencing. Derived from templates by Library of Science and Medicine Illustrations, licensed under CC BY-NC-SA 4.0.

deoxynucleoside triphosphates (dNTPs) as building blocks. This copying process is used because DNA sequencing machines require input sequences to be of a certain length, purity, and concentration in order to work well.

Specific sequencing strategies are required to perform 16S rRNA marker gene surveys effectively. The Illumina MiSeq platform is commonly used for 16S marker gene surveys (Bartram et al., 2011). The platform can generate up to 250 base paired-end reads (Quail et al., 2012). Paired-end reads add an extra layer of error correction, which is a key advantage for the algorithms that identify microbes from 16S data. Partially overlapped paired end sequencing increases sequencing error because error correction is reduced (see Figure 3.7; Kozich et al., 2013). Up to 384 samples can be processed simultaneously, generating over 24 million discrete reads. Sequencing data are typically recorded in a text-based format that stores a biological sequence and the corresponding quality metadata called the FASTQ format. The quality of a base call is measured by the quality value Q, which is given by (Cock et al., 2009):

$$Q_{\text{phred}} = -10 \log_{10} p \tag{3.1}$$

Where p is the probability that the corresponding base call is incorrect. A common default cutoff for Q_{phred} is approximately 13, which approximately corresponds to $p < 0.05$ (see Figure 3.8; Cock et al., 2009). However, considering the millions of bases called during a full sequencing run this is extremely lenient. The algorithms that estimate a base call probability are proprietary, and are kept secret by manufacturers.

3.3.4 Noise and bias

Ideally a table of sequence counts should reflect the true composition of the microbial community present in a sample. Unfortunately biology is a chaotic business. One of the largest challenges while generating accurate microbiome count data is noise. Noise is defined as "random fluctuations that obscure or do not contain meaningful data or other information" (*OED Online* 2017). Noise in microbiome census

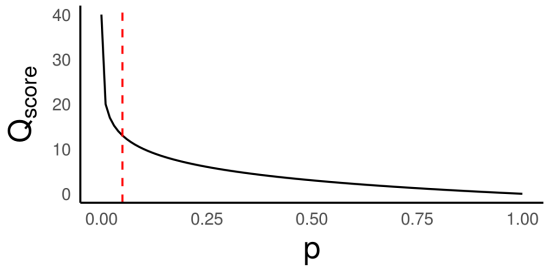


Figure 3.8: Relationship between the probability that a base call was correct (p) and the quality index of the base call (Q), see Equation 3.1. The dashed vertical red line indicates $p = 0.05$ ($Q \approx 13$).

data can be broadly separated into two categories: biological noise and technical noise. Biological noise is noise introduced by stochastic biological processes, and technical noise is noise introduced by the measuring processes that occur while taking a microbial census (Callahan et al., 2016b). Specific strategies can be used to minimise or correct noise (denoise), discussed below.

Biological noise

Microbes can have multiple slightly different copies of the 16S rRNA operon (intragenomic variation; Coenye and Vandamme, 2003). An operon is a unit of DNA containing a cluster of genes controlled by a single promoter (Sadava et al., 2009). Sequence fragments of the operon copies could appear to be from different species despite originating from the same cell. This has been known to affect species identification even during sequence clustering if the sequence fragments are different enough to not meet the 97% similarity criterion (Sacchi et al., 2002). It should be noted that this is distinct from the phenomenon of copy-number variation. The number of 16S rRNA operons can differ significantly across different species. This variance is driven by different ecological strategies (Klappenbach et al., 2000): a high number of 16S rRNA operons is associated with rapid growth (copiotrophs), and a low number of operons is associated with slow growth (oligotrophs). A frequent species in raw microbiome count data could represent a high-copy number taxon of low abundance, or vice versa. Microbiome count data can be adjusted to account for this: public databases of 16S copy numbers exist (Stoddard et al., 2015). Incorporating this information into microbiome count data analysis has been found to improve diversity and composition estimates (Kembel et al., 2012). Despite this, the adjustment is not a common step in microbiome pipelines.

In some cases the entire 16S sequence can be identical across multiple different species. The 1,412 nucleotide sequence of nearly the entire 16S rRNA gene was

found to be identical across six species in the *Brucella* genus (Gee et al., 2004). Despite the ease of differentiating *Brucella* species - most members of the genus were first characterised around the start of the 20th century (Moreno and Moriyón, 2002) and each species infects a different type of host - by a quirk of evolution its 16S sequence remains identical across the genus. This may hold true for other uncharacterised genera. However, technical noise - described below - is by far the largest contributor of systemic bias to microbiome count data.

Technical noise

Bacteria are very hard to kill: in environments which would rapidly kill complex multicellular life, bacteria can happily thrive. The composition of the bacterial cell wall contributes to this toughness. During DNA extraction the goal is to isolate DNA from the cell remnants while preserving the fragile DNA. It has been found that differences in the composition of the cell wall can cause bacterial lysis to be more or less efficient (Carrigg et al., 2007). The composition of the bacterial cell wall differs between microbes. Some bacterial species are easily lysed, while others are extremely resistant. This will introduce a systematic bias in the apparent composition of the microbial community: stubborn bugs will appear to be less abundant.

The process of copying 16S gene fragments with PCR prior to sequencing can introduce many different types of bias. Primer-template mismatches have been found to introduce quantitative biases (Parada et al., 2015). Primer-template mismatches are exacerbated by the popularity of universal primers. The abundance of microbes with 16S gene sequences that do not perfectly match (even by a single base) universal primers will be underestimated, as fewer copies will be made. This can simultaneously overestimate the abundance of bacterial taxa that do match the primers perfectly. The only way to detect this kind of bias is to combine domain expertise with *in silico* and mock community validation of primer pairs. The PCR copying process is not perfect: artificial base changes can be introduced (Brodin et al., 2013), which will inflate the number of unique sequences to be analysed (also confounding diversity and composition metrics). Attempts to denoise PCR amplicons (sequence copies) with proof-reading enzymes have been successful, but the number of chimeras (described below) is significantly increased (Schnell et al., 2015).

The Chimera was a fire-breathing hybrid monster from Greek mythology. It was formed from different parts of multiple creatures, and seeing a Chimera was an omen for disaster (Vogel, 2015). PCR chimeras are also an omen for disaster in microbiome experiments if they are not reduced, identified, and removed. PCR chimeras are sequences formed when two or more biological sequences join together (see Figure 3.9). Although the total number of chimeric reads overall is very low

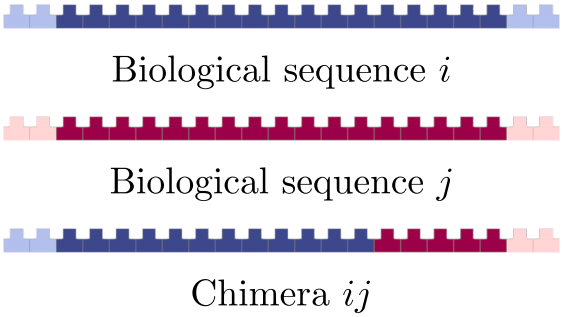


Figure 3.9: Chimera formation. The chimera would appear as a novel organism in downstream analysis unless removed. Derived from "PCR ssDNA" by Library of Science and Medicine Illustrations, licensed under CC BY-NC-SA 4.0.

the clustering stages of downstream analyses cause the chimeric reads to have a much larger impact. If chimeric sequences are not identified and removed then they will often appear as rare or novel bacteria, although the chimeric sequence is a combination of two or more other organisms. Chimeras are believed to form when DNA polymerase incompletely synthesises the new sequence during the extension stage of PCR (Smyth et al., 2010). The partially synthesised DNA sequences bind to different templates with similar sequences. This new sequence can then act as a primer that is extended to create a chimeric sequence.

Tag switching chimeras are an important chimera variant. If amplicons from different samples are pooled during sequencing (multiplexed) tag switching can occur. A key assumption of the multiplexing process is that amplicons can be correctly assigned to the samples from which they originated. If this assumption is violated false positive observations will occur during later analysis. Tag switching occurs when amplicons are copied during PCR. The barcode sections of amplicons can be incorrectly copied, and if faulty sequences match barcodes already used for other samples then amplicons will be misattributed to a different sample (Schnell et al., 2015). A study found between 2.1% and 2.6% of reads were found to have tag combinations (Schnell et al., 2015).

Sequencing error is a major source of noise in a microbiome experiment (Kozich et al., 2013). Sequencing error falls into three main categories: insertions, deletions, and substitutions. Insertions and deletions occur when a nucleotide is added or removed but a nucleotide is not actually present. A substitution occurs when a nucleotide is mistaken for another nucleotide (e.g. `A` is present but `T` is called). Illumina platforms are most commonly affected by substitution miscalls (Schirmer et al., 2015). A useful way to measure sequencing error is to co-sequence a sample with known proportions of bacteria (or more commonly genomic DNA that simulate a number of bacterial species) while sequencing other samples (Bokulich et al.,

2016). The sequenced data are processed through a bioinformatics pipeline, and the number of bacterial species identified is compared to the known quantity of bacteria present in the sample. From this process, the error rate of a sequencing run can be measured. A control sample that consists of a known mixture of microbial cells that mimics a metagenomic sample is known as a mock community. Mock communities are commonly used to benchmark different metagenomic pipelines against each other (Bokulich et al., 2016).

It is common for the number of identified bacteria to be in excess of the amount actually present in the sample (Callahan et al., 2016a). Sequencing error will introduce low abundance unique sequences to the pool of measured sequences. A sequence that is only one or two nucleotides different from another sequence can be treated as a different bacterial species by downstream algorithms. Additionally, the introduction of false novelty increases the difficulty of removing chimeras. A `mothur` (bioinformatics software for microbial ecology) standard operating protocol (Kozich et al., 2013; Schloss et al., 2009) reports that 31 bacterial species were observed from a mock community of 20 control species. This is equivalent to a sequencing error rate of $6.5 \times 10^{-5}\%$, which is the lowest reported in the literature. A more typical sequencing error of an Illumina MiSeq sequencing machine, without steps taken to minimise sequencing error, is a rate of between approximately 0.1-0.8% (Glenn, 2011; Quail et al., 2012).

3.3.5 From sequences to clusters

An operational taxonomic unit (OTU) was originally defined as the group of organisms currently being studied (Sokal and Sneath, 1963). The term OTU is now used to describe clusters of bacteria that have been grouped together via the relative similarity of specific marker genes (e.g. 16S rRNA). Matching OTU clusters to traditional taxonomy (e.g. mapping OTU 4 to *Bacillus subtilis*) is a difficult task. Practically, OTUs are considered to be analogous to a bacterial "species", although an OTU can represent any taxonomic level or may not resolve to any known taxonomy (e.g. uncharacterised microbes). The nomenclature has become popular due to the difficulties in defining what exactly a bacterial species is. Broadly speaking three paradigms are popular for identifying OTU clusters: phylotyping, *de novo*, and open-reference clustering. A clustering approach with hard similarity thresholds was first adopted in an attempt to counteract noise introduced from sequencing error (Amir et al., 2017).

Phylotyping

The phylotyping approach matches sequences to a curated taxonomic database (Stocker et al., 2011). Matches are binned into "phylotypes" at different taxonomic

ranks via similarity score thresholds. The key advantages of the process are its relative insensitivity to sequencing error and its computational efficiency (hundreds of samples can be processed on a laptop; Rideout et al., 2014). However, this is a side effect of the poor resolution that phylotyping offers: genus is the lowest taxonomic rank that phylotyping can assign a sequence to. Other methods can identify species, subspecies, and even specific sequence variants (Callahan et al., 2016a). Noise introduced from sequencing error is only reduced because phylotyping cannot detect it.

Many reference databases are a work in progress, and if a sequence cannot be matched against the reference it is discarded. Phylotyping can only measure what is already known (i.e what has already been characterised in a reference database). Rare or previously unknown organisms missed by phylotyping will affect every aspect of analysis (e.g. diversity estimates or differential abundance tests). Different environments are characterised at varying levels. The human gut is reasonably well characterised, as much work has been done to understand what a healthy human gut is (MetaHIT Consortium, 2011; Aagaard et al., 2013). Phylotyping a sample gathered from a rainforest floor would probably result in the vast majority of sequences being discarded.

During the phylotyping process sequences are not compared with one another. Instead, because sequences are compared only with the reference database, two dissimilar sequences can be binned into the same phylotype (Westcott and Schloss, 2015). For example, two sequences can match to a reference at 97% similarity but only be 94% similar to each other (see Figure 3.10). Furthermore, the sequences present in the reference database must be least 3% dissimilar to other reference sequences over the entire length of the gene (Westcott and Schloss, 2015). It is important to realise that phylotyping only considers gene fragments of the hypervariable regions, which have been shown to evolve at a different rate to the rest of the gene (Kim et al., 2011). Therefore the gene fragment sequences being phylotyped can be at least 97% similar to multiple reference sequences, despite the references being 3% dissimilar to one another over the full length of the 16S rRNA gene.

De novo clustering

Distance-based (Schloss and Westcott, 2011) or *de novo* (Navas-Molina et al., 2013; from the Latin for "of new") clustering uses the distance between sequences to cluster sequences into OTUs. As all sequences must be compared to one another at runtime, the computational complexity of *de novo* clustering algorithms scales approximately quadratically with the number of input sequences. It has been shown that nearly all unique sequences arise from sequencing error (Kozich et al., 2013). The inflated number of unique sequences significantly increases the memory

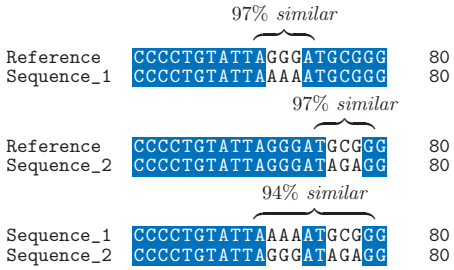


Figure 3.10: Sequences are binned into phylotypes via comparison to a taxonomic reference database but not to each other.

and time requirements of *de novo* clustering. A *de novo* clustering algorithm must be paired with specific noise reducing sequencing strategies, as mentioned in Section 3.3.3, in order to minimise sequencing error: doubling the number of sequences causes a four-fold increase in runtime. A key strength of *de novo* clustering is its independence from reference databases for the clustering process (reference databases are still required to assign human-readable taxonomy to an OTU). An unresolved problem with *de novo* algorithms is that they are sensitive to the input order of sequences (Mahé et al., 2014).

Heuristics which incorporate a pre-clustering strategy offer an alternative to the necessity of specific sequence strategies and strict quality control. `USEARCH` and `VSEARCH` are two microbial ecology bioinformatics software packages that implement two heuristic approaches including distance-based greedy clustering and abundance-based greedy clustering (Edgar, 2010; He et al., 2015). `VSEARCH` is an open source alternative to `USEARCH` (Rognes et al., 2016). A wide variety of *de novo* clustering algorithms are available, including single linkage, complete linkage, average linkage, heuristic-based, and Swarm (Schloss et al., 2009; Rognes et al., 2016; Mahé et al., 2014). A thorough explanation of these algorithms is available in Westcott and Schloss (2015).

Open-reference clustering

Open-reference clustering is a hybrid approach, created by the developers of Quantitative Insights Into Microbial Ecology (`QIIME`), that can scale to up to billions of input sequences (Rideout et al., 2014). Sequences are firstly binned into phylotypes. A small proportion (0.3% by default) of the unmatched sequences are then clustered via *de novo* methods. The motivation behind open-reference clustering is to improve scalability while maintaining consistency versus OTUs generated by other methods. Although this approach is supposed to combine the strengths of phylotyping and *de novo* clustering while minimising the weaknesses, serious problems have been found with the algorithm (Westcott and Schloss, 2015).

Approximately ten thousand false positive OTUs were picked with open-reference clustering from a mock community composed of twenty species (Kopylova et al., 2016). The quality of OTUs picked by `QIIME`, one of the most popular microbiome workflows (cited over 7,000 times on Google scholar to date), has been called into question because of this.

Performance summary

Quantifying the performance of an OTU clustering algorithm is difficult and many different metrics have been used. Some focus on scalability (Rideout et al., 2014; Mahé et al., 2014). While this is a good metric for assessing the performance of a clustering algorithm in isolation, this is unsuitable for measuring the quality of the clustering output (i.e. do the picked OTUs reflect the true bacterial community present?). Some early work tried to compare picked OTUs to simulated data drawn from bacterial taxonomies (White et al., 2010). This approach is flawed: taxonomic schemes are created by humans, and are susceptible to many different kinds of bias (historical problems were described in Section 3.3.3). Comparing the ability of an OTU picking algorithm to generate OTUs similar to other existing methods (Rideout et al., 2014) is also a flawed approach: consistency is of limited use if all the existing approaches generate poor OTUs. A more sensible benchmark is to sequence a mock community of known organisms and to compare the number of picked OTUs with the types of organisms that are known to be present (May et al., 2014; Mahé et al., 2014; Kopylova et al., 2016). Critics of this approach note that mock communities are overly simplistic, and the techniques that perform best on a mock community may not transfer to real-world complexity (Westcott and Schloss, 2015). Put simply, pipelines that produce less OTUs are not always better.

Overall, *de novo* methods are the best approach when clustering sequences into OTUs. There is no single best *de novo* clustering algorithm for all datasets: it is best to assess the performance of the algorithms for new datasets, with a preference towards open source algorithms to aid reproducible research. Given the myriad of problems associated with clustering sequences into OTUs, alternative approaches are appealing. Although similarity thresholds are a computationally convenient way of mitigating artificial variation introduced by sequencing error, there is a growing realisation that OTUs may not be ecologically meaningful and may not represent phylogenetically unmixed units of bacteria. Bypassing the concept of a similarity threshold is key to overcoming the limitations of the OTU approach. Recent proposals suggest that the OTU paradigm should be replaced in favour of denoising strategies that can identify exact sequence variants (Callahan et al., 2017). Exact sequence variants have consistent labels with intrinsic biological meaning and are identified without the use of reference databases, which improves the accuracy, reproducibility, and reusability of microbiome experiments.

3.3.6 The problem with thresholds

Enforcing hard similarity thresholds can cause picked OTUs to be ecologically irrelevant. The clustering process can discard informative sequence variation and can group together ecologically distinct bacteria resulting in a phylogenetically mixed unit (Shapiro and Polz, 2014). A number of algorithms that avoid using a hard global similarity threshold have been developed recently. Eren et al. (2013), defined the process of generating high resolution sequence variants as oligotyping, which outputs a count of high-resolution OTUs called oligotypes. Oligotypes of *Pelagibacter* sampled from Cape Cod that are 99.6% similar were found to have remarkably different fluctuations with seasonal changes in water temperature (Eren et al., 2013), demonstrating the benefit of threshold-free approaches.

Threshold-free approaches share a common goal: to report the exact sequences of gene fragments present in a sample. This avoids the use of arbitrary operational definitions of bacteria. Despite this common goal, the terminology used to describe the output of threshold-free approaches varies considerably (see bottom of Table 3.2). Given that the goal has changed between standard OTU approaches and the new threshold-free paradigm, it is sensible to avoid the use of OTU terminology to avoid confusion. `dada2` labels exact sequences as amplicon sequence variant (ASV). The term ASV will be used throughout the thesis for consistency. Aside from improved resolution, the benefits of using ASV are numerous. ASV labels are consistent across experiments - the label of a bacterial unit is its exact sequence - improving reproducibility. *De novo* OTUs cannot be compared across experiments as sequences must be clustered simultaneously at run time. This makes large scale replication studies computationally infeasible. ASV labels have intrinsic biological meaning, and are identified without using reference databases.The computational costs of generating ASV typically scale linearly with the number of input sequences. Due to these benefits it has been proposed to replace the use of OTUs with ASVs across the field (Callahan et al., 2017). The methods of generating high-resolution amplicon data are diverse, and a full review of every algorithm is outside the scope of this thesis. The `dada2` software package is used throughout this thesis to generate microbiome count data, and is discussed in detail in Section 4.2 (see Figure 3.11).

Table 3.2: Summary of algorithms that assign gene fragments a label. Blank spaces indicate repeating information.

Package	Algorithm	Output	Goal	Reference
`mothur`	`phylotype`	Phylotypes	Match sequences to reference database	Schloss et al., 2009
`QIIME`	`uclust` (closed)			Caporaso et al., 2010; Edgar, 2010
`mothur`	Neighbour joining	OTUs	Cluster sequences independently of reference database	Schloss et al., 2009
`QIIME`	`uclust` (open)		Pick OTUs consistently; scale to billions of sequences	Caporaso et al., 2010; Edgar, 2010
`USEARCH`	`uclust` (*de novo*)		Cluster independently; use heuristics to improve performance	Caporaso et al., 2010; Edgar, 2010
`CD-HIT Suite`	`CD-HIT-OTU`			Huang et al., 2010
`USEARCH`	`UPARSE`		Heuristic clustering with fewer false positive OTUs	Edgar, 2013
`VSEARCH`	`VSEARCH`		Heuristic open source alternative to USEARCH	Rognes et al., 2016
`swarm`	`swarm`		Heuristic local threshold clustering resilient to input order	Mahé et al., 2014
`mothur`	`opticlust`		Improve performance by optimising the Matthews correlation coefficient	Westcott and Schloss, 2017
`oligotyping`	MED	Oligotypes	Report exact sequences	Eren et al., 2013; Eren et al., 2015
`dada2`	DADA	ASV		Callahan et al., 2016a
`USEARCH`	`UNOISE2`	zOTU		Edgar, 2016
`deblur`	`deblur`	subOTUs		Amir et al., 2017

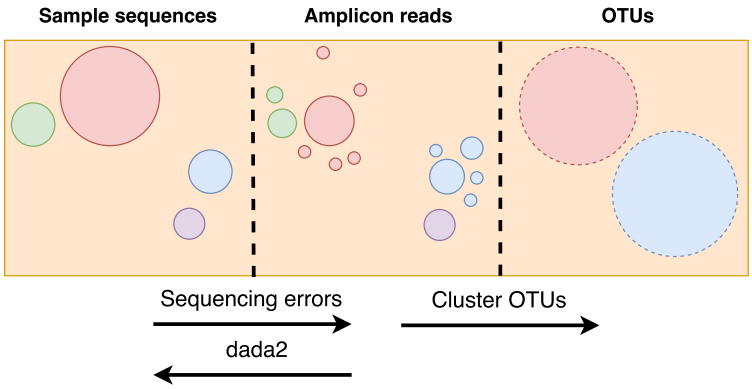


Figure 3.11: Circles represent sets of identical sequence reads. Colours represent the true biological sequences present in the sample. OTU methods cluster similar reads together to counteract sequencing error. `dada2` infers the exact sequence variants truly present in the sample. Adapted from supplement of Callahan et al. (2016a).

3.3.7 It's hard to be normal

Normalisation is essential to remove bias and variation introduced during sampling and sequencing. Normalisation is defined as the process of transforming data to enable fair comparison of measurements gathered from different samples by eliminating artefacts that arose during the measuring process (Weiss et al., 2017). Many normalisation processes widely applied to microbiome count data have been found to introduce errors and bias. Proper normalisation techniques are essential for experimental results to be considered valid.

Microbiome count data consists of discrete counts of bacterial units or specific DNA sequences (Weiss et al., 2017). The total number of reads per sample (also known as the depth of coverage or library size) will often vary by orders of magnitude within a single sequencing run (see Table 3.3; Caporaso et al., 2011). This variation is introduced as a technical artefact from the high-throughput sequencing process and does not reflect true biological variation. In order to compare microbiome samples to each other they must often be normalised to take into account uneven library size. If this is not corrected for then correlations between taxa will become distorted (Weiss et al., 2017).

Heteroscedasticity describes the situation in which a dataset has unequal variance over a second predictor variable (McMurdie and Holmes, 2014). Figure 3.12 demonstrates that the mean-to-variance ratio of microbiome count data is not stable (i.e. the data are heteroscedastic). This results in overdispersion being observed when traditional statistical models based on the Poisson distribution are

Table 3.3: Distribution of sample library sizes in the Global Patterns dataset (Caporaso et al., 2011). The sample with the smallest library size is two orders of magnitude smaller than the largest.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
58688	567103	1106849	1085257	1527330	2357181

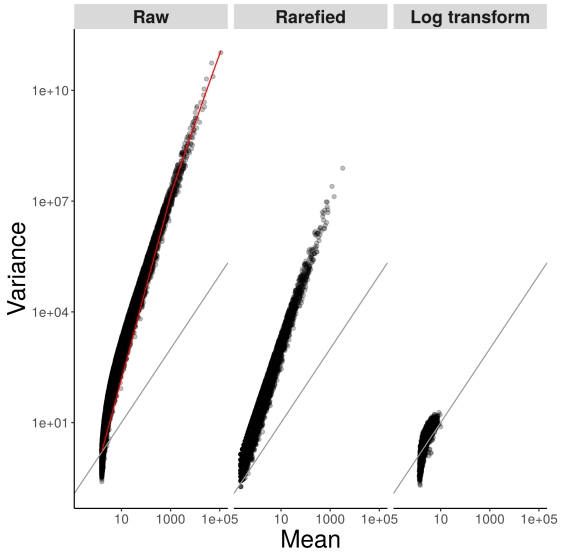


Figure 3.12: Overdispersion in microbiome count data can be observed by comparing the common-scale variance versus mean (McMurdie and Holmes, 2014). Each point shows the estimated mean and variance of an OTU across all biological replicates in the dataset. The red curve shows the fitted variance estimate calculated by `DESeq` (Anders and Huber, 2010).

applied to microbiome count data (McMurdie and Holmes, 2014). Overdispersion is the presence of greater statistical variability than is expected by a given model. There are biological reasons for heteroscedasticity occurring in microbiome count data (e.g. the exponential growth of bacteria; Bálint et al., 2016).

Failing to address heteroscedasticity can lead to several problems. Using heteroscedastic data with standard parametric statistical tests of significance is not appropriate, as parametric tests assume that variance is equal across samples (McMurdie and Holmes, 2014). Heteroscedasticity decreases the ability of downstream tests to detect a multivariate effect present in low-variance taxa and makes it difficult to detect taxa that drive an effect. In addition, heteroscedasticity confounds location and dispersion effects (Warton et al., 2012).

Scaling counts with total sum scaling (proportions), where the sum of all bacteria is 1 for each sample, is a simple way of standardising library size (see Equation 3.2). Proportion $p_{i,j}$ is calculated from $x_{i,j}$, where $x_{i,j}$ denotes the count of OTU i in the j^{th} sample).

$$p_{i,j} = \frac{x_{i,j}}{\sum_{k=1}^{n} x_{i,k}} \tag{3.2}$$

Although this naïve approach is simple to apply and resolves the effect of different library sizes it is inappropriate because it does not resolve the heteroscedasticity present in the data. Additionally, total sum scaling transforms microbiome count data into compositional data (explained below), bringing additional challenges to analysis. Compositional data are vectors of positive elements constrained to sum to a constant (Aitchison et al., 1994). Friedman and Alm were the first to recognise that common normalisation processes cause microbiome count data to become compositional, and to highlight the problems associated with this (Friedman and Alm, 2012). This observation has been extended to essentially all data derived from high throughput sequencing such as RNA-seq (Fernandes et al., 2014). Compositional data are afflicted by the closure problem: elements compete to form the constant sum constraint (Aitchison, 1986). In practice this means that large changes in the absolute abundance of one element of the vector will artificially suppress the abundance of other elements. This violates assumptions of sample independence and introduces bias. Standard statistical tests will produce spurious correlations and false positive or negative results when applied to compositional data. For example, if a particular taxon increases in abundance spurious negative correlations will be introduced for less abundant taxa if measured with standard tests (e.g. the Pearson correlation coefficient; Friedman and Alm, 2012). In addition, as the library size of a collection of samples is determined by the capacity of the sequencing instrument even unnormalised sequencing data are compositional (Gloor and Reid, 2016).

I didn't like that data anyway: The rarefying approach

Rarefying, or random subsampling without replacement, is an approach that is present across all major microbiome and microbial ecology toolkits (Caporaso et al., 2010; Schloss et al., 2009; Oksanen et al., 2015) that corrects uneven library sizes across samples (see Table 3.4). This normalisation process is sometimes mistakenly referred to as rarefaction across literature and toolkits. It is important to distinguish between the data normalisation process and the technique used by ecologists to generate taxon re-sampling curves in order to assess the coverage or richness of a sample. The data normalisation process will be exclusively referred

Table 3.4: The effect of rarefying. Left: raw abundances. Right: after normalisation with rarefying. Data are hypothetical. It should be noted that due to the random nature of the process the rarefied count of sample B will not necessarily be even.

	Sample A	Sample B
OTU i	60	500
OTU j	40	500
Total	100	1000

	Sample A	Sample B
OTU i	60	50
OTU j	40	50
Total	100	100

to as rarefying. The latter procedure is discussed in Section 5.2, and will only be referred to as rarefaction.

Rarefying does not correct heteroscedasticity (see middle panel of Figure 3.12), transforms microbiome count data to compositional data, and decreases statistical power by discarding observations. It has been suggested that a great deal of work that has used rarefied counts is *statistically inadmissible* (McMurdie and Holmes, 2014). Despite this, rarefying is still recommended by popular standard operating procedures, reviews (Kozich et al., 2013; Weiss et al., 2017), and is incorporated into many automated workflows. Rarefying involves the following steps:

1. Set a minimum library size $N_{L,\text{lim}}$
2. Discard samples that have fewer reads than $N_{L,\text{lim}}$
3. Randomly subsample remaining libraries without replacement to match size $N_{L,\text{lim}}$

Rarefying poses problems for transparency and reproducibility. $N_{L,\text{lim}}$ is usually chosen to be the size of the smallest library that meets a specified cut-off. For example the Forsyth Institute, the centre that sequenced the microbiome count data discussed in Chapter 5, recommends removing samples that have less than 5000 discrete reads. The arbitrary nature of this cut-off is vulnerable to subjectivity and bias. Additionally, the random portion of the subsampling procedure adds noise while failing to add anything of value to the process. Often the seed used to initialise the pseudorandom number generator is not recorded. Without this seed rarefied microbiome count data cannot be reproduced from raw sequencing data.

The largest problems with rarefying are related to its reliance on discarding data, a process which makes statisticians very angry. McMurdie and Holmes, the first to highlight the problems described above, noted that both Type-I (decreased specificity) and Type-II error (loss of power) is increased after rarefying. When clustering samples the loss of power manifested in two ways: samples that were

not classified because they were discarded, and samples that could not be clustered because of the discarded data that were present prior to rarefying. During differential abundance testing, the loss of power manifested as the inability of tests to correctly identify significantly different rare to moderate taxa (McMurdie and Holmes, 2014).

Two out of three ain't bad: Log transformations

Log transformations are often used to correct data with unequal variances and positive skew. Additionally, microbiome count data transformed with a log transformation is not compositional. Microbiome count data are often sparse (Paulson et al., 2013), with many zeroes present in the data. $log_2(0)$ is undefined, so a log transformation cannot be applied to sparse data. In order to log transform sparse data a small positive constant is added to the data, also known as a pseudocount. The generalised log transform of $x_{i,j}$ is given by:

$$y_{i,j} = log_2(x_{i,j} + x_0) \tag{3.3}$$

where $y_{i,j}$ gives the transformed value, $x_{i,j}$ gives the count of the i-th OTU from the j-th sample, and x_0 gives a positive constant (usually 1; Paulson et al., 2013). The theoretical justification for a pseudocount is that it represents a value below the detection limit of the sequencing process (Paulson et al., 2013). Applying a log transformation to microbiome count data acts as an approximate variance stabilising transformation (Callahan et al., 2016b). The approximate nature of the transformation can sometimes fail to stabilise the variance, and generally does not fully resolve the problem. The right panel of Figure 3.12 shows that the transformation did not completely remove the trend, particularly for less abundant taxa. Additionally, the log transforming can crush the data at low to medium abundances (see Figure 3.13) compared with other transformations that measure and incorporate the mean-variance relationship during the transformation process, as implemented in the DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) software packages. An unresolved problem with the use of pseudocounts is selecting the value of x_0. The choice of x_0 can change the results of downstream analyses dramatically (Costea et al., 2014; see Figure 3.14). It is also important to note that log transformations do not resolve uneven library sizes. One proposed approach to counteract this is to include the library size of a sample into later multivariate analysis (Bálint et al., 2016).

Although the log transformation does not normalise uneven library sizes across samples, it is simple to apply and can be considered to be *good enough* for many analysis tasks, despite the troubling use of pseudocounts. Log transformations are often used in downstream applications that cannot tolerate negative numbers

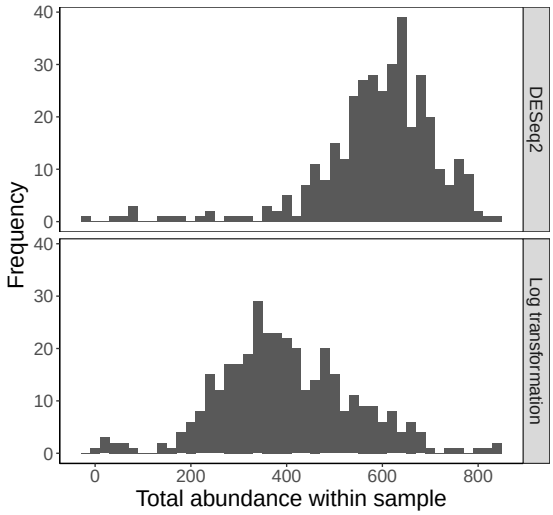


Figure 3.13: Common-scale DESeq2 transformed abundance of mouse gut microbiome count data (top; Schloss et al., 2012) and log transformed abundance ($log_2(x+1)$; bottom). Modified from Callahan et al., 2016b.

produced by these theoretically superior procedures. For example, the Bray-Curtis dissimilarity measure (Bray and Curtis, 1957) requires non-negative counts.

3.4 Computational intelligence in microbial ecology

Computational Intelligence (CI) approaches have not been widely applied specifically in microbiome research to date, but efforts have been made to apply CI strategies in both biomedical and bioinformatics applications. ANN variants have been most widely applied to microbiome count data. Self-organising maps - a type of unsupervised neural network that can cluster multidimensional data and visualise it in a two-dimensional map - have been used to cluster genome signatures and model the way environmental factors impact their distribution in a microbial community (Dick et al., 2009). Deep learning approaches have been applied to metagenomic data in order to learn hierarchical representations of the dataset (Ditzler et al., 2015). Deep learning is the study of neural networks with more than one hidden layer (Deng, Yu et al., 2014). Two datasets were analysed: in the first, metagenomic samples were classified according to what area of the human body they were sampled from. In the second dataset, metagenomic samples were classified according to the pH of the environment they were sampled from (originally continuous; binned into

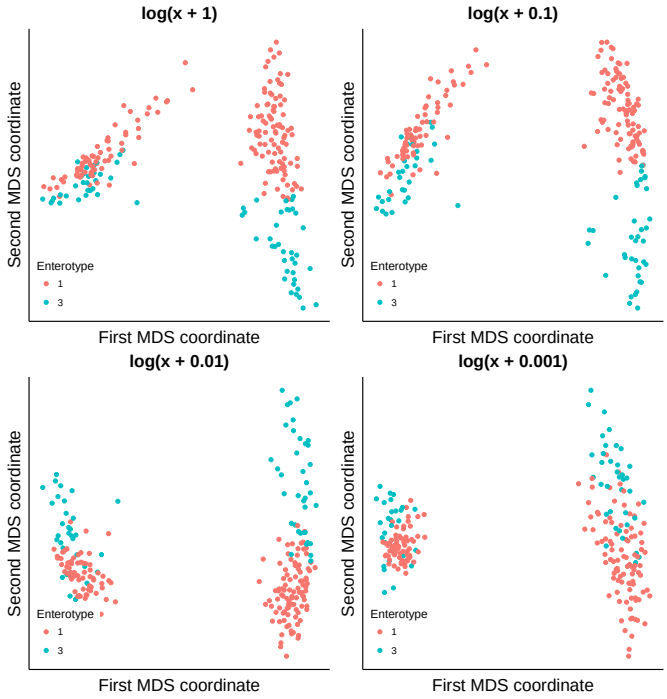


Figure 3.14: Clustering analysis via multidimensional scaling (data are from Arumugam et al., 2011). Pseudocounts exponentially decrease from 1 to 0.001, which significantly changes the clustering output.

high, medium, or low). Although a deep learning approach did not significantly improve the accuracy of classification, it did allow a tree structure to be generated and analysed via a recursive neural network.

Genetic and evolutionary feature selection has been used to identify a subset of OTUs present in the vaginal microbiome that can classify bacterial vaginosis (Carter et al., 2014). The algorithm used the relative abundance of OTUs and patient metadata as input. The aim of the experiment was to develop a more accurate and objective diagnostic test compared with current clinical practice. Two diagnostic tests are currently used clinically: the Amsel criteria ("any symptom approach") and the Nugent score. The Amsel criteria diagnoses bacterial vaginosis if at least one of the following symptoms occurs: the presence of discharge, a positive "whiff test", or a pH greater than 4.5 (Amsel et al., 1983). The Nugent score uses a scale from 0-10. Bacterial cells are imaged with a microscope, and cells similar in size and shape to *Lactobacillus* species are counted (Nugent et al., 1991): for disease to be diagnosed the score must be greater than 7. Up to a third of bacterial vaginosis diagnoses using these two tests are false positives (Forney et al., 2006). Treating healthy women with broad-spectrum antibiotics can cause long term damage to the

microbiomes across the body (Zaura et al., 2015), increasing the risk of infection by pathogenic species such as *Clostridium difficile* (Theriot et al., 2014). Carter et al. (2014), achieved an accuracy of 99.5% after 8000 iterations with Genetic and Evolutionary Feature Selection.

A fuzzy alternative to traditional distance matrices has been proposed specifically for high-throughput metagenomic sequencing data (Krachunov et al., 2015). The pairwise distance between gene fragment sequences is often calculated (e.g. prior to multiple sequence alignment) using the Hamming distance (Pinheiro et al., 2005). The distance H between sequences j and k of length n is given by:

$$H(j,k) = \sum_{i=1}^{n} \frac{[j_i \neq k_i]}{n} = \frac{\sum_i [j_i \neq k_i]}{\sum_i 1} \tag{3.4}$$

Krachunov et al. implemented a fuzzy distance by taking into consideration the confidence score $s(j,i)$ of position j in base call i (equivalent to Q_{phred} described earlier). With this established, Krachunov et al. introduced fuzzy pairwise alignment to counteract the effects of sequencing error. This theoretical approach would be more beneficial for shotgun sequenced metagenomic data, as sequences generated by marker gene surveys are usually already clustered into OTUs.

3.5 Summary

The microbiome has been shown to influence many different diseases. This thesis begins with a focus on IBD, because many public datasets are available, which helped to develop the bioinformatics pipeline used in the rest of the thesis. After IBD the thesis transitions to a focus on depression, specifically major depressive disorder. Depression is a complex disease with heterogeneous aetiology, diagnosis, treatment, and prognosis. Possible mechanisms for the modulation of behaviour by the microbiome are explained by the microbiome-gut-brain axis; the potential mechanisms of action are numerous and varied. The microbiome has been repeatedly shown to program the HPA axis; HPA axis dysfunction is one of the most consistent markers of depression. Pathogenic bacteria have been found to activate stress circuits via stimulation of the vagus nerve, which directly impacts the central nervous system. Other mechanisms include the epigenetic effects of short chain fatty acid byproducts, and the microbial synthesis of neurotransmitters.

Standard methods for generating microbiome count data rely on clustering for two reasons: to compensate for sequence variation that is introduced by technical noise, and because there is no universal definition of a bacterial species. The vast majority of unique sequences generated by standard methods arise from technical noise such as sequencing error. If a clustering approach was not used, biological variation would quickly be hidden by technical artefacts such as artificial base

changes, chimeras, and sequencing error. However, the 97% similarity threshold widely used across the scientific community has been shown to generate ecologically mixed units of bacteria, which can confound analysis. Thus, the popularity of similarity thresholds stems partly from computational convenience rather than good practice. The process of clustering sequences into OTUs is complex, and subtle changes can generate wildly inaccurate microbiome count data. It is common for ten thousand OTUs to be observed from a mock community sample of tens of bacteria. Even the best performing clustering algorithms come with caveats: the algorithms tend to scale poorly with study size, and OTUs cannot be compared across studies, hindering reproducible research.

A better approach is to increase the resolution of OTUs by grouping bacteria into exact sequence variants. The algorithms that achieve this vary widely in methodology. Some model sequencing errors and infer denoised sequences (divisive amplicon denoising algorithm (DADA)), while others use information theory to iteratively minimise entropy (minimum entropy decomposition (MED)). Experimental evidence has shown the benefits of this approach: for example, units of bacteria 99.6% similar to each other have been shown to have vastly different abundances in relation to ocean temperature. Intuitively members of a single ecological unit should have the same response to an environmental condition. This approach also increases the dimensionality of the data, rendering it more challenging to analyse compared with the output of standard clustering methods.

Microbiome count data are difficult to analyse. Technical and biological noise must be accounted and compensated for. Even perfect microbiome count data will have uneven library sizes across samples (which infers a level of certainty of the sampling process) and be heteroscedastic. Common methods for mitigating uneven library sizes introduce a different kind of challenge by converting the data to compositional data, which breaks the assumptions of many standard analysis protocols. Data driven CI algorithms offer a way to compensate for these issues. This thesis will investigate the following problems, which have not been answered to date:

- Given the high variability of microbiomes across subjects, can a robust set of microbial markers be found that can predict Inflammatory Bowel Disease that can generalise well?
- Are significant differences present in the structure and composition of the oral microbiome in depressed subjects compared with control subjects?
- Can data driven CI algorithms be used to analyse microbiome count data in order to compensate for microbiome count data properties such as uneven library size and heteroscedasticity?

ROBUSTLY PREDICTING INFLAMMATORY BOWEL DISEASE

> Never trust a computer you can't throw out a window.
>
> Steve Wozniak

4.1 Introduction

The first projects that aimed to characterise the composition of a healthy human microbiome also aimed to identify associations between the composition of the gut microbiome and Inflammatory Bowel Disease (IBD) and obesity (MetaHIT Consortium, 2011; Turnbaugh et al., 2007). IBD caused 53,000 deaths worldwide in 2013 and its prevalence has been increasing throughout the developed world for decades (Molodecky et al., 2012). IBD symptomatology is generally non-specific and diagnosis is usually confirmed via invasive colonoscopy, with consequent delays. Delayed paediatric IBD diagnosis can reduce growth and is linked to poor treatment outcomes. First attempts to model the data generated by these projects relied on analytic approaches borrowed from ecology and the application of simple classification algorithms. Machine learning algorithms quickly grew popular due to the complexity of the data. Benchmarks found that support vector machines (SVMs) and Random Forests generally performed well on microbiome census data (Statnikov et al., 2013), and these models were quickly integrated as a standard step in microbiome analysis workflows. Both SVMs and Random Forests are useful for their ability to perform well on highly dimensional data and to generate feature subsets to identify bacterial species that are associated with disease. However, current models use simple labels for classification tasks (e.g. single label health or disease) and the robustness of feature selector algorithm output has not been considered to date. Additionally, models for IBD prediction have for the most part relied on on taxonomic data (i.e. what species are present?) rather than functional data (i.e. what are the species doing?). IBD was chosen for analysis throughout this chapter due to the large amounts of public data available and to gain experience with bioinformatics algorithms for processing the 16S ribosomal ribonucleic acid (16S rRNA) marker gene survey data. 16S rRNA data provides taxonomic data that describes the structure and composition of microbial communities and is cost-effective to collect, process, and analyse: sequenced 16S rRNA data is small enough to be processed by most university laboratories and analysed on standard

desktop workstations (other types of sequence data can require larger laboratories and high performance clusters of computers).

The structure of this chapter is as follows. A brief background is provided in section 4.2 to explain the bioinformatics algorithms applied throughout this chapter and biological aspects of IBD that are relevant to later analysis. A hybrid model is then implemented to decompose a complex problem into a series of simpler classification tasks. The resulting hybrid model — described in section 4.3 — is capable of diagnosing the presence of IBD, identifying the subtype of IBD if it is present, and predicting the current severity of IBD if the disease is in its active state. Furthermore, the concept of aggregating ensemble feature selection (EFS) will be applied to high-resolution microbiome census data to improve the power of non-invasive IBD prediction and for knowledge discovery. Section 4.4 will outline how EFS can be used to create a robust subset of bacterial species that can be used to classify IBD subtypes with the highest performance described in the literature to date. Biologically plausible novel bacterial species are shown to be implicated in the aetiology of IBD by the EFS procedure in section 4.4.4. This chapter concludes with a summary in section 4.5.

4.2 Background

A deregulated immune response to changes in the composition of the gastrointestinal microbiome (dysbiosis) implicates the microbiome in the aetiology of IBD (Halfvarson et al., 2017). Each subtype of IBD (e.g. ileal crohn's disease (CD)) has been associated with distinct microbial signatures. Industrialised western nations have the highest IBD incidence and prevalence - approximately 261,000 people suffer from IBD in the United Kingdom - and this has increased significantly worldwide since the start of the 20th century (Molodecky et al., 2012). IBD symptoms include abdominal pain, weight loss, and diarrhoea. In severe cases surgical intervention is required and the inflamed parts of the gastrointestinal tract are removed. IBD is a complex disease with uncertain aetiology (Hanauer, 2006). IBD has two major subtypes: ulcerative colitis (UC) - the effects of which are limited to the gut - and CD, which can affect the entire gastrointestinal tract. IBD is usually episodic and severe inflammation is considered to be active IBD. IBD can enter remission during periods in which limited or no symptoms occur. IBD diagnosis is slow in children because IBD has non-specific symptoms; Colonoscopy is a specialised procedure and IBD symptoms are required before colonoscopy will be used for confirmation. Thus further development of non-invasive tests for IBD would be valuable to improve treatment outcomes.

$$\begin{array}{ccccc} \text{Microbe}_1 & \dots & \dots & \dots & \text{Microbe}_M \\ \left[\begin{array}{ccccc} 0 & 9 & 8 & 0 & 7 \\ 3 & 5 & 6 & 5 & 1 \\ 1 & 0 & 0 & 3 & 2 \end{array}\right. & & & & \left.\begin{array}{l} \text{sample}_1 \\ \dots \\ \text{sample}_N \end{array}\right] \end{array}$$

Figure 4.1: Example unnormalised community data matrix.

4.2.1 Data used throughout this chapter

Two publicly available datasets are analysed throughout this chapter:

- The hybrid model presented in section 4.3 uses a dataset of 158 children (control n=37, IBD n=*122*, Papa et al., 2012, see Table 4.1);
- The ensemble feature selection approach presented in section 4.4 uses a dataset of 1485 samples gathered from the gastrointestinal tract of treatment-naïve children and adults (see Table 4.2, Gevers et al., 2014).

It is important to note that subjects in the Papa et al. dataset were not treatment naïve: many had a range of treatments including antibiotics and steroids prior to sampling. Adults were discarded from the Gevers et al. dataset due to insufficient numbers, and children were defined as being $\leq$ 16 years old (per the A1 Montreal classification of IBD; Silverberg et al., 2005). Samples were collected at disease onset at the time of diagnosis in the Gevers et al., so IBD was in an active state. The Gevers et al. dataset included samples collected via biopsy and stools. This chapter focused on stool samples in order to develop a set of robust markers that can be used to non-invasively predict IBD. Only stool samples ($n = 311$) remained after discarding the biopsy samples.

The publicly available data were available in the form of sequenced 16S rRNA DNA from the Sequence Read Archive (SRA). The sequenced DNA data were processed with various bioinformatics tools, which are described further in the following subsections. The output of the bioinformatics tools is microbiome census data. Microbiome census data form an $N \times M$ matrix of integers where N is the total number of samples and M is the total number of unique sequences or operational taxonomic units (OTUs) observed across all samples (see Figure 4.1). Microbiome census data are highly dimensional: approximately 4500 amplicon sequence variants (ASVs) were identified from the Gevers et al. data.

Table 4.1: Demographic data of Papa et al. dataset.

		CD	UC	Control
n		48	73	37
Gender	Male	29	39	16
	Female	19	33	21
Age	Median	14.6	13.4	11
	Range	3 — 23	4 — 24	3 — 21
Montreal class	L1	4		
	L2	1		
	L3	22		
	L4	7		
	B1	40		
	B2	6		
	B3	2		
	E1		25	
	E2		12	
	E3		36	
Disease activity	Control	0	0	37
	Inactive	29	26	
	Mild	11	22	
	Moderate	5	15	
	Severe	3	10	

4.2.2 Generating operational taxonomic units with `uclust`

For the development of the hybrid model a standard OTU approach was used to identify bacterial species. The open reference OTU picking method was used as it was the default algorithm recommended by the developers of Quantitative Insights Into Microbial Ecology (`QIIME`) (Navas-Molina et al., 2013), which was used to process the Papa et al. data that was input to the hybrid model. Open reference OTU picking is a combination of closed reference OTU picking (database matching, see Algorithm 4.1) and *de novo* OTU picking (see chapter 3.3.5 and Algorithm 4.2). The open reference algorithm is scalable to billions of input sequences but is still capable of identifying novel bacterial sequences that are not present in reference databases (see Algorithm 4.3).

Table 4.2: Demographic data of Gevers et al. data set. Only stool samples were used for analysis, and montreal class information was only available for around two thirds of the data.

		Gevers et al.
Number of subjects		1485
Disease status	Control	19%
	CD	58%
	UC	18%
	Indeterminate Colitis (IC)	5%
Age (mean)		23
Disease duration (mean)		0
Total samples		2308
Stool samples		28%
Biopsy samples		72%
Montreal class (CD)	L1	24%
	L2	23%
	L3	53%
	B1	90%
	B2	6%
	B3	2%

4.2.3 Inferring a functional profile

A number of algorithms have been developed that can infer the functional composition of a metagenome using marker gene data and reference genomes such as Tax4Fun and PICRUSt (Langille et al., 2013). In many environments, such as the human gut, the majority of the bacterial species present are well characterised and have had their full genomes sequenced. The first approaches that attempted to predict functional content from marker gene surveys used a relatively simple procedure to map a subset of abundant 16S rRNA gene sequences to closely related reference genomes (Morgan et al., 2012). PICRUSt formalises this approach into an automated algorithm and extends the concept to include a modified ancestral state reconstruction (ASR) approach (see Figure 4.2). The core concept behind PICRUSt is that phylogeny and function are strongly correlated. It is common for microbial ecologists to predict the function of novel bacterial species from closely related cultured organisms. This allows PICRUSt to "fill in the gaps" in well-characterised environments and produce accurate estimates of functional metagenomic content.

Algorithm 4.1 `USEARCH` algorithm (default implementation of closed-reference OTU picking in QIIME 1.X)

```
dereplicate query sequences (Q)
for all Q do
    Identify a small set of database sequences (D) that have many k-mers in
common with Q                                  ▷ Heuristic to improve search speed
    Count number of shared k-mers between Q and D (U)
    Order D by decreasing U
    for all D_i do
        Compute optimum global alignment A between D_i and Q
        if A > identity threshold then
            Accept D_i and terminate search
        else if A < identity threshold then
            Reject D_i
            if number of rejections >32 then
                Terminate search: no match found
            end if
        end if
    end for
end for
```

However, the PICRUSt approach should not be applied to poorly characterised environments. PICRUSt benefits from reference genomes that are phylogenetically similar to the input data.

4.2.4 Generating amplicon sequence variants with `dada2`

Raw 16S data typically consist of millions of short sequences (typically less than 400 nucleotides long). Conventionally the sequence reads are clustered according to fixed similarity thresholds; typically sequences that are more than 97% similar are binned into an OTU, which approximates a bacterial species (Caporaso et al., 2010; Schloss et al., 2009). A clustering strategy is required because during amplification and sequencing significant noise is introduced into the set of sequence reads (Callahan et al., 2017) (e.g. insertion, deletion, or substitution sequencing errors). A range of new methods (Eren et al., 2013; Eren et al., 2015; Callahan et al., 2016a) have been developed that are capable of removing this noise from the set of sequence reads. These methods are capable of resolving ASVs to a single-nucleotide resolution, which removes the need for arbitrary similarity thresholds. These high-resolution methods have better specificity and sensitivity compared with OTU clustering algorithms (Callahan et al., 2016a), and are better at identifying patterns

Algorithm 4.2 `UCLUST` algorithm (default implementation of *de novo* OTU picking in QIIME 1.*X*)

```
dereplicate query sequences Q
sort Q by length
initialise empty database of centroid sequences C
for all Q do
    for all C do
        if Q_i matches C_i then                          ▷ Using USEARCH
            Add Q_i to C_i
        else if no match then
            Q_i becomes new centroid of new cluster, add Q_i to C
        end if
    end for
end for
```

Algorithm 4.3 Open-reference OTU picking (algorithm recommended by QIIME 1.*X* developers)

```
dereplicate query sequences Q
for all Q do
    match Q to reference database                        ▷ Using USEARCH
    subsample Q (Q_subsample) that do not return matches (default: 0.001%)
    for all Q_subsample do
        match Q to centroids                             ▷ Using UCLUST
    end for
end for
```

of community similarity because ASVs are much less likely to be ecologically mixed units (Eren et al., 2013). It is important to note that although the term OTU can apply to ASVs — the definition of OTU is intentionally vague and simply means "the thing(s) being studied" — for example, one proposed term for ASVs is zero-radius operational taxonomic unit (zOTU) (Edgar, 2016). However, it is useful to consistently use different terminology to avoid confusion as the underlying paradigms are so different (clustering versus denoising). ASVs offer increased taxonomic resolution, are defined independently of any reference database, have consistent labels, and can be reused across studies (Callahan et al., 2017). In a conventional clustering approach, units are defined according to a reference database (reuse is possible but uncharacterised organisms will be omitted) or in a *de novo* fashion (*de novo* OTUs can include uncharacterised organisms but lack consistent labels and cannot be reused across studies). OTUs must be mapped to a taxonomy in order to provide consistent labels, while ASVs are independent of taxonomy and

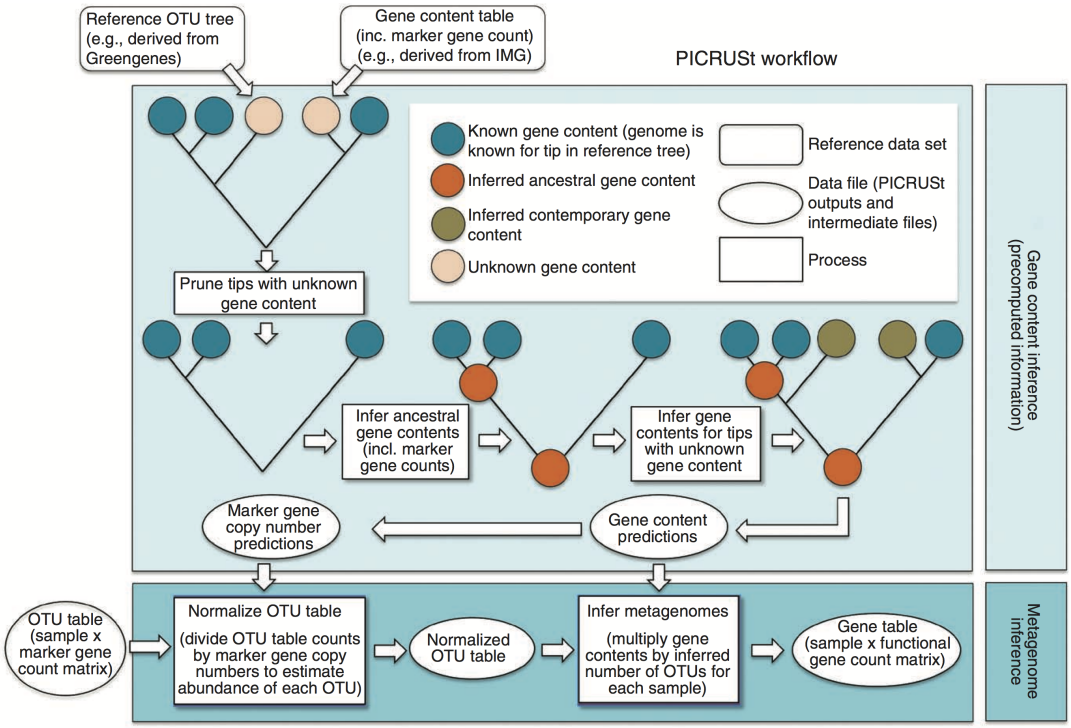


Figure 4.2: The PICRUSt workflow by Langille et al., 2013

represent true biological variation. In this work, the divisive amplicon denoising algorithm (DADA) was used to generate high-resolution microbiome census data.

DADA uses a statistical model to learn the types of amplicon errors present in a set of sequence reads. The exact sequence variants truly present in the samples are inferred from the model; these are called ASVs. Heuristic pairwise sequence alignments are performed for sequences that are closely related. The DADA error model measures the rate λ_{ji} at which sequence i is produced from sample sequence j as a function of sequence composition and quality. λ_{ji} is the product over the transition probabilities between the L aligned nucleotides (Callahan et al., 2016a):

$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \mapsto i(l), q_i(l)) \tag{4.1}$$

Generalised transition probabilities are included with the model, but the transition probabilities can also be learned from the data. An example transition probability is $p(\mathtt{G} \mapsto \mathtt{T}, 40)$, which describes the probability that a guanine nucleobase to thymine nucleobase substitution error has occurred in a base call with a

Algorithm 4.4 Divisive partitioning algorithm

```
dereplicate sequences                          ▷ store abundance and quality data
for all unique sequences do
    assign sequences to single partition
    set most abundant sequence as centre of partition
    calculate p_A
    while p_A,min < ω_A do                     ▷ user chooses ω_A
        create new partition
        set centre of partition to sequence p_A,min
        allow sequences most likely to have originated from partition to join
        calculate p_A
    end while      ▷ all sequences now copied from the centre of their partition
end for
```

quality score of 40. The probability depends on all three factors. Substitution errors are the most common kind of sequencing error on Illumina sequencing platforms (see chapter 3.3.2). λ_{ji} is estimated for all aligned sequences. Unaligned sequences are assigned a λ_{ji} of 0.

The abundance p-value (p_A) measures the likelihood that sequence i is too abundant to be explained by sequencing error alone, and is defined by (Callahan et al., 2016a):

$$p_A(j \mapsto i) = \frac{1}{1 - p_{\text{pois}(n_j\lambda_{ji},0)}} \sum_{a=a_i}^{\infty} p_{\text{pois}}(n_j\lambda_{ji}, a) \tag{4.2}$$

where, if sequencing errors are independent across reads, the abundance of sequences with sequence i that will be produced from the sample sequence j is Poisson distributed (p_{pois}) with expectation equal to error rate λ_{ji} multiplied by sample sequence j's expected reads. Let unique sequence i with abundance a_i be in partition j containing n_j reads. To generate exact sequence variants, sequences are processed with the divisive partitioning algorithm (see Algorithm 4.4). The centre of the generated partitions represent the denoised exact sequence variants.

4.3 Development of a hybrid model

4.3.1 IBD supervised classification

Non-invasive classification of IBD via the microbiome has been attempted many times across paediatric and adult cohorts (Papa et al., 2012; Tong et al., 2013; Gevers et al., 2014). Although IBD is a dynamic disease the majority of work done to date has focused on samples taken at a single time point with few exceptions

(Halfvarson et al., 2017). All of the models used are supervised classification algorithms and use stool samples as a proxy to map the intestinal microbiome (Gevers et al., 2014 also studied invasive biopsy samples). However, mapping the intestinal microbiome to predict IBD has to date been limited to analysing the relative abundances of bacterial taxonomic groups (i.e. *what* is present in the gut) as all of the models use a bacterial census generated with an OTU clustering paradigm as a starting point for analysis. Papa et al. and Tong et al. use the relative abundance directly as input to classification algorithms to predict disease status. Halfvarson et al. use summary statistics (distance to healthy samples generated from taxonomic data) that represent the microbiome to predict disease status. Gevers et al. use the relative abundance as input for regression, predicting the prognosis of IBD severity using the paediatric crohn's disease activity index (PCDAI).

Understanding what the gut microbiome is doing (e.g. gene functions) is difficult and cannot be measured directly via 16S rRNA marker gene surveys. Metagenomic shotgun sequencing is required to directly measure gene functions. In shotgun sequencing DNA molecules are broken into many pieces and all of the DNA molecule fragments are directly sequenced. Metagenomic shotgun sequencing of dozens or hundreds of samples is often cost prohibitive. A variety of algorithms have been developed that can infer gene functions from a bacterial census via a reference database, including PICRUSt (Langille et al., 2013). There has been debate as to whether taxonomic is appropriate for classification at all: it has been proposed that classifying samples into groups from gene functions could enable greater classification performance and be more biologically meaningful (Xu et al., 2014). Taxonomic profiles are extremely variable across samples, while functional profiles are more stable: decreased noise could improve analysis. However, decreased variation could make it harder to stratify samples. Initial benchmarks have found few differences between functional classification and taxonomic classification, except for a single classification task (the Costello body habitats dataset). However, none of the classification problems involved disease stratification.

Boruta — an all-relevant feature selection algorithm based on a Random Forest (Kursa, Rudnicki et al., 2010) — has been widely applied to taxonomic classification tasks to identify members of microbial communities that are associated with disease stratification. Standard feature selection algorithms are minimal optimal: they try to minimise the size of the feature subset while maximising classification accuracy. Boruta is better suited for biological data analysis as the algorithm will retain all features that carry information useful for prediction. Retaining all relevant features is an important first step to gaining a better understanding of the underlying biological phenomenon. Boruta can be used to measure the relevance of functional features for disease stratification.

Other machine learning algorithms that have been widely applied to the prediction of disease from different types of biological data (e.g. DNA microarray data) have been much less frequently applied to metagenomic classification problems, including SVMs and multilayer perceptrons (MLPs). SVM classifiers can process a large number of irrelevant features and high feature-to-sample ratios, and use regularisation techniques to avoid overfitting (Statnikov and Aliferis, 2007). MLPs and deep learning have rarely been applied to metagenomic classification, but show good initial results (Ditzler et al., 2015). Classifying IBD thoroughly requires a more complex strategy than predicting simply presence or absence. Aspects that can be considered include the presence of the disease, disease subtype, disease severity, and predicting response to treatment. The first three aspects have been considered previously in isolation, but not as a unified decision (data are not available for predicting prognosis). This approach generates a highly complex multiclass classification problem, including the following classes:

- Presence: Healthy, active IBD, or IBD in remission
- Subtype: Healthy (control), CD, or UC
- Severity: mild, moderate, or severe

Attempts were made to model this complex multiclass problem using a single model. However, these attempts were unsuccessful (see Figure 4.3). One possible reason for this is that different types of models perform well on different classification problems (there is no single optimum model according to the no free lunch theorem; Wolpert and Macready, 1997). A model can be designed to classify only a subset of possible class labels in a multiple classifier system if the outputs are combined to restore the whole label (Woźniak et al., 2014). Section 4.3.2 outlines how the standard single-classifier approach to IBD prediction can be replaced with a hybrid system. This approach will simultaneously: i) determine the relevance (via the Boruta algorithm) of microbial functions present in the intestinal microbiome for IBD prediction; ii) enable the concurrent prediction of IBD presence, subtype, and severity by decomposing IBD diagnosis into a series of more easily solved classification problems.

4.3.2 A metagenomic hybrid classifier

Multiple classifier systems (hybrid intelligent systems) have many advantages: combined classifiers can outperform the best individual classifiers, multiple classifiers are more likely to find an optimal model, and multiple classifiers can be efficiently implemented in a multi-threaded environment in a parallel manner (Woźniak et al., 2014). The topology of multiple classifier systems is usually parallel or

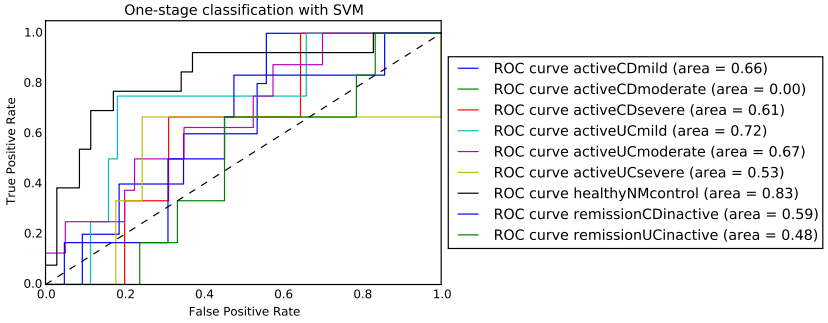


Figure 4.3: AUROC analysis shows that standard multiclass classification is unable to model thorough IBD prediction (0.5: prediction is equal to random chance).

conditional (serial). In parallel topology each classifier has identical inputs, and the final decision is made from the combined outputs of each classifier. In conditional topology, classifiers are used in a serial manner. Input is only passed to the classifier next in the sequence if some condition is met. The hybrid model implemented in this chapter used a serial approach. By returning a reduced set of classes at each stage of the serial classifier a complex problem can be iteratively decomposed into a series of simpler problems that are easier to classify.

The topology of the serial multiple classifier system (the hybrid model approach; see Figure 4.4) was designed so that a complex problem (thorough IBD diagnosis) could be reduced to a set of simpler, but clinically important, problems. The accurate identification of IBD presence (i.e. IBD or control) is important to guide treatment options. Some IBD treatments are contraindicated for subjects with conditions that can be misdiagnosed as IBD (i.e. prescribing immunosuppressant drugs for amoebic dysentery). The subtype of IBD and current IBD activity is important to guide the treatment course (e.g. severe Crohn's disease may require surgical intervention).

In 16S rRNA data a sample is defined as all DNA sequences identified by the 16S marker gene survey per faecal sample. The DNA sequences identify hundreds of different bacterial groups per sample. The DNA sequences were mapped to vector representations in order to input them into supervised learning classification algorithms. An OTU approach was used to generate these vector representations from DNA sequences (see Figure 4.5). Bacterial taxonomic groups were identified and clustered from the similarity of the DNA sequences present per sample (97% similarity was used to approximate a bacterial species). The `QIIME` software package (Caporaso et al., 2010) was used to generate OTU tables (a community data matrix), with the open-reference subsampled OTU picking algorithm. The OTU table recorded how many times an identified OTU occurred for each sample.

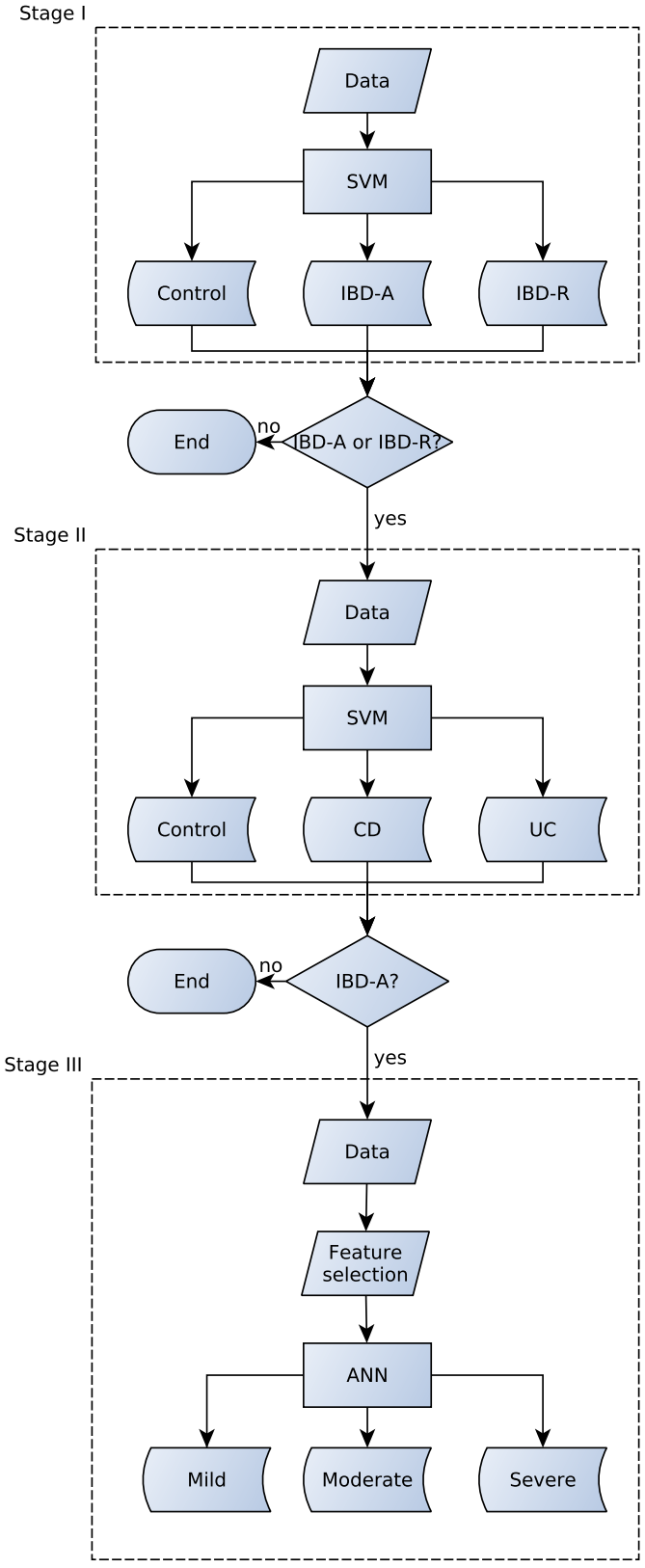


Figure 4.4: Conditional multiple classifier system topology. IBD-A: IBD in its active state, IBD-R: IBD in remission. First and second stage use a support vector machine (SVM) for classification, third stage uses an Artificial Neural Network (ANN).

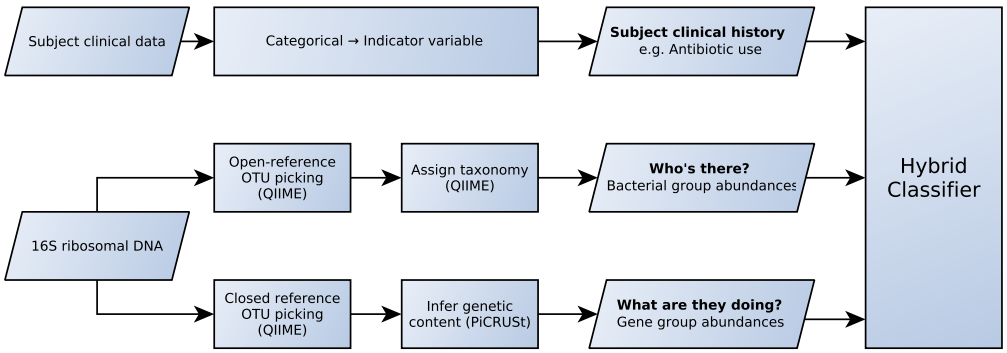


Figure 4.5: Feature engineering pipeline.

OTU abundances were scaled to be in the range $[0, 1]$.

The PICRUSt algorithm (Langille et al., 2013) was used to infer functional content from the marker gene survey. PICRUSt infers genetic content from bacterial phylogenies via comparison to a database of reference genomes. The presence and abundance of gene functions present per sample were binned into categories and abundances were scaled to be in the range $[0, 1]$. Subject clinical history was converted from categorical variables to indicator variables (e.g. has the patient been prescribed immunosuppressant drugs). Immunosuppressant drugs and antibiotics (Zaura et al., 2015) have been shown to cause large long term changes to microbiomes across the human body and it is thus essential to record their application. Other clinical data includes subject ethnicity and family history of IBD.

Feature selection was applied where appropriate because high dimensional learning with traditional artificial neural networks is difficult (Verleysen et al., 2003). 3-fold cross-validated SVM Recursive Feature Elimination (RFE) (Guyon et al., 2002) was used to automatically identify the optimal number of features. SVMs show good performance on high dimensional classification problems (Statnikov and Aliferis, 2007). SVM-RFE repeatedly eliminates the features least important (measured by SVM feature weights) to classification performance until the optimum is reached.

4.3.3 Evaluating the hybrid model

Determining the relevance of the different feature types was performed with the Boruta algorithm. The predicted gene functions generated with PICRUSt were the most relevant type of feature across all three stages of the hybrid classifier (see Table 4.3). To date no other metagenomic classifiers of host health status have used predicted metagenomes as a feature. However, classification of microbial communities with predicted metagenomes has occurred with good results (Xu et al.,

Table 4.3: Distribution of relevant features per stage. Feature relevance calculated with the Boruta algorithm. See Figure 4.5 for a description of taxonomic, functional, and clinical features.

Stage	Taxonomic features	Functional features	Clinical features
I: IBD presence	27%	64%	9%
II: IBD subtype	34%	53%	13%
III: IBD activity	20%	80%	0%

2014). The aim of Boruta is to understand the mechanisms of action that created the dataset.

As each stage of the hybrid classifier attempts to model different problems, it is useful to analyse the feature ranks of each stage individually to gain an insight into different phenomena. Relevant features associated with IBD presence, subtype, and severity may be significantly different. Therefore analysing the relevant features identified by Boruta could generate new insights into the aetiology and pathophysiology of IBD. Carotenoid biosynthesis was a relevant feature in the first and second stage. Carotenoids are a group of organic pigments synthesised by plants and bacteria, and are the pigments that produce attractive colours in plants. They are sourced mainly from fruit and vegetables and are antioxidants. The pathogenesis of IBD is thought to involve oxidative stress. In IBD patients antioxidants that circulate in blood plasma - including carotenoids - are present at significantly lower concentrations than controls (D'Odorico et al., 2001). This pattern is also found in this analysis but in this work the carotenoid biosynthesis is only measured from bacteria (plants do not have 16S rRNA). The intestinal microbiome synthesises a variety of important vitamins that are required by host metabolism such as vitamin B12. However, limited work has been done in assessing the role of the microbiota in carotenoid biosynthesis (e.g. vitamin A). Carotenoid synthesis by commensal bacteria could contribute to overall host health in previously undiscovered ways, and imbalances in the intestinal microbiome could reduce the amount of carotenoid biosynthesis occurring. Genes associated with bacterial infections were found to be relevant features in the first stage. There is evidence that conserved genes associated with *Vibrio cholerae* can be acquired by *Campylobacter concisus*, leading to the pathogenesis of IBD (Zhang et al., 2014). *Vibrio cholerae* can increase the permeability of the intestine, triggering the onset and relapse of IBD. However, it is important to note that lateral gene transfer cannot be modelled by approaches that infer functional content from taxonomic data such as PICRUSt.

The second aim of the hybrid classifier was to deliver good predictive performance. It is useful to monitor the performance of each individual stage in order to

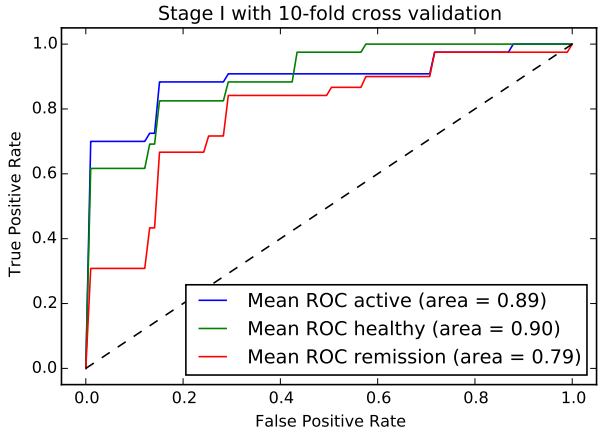
Table 4.4: Cross validated classification performance of the hybrid model.

Stage	Average precision score	Support (classes balanced)
I: IBD presence	0.71	111
II: IBD subtype	0.65	111
III: IBD activity	0.61	45

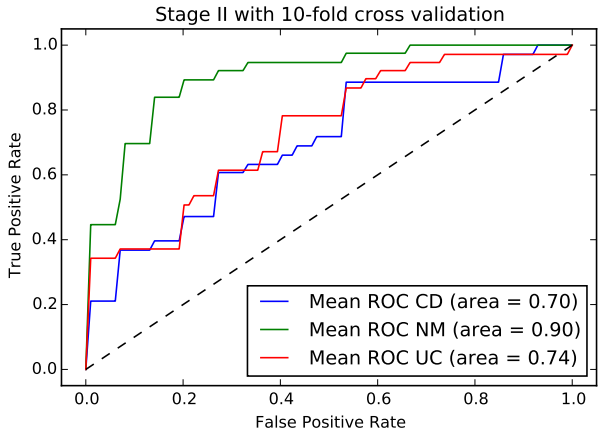
make fair comparisons to current work in the literature and to gain a global view of classification quality. The predictive performance of a classifier as measured by a Receiver Operating Characteristic (ROC) analysis is often measured via the area under the curve (AUC). The AUROC is often a better indicator of classifier performance than the misclassification rate or a loss matrix (Downey Jr et al., 1999). The first stage of the hybrid classifier showed good classification performance for the IBD remission class (see Figure 4.6). The IBD active and control classes showed excellent performance. The second stage of the hybrid classifier had excellent performance for the failsafe control class and good performance for the CD and UC classes. The third stage of the hybrid classifier had good performance for all classes. The average precision score of the third stage of the hybrid classifier shows the worst performance across all stages. This could be contributed to the lack of training data as IBD severity information was not recorded for all subjects (see Table 4.4) compared with the other two training stages. Additionally, classification of a subjective criteria (i.e. a class arising from a rating rather than a biological test) is a difficult problem. The hybrid classifier shows superior performance as measured by the AUROC to the standalone Random Forest classifier reported in Papa et al., 2012 (0.83 AUROC for IBD prediction versus up to 0.90 in Stage I of Figure 4.6).

Despite roughly tripling the amount of features when compared with the original analysis by Papa et al. of a bacterial census (643 features were used including a bacterial census, predicted gene abundances, and clinical features in the final model) the SVMs used in the first two stages performed well. SVMs are insensitive to high feature-to-sample ratios. Random Forests and SVMs performed poorly on the third stage of the hybrid classifier. Both could consistently identify mild and severe classes but were unable to classify moderate classes. A MLP showed good performance for all classes despite the nonlinearity of the data.

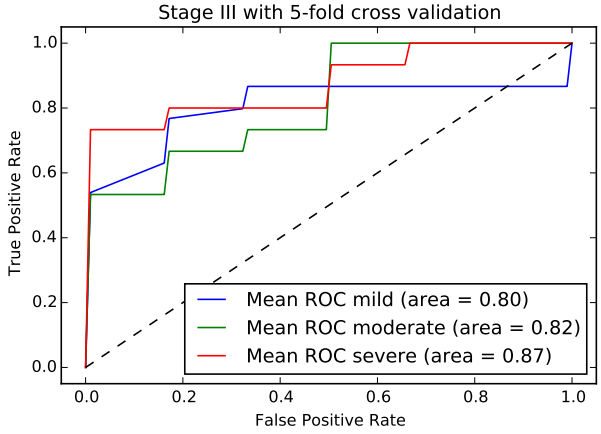
An analysis of the relevant features across all stages of the hybrid classifier showed that predicted genetic content was a valuable feature type, forming the majority of relevant features (see Table 4.3). The sensitivity and specificity reported by Papa et al. of the Random Forest classifier matched or surpassed alternative clinical methods (i.e. non-colonoscopy tests) for detecting IBD. The hybrid classifier



(a)



(b)



(c)

Figure 4.6: a) Stage I: Classification of IBD presence with a SVM. (b) Classification of IBD subtype with a SVM. (c) Classification of IBD severity with a MLP. A predictive model with an AUROC of 0.5 is no better than random chance (shown as the dotted diagonal line).

presented in this chapter shows superior performance as measured by the AUROC to the standalone Random Forest classifier reported in Papa et al., 2012.

An advantage of using a conditional multiple classifier system is its ability to maintain good performance for three different classification problems across nine classes. A MLP was the only algorithm capable of reliably classifying all classes for stage III, while SVMs showed superior performance for stages I and II. The three stages were designed to provide relevant information to a clinician which could guide treatment minimising the need for invasive colonoscopy. SVMs have rarely been used for metagenomic classification problems but show good performance in this case. Very little work has been done on applying MLPs to metagenomic classification problems. MLPs have shown value here for the classification of disease properties determined by subjective criteria. This is commonly done in many diseases with uncertain aetiologies, including depression.

The Boruta algorithm validated the use of predicted metagenomes as a novel feature set for the classification of IBD from marker gene surveys. Boruta revealed relevant features involved in biological mechanisms behind the pathogenesis of IBD. Significantly reduced abundance of antioxidant carotenoids in IBD subjects has been previously measured from blood plasma but not from the intestinal microbiome (D'Odorico et al., 2001). Carotenoids are typically sourced from fresh fruit and vegetables: they provide plants with bright pigments. The role of the intestinal microbiome providing the human host with nutrients and vitamins has been well documented. No work to date has described the role of the intestinal microbiome in providing carotenoids to the host.

Genes associated with the lifecycle of pathogenic bacteria were also detected as relevant by the Boruta algorithm. Of note is the *Vibrio cholerae* lifecycle which is relevant for the first stage of the hybrid classifier that identifies IBD in remission, active IBD, and control classes. Evidence has been found that a combination of *Vibrio cholerae* and *Campylobacter concisus* is implicated in altering the permeability of the intestine, leading to IBD relapse into an active state (Zhang et al., 2014). This was previously missed by the bacterial census. Identifying bacterial species with a 16S marker gene survey is difficult due to technical limitations of the protocol. Species that are identified are typically present in very low abundances. This creates highly sparse feature vectors. Sparse data are challenging to learn from because it can increase the hypothesis space through which the learning algorithm must search.

4.4 Generation of robust microbial markers

The approaches to feature selection to microbiome census data described in section 4.3 have not considered the robustness of feature selector output. Domain

experts are often interested in experimentally validating feature subsets, which is an expensive proposition for biological data. Feature selection algorithms can return different feature subsets from the same input data; different feature subsets can be equally optimal, particularly if a high degree of redundancy is present in the dataset (Kalousis et al., 2007). Feature selection algorithms can also return significantly different feature subsets from input data that has been changed slightly (e.g. by removing a sample or after adding noise to a feature). Domain experts will have more confidence in feature selection algorithms that generate consistent (robust) feature subsets.

Current studies that aim to identify associations between IBD and the microbiome, including the work reported in the previous section, have used fuzzy OTU approaches to generate a bacterial census. From this body of work a wide array of bacterial genera have been implicated in the pathogenesis of IBD (fuzzy OTU algorithms are typically limited to identifying bacteria at the genus level). The rationale for applying aggregating EFS to high-resolution microbiome census data was two-fold: i) to enable knowledge discovery from the increased resolution of the input data ii) to improve the clinical utility of any identified feature subsets by increasing confidence in feature selector output. Firstly, it is possible to match exact DNA sequences up to a sub-species level. Secondly, exact DNA sequences can be measured *in vivo* via a variety of methods (e.g. quantitative polymerase chain reaction (qPCR)) whereas fuzzy clusters of DNA sequences are much more difficult to measure.

The data analysed in this section originate from a publicly available dataset (Gevers et al., 2014) which consists of 1643 samples collected from treatment-naïve children and adults diagnosed with IBD and controls (see Section 4.2.1). This chapter focused on stool samples in order to develop a set of robust markers that can be used to non-invasively predict IBD, so all biopsy samples were discarded, leaving 311 stool samples. Classes were defined according to an IBD subtype: control versus UC or control versus CD. Although they fall under the umbrella term IBD the subtypes have significant biological differences (Ananthakrishnan, 2015; Sartor, 2006), which is the rationale for choosing an IBD subtype to define classes.

A reproducible computational workflow was implemented with `Docker` and `nextflow` (Di Tommaso et al., 2017). `Docker` is an open source container platform. A container bundles together all of the data, software, and library dependencies necessary to run a piece of software into an image, similar to a very efficient virtual machine. `Docker` helps to improve reproducible research by solving "dependency hell", poor documentation (`docker` images are self-documenting), and code rot (Boettiger, 2015). The dataset was downloaded using `esearch` (Kans, 2013), `sra-tools` (Leinonen et al., 2010), and `GNU Parallel` (Tange et al., 2011).

Microbiome count data were generated with `dada2` (Callahan et al., 2016a) and processed with `phyloseq` (McMurdie and Holmes, 2013) according to a standard operating protocol (Callahan et al., 2016b). A variance stabilising transformation (Love et al., 2014) was applied to the microbiome count data to normalise the uneven library sizes and heteroscedasticity in the data, which has been recommended for machine learning applications (McMurdie and Holmes, 2014). Aggregating EFS was implemented using the `OmicsMarkeR` package (Determan Jr, 2015). The Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) was used to mitigate the class imbalance present in the dataset. SMOTE is a powerful synthetic sampling technque that has been successfully applied for a variety of applications (including biomedical data) (He and Garcia, 2009). Imbalanced data can be significantly more difficult to learn, decreasing model performance (Japkowicz and Stephen, 2002; He and Garcia, 2009). The distribution of microbial markers was visualised with `Venny` (Oliveros, 2015).

4.4.1 Aggregating Ensemble Feature Selection

The robustness of a feature selector can be defined by the variation of feature subset output caused by small changes to the input (Saeys et al., 2008). EFS can generate robust feature subsets (Abeel et al., 2010). EFS is inspired by ensemble learning, where the output of multiple weaker classifiers can be combined to outperform a single strong model. It has been shown that combining the output of multiple unstable feature selectors can create a robust consensus feature ranking (Abeel et al., 2010). Typically filter, wrapper, and embedded feature selection methods that do not consider the robustness of output have been previously applied to microbiome data (Statnikov et al., 2013). Random Forests have been widely applied for supervised classification of IBD from microbiome data and the feature rankings have been reported for knowledge discovery purposes (Tong et al., 2013; Papa et al., 2012; Gevers et al., 2014); rankings are often combined with a RFE procedure to generate a feature subset. Recently an EFS approach was used to generate a feature subset for the non-invasive prediction of advanced fibrosis in non-alcoholic fatty liver disease (Loomba et al., 2017). However, this approach does not employ an aggregation paradigm to measure the robustness of the derived features.

The microbiome census data (input data) were modified by instance perturbation (removing or adding features) via resampling with replacement (bootstrapping). Modification can also be done at the feature level (e.g. by adding random noise to a feature or group of features) or by a combination of instance and feature level perturbation. To measure the overall effect of bootstrapping on feature stability, Saeys et al. proposed a similarity measure based approach. In this approach the stability was measured by averaging the pairwise similarity comparison of feature

subset output for k bootstraps, which was defined as (Saeys et al., 2008):

$$S_{\text{global}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S\left(f_i, f_j\right)}{k\left(k-1\right)} \tag{4.3}$$

where f_i is the feature selector output applied to bootstrap i, and $S\left(f_i, f_j\right)$ is a similarity measure between f_i and f_j. In this work the Jaccard Index was used as similarity measure $S(f_i, f_j)$ (Saeys et al., 2008):

$$S\left(f_i, f_j\right) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} = \frac{\sum_l I(F_i^l = f_j^l = 1)}{\sum_l I(F_i^l + f_j^l > 0)} \tag{4.4}$$

where the function I returns 1 if its argument is true and 0 if its argument is false.

It has been shown that an aggregating EFS approach can improve the robustness of feature selectors (Saeys et al., 2008) for the prediction of cancer from gene expression data. Ensemble models are capable of outperforming single models because if a group of different but equally good hypotheses exist it is less likely that an ensemble will pick the wrong hypothesis. Furthermore, algorithms can end up in different local optima enabling an ensemble to better approximate a true function. Finally it is known that EFS can achieve greater robustness because it expands the hypotheses space (Dietterich et al., 2000).

EFS has two stages: choosing a set of feature selection algorithms, and combining the feature subsets into a final consensus ranked list. Let ensemble E contain s feature selectors $F_1, \ldots, F_s$. Each feature selector outputs a feature ranking $f_i = f_i^1, \ldots, f_i^N$. In this work a consensus ranking f was formed by combining feature subsets with complete linear aggregation (Saeys et al., 2008):

$$f = \sum_{i=1}^{s} w(f_i^l) \tag{4.5}$$

where weighting function w is set to $w(f_i^l) = f_i^l$. Feature selection must always be combined with an evaluation of classification performance: domain experts will not be interested in a stable feature subset that has poor predictive performance. In this work embedded feature selection algorithms were applied, such as Random Forests (Breiman, 2001) and linear SVM. Embedded feature selection algorithms provide feature ranking during training which decreases the computational complexity of the EFS process. Random Forests are an ensemble of decorrelated decision trees (Qi, 2012); feature rankings are calculated by randomly permuting a feature in the out-of-bag samples and calculating the mean change in impurity or accuracy compared with the out-of-bag rate with unpermuted features. Linear SVMs can rank features from the absolute value of the weight vector of the hyperplane (Guyon et al., 2002); RFE is used to reduce the size of the feature subsets by iteratively removing the poorest 10% of features until the subset is empty. In

order to effectively evaluate which feature selector should be chosen for a particular classification problem it is necessary to use a metric that balances the classification performance of a feature aggregation and the stability of the aggregated features. The Robustness-Performance Trade-off (RPT) (Abeel et al., 2010) is a metric that does this. The RPT is a variant of the widely used F1-score which is the harmonic mean of the precision and recall (Van Rijsbergen, 1979), RPT is defined as (Saeys et al., 2008):

$$\mathrm{RPT}_{\beta} = \frac{(\beta^2 + 1) \cdot S_{\text{global}} \cdot P}{\beta^2 \cdot S_{\text{global}} + P} \tag{4.6}$$

where P is the prediction accuracy of the classification model trained on the robust feature subset. β is a parameter used to weight the relative importance between robustness and classification performance. In this work $\beta = 1$ to give equal importance to classification performance and robustness.

Prior to applying EFS a simple filter was applied to remove extremely rare ASVs. ASVs present in less than 5% of samples were removed, as this study aims to find microbial markers that are present across a broad population. Prior to EFS 20% of data were retained from the dataset for independent validation of the final model. In the first stage of EFS, a portion of the data (20%) is retained in order to test the performance of the model trained on the remainder of the data (see Figure 4.7). The training data were repeatedly sampled with replacement (bootstrapped). For each bootstrap bag a SVM and Random Forest were fit, and recursive feature elimination was applied to each bag. Feature ranks were extracted across all of the bags, and merged via complete linear aggregation (Abeel et al., 2010) to form a single feature ranking list. Each ranked list was combined across all of the bootstraps to form a final feature subset, along with frequency and consistency measurements. The RPT was calculated for both models from the classification performance of the model on the test data and the global similarity measure across all feature lists. Random Forests were used to validate the generalisation ability of the microbial markers as they had the highest RPT for both CD and UC. All classification results reported are from the Random Forest model.

4.4.2 Robust microbial markers of IBD

Approximately 0.5% of ASVs were retained after a two-stage filter and aggregating EFS feature selection strategy detailed in section 4.4.1 (see Table 4.7). Nearly 4500 ASVs were identified from the stool samples: a simple filter was applied to remove any ASVs that were not present in at least 5% of samples. After this process, aggregating EFS was successfully applied to the remaining features (around 250 prevalent ASVs). The overlap of ASVs across IBD subtypes is low - 12.1% of ASVs

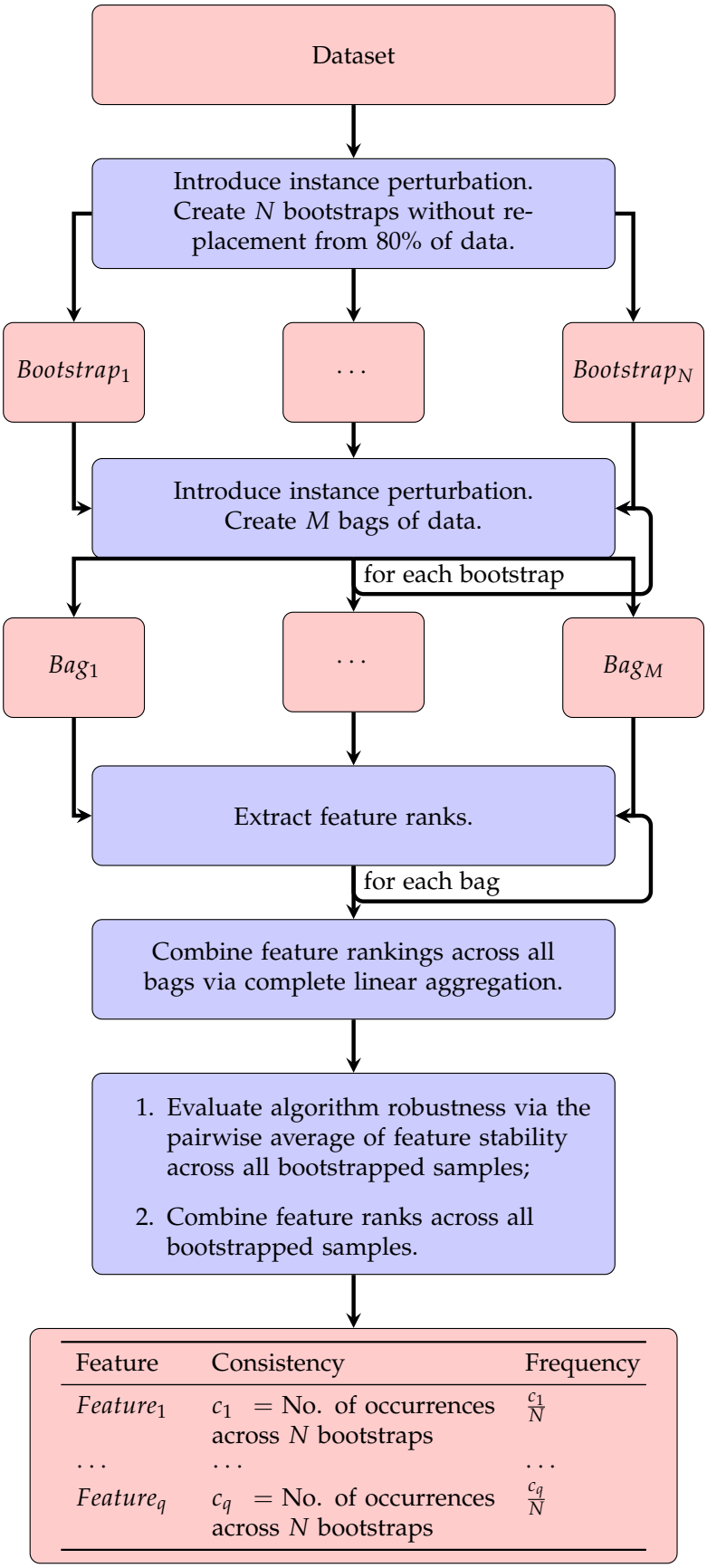


Figure 4.7: Ensemble feature selection workflow, $N = 15, M = 40$.

were shared across the CD and UC subsets (see Figure 4.8) - which reflects the distinct biological differences between the two subtypes.

The stability of a selected feature can be measured by its frequency, which is the number of times a feature appears in each bootstrap divided by the total number of bootstraps (see Figure 4.7). Perfectly robust features have a frequency of 1 while the least robust features will only be present in a single bootstrap; in this work 5 bootstraps were used so features with a frequency of 0.2 are the least stable. In the CD cohort 3 ASVs had a perfect frequency (they were present in every bootstrap), and in the UC cohort 4 features had a perfect frequency (see Tables 4.5–4.6). It is important to note that the ASV paradigm reveals greater differences than would otherwise be reported by a clustering approach (Callahan et al., 2017). OTUs are generally capable of being matched to taxonomic databases at the level of family or genus (Gevers et al., 2014); all other IBD classification work agglomerated OTUs into genus-level relative abundance data to represent the microbiome. ASVs are capable of resolving separate bacterial strains (i.e. at a level higher than species). However, because ASVs are relatively short fragments of the full 16S rRNA gene, taxonomic assignment is sometimes limited to higher ranks. The agglomeration process will discard bacteria that do not meet a defined phylogenetic or taxonomic threshold. For example, if a genus-level agglomeration is chosen then OTUs or ASVs that only match to the family level or higher will be discarded. In this work ASVs are not agglomerated into specific taxonomic ranks as biological phenomenon (e.g. IBD subtype) may not be accurately modelled according to human-defined taxonomic hierarchies. ASVs have been shown to accurately represent true biological variation independently of any taxonomic reference database (Callahan et al., 2017). Of the most robust features for CD prediction two could be mapped to genus (*Bacteroides* and *Haemophilus*) and one to family (*Lachnospiraceae*). One of the most robust features for UC prediction could be mapped to species (*Bacteroides vulgatus*), two to genus (*Pediococcus* and *Ersyipelotrichaceae*), and one to family (*Ruminococcaceae*).

4.4.3 Evaluating the microbial markers

Robust microbial markers should have strong predictive power to be valuable for knowledge discovery and further investigation by domain experts. The classification and feature selection ability of Random Forests and SVMs were tested. Random Forests were chosen as the final model for both IBD subtypes as they had the highest RPT (a balanced metric of classification performance and aggregated feature robustness, see Table 4.8). The final models were used to validate the feature subsets against independent validation data. The dataset was split into two cohorts according to IBD subtype; the classification task was to distinguish between control and disease subjects (two class classification). The rationale for this approach

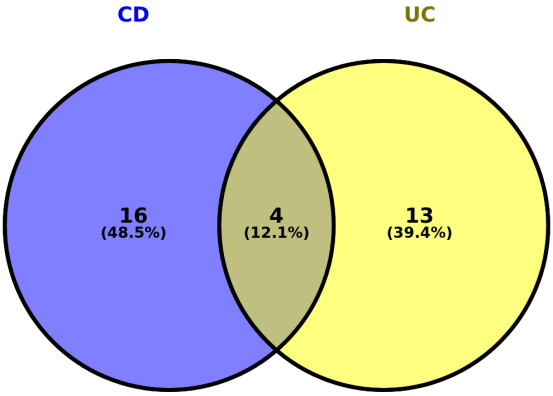


Figure 4.8: Venn diagram of ASV microbial marker distribution by cohort (CD: Crohn's disease, UC: ulcerative colitis).

Table 4.5: Taxonomy of Robust Microbial Markers of Crohn's disease.

Frequency	Family	Genus	Species	Previously reported?
1	Bacteroidaceae	Bacteroides		Gevers et al., 2014
1	Pasteurellaceae	Haemophilus		Gevers et al., 2014
1	Lachnospiraceae			Gevers et al., 2014
0.8	Actinomycetaceae	Actinomyces	graevenitzii	✗
0.8	Lachnospiraceae	Roseburia		Gevers et al., 2014
0.8	Peptostreptococcaceae	Intestinibacter	bartlettii	✗
0.8	Ruminococcaceae	Ruminococcaceae UCG-002		Gevers et al., 2014
0.6	Erysipelotrichaceae	Erysipelatoclostridium		Gevers et al., 2014
0.4	Lachnospiraceae	Roseburia	inulinivorans	✗
0.4	Bacteroidaceae	Bacteroides	vulgatus	Gevers et al., 2014
0.4	Alcaligenaceae	Parasutterella	excrementihominis	Ricanek et al., 2012
0.4	Pasteurellaceae	Actinobacillus		Gevers et al., 2014
0.2	Veillonellaceae	Megamonas	funiformis	✗
0.2	Fusobacteriaceae	Fusobacterium		Gevers et al., 2014
0.2	Bacteroidaceae	Bacteroides		Chen et al., 2014
0.2	Pasteurellaceae	Haemophilus	influenzae or parainfluenzae	Gevers et al., 2014
0.2	Ruminococcaceae	Ruminiclostridium 5		Gevers et al., 2014
0.2	Enterobacteriaceae	Escherichia /Shigella		Gevers et al., 2014
0.2	Ruminococcaceae	Ruminococcus 2	bromii	Swidsinski et al., 2005
0.2	Lachnospiraceae	Blautia		Gevers et al., 2014

Table 4.6: Taxonomy of Robust Microbial Markers of ulcerative colitis.

Frequency	Order	Family	Genus	Species	Previously reported?
1	Clostridiales	Ruminococcaceae			Gevers et al., 2014
1	Bacteroidales	Bacteroidaceae	Bacteroides	vulgatus	Gevers et al., 2014
1	Lactobacillales	Lactobacillaceae	Pediococcus		Wang et al., 2014
1	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae UCG-003		Gevers et al., 2014
0.8	Clostridiales	Lachnospiraceae	Anaerostipes	hadrus	✗
0.8	Clostridiales	Peptostreptococcaceae	Intestinibacter	bartlettii	✗
0.8	Lactobacillales	Streptococcaceae	Streptococcus		Gevers et al., 2014
0.6	Enterobacteriales	Enterobacteriaceae			Gevers et al., 2014
0.6	Clostridiales	Lachnospiraceae			Gevers et al., 2014
0.6	Lactobacillales	Streptococcaceae	Lactococcus		Gevers et al., 2014
0.6	Lactobacillales	Lactobacillaceae	Lactobacillus		Gevers et al., 2014
0.2	Bacillales	Family_XI	Gemella		Gevers et al., 2014
0.2	Clostridiales	Lachnospiraceae			Gevers et al., 2014
0.2	Bacteroidales	Bacteroidaceae	Bacteroides		Gevers et al., 2014
0.2	Clostridiales	Ruminococcaceae	Faecalibacterium	cf./prausnitzii	Sokol et al., 2008
0.2	Bacteroidales	Bacteroidaceae	Bacteroides		Gevers et al., 2014
0.2	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium		Gevers et al., 2014

stems from the important biological differences present in the pathophysiology of UC and CD, which represents an interesting area for knowledge discovery to be derived from consensus feature subsets.

Non-invasive prediction of both IBD diagnosis and IBD subtype from stool samples has been previously attempted, including from this dataset (Gevers et al., 2014). In (Gevers et al., 2014) IBD was predicted from biopsies of the terminal ileum (mean AUC: 0.85) and rectum (mean AUC: 0.78) with good performance. Prediction from stool samples was less successful (mean AUC: 0.66 with much lower consistency). The models used relative microbial abundance data agglomerated to a genus level. In Tong et al. IBD was predicted with an accuracy of up to 70% (Tong et al., 2013) using nearest shrunken centroid classification from biopsy samples. In (Papa et al., 2012) classification performance was reported at two different thresholds: in the first, a sensitivity of 80.3% and a specificity of 69.7% was reported. The second reported a sensitivity of 45.8% and a specificity of 92.4%. It is important to note that the patient cohort used in Papa et al. had a mean disease duration of 34.8 months, while the publicly available dataset used in this work consists of samples collected at time of diagnosis. Due to this lengthy disease duration many of the patients in the cohort had been treated with anti-inflammatory drugs or other pharmacological interventions which may have impacted the composition of the microbiome — the data used in this work do not suffer from this limitation.

The non-invasive IBD classification described in this section is the highest performance described in the literature to date. The classification performance of both feature subsets was excellent. CD was classified with a Positive Predictive Value (PPV) of 87.6% in the testing set and 96.4% in the validation set, and a Negative Predictive Value (NPV) of 97.1% in the testing set and 100% in the validation set. UC was predicted with a PPV of 94.5% in the testing set and 100% in the validation set, and a NPV of 100% in the testing set and 92.6% in the validation set (see Table 4.7). This is significantly better than performance metrics reported in Gevers et al., 2014; Papa et al., 2012; Tong et al., 2013.

4.4.4 Knowledge discovery from high resolution microbiome census data

Every described denoised microbial sequence marker that has been implicated in the pathogenesis of IBD by the aggregating EFS procedure is novel, as previous work has relied on analysis of fuzzy clusters (see Tables 4.5–4.6). The reported set of 16S ASVs can non-invasively predict IBD with the highest reported accuracy to date, have innate biological meaning and do not rely on reference databases or taxonomic assignments. The behaviour of ASVs that match the same species can be markedly

Table 4.7: Classification performance of feature subset

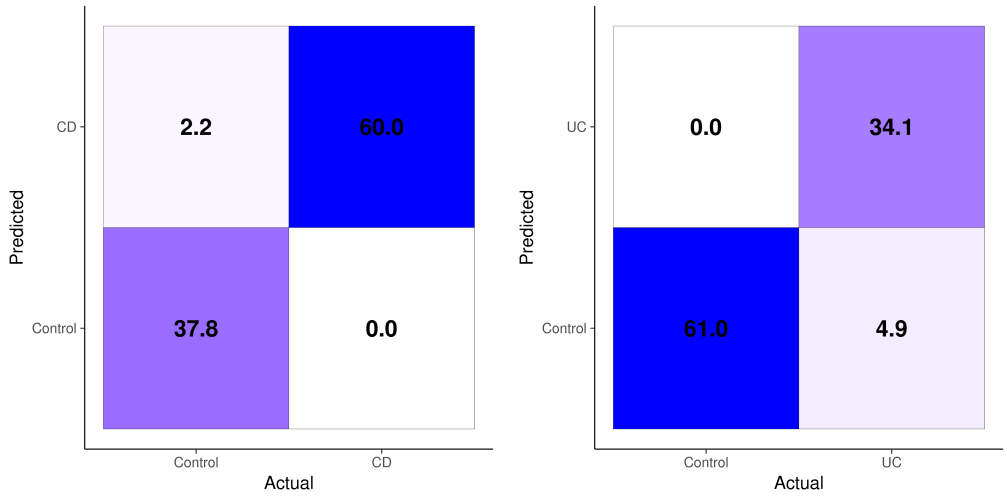
Classification problem	Data split	Sensitivity	Specificity	PPV	NPV	Other
Crohn's disease	Testing	94.5%	90.9%	87.6%	96.1%	
Ulcerative colitis		100%	94.5%	94.5%	100%	
Crohn's disease	Validation	100%	94.4%	96.4%	100%	
Ulcerative colitis		87.5%	100%	100%	92.6%	
Papa et al. stool		80.3%	69.7%			
Papa et al. stool		45.8%	92.4%			
Gevers et al. stool						AUROC 0.66
Tong et al. biopsy						70% accuracy

Table 4.8: An ensemble of Random Forests were chosen for both classification problems as they had the highest Robustness-Performance Tradeoff (RPT) measure.

Classification task	Model	RPT	No. features retained
Crohn's disease	**Random Forest**	**0.60**	**20**
	SVM	0.58	20
Ulcerative colitis	**Random Forest**	**0.70**	**17**
	SVM	0.48	17

different (Eren et al., 2013), which demonstrates the limitations of human-defined taxonomic systems. In order to compare the ASVs to previous work the ASVs were mapped to the SILVA taxonomic database (Quast et al., 2012). Elements of the robust microbial marker set that have been found previously in the literature are described below. In addition, several novel bacterial species that have not been previously implicated in IBD pathogenesis are also described below. It is important to note fuzzy clusters, under normal circumstances, are limited to resolving bacteria at high taxonomic ranks such as Order, Family, or Genus. All of the identified ASVs have been previously reported in the literature as biomarkers for IBD at high taxonomic ranks which confirms that the aggregating EFS process has selected biologically plausible markers. One of the many advantages of the denoising ASV paradigm is increased taxonomic resolution; as the resolution increases, previously undescribed microbial markers emerge. The previously described markers below are gathered from differential abundance statistical tests and machine learning algorithms (e.g. Random Forest ranks). The reported biomarkers are from samples gathered from the entire gastrointestinal tract, including stool, rectal or ileal biopsies.

Blautia, *Ruminoccous*, Pasteurellaceae, Erysipelotrichales, and Veillonellaceae



(a) (b)

Figure 4.9: Confusion matrices of models fitted to feature subsets on paediatric validation data; Crohn's disease (left) and Ulcerative colitis (right). Each cell contains a percentage of samples assigned to it: light colours represent a small percentage, and darker colours represent a large percentage.

are repeatedly observed in the set of robust markers in line with current work (Gevers et al., 2014). Enterobacteriaceae, Bacteroidales, and Clostridiales have been repeatedly identified across the literature as IBD biomarkers (Morgan et al., 2012; Papa et al., 2012; Gevers et al., 2014), and all are strongly represented in the set of microbial markers. Fusobacterium has been previously reported as a biomarker for a number of conditions including IBD (Strauss et al., 2011) and colorectal cancer (Kostic et al., 2012); the risk of developing colorectal cancer in IBD patients is significantly increased (Triantafillidis et al., 2009). Lachnospiraceae, including the Roseburia genus specifically, is differentially abundant in IBD subjects (Morgan et al., 2012). *Faecalibacterium prausnitzii* is an anti-inflammatory organism and is associated with health (Sokol et al., 2008). *Parasutterella excrementihominis* has been observed to be unique to a cohort of treatment-naïve children (Ricanek et al., 2012). Bacillales (Hourigan et al., 2015) and Bifidobacterium (Wang et al., 2014) have also been found to be IBD biomarkers.

When the taxonomic resolution is increased, bacterial species previously unassociated with IBD begin to emerge. *Actinomyces graevenitzii* is capable of infecting humans in combination with other bacterial species. Copathogens such as *A. graevenitzii* rely on other bacterial species to inhibit the host immune system or to reduce the amount of oxygen in the local environment before infection can occur

(Tietz et al., 2005); *A. graevenitzii* has been implicated in coinfection with tuberculosis (Tietz et al., 2005). In active IBD localised areas of the gut are hypoxic due to metabolic demand outpacing supply (Colgan et al., 2013): the IBD gut appears to provide ideal conditions for *A. graevenitzii* to grow. *A. graevenitzii* is a strong biomarker for CD, with a frequency of 0.8 (see Tables 4.5–4.6). *Intestinibacter bartlettii* has only been very recently defined, and its role in the human gut and human health is uncertain; recent work shows that *I. bartlettii* is thought to be resistant to oxidative stress and is involved with mucus degradation (Forslund et al., 2015). Oxidative stress is significantly increased in areas of mucosal inflammation in IBD (Colgan et al., 2013). *I. bartlettii* is a robust biomarker for both of the CD and UC cohorts, with a frequency of 0.8. Both *Anaerostipes hadrus* and *Roseburia inulinivorans* are lactate utilising butyrate-producing bacteria, which have been proposed as potential probiotics because butyrate promotes gut health (Duncan and Flint, 2013). *A. hadrus* is a strong biomarker for UC only with a frequency of 0.8, and *R. inulinivorans* is a moderate marker for CD with a frequency of 0.4. *Megamonas funiformis* is a weak biomarker for CD (with a frequency of 0.2) and was originally isolated from human faeces. Its role in the human gut or health is currently unclear (Sakon et al., 2008). In summary, a group of previously undescribed biologically plausible bacterial species that are robust microbial markers for IBD is presented. The group includes gut health promoting bacteria, bacterial species that thrive in the inflammatory environment of an IBD gut and possibly exacerbate the disease, and other bacterial species with unclear roles in human health.

4.5 Summary

Modelling the microbiome with supervised learning and feature selection approaches have implicated specific groups of bacterial genera in the pathophysiology of IBD. However existing models have only considered taxonomic data, whereas functional data could illuminate potential mechanisms that underlie IBD aetiology. In addition, existing feature selection algorithms have not considered the robustness of the feature selector output, which is important when planning clinical validation of *in silico* work. The hybrid model presented in this chapter merges the output of three different types of data (taxonomic, functional, and clinical) and measures the relevance of each data type to decompose non-invasive diagnosis of IBD into a series of simpler classification tasks. After using the Boruta algorithm to measure feature relevance, the concept of feature robustness was explored with the aggregating EFS model. Current work in non-invasive IBD prediction has not progressed beyond predictive model prototypes. The introduction of measuring feature robustness for microbiome census data is a valuable contribution that could assist clinical

validation in the future.

There has been debate as to whether functional data (inferred or measured directly) are useful for making predictions from the microbiome. Preliminary results suggested that there was little advantage to be gained from using functional data instead of taxonomic data for many classification tasks. However, there are compelling theoretical justifications to favour functional data over taxonomic data: the composition of the microbiome is often highly inconsistent across samples, but the functional content of the microbiome is typically more homogenous. An environmental niche will define the functional characteristics of a microbiome. For example, a highly saline environment will probably have a microbiome that includes many halotolerant and halophile species, and genes associated with salt tolerance. Many different halotolerant or halophile species can be present in a highly saline environment, but the genes present will remain similar. Theoretically this could decrease noise associated with natural fluctuations of the microbiome and improve the predictive power of models. However, a systematic analysis of functional data's utility has not been attempted with regards to IBD classification. To measure the relevance of functional data for disease classification the Boruta algorithm was applied to a variety of different classification tasks that formed a serial hybrid model. Functional data were the most relevant feature type for all classification tasks.

After applying Boruta feature selection to the hybrid model the concept of feature selector robustness was investigated and found to be absent in current work on predicting disease from the microbiome. Aggregating EFS was applied for non-invasive IBD prediction from high-resolution microbiome data. A set of robust novel bacterial species were implicated in the pathogenesis of IBD. The novel bacterial species were biologically plausible and have been associated with broad changes to gut health (via butyrate production) in the past. Due to the robust nature of the microbial markers the route towards potential clinical applications is made simpler. Due to the expense of marker gene surveys a simpler test would need to be developed. The use of ASV specific PCR probes to measure the relative abundance of the microbial markers is one cost-effective option. However, new models would need to be developed and validated as the process of generating the data would be significantly different compared with marker gene surveys, which could impact the predictive potential of the data.

To further illustrate the power of Computational Intelligence (CI) algorithms in the analysis of genomic data, high-resolution microbiome denosing algorithms will be extended too the analysis of the oral microbiome in subjects with depression in chapter 5. Furthermore, the concept of combining different types of data for classification is extended in chapter 5 for the prediction of depression from a saliva sample.

Publications arising from this work

The basis of this work has been published in:

Wingfield, B., S. Coleman, T. M. McGinnity and A. Bjourson (2018). 'Robust Microbial Markers for Non-Invasive Inflammatory Bowel Disease Identification'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Accepted for publication, in early access., pp. 1–1. ISSN: 1545-5963. DOI: 10.1109/TCBB.2018.2831212.

Wingfield, B., S. Coleman, T. McGinnity and A. J. Bjourson (2016). 'A metagenomic hybrid classifier for paediatric inflammatory bowel disease'. In: *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, pp. 1083–1089.

ALTERED ORAL MICROBIOTA IN YOUNG ADULTS WITH DEPRESSION

> That's a fool's experiment. But I love fools' experiments. I am always making them.
>
> Charles Darwin

5.1 Introduction

Depression is diagnosed with subjective criteria (e.g. the Hamilton depression scale) and no diagnostic tests are in widespread clinical use despite decades of work (Mössner et al., 2007) because of depression's unclear and complex pathophysiology. Studies attempting to identify the pathogenesis of depression have identified a number of candidate mechanisms including neurotransmitter deficiencies (Luscher et al., 2011), changes to neurotrophic levels (Brunoni et al., 2008), structural brain abnormalities (Lorenzetti et al., 2009), immune system dysregulation (Dantzer et al., 2008), and circadian rhythm disruption (Wulff et al., 2010). However, none of the theories have been fully accepted as a definitive model. Pharmacological interventions which target these candidate mechanisms remain the fastest and most effective way of treating the most severe forms of depression (Kirsch et al., 2008), but up to 50% of patients do not respond to first round treatment with antidepressant drugs (Mrazek et al., 2014). Due to the low response rate, multiple treatment cycles are a common approach to identify an effective antidepressant drug, which worsens patient outcomes (Mrazek et al., 2014). For patient outcomes to improve, new aetiological theories must be developed and explored in combination with a precision medicine approach.

The gut microbiota — the complex community of microorganisms that inhabit the human gastrointestinal tract – have been implicated in the pathophysiology of many diseases, including Inflammatory Bowel Disease (IBD) (Gevers et al., 2014), obesity (Turnbaugh et al., 2006), and diabetes (Qin et al., 2012). A growing body of evidence supports the view that the gut microbiota play a key role in the aetiology of depression via regulation of the central nervous system known as the gut-brain axis (Cryan and Dinan, 2012; Foster and Neufeld, 2013). Broadly speaking three mechanisms have been proposed to explain this regulation by the microbiota: by altering neurotransmitter signalling, modulating the hypothalamic-pituitary-adrenal (HPA) axis, and inflammation. Inflammation may lead to a

dysfunctional intestinal epithelium barrier or "leaky gut" due to the opening of intercellular tight junctions (Kelly et al., 2015), driving a chronic low-grade pro-inflammatory state and subsequent activation of the HPA-axis, via the exit of bacterial organisms and their products, as well as inflammatory mediators (Kelly et al., 2016). Bacterial products and inflammatory components can cross the blood brain barrier and cause an inflammatory response via the activation of microglia cells (Yirmiya et al., 2015), which are the primary immune cells of the central nervous system. Regular microglial activation can cause chronic brain inflammation, which potentially contributes to the structural and functional brain differences associated with mental health disorders (Stein et al., 2017). The gut microbiota can also modulate the concentration of neurotransmitters present in the host (O'Mahony et al., 2015). In addition to helping to understand the pathophysiology of depression, the gut microbiota can also provide an opportune location for the development of novel treatments. psychobiotics — probiotics with potential mental health benefits — could be therapeutically useful for treating mental illnesses in humans. Animal models have shown the psychobiotic potential of species such as *Lactobacillus rhamnosus* (JB-1) (Bravo et al., 2011). Although this candidate psychobiotic has to date failed to translate to humans (Kelly et al., 2017) recent work has shown administering *Bifidobacterium longum* 1714 can reduce stress and improve memory in a human cohort (Allen et al., 2016).

The majority of research to date has focused on the role of the microbiome-gut-brain axis in brain physiology and neurochemistry (Naseribafrouei et al., 2014; Jiang et al., 2015; Zheng et al., 2016). Although the oral microbiome is one of the most diverse microbiomes in the human body, has a significant influence on microbiomes found across the rest of the gastrointestinal tract, and plays a key role in health and disease (Wade, 2013) it has received little attention to date. Saliva is a cost effective non-invasive biomarker source that offers collection, handling and economic advantages compared with methods that require stool or biopsy samples (Yoshizawa et al., 2013). Saliva is a heterogeneous fluid made up of water, proteins and small inorganic substances. Saliva is essential for digestion, lubrication, and acts as a barrier to pathogens (Humphrey and Williamson, 2001). Three major salivary glands create approximately 90% of saliva fluid and these glands are surrounded by blood capillaries. Therefore saliva glands have the potential to absorb blood based biomarkers of disease, which suggests that saliva fluid may contain comprehensive disease information (Liu and Duan, 2012). Oral dysbiosis has been linked to systemic diseases (i.e. affecting the entire body, not just the oral cavity) with an underlying inflammatory aetiology such as rheumatoid arthritis (Said et al., 2013) and Alzheimer's disease (Shoemark and Allen, 2015). In this Chapter bacterial 16S ribosomal ribonucleic acid (16S rRNA) high-throughput gene sequencing will be used to compare the oral microbiota of 87 young adults (44

depressed and 43 controls) to evaluate if changes to the structure and composition of the oral microbiota are associated with depression status. Section 5.2 will outline the theory that underpins microbial ecology, the study of the relationships between microorganisms and their environment. Section 5.3 will describe the process of applying the microbial ecology theory to the 16S rRNA data to identify alterations induced by depression. Section 5.4 presents a set of alterations to the oral microbiome associated with depression for the first time. Section 5.5 implements a data-driven Computational Intelligence (CI) algorithm known as a Super Self-Organising Map (sSOM) (Wehrens, Buydens et al., 2007) to perform multimodal classification and enable the prediction of depression from microbiome census data with the highest reported performance in literature to date. This Chapter finishes with a summary in Section 5.6.

5.2 Microbial ecology theory

Many methods of analysing microbiome census data are taken from ecology. Ecology is the scientific analysis of the interactions between organisms and their environment. Ecology includes studying interactions between members of the same species, interactions across different species, and interactions with abiotic (e.g. physical or chemical) factors of the environment (Stauffer, 1957). Microbiome census data can be used to investigate the ecology of microorganisms (also known as microbial ecology or environmental microbiology).

Ecological terms and processes that describe the diversity and structure of a site are widely applied to microbiome census data. Whittaker introduced the terms alpha diversity (α-diversity), beta diversity (β-diversity), and gamma diversity (γ-diversity; Whittaker, 1972). He proposed that the total species diversity of an environment (γ-diversity) could be estimated from the mean species diversity across local sites (α-diversity) and the changes among these sites (β-diversity). The questions these terms and processes can help answer about the richness and structure of microbiomes is an important stage of most microbiome experiments. However, it is important to be cautious when applying widely used procedures from ecology to microbiome count data derived from high throughput sequencing, as many assumptions (e.g. regarding heteroscedasticity, sparsity, and compositionality; described fully in Chapter 3) do not hold true.

5.2.1 Estimating diversity

In microbial ecology α-diversity is used to measure the *within-community* diversity of samples (i.e. considering the diversity of samples individually). Methods of measuring diversity are usually split into two categories: presence-absence metrics

and relative abundance indices. Measuring α-diversity with species richness falls into the former category. Most diversity indices fall into the latter category. A diversity index is a mathematical measure of the diversity of a site. By taking into account the relative abundance of species, diversity indices offer more information about the true diversity of a site.

Alpha diversity indices

Table 5.1: Species richness of two example sites.

	Site A	Site B
Species 1	25	90
Species 2	25	10
Species 3	25	5
Species 4	25	5
Total	100	110

Consider the simple example in Table 5.1. Which site is more diverse? Although site B has a greater richness than site A, the distribution of species in site A is much more even. Biological diversity stabilises complex ecological systems in response to environmental changes (Cleland, 2011). Experimental evidence has shown that environments such as site A are positively correlated with ecosystem-level stability but negatively correlated with species-level stability, as smaller species populations are more likely to go extinct from random environmental changes (Cleland, 2011). Therefore for a more comprehensive view of biological diversity, evenness — which measures the relative abundance of the different species present in a site — must be incorporated into diversity estimates. A diversity index can take into account the distribution of species in a site. Dozens of diversity indices exist, and three examples will be described further: Simpson's Diversity Index (Simpson, 1949), the Chao index (Chao, 1984), and Faith's Phylogenetic Diversity Index (Faith, 1992). The first is important because it describes both richness and evenness. The second is important because in addition to describing richness and evenness the index takes into account uneven sampling depths across samples. The third incorporates phylogenetic differences between species. Simpson's Index is defined as:

$$d_{\text{simpson}} = \frac{\sum_{i=1}^{S_{\text{obs}}} n_i \left(n_i - 1\right)}{N \left(N - 1\right)} \tag{5.1}$$

where S_{obs} is the number of observed operational taxonomic units (OTUs), n_i is the number of individuals present in the i-th OTU, and N is the total number of

individuals in a site. When calculating d_{simpson}, 0 represents infinite diversity, and 1 no diversity. This is somewhat counterintuitive, and it is common to calculate the reciprocal of d_{simpson} to obtain the inverse Simpson's index in which higher diversity is represented by a larger value.

A sampled site will always have undetected species present. This problem is worse for microbiome census data where library sizes can differ by orders of magnitude across samples. The Chao index offers a method for estimating and incorporating the number of unseen species present in a site. The Chao 1 index, a widely used variant of the Chao index, is defined as:

$$d_{\text{chao}} = S_{\text{obs}} + \frac{F_1^2}{2F_2} \tag{5.2}$$

where S_{obs} is the number of observed OTUs, F_1 is the number of species with exactly one observation (singletons), and F_2 is the number of species with exactly two observations (doubletons). The theoretical justification for the Chao index is that if rare species (singletons) are still being discovered during sampling then more undiscovered rare species are likely to be present. If all species are observed at least twice then it is likely no more species will be discovered. It has been shown that the Chao 1 estimator performs well on standard ecological datasets, and the degree of certainty can be measured with confidence intervals (Colwell and Coddington, 1994).

Although calculating Chao 1 is supported by all microbiome workflow software packages it is not conceptually valid to apply non-parametric estimators reliant on singletons to microbiome census data. It is currently impossible to distinguish true singletons from sequencing error. If singletons are included in microbiome census data, the majority are likely to be false positives. Chao 1 is extremely sensitive to singletons (singleton abundance is squared), so including false positive singletons artificially inflates the apparent richness of a sample. Due to the difficulties described above, denoising pipelines such as `dada2` do not attempt to call singletons. This process would result in Chao 1 not providing a proper estimated richness. The diversity of microbiome census data should instead be estimated with indices that are not reliant on singletons, despite these measures not taking into account uneven library sizes across samples.

Faith's phylogenetic diversity is a diversity estimator that incorporates phylogenetic differences between species, and is defined by "the sum of the lengths of all those branches that are members of the corresponding minimum spanning path" (Faith, 1992), where a branch is defined as a segment of a cladogram (a phylogenetic tree), and the minimum spanning path is defined as the minimum distance between two nodes (Faith, 1992). Capturing phylogenetic information can improve the ability of an alpha diversity estimator to capture the true level of diversity in an environment. For example, a site containing many closely related

species could be considered less diverse than a site containing fewer but highly unrelated organisms.

Beta diversity indices

The Bray-Curtis dissimilarity index is a beta diversity index that quantifies the dissimilarity between different sites, and is defined as (Bray and Curtis, 1957):

$$\mathrm{BC}_{i,j} = 1 - \frac{2C_{i,j}}{S_i + S_j} \tag{5.3}$$

where $C_{i,j}$ is the sum of the lesser counts for each species found in both sites, and S_i and S_j are the total number of species counted at both sites. The Bray-Curtis dissimilarity is bounded to between $0 \leq \mathrm{BC}_{i,j} \leq 1$, where 0 represents total dissimilarity (no species are shared between the two sites) and 1 represents total similarity.

Taxon resampling curves

Taxon resampling curves can be used to determine if enough observations have been made so that a quantity (e.g. an estimate of biological diversity; R) can be estimated from the sampling process. Further biological replicates or deeper sequencing may be required to get a true understanding of the structure of a microbial community if a site has been insufficiently sampled. R is typically measured via species richness (the raw number of different species in a defined environment) or other measures of α-diversity. Taxon resampling curves are also known as rarefaction curves, and the process of generating rarefaction curves is sometimes called rarefaction. However, it is important not to confuse generating rarefaction curves with the normalisation procedure that is also called rarefaction. The rarefaction normalisation technique randomly discards observations from samples until a defined threshold is met. Rarefaction normalisation effectively reduces the library size of samples to a common threshold (typically set to be the library size of the sample with the fewest observations).

Taxon resampling curves plot the value of R against the number of observations used to calculate R (see Figure 5.1). R is estimated for fewer observations by random undersampling. If R is horizontally asymptotic it is reasonable to assume that sufficient sampling has been done. If R has not converged, then it is possible that a good estimate of R cannot be made for the true population. Taxon resampling curves are only suggestive and cannot be interpreted in a strict way. It is possible that rare species have been missed by chance and that the sampling effort has been insufficient even if R has converged. Additionally, applying taxon resampling curves to data gathered via high-throughput sequencing raises an additional problem: it

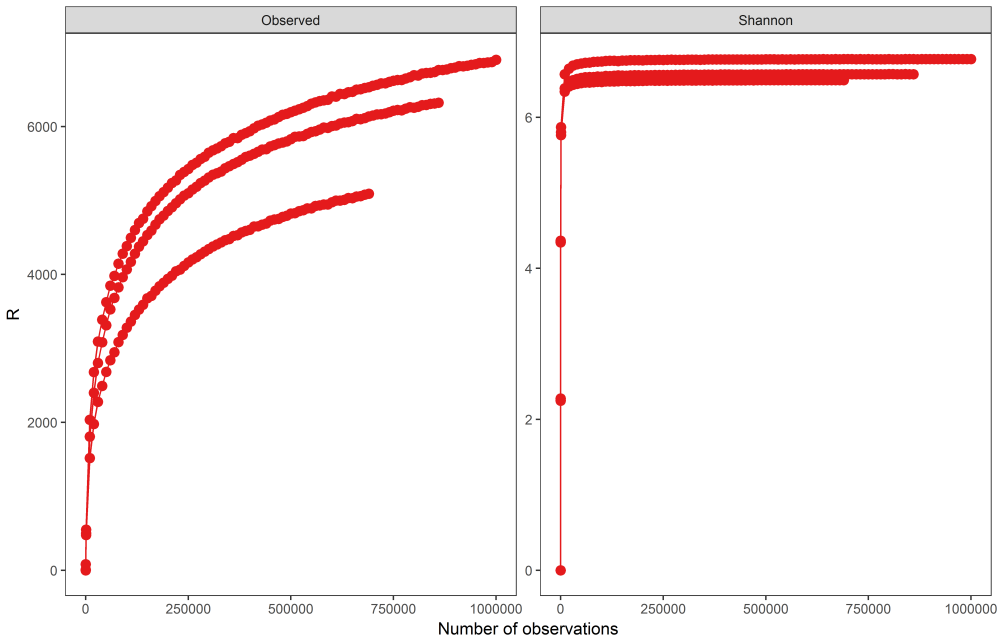


Figure 5.1: Taxon resampling curve (example data taken from the Global Patterns dataset). Left: Number of OTUs. As sampling increases R increases, falsely indicating sampling is insufficient. Right: α-diversity measured by the Shannon diversity index. As sampling increases R has reached a horizontal asymptote, indicating the value of R is a reasonable estimate of diversity and sufficient sampling has taken place.

is impossible to determine if rare species are truly present or if they are observed because of sequencing error.

High-throughput sequencing will have a constant error rate above 0%. As the library size (the number of discrete sequence reads) of a sample increases, the number of observed species will also increase because of sequencing error, even if all of the true species have already been observed. This will cause R to never converge (see left side of Figure 5.1). Taxon resampling curves must be therefore interpreted with this caveat in mind.

5.2.2 Analysing composition

One of the most widely applied approaches to analysing the composition of the microbiome is differential abundance tests. Differential abundance tests measure if the mean abundance of a taxa is significantly different across multiple sample classes defined by the experimental design. `DESeq` is a complex software package

Table 5.2: Poisson distribution example

Expected number of red balls	Standard deviation of number of red balls	Relative error in estimate
10	$\sqrt{10} = 3.16$	31.60%
100	$\sqrt{100} = 10.00$	10.00%
1,000	$\sqrt{1,000} = 31.62$	3.20%
10,000	$\sqrt{10,000} = 100.00$	1.00%

designed to accurately analyse the differential expression of sequence count data (Anders and Huber, 2010). Although `DESeq` was originally developed for RNA-Seq count data, it has been applied to microbiome count data, as they share many properties (McMurdie and Holmes, 2014). Before `DESeq` can identify differentially abundant taxa the library size for each sample must be normalised. `DESeq` creates a "virtual reference sample" by taking the geometric mean of each taxa abundance for all samples. Each sample is then normalised to the reference sample to identify a scaling factor (called a size factor) for each sample. The key challenge for a differential abundance test is to determine if the variance in sequence read counts across biological replicates arises from random noise or from true biological variation.

Noise (variance) is correlated with abundance in microbiome census data. Therefore statistical power is also correlated with abundance in microbiome count data. Consider an experiment with a bag containing small white and red balls. The task of the experiment is to determine the fraction of red balls present in the bag (e.g. 20%). Each subject is permitted to withdraw a certain number of balls from the bag without looking. Variation in the number of balls sampled implies different levels of uncertainty about the estimated fraction of red balls present in the bag (see Table 5.2).

The negative binomial distribution is a probability distribution that is a generalisation of the Poisson distribution and includes two parameters (mean μ and variance $q + v$). It has been found that the Poisson distribution can be used to accurately model noise between technical replicates but is unable to accurately model the variance introduced from biological replicates for RNA-Seq data (Marioni et al., 2008). The differential abundance test implemented by `DESeq` assumes that the count $K_{i,j}$ for gene i in sample j is generated by the negative binomial distribution with mean $s_j\mu_j$ and dispersion α:

$$K_{i,j} \sim \text{NB}(s_j\mu_i\alpha_i) \tag{5.4}$$

where s_j is a scaling factor that accounts for the library size of sample j. Estimating

α is difficult for each gene with sample sizes that are typical in biological experiments. Therefore α is estimated by assuming that genes with similar abundances have similar variances across samples. By sharing this information across samples the mean-dispersion relationship can be accurately estimated. This is the key advantage of applying `DESeq` to microbiome census data: the information sharing process increases the power of the test to detect differential abundance whilst controlling false positives. The null hypothesis is that all samples have the same μ_j (and that any differences are generated only by noise). The alternative hypothesis is that μ_j is the same only within groups (Love et al., 2014):

$$\log \mu_j = \beta_0 + x_j \beta_T \tag{5.5}$$

where $x_j = 0$ if j is a control sample and $x_j = 1$ if j is a treatment sample (i.e. the phenomenon being investigated by the experimental design). The data are fitted to a generalised linear model and the coefficients β are estimated. A Wald test is used to determine the probability that the difference between control and treatment is observed if there is no true effect (i.e. low probability indicates there is true biological variation; Love et al., 2014).

`DESeq` also offers a method of transforming count data so that the variance is approximately independent of the mean, called a variance stabilising transformation. This is useful for downstream applications such as machine learning and clustering. The variance stabilising transformation is given by (Anders and Huber, 2010):

$$\tau(\kappa) = \int^{\kappa} \frac{dq}{\sqrt{w(q)}}. \tag{5.6}$$

where $w(q)$ is the variance-mean dependence estimated by `DESeq` (see red line on the left panel of Figure 3.12). Applying transformation τ to count data $\frac{k_{i,j}}{s_j}$, where $k_{i,j}$ is the count of the i-th sequence of the j-th subject and s_j is the size factor (depth of coverage) of the j-th sample, returns values that have approximately similar variances. The returned values are normalised with respect to library size and are on the $\log_2$ scale.

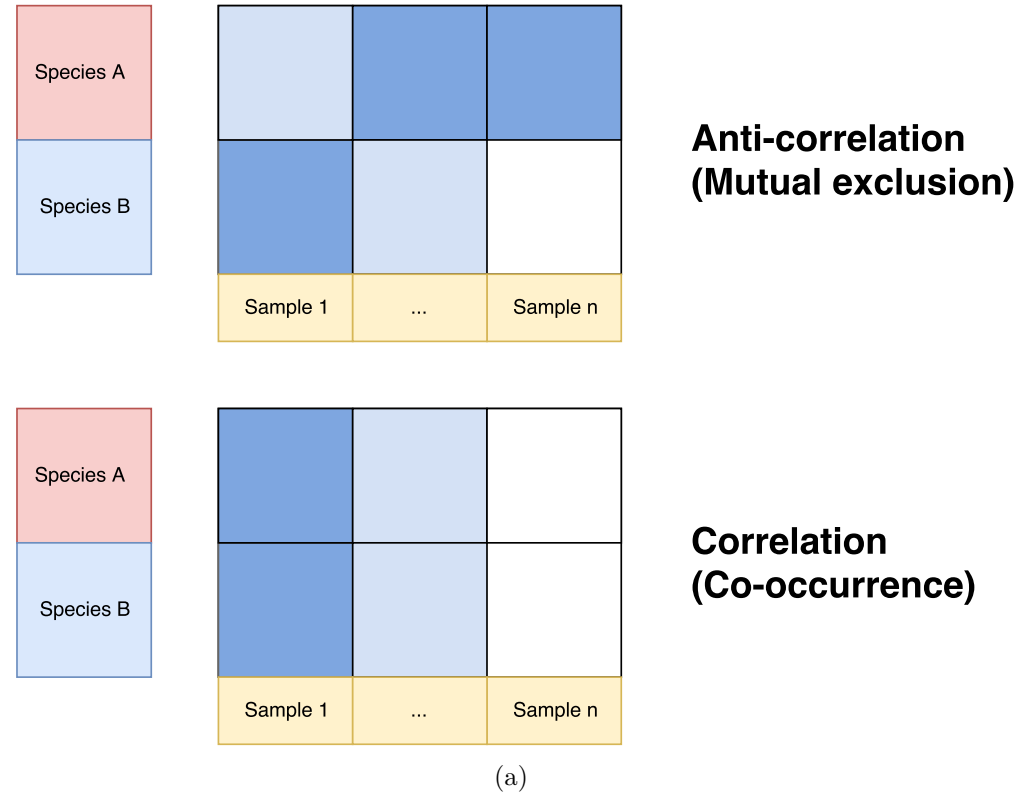
The variance stabilising transformation can produce negative numbers which can break downstream techniques that require positive numbers. A negative count is equivalent to the abundance of a particular sequence being lower than the detection limit of the sampling method. This is a particular problem for many techniques which have been developed and widely applied in ecology. Ecologists cannot be faulted for not considering this as, for example, it would be odd to find less than 0 sheep in a surveyed field.

5.2.3 Inferring microbial interactions

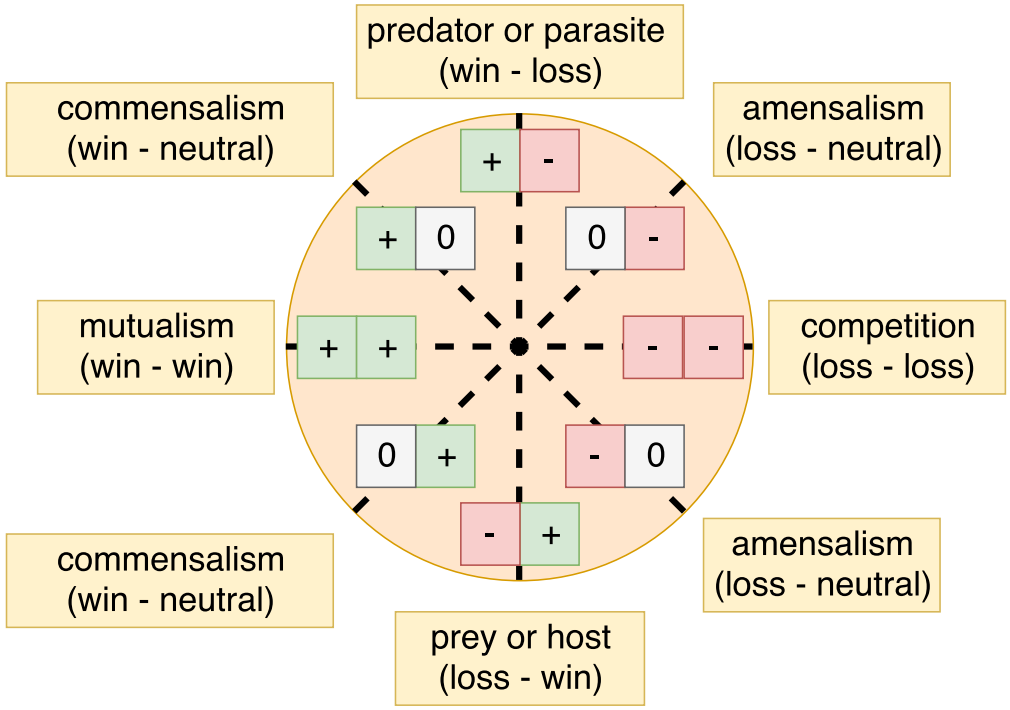
Interactions between species have been regularly inferred from abundance patterns. Diamond first suggested ecological relationships could be inferred from the presence or absence of species across habitats (known as a "checkerboard pattern"; see Figure 5.2(a); Diamond, 1975). Similar patterns have also been observed in microorganisms (Horner-Devine et al., 2007): microbial interaction analysis detects patterns of co-occurrence and mutual exclusion across different samples, which are thought to represent a range of ecological relationships such as mutualism or commensalism (see Figure 5.2(b)). Complex ecological relationships are responsible for driving a number of natural phenomena, including dental plaque formation (Kolenbrander et al., 2010) and algal blooms. Microbial interaction analysis is an instance of network inference, which attempts to identify relationships from count data.

Generally there are two reasons why a pair of species consistently correlate or exclude one another: because of an ecological relationship (e.g. commensalism, amensalism, mutualism, competition, and predator or prey) or due to an ecological niche overlap or alternative preference. An ecological niche is defined as the role and position a species has in its environment. For example, a halophile will have a co-exclusion relationship with any species that cannot tolerate salinity in a saline environment. In contrast to ecological relationships there is no direct interaction between the species in this example, the pattern of co-occurrence is caused by the environment.

Microbiome count data are sparse and compositional, which creates problems for standard similarity measures that are used in similarity-based network inference. Computing a correlation score for a pair of species with an abundance of zero is particularly problematic (known as the "double zero problem"), as it is impossible to know if the species are below the detection threshold or if the species is truly absent. It is therefore sensible to avoid giving sparse pairs a high correlation score. Similarity measures are also often severely distorted by compositional data, because if one species has a particularly high abundance other species will appear to have a lower abundance as the sum is constrained to an arbitrary limit (1 in the case of data normalised to be a proportion). Sparse Correlations for Compositional data (`SparCC`) (Friedman and Alm, 2012) is a network inference algorithm that is designed to not be affected by compositionality or sparsity, and is applied in this Chapter. Once an accurate similarity score is generated the significance of the correlation for each species pair can be assessed via a bootstrap significance testing procedure. A network with edges as relationships and nodes as taxa can be created to visualise significant ecological relationships (after discarding edges with a p value above 0.05, and removing nodes with no edges).



(a)



(b)

Figure 5.2: (a) Example of checkerboard patterns arising from abundance data that show co-occurrence relation (b) Ecological relationships

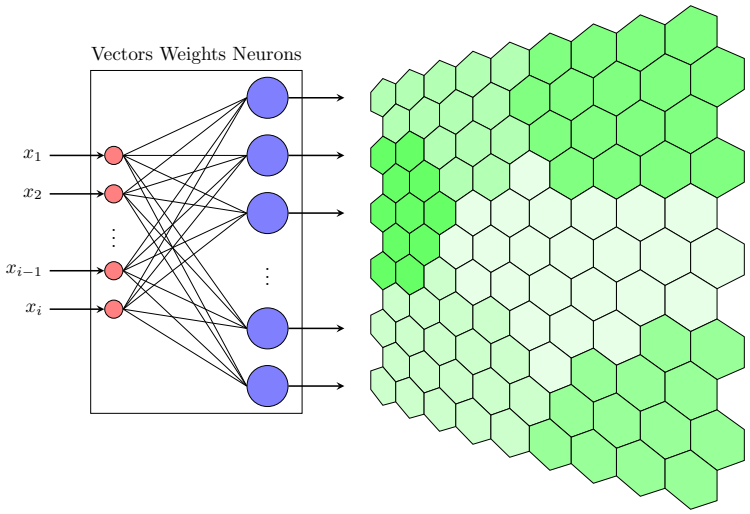


Figure 5.3: A self-organising map. Adapted from "Self-organising map" (Wilbrow, 2013).

5.2.4 Self-Organising Maps for microbial ecology

Data-driven approaches have minimal priors or constraints. Standard statistical tests and models make many assumptions about input data. For example, parametric statistical tests assume that input data are normally distributed, have a homogenous variance, have a linear relationship, and are independent. As discussed in Chapter 3, microbiome census data violate many of these assumptions. In this chapter an Artificial Neural Network (ANN) variant called a Super Self-Organising Map (sSOM) (Kohonen, 1998; Melssen et al., 2006) is applied for the purpose of supervised classification as it is a data-driven algorithm and can tolerate highly dimensional input data. ANNs are inspired by biological nervous systems and have been widely used for supervised regression and classification. ANNs can model complex nonlinear relationships between an input feature space (i.e. the composition of a microbiome) and an output feature space (class membership).

Self-Organising Maps (SOMs) (Kohonen, 1990) make no assumptions about the distribution or properties of the input data, and can easily scale to very large data sets, which makes them appealing for analysing highly complex biological data. SOMs are often used for unsupervised learning; an XY-Fused (XYF) SOM consists of two layers (one for input, X, and one for output, Y) and is capable of supervised classification. sSOMs are based on XYF maps, but expand the concept to include a set of input maps (one for each data type) to enable multimodal data fusion (Melssen et al., 2006).

Self-Organising Maps

A SOM is a set of unconnected units that are ordered according to a topology parameter (often a two-dimensional hexagonal grid). The units are connected to the vertices of the topology. Each unit is assigned a weight vector for each input instance (e.g. a vector of bacterial counts). Input instances are randomly presented to all units in the network. The unit with a weight vector closest to the presented instance is deemed the winner. After the winner is chosen the weight vectors of the winning unit and its closest neighbours are updated to be more similar to the presented instance. The updating process is done by calculating the difference between the input instance and the weight vector of the respective unit and modifying the difference by a (the learning rate). The modified difference is then added to the original weight vector, making the winning unit and its neighbours more similar to the input instance. This process is iterated until every input instance has been presented to the network a sufficient number of times, which is a parameter set by the user (epoch number).

Super Self-Organising Maps

The XYF network (the sSOM) uses a fused similarity measure that relies on a combination of similarities between an input instance X and all units in the X network, and the similarities between the output instance Y (class membership) and all units in the Y network. The fused similarity measure for X_i, Y_i is defined as:

$$S_{\text{fused}}(i,j) = \alpha(t)\, S(X_i, \text{Xmap}_x) + (1 - \alpha(t)\, S(Y_i, \text{Ymap}_x)) \tag{5.7}$$

where $\alpha(t)$ is the relative weight between similarities $S(X_i, \text{Xmap})$ and $S(Y_i, \text{Ymap})$ in the t-th epoch (Melssen et al., 2006). One epoch occurs after all samples in the training set have been presented to the network. The fused similarity measure is used to determine the winning unit that is best across both maps (i.e. incorporating input data as well as class membership data). More in depth explanations of the algorithm are available (Melssen et al., 2006).

In many scientific fields information about a phenomenon can be recorded from different types of detectors, across multiple experiments, and in different conditions. Each of these different recording methods is referred to as a modality. Multimodality is particularly important for biological data because a single modality will rarely provide complete knowledge about a complex system. To test the hypothesis that a multimodal paradigm would benefit modelling the oral microbiome, a multimodal approach was implemented by fusing microbiome census data, sequencing metadata (the library size represents a measure of certainty about the sequencing process), and host (environmental) data.

Table 5.3: Sample demographics Cases; n=44 and controls; n=43 controls. Age, gender, smoking status and depression severity score based on participant response to CIDI depression section. Maximum depression score for inclusion in healthy group = 15, and minimum depression score for inclusion in depression group = 30.

Demographics	Controls (n = 43)	Cases (n = 44)
Age (mean)	21	22
(Range ± SD)	(18 – 36 ± 3.9)	(18 – 38 ± 5.3)
Gender		
Male	13 (30.2)	11 (25.0)
Female	30 (69.8)	33 (75.0)
Smoking status		
Past (%)	0 (0.0)	7 (29.5)
Daily (%)	3 (7.0)	13 (29.5)
Occasional (%)	6 (14.0)	9 (20.5)
Never (%)	34 (79.1)	11 (25.0)
Missing (%)		4 (9.12)
Depression score (mean)	34.6	10.1
(Range ± SD)	(32 – 35 ± 0.9)	(7 – 14 ± 2.5)

5.3 Modelling the oral microbiome

Samples for this study were utilised from the Ulster University Student Wellbeing Study (UUSWS), conducted as part of the WHO World Mental Health International College Student Project (WMH-ICS), with Ulster University representing Northern Ireland in this global initiative (McLafferty et al., 2017). Ethical approval was obtained from Ulster University Research Ethics Committee (REC/15/0004). First year students were recruited during registration where they gave written consent, provided a saliva sample and were given a unique, anonymous number to complete an online mental health survey clinically validated against the DSM-IV.

Saliva samples were collected using Oragene OG-500 kits (DNA Genotek, Ontario Canada), enabling the self-collection and stabilisation of DNA at room temperature. Cases of depression (n=43) were selected based on survey responses to seven questions corresponding to DSM-IV criteria for depression using a Likert scale response, and controls matched where possible for age, gender, ethnicity and smoking status (see Table 5.3). After quality control checks 83 samples remained for analysis.

Microbiome DNA purification was carried out using MasterPure™ DNA Purification Kit and Ready-Lyse™ Lysozyme from the MasterPure™ Gram Positive DNA Purification Kit (Epicentre, Madison, US) according to the manufacturer's

instructions. The quantity of DNA was measured on a Nanodrop spectrometer (Fisher Scientific, Loughborough, UK) and the quality measured using the 260/280 ratio and 1.5% gel electrophoresis. To confirm the presence of bacterial DNA, broad range 16S PCR was carried out. Finally, 50µl of 22ng/µl of good quality DNA was sent to The Forsyth Institute for 16S high-throughput sequencing (Duran-Pinedo and Frias-Lopez, 2015).

To prepare for sequencing, PCR amplification of 10—50ng of sample DNA was carried out using V3 – V4 primers and 5 Prime Hot Master Mix. The amplicon product was then purified using Solid Phase Reversible Immobilization with AMPure beads, and 100ng of each amplicon library was pooled, gel-purified, and quantified using a bioanalyser and subsequent qPCR. Finally, 12 pM of the library mixture was then spiked with 20% PhiX (Illumina, San Diego, CA), and sequenced on Illumina MiSeq (Belstrøm et al., 2016). The *in vitro* work described above was performed by Elaine Murray and Coral Lapsley at the Northern Ireland Centre for Stratified Medicine.

The resulting sequence data were denoised with the `R` v3.4.2 package `dada2` (v1.4.0; Callahan et al., 2016b) using a standard operating protocol (Callahan et al., 2016b). In brief quality-filtered paired end sequences reads were trimmed, denoised, and joined into contigs. Chimeric sequences were removed and taxonomy was assigned to the denoised sequence reads using the Ribosome Database Project's naïve Bayesian classifier (Wang et al., 2007a) and the SILVA 16S rRNA gene reference database (Quast et al., 2012). The denoised sequences represented exact 16S rRNA gene sequence variants. These sequence variants were not binned into fuzzy operational taxonomic units as the exact sequence variant paradigm is superior to a sequence similarity cutoff approach (Callahan et al., 2017). A *de novo* phylogenetic tree was generated from the amplicon sequence variants (ASVs) with the `R` package `phangorn` v2.3.1 (Schliep, 2010). The abundance of 16S rRNA gene sequence variants, taxonomy data, phylogenetic tree, and sample information (e.g. depression status) were combined into a `phyloseq` v1.20.0 (McMurdie and Holmes, 2013) object for statistical analysis. Exact sequence variants of interest were further analysed (e.g. differentially abundant exact sequence variants) by matching sequences against the Human Oral Microbiome Database (Chen et al., 2010); ASVs were matched to a species level to identify possible mechanisms of action.

5.3.1 Statistical analysis

The microbial community composition (β-diversity) was estimated using Bray-Curtis dissimilarity with the `R` package `vegan` (v2.4.3; Oksanen et al., 2007). The Bray-Curtis dissimilarity was estimated from normalised copy number compensated microbiome census data. To detect statistical differences in β diversity between

groups a Permutational multivariate analysis of variance (PERMANOVA) implemented in the vegan package was used. A β-dispersion test (`vegan::betadisper`) was used to verify that statistically significant groups identified by PERMANOVA had the same dispersions. The community structure of the oral microbiome was visualised with a canonical correspondence analysis (CCA) biplot; statistically significant environmental terms (determined by the PERMANOVA test) were included on the ordination. The significance of the CCA ordination solution was confirmed with a permutation test (`vegan::anova.cca`).

Differential abundance of ASVs was tested using the R package DESeq2 (v1.18.1; Love et al., 2014). To preserve statistical power very rare ASVs (present in less than 10% of samples) were removed prior to testing. DESeq2 implements a generalised linear model (GLM) based on the negative binomial distribution to detect differential expression in count data while accounting for differences in library size and biological variation. Although DESeq2 was originally developed for RNASeq data recent work has shown that it is well suited for application to microbiome census data compared with other widely used statistical techniques that rely on destructive normalisation techniques (McMurdie and Holmes, 2014). Raw reads from both the microbiome count data and functional profiles were fitted to a negative binomial GLM and a Wald test was used to determine the significance of GLM coefficients. DESeq2 corrects for multiple testing with the Benjamini-Hochberg adjustment; statistical significance was determined at the 5% level. Differential abundance was expressed as $\log_2$ fold change in depressed subjects relative to control subjects. Differential abundance was determined for both microbiome census data and functional profiles with a design blocking variation introduced by smoking and gender (i.e. only considering the potential effects of depression on abundance).

The SparCC algorithm (Friedman and Alm, 2012) implemented in the fastspar (v0.0.3) software package was used to calculate the correlation (co-occurrence) of ASVs. The co-occurrence matrix is a symmetrical $N \times N$ matrix (where N gives the total number of ASVs). Exact p-values were calculated for the co-occurrence matrix via permutation tests (1000 iterations). The original SparCC algorithm estimates pseudo p-values, which can be zero. Permutation p-values should never be zero, as zero values cause multiple testing correction procedures to be overly lenient (Phipson and Smyth, 2010). The fastspar implementation reports exact p-values. GNU Parallel (v20141022; Tange et al., 2011) was used to parallelise fastspar to decrease the execution time of the process. A correlation matrix and exact p-value matrix were estimated for non-smoking depressed subjects and non-smoking healthy subjects. The p-value matrix was false discovery rate adjusted (Benjamini and Hochberg, 1995). The R package `igraph` (v1.1.2; Csardi and Nepusz, 2006) was used to build an undirected graph from each co-occurrence matrix, in which nodes

are exact sequence variants and edges are the interaction type (e.g. co-presence or co-absence). Edges with $p > 0.05$ were removed and nodes with no edges after filtering were also removed. This resulted in a graph subset for both the healthy and depressed cohort. The set difference of the graphs was taken to identify statistically significant microbial interactions that were unique to the depressed cohort.

16S rRNA gene copy number compensation and prediction of functional content with Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt)

Different bacteria have a different amount of 16S rRNA gene copies (16S copy number), which can bias estimates of abundance and diversity (a bacteria with a very high 16S copy number will have an artificially inflated abundance). The 16S copy number of ASVs was estimated from the ribosomal RNA database (v5.1; Stoddard et al., 2014). Approximately 50% of the ASVs were not present in the database. The copy number for unknown ASVs was estimated using the copy number of the known DSVs and a phylogenetic ancestral state reconstruction algorithm (the `R` package `picante` 1.6-2 Kembel et al., 2010). The compensated abundance ASV $y_{i,j}$ was calculated by $y_{i,j} = \frac{x_{i,j}}{z_i}$ where $x_{i,j}$ gives the count of the i-th amplicon sequence variant from the j-th sample, and z_i gives the copy number. ASVs with an abundance less than 1 for every sample after this transformation were removed. The compensated counts were used for every stage of the analysis, except differential abundance testing and functional prediction.

PICRUSt (Langille et al., 2013) was used to identify differences in inferred functional content between depressed and control groups. In brief: ASVs were added to the GreenGenes version 13.5 reference database. ASVs that diverged by more than 3% were discarded according to a standard operating protocol (Maffei, 2018). New PICRUSt precalculated files were created from the new reference database. ASV abundance was normalised by 16S copy number and the bacterial composition was used to predict KEGG Ortholog (KO) from the new precalculated files. KOs were collapsed into Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using the `categorize_by_function.py` command provided by PICRUSt. Linear discriminate analysis effect size (LEfSe) was used to identify differentially abundant functional pathways in the depressed cohort (Segata et al., 2011).

5.3.2 Multimodal classification of depression

The `kohonen` package (v3.0.4; Wehrens, Buydens et al., 2007) in `R` was used to implement a sSOM with separate layers for each data type. The sSOM was used to

perform two-class supervised classification (healthy or depressed). Three types of microbiome data were used to train a map (four including class memberships) i.e. untransformed raw microbiome census data, the library size for each sample, and environmental data. The cohort was randomly divided into a training set (80% of samples) and a testing set (20% of samples). ANNs are sensitive to feature scaling (i.e. extreme ranges), so each data type was centered and scaled for both partitions. After training the first three layers of the sSOM were used to predict the class of unseen data. The predictions were compared against the true class memberships to evaluate the performance of the model.

Benchmarks were implemented to verify that the multimodality improved classification performance, and to check the performance of non-fusing alternative algorithms. Random Forests (an ensemble of decorrelated decision trees) were chosen as a neural network alternative. Random Forests are capable of modelling nonlinear class boundaries and have been found to be one of the most effective machine learning algorithms for microbiome count data (Statnikov et al., 2013). Random Forests were benchmarked with microbiome census data that had been normalised with popular techniques, including total sum scaling (proportions), random subsampling (rarefying), and a variance stabilising transformation provided by `DESeq2`.

5.4 Markers of depression in the oral microbiome

Sequencing the V3 – V4 regions of the 16S rRNA gene generated a total of approximately 12.5 million sequence reads (median ± MAD): $\approx 66,000 \pm 28,000$ sequence reads per subject. Sequence reads were denoised into ASVs, and assigned taxonomic classifications to the highest resolution possible. The denoised dataset that was analysed included the abundance of 2883 unique sequences covering 9 phyla, 18 classes, 33 orders, 53 families, 84 genera, and 133 species. The dominant phyla present in the oral microbiota across the entire cohort were Bacteroidetes ($42.18 \pm 13.87\%$), Proteobacteria ($24.57 \pm 17.29\%$), and Firmicutes ($26.62 \pm 9.93\%$) (see Figure 5.5(a)). The most prevalent families in the oral microbiota for all subjects were Prevotellaceae (37.22%), Pasteurellaceae (15.60%), Streptococcaceae (10.59%), Veillonellaceae (5.46%), and Neisseriaceae (5.50%) (see Figure 5.5(b)). A taxon resampling curve shows that the oral microbiome of the depressed and healthy cohort was sampled thoroughly enough to get an accurate representation of the composition of the oral microbiome (see right side of Figure 5.4).

The structure and composition of the oral microbiome was characterised with a range of techniques, beginning with ecological measures such as richness (the

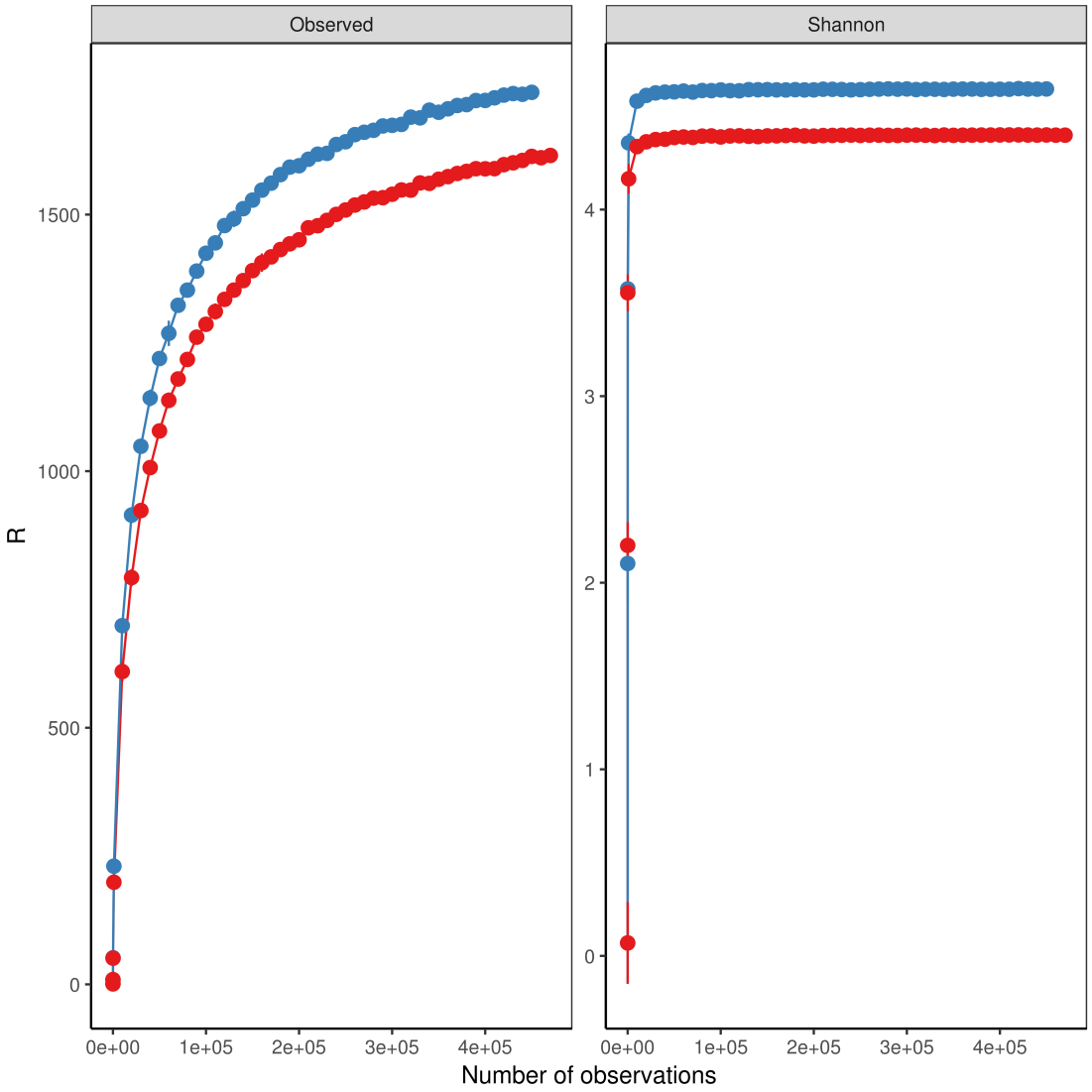
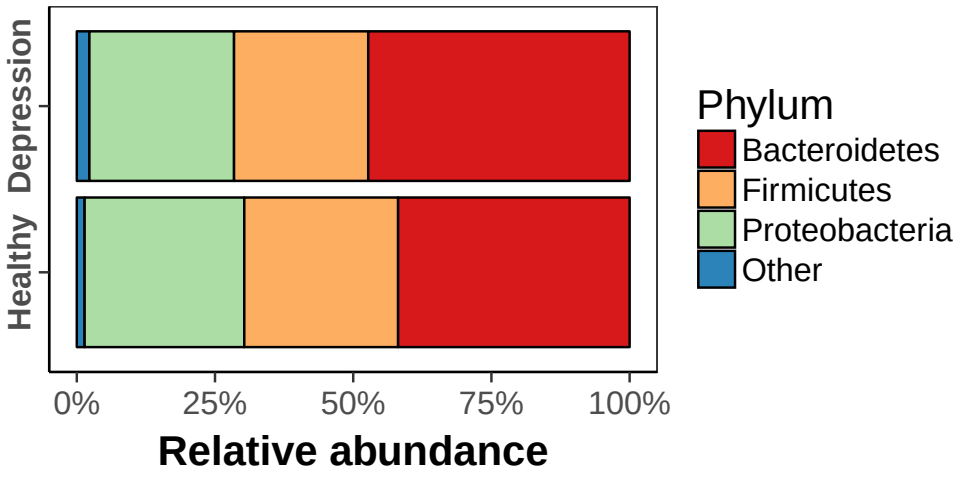
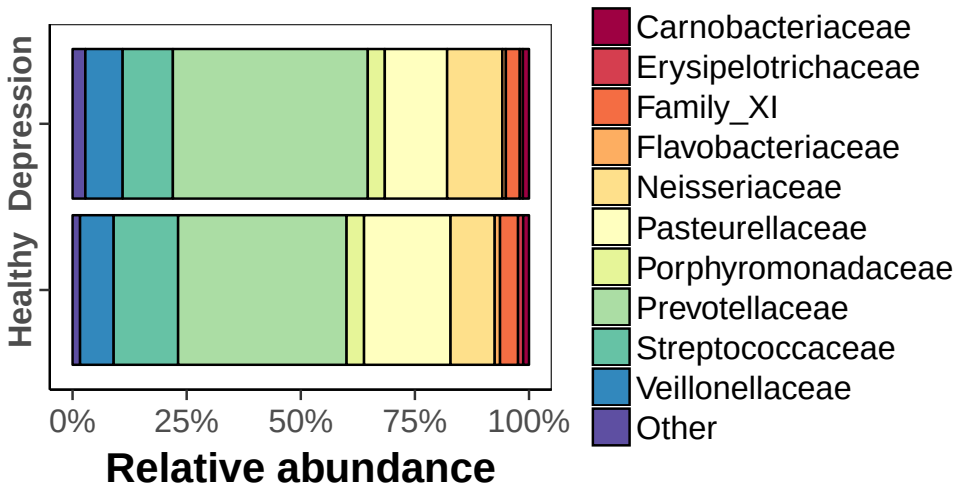


Figure 5.4: Taxon resampling curve (red: control, blue: depression). Left: Number of ASV. Right: α-diversity (Shannon diversity index). Sufficient sampling has been done to get a reasonable measurement of microbial community composition as R has converged.

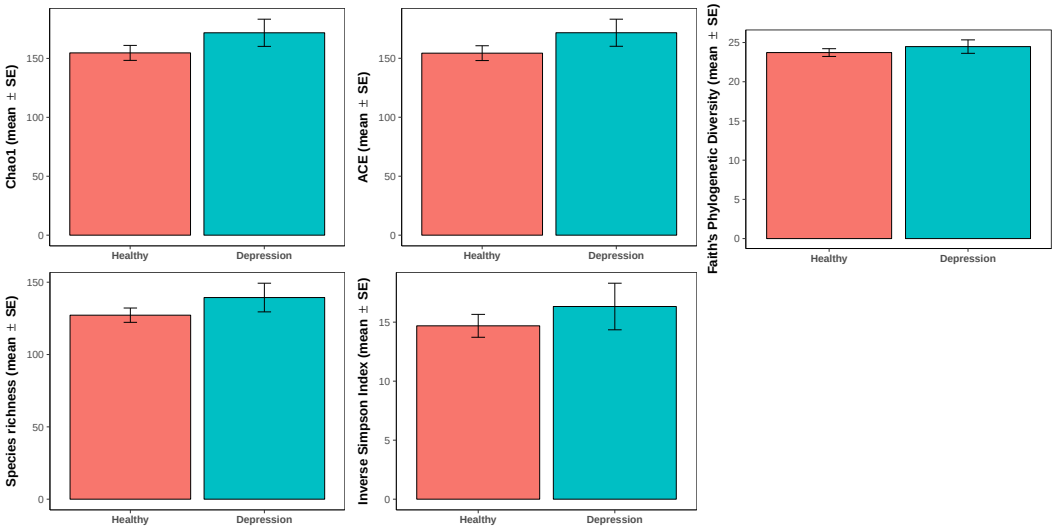


(a)

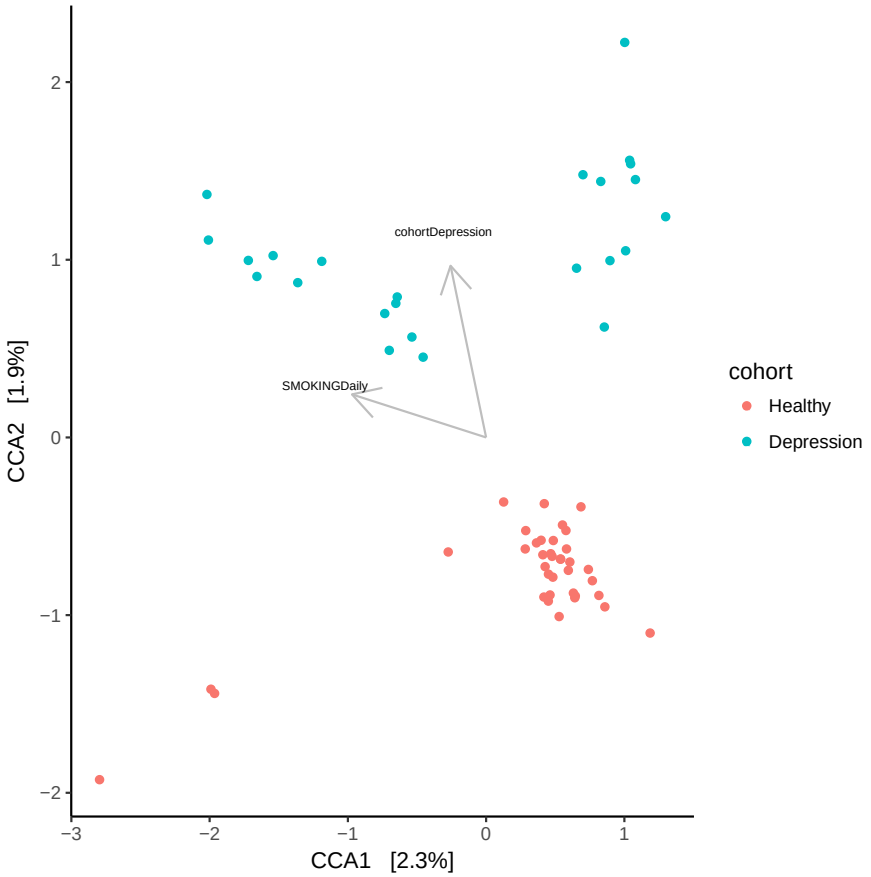


(b)

Figure 5.5: Visualisations of microbial community composition. (a) Phyla level (b) Family level

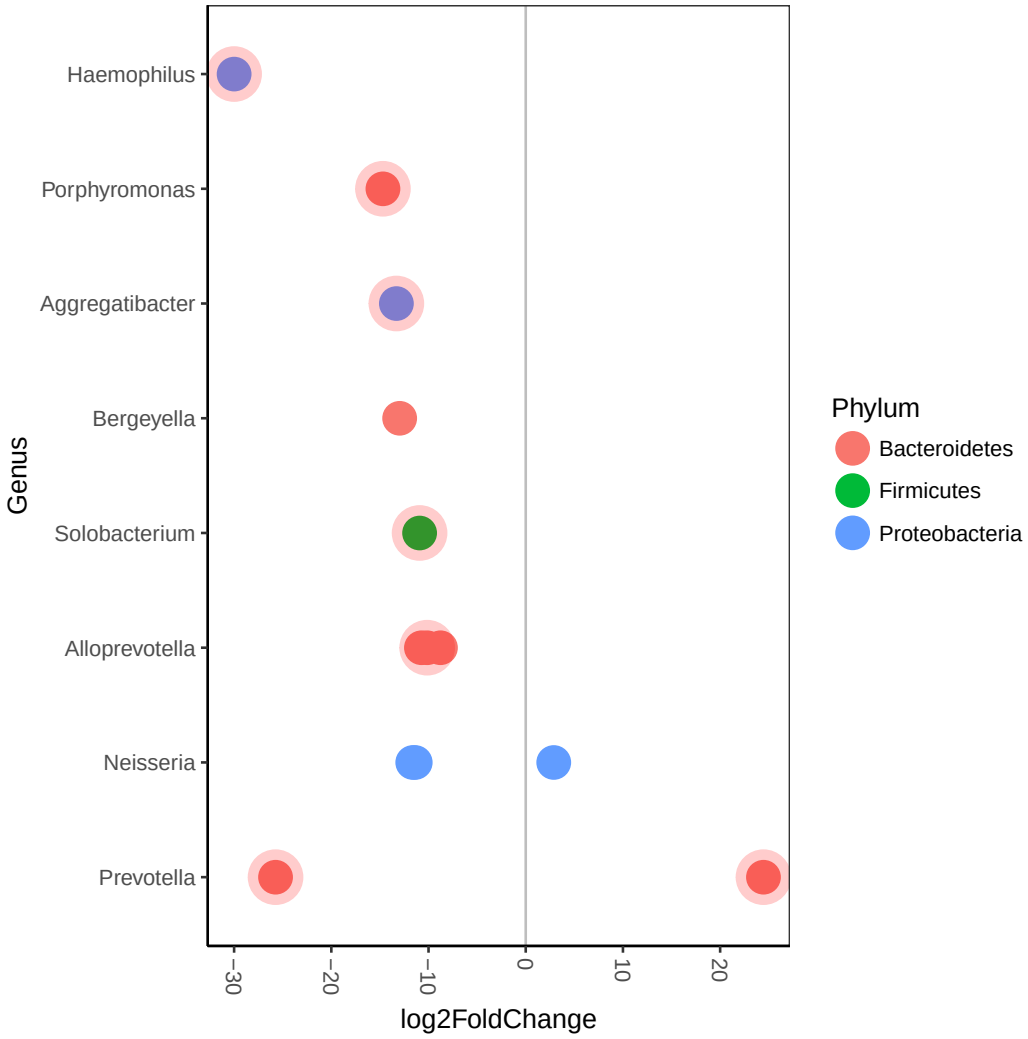


(a)

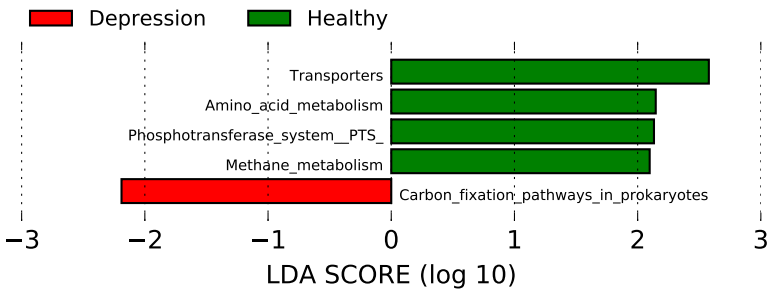


(b)

Figure 5.6: Visualisations of microbial community structure. (a) α-diversity (b) β-diversity



(a)



(b)

Figure 5.7: Visualisations of differential abundance: (a) Taxonomic (b) Functional.

Table 5.4: Amplicon sequence variants were further classified by matching against the Human Oral Microbiome Database.

Change	Classification	p	Notes
Increased	*Prevotella nigrescens*	<0.001	Associated with periodontitis (Stingu et al., 2013)
	Neisseria sicca[a]	0.023	Commensal with pathogenic potential (Johnson, 1983)
Decreased	*Alloprevotella rava*	0.031	
	Alloprevotella tannerae	<0.001	Associated with endodontic infections (Xia et al., 2000)
	Solobacterium moorei	<0.001	Associated with halitosis (Kazor et al., 2003) and endodontic infections (Munson et al., 2002)
	Neisseria subflava	<0.001	Commensal
	Aggregatibacter segnis	<0.001	Can cause infective endocarditis (Nørskov-Lauritsen, 2014)
	Porphyromonas endodontalis	<0.001	Pulpal pathogen (Mirucki et al., 2014)
	Prevotella nanceiensis	<0.001	First isolated from healthy subgingival oral biofilm
	Haemophilus parainfluenzae	<0.001	Can cause infective endocarditis (Nørskov-Lauritsen, 2014)

[a] Also matches *N. flava* and *N. mucosa* at equal identity

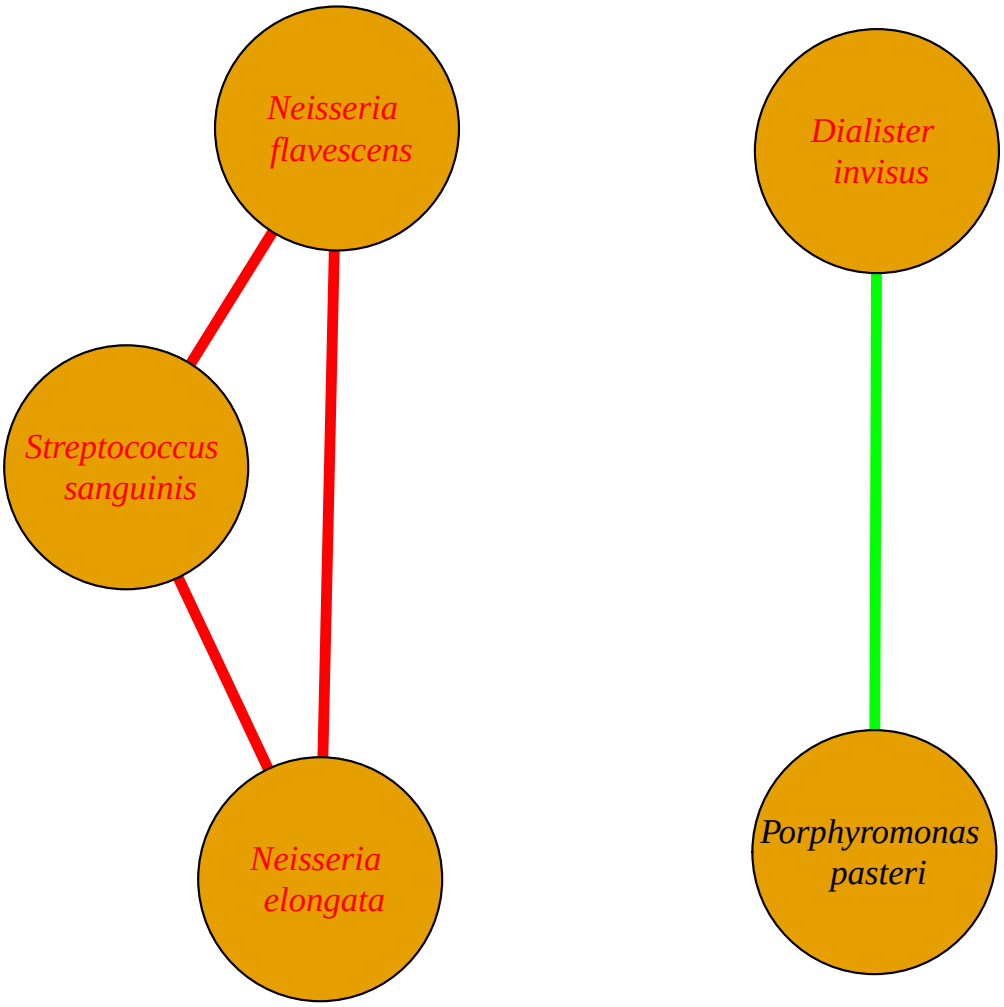


Figure 5.8: Network of statistically significant pairwise microbial interactions unique to the depressed cohort. Nodes are bacterial species, edges are interactions (green: positive co-occurrence, red: negative co-exclusion). Opportunistic pathogens are labelled in a red font.

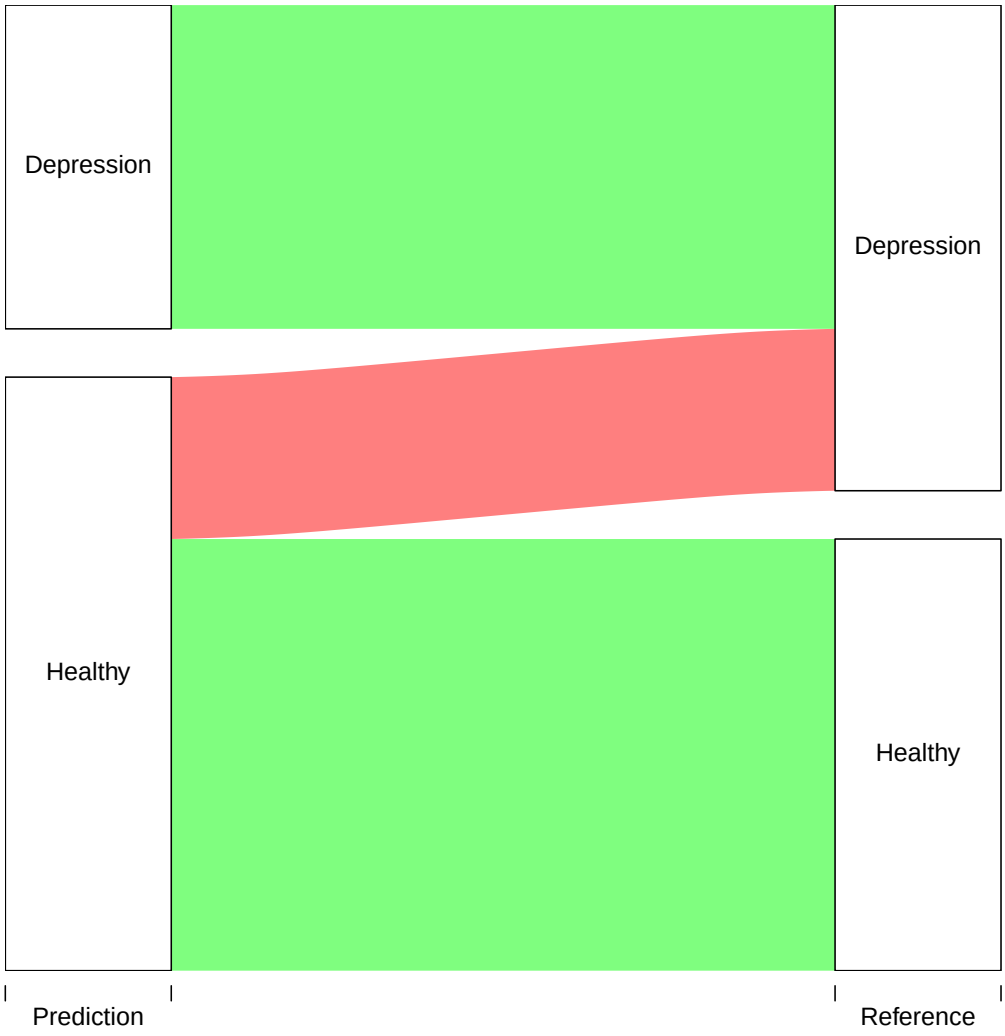


Figure 5.9: Alluvial diagram of classification performance

Model	Data transformation	Multimodal	Balanced accuracy
Random Forest	Proportion	✗	49.3%
	Rarefied	✗	37.5%
	Variance stabilised	✗	49.3%
Self Organising Map	None	✗	28.6%
	None	✓	83.3%

Figure 5.10: Classification performance

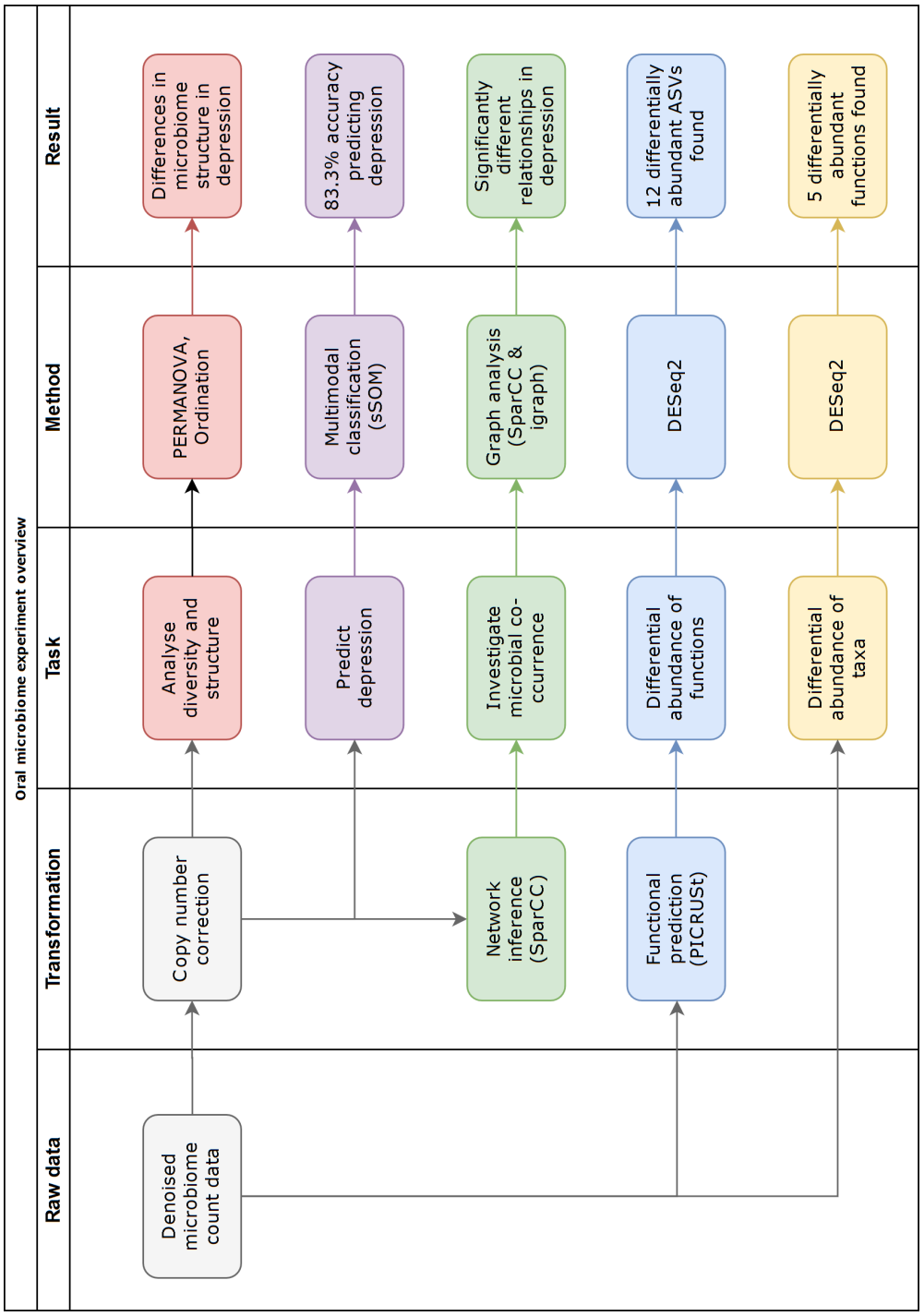


Figure 5.11: Methodology and results high-level overview

number of unique ASVs present in a sample), alpha diversity and beta diversity. To calculate alpha diversity simple estimators such as the Shannon diversity index and the Inverse Simpson diversity index were initially used, and then moved on to non-parametric species estimators such as the Abundance-based coverage estimator (ACE) and Chao1 which provide a measure of richness while compensating for differing sampling intensity across samples. Faith's Phylogenetic Diversity index was used to measure richness while incorporating data about phylogenetic relationships. Depression was not associated with significant changes to richness or alpha diversity for any of the tested metrics (see Figure 5.6). The Bray-Curtis dissimilarity statistic was used to measure beta diversity, and significant differences were found in the composition of the oral microbiota between depression and control groups (PERMANOVA: $p = 0.038$). Smoking was also associated with significant differences in composition of the oral microbiota (PERMANOVA: $p < 0.001$). Canonical Correspondence Analysis (CCA) was used to test and visualise the affect that statistically significant environmental variables had on the structure of the oral microbiota. The CCA biplot shows clear clustering between depressed and healthy cohorts into distinct groups, also, clustering between smokers and non-smokers (see Figure 5.5). The first canonical axis was negatively correlated with smoking daily, and the second canonical axis was positively correlated with depression and slightly positively correlated with smoking daily.

Differential abundance testing of prevalent ASVs found that 12 bacterial species were differentially abundant in the depressed cohort relative to the controls (Figure 5.7). From these sequence variants, 2 were significantly more abundant in depressed subjects, and 10 were significantly less abundant in depressed subjects. These differentially abundant sequences were matched against the Human Oral Microbiome Database (Chen et al., 2010) in order to gain an understanding of possible underlying mechanisms of action. The majority of identified organisms were opportunistic pathogens (i.e. under normal conditions they are commensal) or normal commensal organisms. Opportunistic pathogens that are decreased in depression have been associated with endodontic infections, halitosis, infective endocarditis, and pulpal pathogens (see Table 5.4). Opportunistic pathogens that have been found to be increased in depression include *P. nigrescens* and *N. sicca*. *P. nigrescens* is associated with periodontitis, while *N. sicca* is a commensal with pathogenic potential (Stingu et al., 2013; Johnson, 1983).

Inferred metagenome analysis with PICRUSt was used to identify possible functional changes in the oral microbiome of depressed subjects. These observed changes include a decrease in carbon fixation pathways and increases in amino acid metabolism, methane metabolism, transporters, and phosphotransferase system (see Figure 5.7).

An analysis of microbial interactions from estimated microbial co-occurrence

patterns (inferred with SparCC) found a group of statistically significant interactions unique to the depressed cohort (see Figure 5.8). A co-exclusion relationship was found between *Neisseria flavescens*, *Streptococcus sanguinis*, and *Neisseria elongata* in the depressed cohort; a co-presence relationship was found between *Dialister invisus* and *Porphyromonas pasteri*.

5.5 Multimodal classification of depression

To determine if the observed microbiome alterations were significant enough for stratification of depression status, a multimodal data-driven supervised learning classification algorithm called a sSOM was applied to the microbiome census data. The classification task was to distinguish between control and depressed subjects (two-class classification). Models were trained on 80% of the data. The generalisation ability of the models was validated by making predictions on unseen data (the remaining 20%). To measure the performance of the classification models a variety of metrics was used, including balanced accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). Balanced accuracy is defined as:

$$\text{Accuracy}_{\text{bal}} = \frac{\text{Specificity} + \text{Sensitivity}}{2} \tag{5.8}$$

Specificity is the true positive rate (i.e. the percentage of depressed subjects that are correctly identified as having depression). Sensitivity is the true negative rate (i.e. the percentage of control subjects that are correctly identified as not being depressed). A multimodal sSOM was able to predict depression with a balanced accuracy of 83.3% on unseen data (see Table 5.5 and Figure 5.9).

Table 5.5: Performance of classification algorithms applied for depression prediction from microbiome census data

Implementation	Accuracy	Sensitivity	Specificity	PPV	NPV
Wingfield et al.	82.35%	66.77%	100.00%	1.00	0.73
Naseribafrouei et al.	66.50%	86.00%	47.00%		
Zheng et al.[a]					

[a] Zheng et al. implemented feature ranking with a Random Forest and reported no classification metrics.

5.6 Summary

The intestinal microbiome has been implicated in the aetiology of depression in a variety of animal models. This includes the ability to transplant a "depressed microbiome" from a depressed individual into a control individual to induce depression (Zheng et al., 2016). Recently a limited amount of work has been done in human intestinal microbiome. However, the oral microbiome has not been analysed for links with depression to date. The oral microbiome presents a compelling target: salivary glands are surrounded by capillaries, and can absorb blood based biomarkers of disease, suggesting saliva fluid can contain vital disease information (Liu and Duan, 2012). Oral microbiome dysbiosis have been identified for diseases including pancreatic cancer (Fan et al., 2016), rheumatoid arthritis (Zhang et al., 2015), and neurological conditions such as Alzheimer's disease (Shoemark and Allen, 2015). Therefore an experiment to determine if any alterations are present in the oral microbiome of a depressed cohort is presented in this Chapter. In addition, most classification tasks that use microbiome census data are not explicitly multimodal. It is important to remember that the microbiome does not exist in isolation: it interacts constantly with its environment (the human host). Thus it would be valuable to analyse the microbiome in a holistic manner, by incorporating information from the human host. Therefore a multimodal sSOM was applied to the data gathered in the experiment to determine if depression status can be predicted from a saliva sample and to evaluate the effectiveness of a multimodal paradigm for classification. The data analysed in this chapter includes microbiome census data and basic clinical information such as smoking use and gender, although theoretically the approach could apply to multi-omic data (e.g. whole genome sequencing).

The experiment found a variety of alterations to the composition and structure of the oral microbiome in a depressed cohort using a range of ecological measures. The structure of the oral microbiome in the depressed cohort can be clearly clustered using CCA. Twelve ASVs were found to be differentially abundant in the depressed oral microbiome: the majority of which were opportunistic pathogens that were less abundant compared with control samples. Five inferred functional pathways were also found to be differentially abundant in the depressed cohort. In addition, a set of unique microbial interactions were found to be present in the depressed microbiome, the majority of which included interactions between opportunistic pathogens. The results directly implicate the oral microbiome in the pathogenesis of depression for the first time, and provide preliminary evidence that depression can be predicted from a saliva sample. These results have significance for both depression diagnosis and depression pathophysiology: the reliability of diagnostic criteria is inherently limited by the response of subjects to questionnaires, and the predictive power of responses can differ significantly across genders and age groups (Aben et al.,

2002). Furthermore, co-morbidities have been shown to decrease the performance of standard depression diagnostic criteria. For example, symptoms such as insomnia and loss of appetite can increase the risk of false positive depression diagnosis (Freedland et al., 1992; Fedoroff, Starkstein et al., 1991). An analysis of diagnostic criteria has shown that for screening purposes the Hamilton depression scale has a sensitivity of 78.1% and a specificity of 74.6% when the threshold was set to 12 (Aben et al., 2002). An approach that predicts depression from oral microbiome census data has the potential to alleviate some of the limitations listed above. However, future work will need to replicate these findings in an independent cohort to confirm the role that the oral microbiome plays in the microbiome-gut-brain axis and the predictive power of the oral microbiome for depression diagnosis.

Chapters 4 and 5 have extensively modelled microbiomes for both knowledge discovery and the prediction of disease. However, none of the models applied in either Chapter are transparent. An understanding of *what* is important in IBD and depression pathophysiology has been gained, but not *why* the models have arrived at their result. Transparency is key to enable trust in the output of models, which is critical for potential clinical applications. Additionally, predictive performance has been the focus of both chapters. Biologists are often interested in qualitatively describing phenomena, (i.e. not doing prediction). Chapter 6 will determine if rough set theory can be used to identify a subset of key organisms according to an experimental design and to describe why the model has arrived at its conclusion using transparent `IF-THEN` rules. Beginning with a proof of concept application to benchmark data the approach will be scaled up for the purpose of describing the oral and intestinal microbiome of a depressed cohort.

Publications arising from this work

The basis of this work is under preparation for submission:

Wingfield, B., C. Lapsley, S. Coleman, T. McGinnity, A. J. Bjourson and E. Murray (2019). 'Altered oral microbiota in a young adult cohort'. In: *Nature Scientific Reports*. Note: manuscript under preparation.

> I was just a chap who messed about in his lab.
>
> Frederick Sanger

6.1 Introduction

The ensemble feature selection (EFS) and Self-Organising Map (SOM) approaches applied in chapters 3 and 4 respectively were black box models. Although both approaches can lend insights into the decisions they have made (e.g. via feature selection) it is impossible to interpret the processes which led to the output of the model (i.e. why has this algorithm classified this example to class x?). One method of improving the process of transforming data into knowledge is by making models interpretable. The data driven SOM approach was applied in chapter 5 in part to overcome the properties of high-throughput sequencing data that violate the assumptions of standard models. Rough Set Theory (RST) is a data driven paradigm that provides many tools for interpretable data analysis. These tools include the concepts of discernibility, rough sets, minimal knowledge representations (reducts) that remove superfluous or irrelevant features, and rule induction. As was observed in chapter 5, evaluating the predictive power of the microbiome is only one aspect of a microbiome experiment. The goal of many experiments is to characterise (describe) the microbiome. The Computational Intelligence (CI) approaches that have been applied throughout this thesis have focused on classification (approximating a categorical variable from input data). However, describing events and patterns in data is a major part of data mining and knowledge discovery (see chapter 2.2). In addition, description is a valuable process for experimental scientists and many microbiome experiments focus solely on describing novel microbial environments. Therefore this chapter aims to demonstrate that RST can be used to characterise (describe) microbiomes. The application of RST provides a solution to an open research question regarding identifying an optimal normalisation technique for microbiome census data. Section 6.2 will provide a description of the RST concepts applied throughout this chapter, a brief summary regarding normalisation practices for microbiome census data, and a description of related work. Section 6.3 introduces the rough characterisation process and provides a demonstrative application of rough characterisation to a benchmark dataset that is widely used to evaluate the ability of algorithms to model microbiome census data. Section 6.4 applies the rough characterisation process to the oral microbiome

dataset described in chapter 5 and a publicly available gut microbiome dataset gathered from a depressed adult cohort to enable knowledge discovery and generate new insights into the microbiome-gut-brain axis.

6.2 Rough Set Theory

A brief explanation of RST was provided in chapter 3. This section provides a more detailed explanation of some key RST tools that are used to extract knowledge from data. In addition, the benefits of a data-driven paradigm with minimal priors for characterising microbiome census data, first introduced in chapter 4, are more fully explored.

6.2.1 Rationale

Many models make assumptions about input data. For example, a naïve Bayesian classifier assumes that feature values are independent of the value of any other feature for a given class. A person may be considered sick if they have a high temperature, a headache, and are shivering. A naïve Bayesian classifier assumes that each feature contributes independently to the probability that the person is sick. In this case, a naïve Bayesian classifier would be an inappropriate choice as there are probable correlations between high temperature and shivering. In contrast, the only assumption required in RST is that each object has an associated set of attributes used to describe the object, and that the data are a true and accurate reflection of reality (Jensen and Shen, 2008). The accuracy assumption can even be relaxed when applying fuzzy RST, which can incorporate different levels of uncertainty.

Data produced via high-throughput sequencing are extremely challenging to analyse. After initial quality control and clustering pre-processing steps (Kozich et al., 2013) (or alternatively denoising; (Callahan et al., 2016b)) microbiome census data are typically organised into large matrices where rows represent samples and columns represent counts of clustered sequence reads that constitute different types of bacteria (see Figure 6.1). The number of discrete sequence reads per sample (the sum of each row) can differ by orders of magnitude (see Table 6.1). This uneven sampling effort does not reflect true biological variation and is an artefact of the sequencing process. The uneven sampling effort will bias the estimates of bacterial abundance and should be normalised to allow fair comparison between samples. Normalisation procedures can also be used to mitigate other types of bias present in microbial community sequencing data introduced by sparsity (Paulson et al., 2013) or heteroscedasticity (McMurdie and Holmes, 2014). However, recommended normalisation procedures that aim to mitigate such complex problems are often

difficult for microbiologists to incorporate (e.g. applying a variance stabilising transformation based on Gamma-Poisson mixture models; (Love et al., 2014)) and can destroy the semantics of the original data.

A widely used normalisation strategy is to convert counts into relative abundances per sample (simple proportions). However, as relative abundances are constrained by an artificial limit (1) they represent compositional data. In addition, as the library size of a collection of samples is determined by the capacity of the sequencing instrument (e.g. an Illumina MiSeq DNA sequencer described in chapter 3 will create approximately 2×10^7 reads) even unnormalised sequencing data are compositional (Gloor and Reid, 2016). Compositional data have an arbitrary or non-informative sum (known as the constant-sum constraint problem; Aitchison and Egozcue, 2005). In recent years the microbiome research community has found that compositionality renders both univariate and multivariate data analysis methods invalid, increasing the popularity of compositional data analysis tools (Gloor and Reid, 2016). In addition, in 2014 it was observed that applying standard statistical tests to microbiome census data that violated the assumptions of the tests had rendered the results of very many microbiome experiments inadmissible (McMurdie and Holmes, 2014). The application of inappropriate statistical tests to large and complex biological datasets has caused similar problems in other fields, including neuroscience. In a famous example, a dead salmon was placed inside an functional Magnetic Resonance Imaging (fMRI) machine and shown photographs of humans in social situations. The salmon was asked what emotion the individuals in the photographs were experiencing while recordings of the salmon's brain were made. The application of inappropriate statistical techniques made areas of the salmon's brain appear to be active during questioning, raising serious questions regarding the reliability of fMRI studies (Bennett et al., 2011). The application of RST resolves a major problem associated with microbiome census data analysis (e.g. normalisation). As it is almost impossible to violate the assumptions of RST, researchers do not need to perform extensive checks before beginning their analysis of microbiome census data. Additionally, it is unlikely that microbiome researchers are fully aware of the advantages and disadvantages of each type of normalisation procedure, as normalisation is often automatically performed by software packages. RST makes redundant the requirement for more complex normalisation algorithms; the semantics of easily intuited relative abundance microbial sequencing data are maintained — aiding interpretation by domain experts and providing a possible solution to an open question in the microbiome research community regarding the choice of an optimal normalisation algorithm (which can differ depending on data and analysis task; Weiss et al., 2017).

$$
\begin{array}{cccccc}
\text{Microbe}_1 & \dots & \dots & \dots & \text{Microbe}_M & \\
\left[\begin{array}{ccccc}
0 & 9 & 8 & 0 & 7 \\
3 & 5 & 6 & 5 & 1 \\
1 & 0 & 0 & 3 & 2
\end{array}\right] & & & & & \begin{array}{l} \text{sample}_1 \\ \dots \\ \text{sample}_N \end{array}
\end{array}
$$

Figure 6.1: Example unnormalised community data matrix.

Table 6.1: Library size (row sum) summary statistics of Global Patterns (Caporaso et al., 2011) dataset.

Minimum	1^{st} Quartile	Median	3^{rd} Quartile	Maximum
5.9×10^4	5.7×10^5	1.1×10^6	1.5×10^6	2.4×10^6

6.2.2 Characterisation

Throughout this thesis and in the field of microbiome research generally machine learning and CI approaches have been applied to data for the purpose of predicting a categorical or numeric variable from a set of input data (classification). These experiments evaluate the performance of this process by measuring a series of predictive metrics. However, classification and regression are only a subset of data mining and knowledge discovery. Popular tasks for data mining and knowledge discovery include (Larose and Larose, 2014):

- Describing patterns and trends in data;
- Approximating a categorical target variable from a larger data set (classification);
- Approximating a numeric target variable from a larger data set (regression);
- Predicting future events (e.g. the share price of a company in 3 months);
- Clustering observations into similar groups;
- Identifying association rules (finding features that co-occur).

Describing pattern and trends in data is the most common aim of microbiome experiments. Many microbiome experiments aim to identify correlations between the characterised microbial community and disease. Only a small part of this process is concerned with evaluating predictive power: the process of determining elements of a microbial community that have predictive power is described as

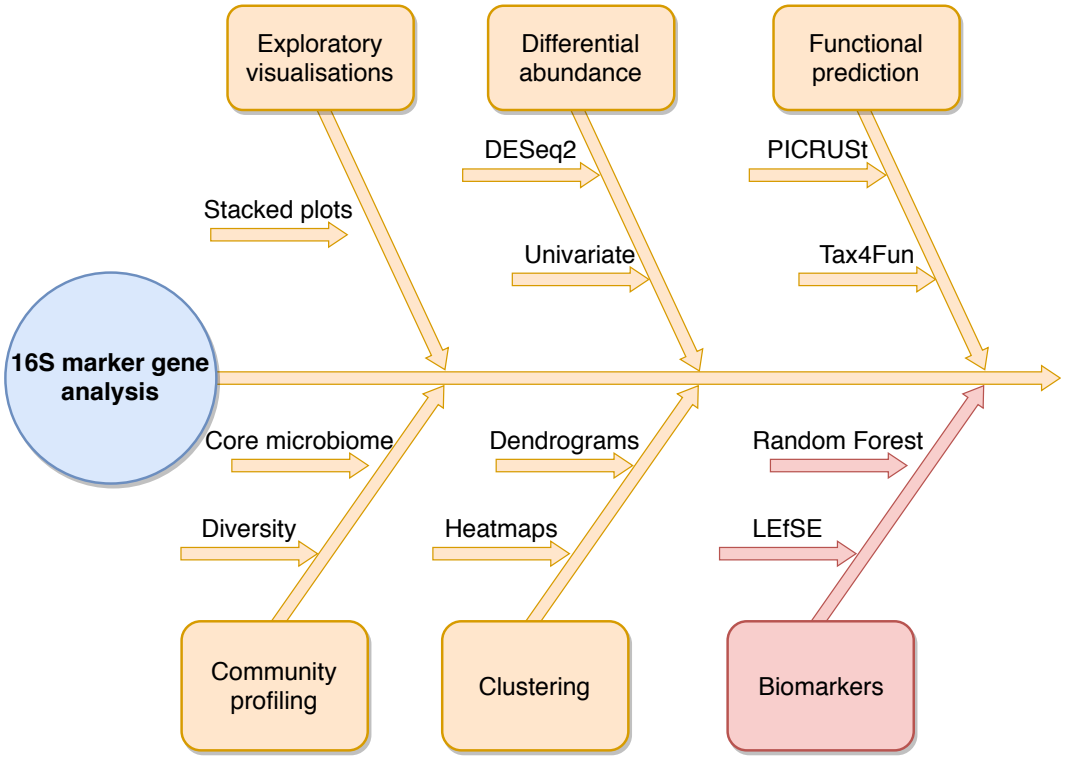


Figure 6.2: Evaluating predictive power (highlighted in red) is only a small part of 16S marker gene analysis. A much greater focus is placed on characterising microbial communities.

biological marker (biomarker) analysis by molecular biologists (see Figure 6.2). RST offers a suite of tools, described in section 6.2.3, that enables the comprehensive description of data. Data represented as a decision table can be stored in a concise form as a minimal knowledge representation, and the data can be transformed into knowledge via the generation of a set of `IF-THEN` rules.

6.2.3 Core concepts

A microbiota profile can be represented by a $M \times N$ decision table. The rows of a decision table correspond to the universe of discourse, X (Jensen and Shen, 2008):

$$X = \{x_1, x_2, \ldots, x_N\} \tag{6.1}$$

The columns of a decision table correspond to the set of features A (the set of

microbes) (Jensen and Shen, 2008):

$$A = \{a_1, a_2, \ldots, a_M\} \tag{6.2}$$

Decision table DT consists of a subset of condition attributes (input features, different microbial species) and decision attributes (class labels e.g. disease or healthy; $DT = C \cup D$). Each attribute has an associated value set, which represents the abundance of the microbial species:

$$V_a = \{v_1^a, v_2^a, \ldots, v_p^a\} \tag{6.3}$$

where $a \in A$. The value set must be discrete (continuous variables must be discretised). Although microbiome census data are discrete counts of sequences, they are typically converted into continuous variables by a normalisation process to mitigate uneven library size bias. Therefore relative abundance microbiome census data, which is used as input data throughout this chapter, must first be discretised. The maximal discernibility heuristic was used to discretise the microbiome census data throughout this paper (Bazan et al., 2000). Any condition or decision attribute subset $P \subseteq C$ or D can induce a partition in X (Petit et al., 2014):

$$X \xrightarrow{P} X(P) = \{X_1^P, \ldots, X_q^P\} \tag{6.4}$$

where X_l^P is the partition of X induced by P. The subsets (Petit et al., 2014):

$$X = X_a^P \cup \ldots \cup X_Q^P \tag{6.5}$$

correspond to the set of equivalence classes, called indiscernibility classes in RST. Discernibility is the core concept of RST: if $(x, y) \in \text{IND}(P)$ (where $\text{IND}(P)$ is the indiscernibility relation induced by attribute subset P) then x and y are indiscernible by attributes from P. For example, if two bacterial species have the same abundance in both healthy and sick subjects, then using only the abundance of the bacterial species it is impossible to discern between the two subjects. In RST a set is approximated by two sets known as the lower and upper approximations (Jensen and Shen, 2008):

$$\underline{P}S = \{x : [x]_P \subseteq S\} \tag{6.6}$$

$$\bar{P}S = \{x : [x]_P \cap S \neq \emptyset\} \tag{6.7}$$

where $S \subseteq X$ and $[x]_P$ are the equivalence classes of the P-indiscernibility relation. The tuple $\langle \underline{P}S, \bar{P}S \rangle$ is known as a rough set. P and Q are sets of attributes inducing equivalence relations over U. The region between the upper and lower approximation sets is called the boundary region. The boundary region represents

the set of objects that can possibly be predicted to be from a specific decision class (non-deterministic; see Figure 6.3). For example, the relative abundance of a set of species may include sick or healthy samples, and from the relative abundance data it is impossible to distinguish between the two (Jensen and Shen, 2008):

$$\mathrm{BND}_P(Q) = \bigcup_{X \in U/Q} \bar{P}S - \bigcup_{X \in U/Q} \underline{P}S \tag{6.8}$$

The positive region, in which objects can be predicted to belong to a decision class with certainty, is given by:

$$\mathrm{POS}_P(Q) = \bigcup_{X \in U/Q} \underline{P}Y \tag{6.9}$$

$$\tag{6.10}$$

The negative region represents the set of objects that cannot be predicted to a decision class (e.g. are definitely healthy):

$$\mathrm{NEG}_P(Q) = X - \bigcup_{Y \in X/Q} \bar{P}Y \tag{6.11}$$

Attributes that cannot be removed without changing the partitioning of objects amongst the indiscernibility relations are indispensable. A minimal set of indispensable condition attributes is known as a reduct.

Minimal knowledge representations

To identify important features the dependence and significance of features must first be measured. A set of features Q can be said to depend on a set of features P if all feature values from Q are determined only by feature values from P. For $(P, Q \subset A), Q$ depends on P (given as $\gamma_p(Q)$) to degree k $(0 \leq k \leq 1)$ if:

$$k = \gamma_p(Q) = \frac{|\mathrm{POS}|}{|X|} \tag{6.12}$$

where $|S|$ gives the cardinality of set S. When $\gamma_p(Q) = 1$ Q depends completely on P. Measuring feature dependence is important because by evaluating the change of feature dependence after removing a feature, feature significance can be computed. The significance of feature $x \in P$ upon Q is given by (Jensen and Shen, 2008):

$$\sigma p(Q, a) = \gamma p(Q) - \gamma p -_a (Q) \tag{6.13}$$

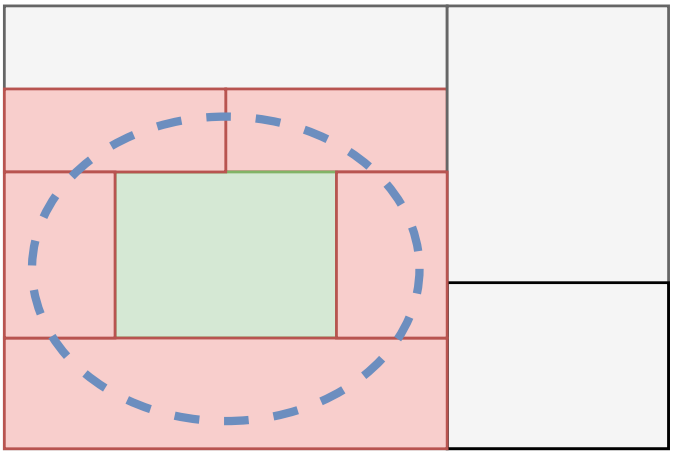


Figure 6.3: Rough set example. The universe of discourse is partitioned into 9 indiscernibility classes by a set of attributes. The blue line represents the set being approximated (e.g. sick subjects). The green section is the lower approximation, and the red sections are the upper approximations of the rough set. In the complement of the upper appromimation (grey) it is certain that no objects in the rough set will be present (e.g. a healthy subject could be in the grey section).

A feature with a significance greater than 0 is indispensable. It is often useful to calculate a minimal form of a decision table (a minimal knowledge representation), known as a reduct. A reduct is defined as a minimal subset R of an initial feature set C, that if given a set of features D, $\gamma_R(D) = \gamma_C(D)$. R is considered a minimal subset if $\gamma_{R-\{a\}}(D) \neq \gamma_R(D)$ for all $a \in R$. In a minimal reduct no features can be removed without affecting the dependency degree. However, this definition shows that minimal reducts are not global. A decision table may have many reduct sets. The collection of all reduct sets is given by (Jensen and Shen, 2008):

$$R_{\text{all}} = \{X | X \in C, \gamma_X(D) = \gamma_C(D); \gamma_{X-\{a\}} \neq \gamma_X(D), \forall a \in X\} \tag{6.14}$$

Discernibility Matrix

Discernibility matrices are used to identify reducts and to induce rules. A discernibility matrix of a decision table $(U, C \cup D)$ is a symmetric $|U| \times |U|$ matrix, where $|U|$ gives the cardinality the decision table U. Each element of the discernibility matrix is defined by:

$$c_{i,j} = \{a \in C | a(x_i) \neq a(x_j)\}, \qquad i, j = 1, \ldots, |U| \tag{6.15}$$

$c_{i,j}$ contains features that differ between instances i and j.

Discretisation

Table 6.2: Discretisation strategies and implementations

	Supervised	Unsupervised
Global	ChiMerge (Kerber, 1992) OneRule (Holte, 1993) Global discernibility heuristic (Bazan et al., 2000)	Equal width interval (Dougherty et al., 1995) Quantiles (Dougherty et al., 1995)
Local	Local discernibility heuristic (Bazan et al., 2000)	k-means clustering (Dougherty et al., 1995)

Attributes with real values must be discretised before RST can be applied. A variety of discretisation strategies are available when working with rough sets, and care must be taken when choosing a discretisation method as the process guarantees information loss. Discretisation strategies can be supervised or unsupervised, and can consider subsets of training samples (local) or the entire instance space (global; see Table 6.2). The simplest discretisation approaches are unsupervised global techniques. For example, equal width binning sorts attribute values and divides the range of observed values into k equally sized intervals. Let x be a feature value bounded by $x_{\min}$ and $x_{\max}$ (Dougherty et al., 1995):

$$\delta = \frac{x_{\max} - x_{\min}}{k} \tag{6.16}$$

where δ gives the interval width, and k gives the chosen number of intervals. Thresholds (interval boundaries) are given by $x_{\min} + i\delta$ where $i = 1, \ldots, k-1$. This approach is applied to each continuous attribute independently and does not incorporate class information.

Although all discretisation methods result in data loss it is likely that unsupervised discretisation methods lose more information compared with supervised discretisation methods (Dougherty et al., 1995). This is because certain attribute values may be associated with a particular class. Unsupervised approaches can cluster attribute values associated with different classes into the same bin. Therefore incorporating class information can help to identify an optimal discretisation strategy. The problem of identifying an optimal set of cuts is extremely computationally complex. Therefore, heuristics are often applied to simplify computation. A heuristic based on rough set theory called the Maximal Discernibility (MD) heuristic (Bazan et al., 2000) is a widely implemented method for supervised discretisation. Let $A = (U, A \cup d)$ be a decision table. Attribute $a \in A$ defines a

Algorithm 6.1 Maximum discernibility discretisation (Bazan et al., 2000)

Input: Decision table A
Output: Set of cuts D
$D = \emptyset$, $C_A =$ initial set of cuts on A
$L = \{(x, y) \in U \times U : d(x) \neq d(y)\}$
while $L \neq \emptyset$ **do**
Choose cut $c_{\max} \in C_A$ that discerns the largest # of instance pairs in L
Input $C_{\max}$ into D, remove from C_A
Remove all instance pairs from L discerned by $c_{\max}$
end while

sequence $v_1^a, \ldots, v_{n_a}^a$, where $\{v_1^a, \ldots, v_{n_a}^a\} = \{a(x) : x \in U\}$ and $n_a \leq n$. The set of all interval cuts on a is given by:

$$C_a = \left\{\left(a, \frac{v_1^a + v_2^a}{2}\right), \ldots, \left(a, \frac{v_{n_a-1}^a + v_{n_a}^a}{2}\right)\right\} \tag{6.17}$$

and the set of all interval cuts on all attributes A:

$$C_A = \bigcup_{a \in A} C_a \tag{6.18}$$

The MD heuristic aims to discern the largest number of pairs of objects (see Algorithm 6.1). The local method computes the quality of cut from a subset of instances, while the global method computes the quality of cut on the whole instance set. The local method was used throughout this chapter for discretisation because the global strategy produced fewer cuts (Bazan et al., 2000). This caused problems for microbiome census data as there were too many features with a single interval, possibly due to the high variability present across different microbiomes.

6.2.4 Current rough set applications to microbiome census data

As far as can be ascertained, there have been no previous attempts to model microbiota profiles using RST described in the literature. However, aspects of RST have been implemented for bioinformatics applications to the wider field of metagenomics. The metagenome is defined as the collection of genomes and genes from the members of a microbiota (Marchesi and Ravel, 2015). Metagenomic analysis requires sequencing all of the DNA present in an environmental sample using shotgun sequencing (in contrast with 16S rRNA marker gene surveys). Classifying short DNA fragments into a phylogeny or taxonomy (e.g. bacterial species) is a standard step in metagenomic workflows. Good quality reference databases exist

for well characterised environments such as the human gut (Quast et al., 2012), and tools such as the Ribosomal Database Project's naïve Bayesian classifier can be trained on reference databases and perform well on novel sequences (Wang et al., 2007a). In poorly characterised environments such as macroscopic bacterial accumulations in Guerrero Negro, Mexico up to 85% of DNA sequence fragments were previously undescribed in reference databases (Ley et al., 2006). Standard pattern matching tools will not work on data gathered from poorly characterised environments. To overcome this challenge some classification algorithms rely on generating a digital signature from DNA fragment characteristics. A common characteristic is the K-mer frequency (DNA words of length k), which is often used to estimate the complexity of a genome (Chor et al., 2009). However, as k increases the number of features used for classification also quickly increases. RST has been applied to remove superfluous K-mers and to improve DNA fragment classification compared with standard bioinformatics tools (Jian et al., 2015). A rough reduction method based on Particle Swarm Optimisation has also been applied to the same problem (Jian et al., 2016).

RST has been used to predict the presence of operons in metagenomic data. An operon is defined as a functioning unit of genomic DNA containing a cluster of genes under the control of a single promoter (Ralston, 2008). Bacteria can adapt to new environments extremely quickly (e.g. antibiotic resistance), partly because clusters of genes can be quickly switched on or off depending on environmental conditions (Ralston, 2008). A decision tree classifier based on the Variable Precision Rough Set Model (VPRSM) was applied to genomic data from *Escherichia coli* to identify if a gene belongs to an operon (Zaidi and Zhang, 2016). The VPRSM had an accuracy of 89.4% using five features: maximum distance, minimum distance, direction, cluster of orthologous groups, and gene order conservation. The use of a decision tree meant that the decisions of the classifier were easy to interpret and could be validated by domain experts.

Both of the described approaches did not implement knowledge discovery from highly dimensional data (they focused on classification performance from a set of summary statistic features). Therefore these approaches were not applied to microbiome census data.

Rule-based systems are not widely applied to microbiome census data, but they offer a number of advantages compared with black-box machine learning algorithms. Transparent `IF-THEN` rules enable the underlying mathematics of RST to be codified into linguistic variables that can be easily understood and transmitted to domain experts (e.g. microbiologists) who lack an understanding of RST. If microbiologists are interested in the predictive power of certain bacterial groups (biomarker analysis), machine learning algorithms are typically applied to microbial sequencing data. The most popular machine learning algorithms are typically black

boxes; for example, ensembles of de-correlated decision trees (Random Forests; (Qi, 2012)) are often recommended for their ease of implementation and performance advantages (Statnikov et al., 2013). Although theoretically the output of decision trees can be interpreted by analysing the structure of a tree, understanding and explaining the combined output of hundreds of trees trained on random feature subsets in an ensemble is an almost impossible task that is rarely attempted (although feature ranks are commonly reported). In contrast, the combined process of generating reducts and inducing rules from essential features offers an elegant way to both describe a microbial community and to determine the biomarker potential of bacterial groups when additional validation data are available. A transparent descriptive or predictive process could aid the understanding and dissemination of important results throughout the microbiome research community and help to improve the reproducibility of research.

6.3 Rough set characterisation of a standard benchmark dataset

To demonstrate the viability of applying rough set theory to microbiome census data, a standard benchmark dataset was chosen to determine the performance of the approach. This initial demonstration uses the Global Patterns dataset (Caporaso et al., 2011) that is distributed as part of the microbiome census data analysis `R 3.4.3` package `phyloseq` (McMurdie and Holmes, 2013). The Global Patterns dataset consists of 25 environmental samples collected across 9 different environments: standard mock community controls, freshwater, creek freshwater, ocean, sediment, soil, human skin, human tongue, and human faeces. The objective of the study was to demonstrate the feasibility of using 16S rRNA gene sequencing to accurately capture microbial diversity. In the microbiome research community the Global Patterns dataset is widely used to benchmark new algorithms or tools (McMurdie and Holmes, 2014; Weiss et al., 2017).The ability of RST to model microbiota profiles was evaluated by testing classification performance on three standard tasks, in ascending order of difficulty:

1. Classify microbial communities from vastly different environments (soil or ocean)

2. Classify microbial communities from closely related environments (lake or creek freshwater)

3. Classify microbial communities from different areas of the human body (tongue, skin, or faeces)

A single reduct was generated from each decision table with the QuickReduct algorithm (Shen and Chouchoulas, 2000) implemented in the `RoughSets R 3.4.3` package (Riza et al., 2014). Due to the number of features in the dataset (4624 types of bacteria in the denoised Global Patterns dataset) it was infeasible to compute all reducts, and a single reduct is useful for this demonstration. The classification performance of the partition in X induced by the set of reduct attributes A_k was evaluated with two measures (Petit et al., 2014):

$$\text{Accuracy}[X(A_k)] = \frac{\sum_{L=1}^{Q} \text{Card}(\underline{A}_k X_L^{A_k})}{\sum_{L=1}^{Q} \text{Card}(\bar{A}_k X_L^{A_k})} \tag{6.19}$$

$$\text{Quality}[X(A_k)] = \frac{\sum_{L=1}^{Q} \text{Card}(\underline{A}_k X_L^{A_k})}{\text{Card}(X)} \tag{6.20}$$

Where Card is cardinality, which represents the number of elements in a set, and L is the total number of upper ($\bar{A}_k X_L^{A_k}$) and lower-approximation ($\underline{A}_k X_L^{A_k}$) set tuples. Accuracy represents the ratio of the size of all lower-approximation sets to the size of all upper-approximation sets ($0 \leq \text{Accuracy}[X(A_k)] \leq 1$). If the family of lower approximation sets is an empty set (i.e. no objects can be said to be certainly predicted) then accuracy is zero. Quality represents the ratio of all objects in the family of lower approximation sets to the total number of objects in the universe of discourse ($0 \leq \text{Quality}[X(A_k)] \leq 1$). It is important to note that classification accuracy and quality are not tested on independent validation data. `IF-THEN` decision rules were generated from the indiscernibility classes defined by the reduct attributes using the `RoughSets` package. The descriptive strength of the rules was evaluated by measuring the support each rule has; support is defined as the number of instances in the dataset that are concordant with the rule. The rules were rationalised to biological phenomena after a thorough literature review. This process involved searching the literature to identify the effects that bacterial species are known to have on humans, and the role they are thought to play as part of the larger microbiome (e.g. butyrate synthesis).

6.3.1 Results of rough set characterisation

Decision tables were created for each of the three classification tasks. The first decision table had 4304 conditional attributes, representing the abundance of different bacterial groups, and 6 samples (3 ocean samples and 3 soil samples). The second decision table had 3893 conditional attributes, and 5 samples (2 lake freshwater and 3 creek freshwater samples). The third decision table had 3878 conditional attributes, and 9 samples (3 skin samples, 3 faecal samples, and 3 tongue samples). A single reduct was generated first for each of the three decision

tables to simplify analysis. For the soil classification task a single feature was present in the reduct: the bacterial Family *Cenarchaeaceae*. The classification ability of the reduct rough set was tested using the accuracy and quality measures described in Section 6.3 (see Table 6.3). The lower approximation set contained all of the samples for each sample type so the accuracy and quality of classification was 1. The freshwater classification task had a single bacterial species present in the reduct: *Nitrososphaera SCA1145*. The human body site classification task also had a single bacterial species present in the reduct: *Propionibacterium acnes*. For both the freshwater and human body classification tasks the lower approximation set also contained all of the samples for each sample type so the accuracy and quality of classification was 1 (creating a crisp set).

Rules were then induced from the D-reduct for each decision table. It is important to note that the generated rules are descriptive and not predictive. By generating a set of `IF-THEN` rules, data can be converted into knowledge. Testing the predictive capability of descriptive rules requires an independent set of validation data. Due to the small size of the Global Patterns dataset it was infeasible to do this. The strength of descriptive rules can be measured by the support that each rule has (the number of instances in the dataset that are concordant with the rule). The first classification task generated three rules for two classes regarding the bacterial Family *Cenarchaeaceae*:

$$\texttt{IF } \textit{Cenarchaeaceae } 0 \texttt{ THEN Soil} \tag{6.21}$$

$$\texttt{IF } \textit{Cenarchaeaceae } (0, 2.81 \times 10^{-6}] \texttt{ THEN Ocean} \tag{6.22}$$

$$\texttt{IF } \textit{Cenarchaeaceae } (2.81 \times 10^{-6}, 1] \texttt{ THEN Ocean} \tag{6.23}$$

It is important to note none of the generated rules have been optimised, and could be simplified by merging the second and third rules. This is simple to do manually for small rule sets but is a complex topic for larger rule sets. Rule optimisation was outside the scope of this application of RST to microbiome census data, which is inteded soley to demonstrate the validity of the technique in this context. The second classification task also generated three rules for two classes regarding the bacterial species *N. SCA1145*:

$$\texttt{IF } \textit{N. SCA1145 } (0, 1.41 \times 10^{-7}] \texttt{ THEN Lake} \tag{6.24}$$

$$\texttt{IF } \textit{N. SCA1145 } (1.41 \times 10^{-7}, 4.56 \times 10^{-7}] \texttt{ THEN Creek} \tag{6.25}$$

$$\texttt{IF } \textit{N. SCA1145 } (4.56 \times 10^{-7}, 1] \texttt{ THEN Creek} \tag{6.26}$$

The third classification task generated three rules for three classes regarding the bacterial species *P. acnes*:

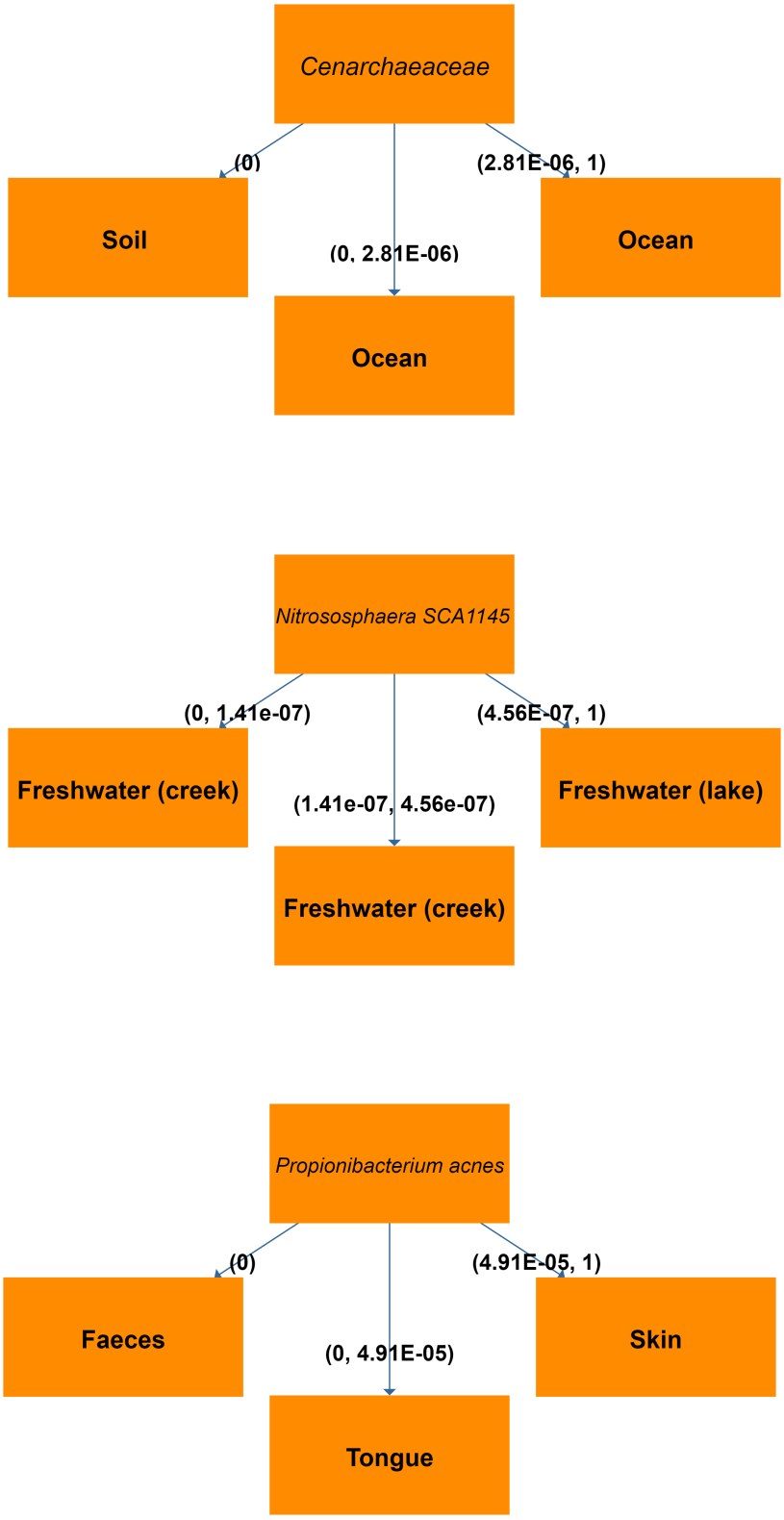


Figure 6.4: Induced rules for the classification of sample types.

IF *P. acnes* 0 THEN Faeces (6.27)

IF *P. acnes* $(0, 4.91 \times 10^{-5}]$ THEN Tongue (6.28)

IF *P. acnes* $(4.91 \times 10^{-5}, 1]$ THEN Skin (6.29)

Rationalising the generated rules to biological phenomena or processes is straightforward as the semantics of the original data were not destroyed by complex normalisation approaches: *Cenarchaeaceae* consist of marine-based anaerobic thermophile archaeons that metabolise sulphur (Preston et al., 1996), and its absence from soil samples is logical (see Equations 6.21–6.23). *N. SCA1145* is an ammonia oxidising archaeon assemblage (candidate species) (Swanson and Sliwinski, 2013). The difference in abundance could be related to the amount of nitrogen available in the respective environments (creek versus lake freshwater, see Equations 6.24–6.26). The most interesting pattern revealed by the induced rules relates to *P. acnes* (see Equations 6.27–6.29). Typically *P. acnes* is a commensal member of the skin microbiome, but it can act as a pro-inflammatory opportunistic pathogen, causing acne (Perry and Lambert, 2011). Its pattern of abundance matches descriptions in the literature: most prevalent on skin, but capable of colonising other areas of the body including the tongue and large intestine (Perry and Lambert, 2011). The absence of *P. acnes* in stool samples could be related to the sensitivity of the sequencing process or the low sample size of the cohort (*P. acnes* is not a major member of the gut microbiome, the most complex of all human microbiomes). Alternatively, as faeces are not a perfect proxy for the large intestine *P. acnes* may be present in the large intestine but be undetectable in stool. The descriptions revealed by the induced rules shows that the RST approach has identified biologically plausible processes that underpin the stratification of samples.

The biggest limitation to the described approach is that only a single reduct is considered, and that the dataset is very small. Due to the high dimensionality of the data sets (approximately 4000 features in each decision table) it was not computationally feasible to identify all possible reducts using the `RoughSets` package. This is a key challenge for scaling this approach for knowledge discovery in health applications. In Section 6.4 the approach is extended to characterise two larger microbiome datasets gathered from depressed adults. The challenge is overcome via the application of the Java library `rseslib` (Bazan and Szczuka, 2000), which is considerably quicker than the `RoughSets` `R` package. `R` is known to be considerably slower than other popular programming languages (Wickham, 2014). `R` is composed of a mix of `C`, `fortran`, and `R`, and the small development team that maintain `R` prioritise stability over rewriting large portions of the code base to improve speed (which would involve breaking compatibility with older versions of `R` in the process; Wickham, 2014).

6.4 Characterising oral and gut microbiomes in depressed adults

6.4.1 Implementation of rough set characterisation

The RST approach described in Section 6.3 was applied to two datasets to enable knowledge discovery:

1. a publicly available gut microbiome depression dataset (Jiang et al., 2015)
2. the oral microbiome depression dataset used in chapter 5 (all subjects that smoked were removed to eliminate confounders)

The ability of RST to model these larger datasets and generate novel insights about microbiomes in depressed adults was evaluated with three tasks:

1. Characterise the depressed and control gut microbiomes;
2. Characterise the gut microbiome present in subjects in remission (in recovery after a depression diagnosis);
3. Characterise the depressed and control oral microbiomes.

In order to apply the RST characterisation to these larger and more complex data the `rseslib` Java library (Bazan and Szczuka, 2000) was used (instead of the `RoughSets R` library from the previous section). Two decision tables were created — one for each microbiome — and all local reducts were computed. The characterisation performance of RST was evaluated using the same accuracy and quality measures described in Equations 6.19 and 6.20. The local maximum discernibility discretisation process generated two intervals for the oral microbiome data, and three intervals for the gut microbiome data. The first interval began at 0 abundance, and the last interval was bounded by 1. To simplify visualisations and analysis these intervals were given the labels low (beginning at 0), medium, and high (bounded by 1).

6.4.2 Results of rough set characterisation

Decision tables were created for each of the two classification tasks. The first decision table (gut microbiome) had approximately 2900 conditional attributes, and 59 samples (30 control, 29 depressed). The second decision table (oral microbiome) had 4400 conditional attributes, and 67 samples (38 control, 29 depressed). All local reducts were computed for both decision tables. The characterisation ability of

Table 6.3: Classification metrics

Classification task	Accuracy	Quality
Oral microbiome	1	1
Gut microbiome	1	1

the reduct rough set was tested using the accuracy and quality measures described in Section 6.3 (see Table 6.3). The lower approximation set contained all of the samples for each sample type so the accuracy and quality of classification was 1. The gut microbiome characterisation task contained 12 amplicon sequence variants (ASVs) covering the bacterial genera *Bacteroides*, *Prevotella*, *Anaerostipes*, *Phascolarctobacterium* and *Odoribacter*. One of the features could not be mapped to a specific genus, and represented the bacterial Family *Ruminococcaceae*. The oral microbiome characterisation task contained 6 features that included the bacterial genera *Selomonas*, *Streptococcus*, *Granulicatella*, *Prevotella* (including *Prevotella* and *P. melaninogenica*), and *Haemophilus*. For both the gut and oral microbiome characterisation tasks the lower approximation set contained all of the samples for both classes (depression and control), so the accuracy and quality of characterisation was 1 (creating a crisp set). This demonstrates that RST can perfectly discern between control and depressed samples. The next step of characterisation is to describe the alterations identified by RST using `IF-THEN` rules.

Different 16S sequence variants can belong to the same genus, and it is important to note that multiple different ASVs were matched to the same genus. When this has occured in the microbiome census data, a number has been appended to the name of the genus to note that although the sequence variant has a shared genus it is in fact different to other 16S sequence variants in the same genus. For example, seven ASVs were in the genus *Bacteroides*, beginning with the ASV labelled *Bacteroides* and ending with the ASV *Bacteroides 6*.

6.4.3 Discussion

More complex rules were generated to characterise both the gut and oral microbiome characterisation tasks (see Figure 6.4 and Tables 6.4– 6.6). For both microbiomes the abundance of bacterial taxa was defined as being low or high in relation to the discretised bins to aid comprehension. At least one of the generated rules will apply to all of the subjects for each dataset, as the rough set characterisation approach had perfect accuracy and quality of characterisation. In the gut microbiome three rules were induced to characterise control samples, and four rules to characterise depressed samples. Control samples are characterised by low abundance of the bacterial genera subset, whilst depressed samples are characterised by a mixture

Table 6.4: Rules that characterise the gut microbiome for depressed and control cohorts

Rule		Antecedent		Consequent
1	IF	*Bacteroides* (3) low AND *Bacteroides* (6) low AND *Prevotella* low AND *Anaerostipes* low AND *Ruminococcaceae* low	THEN	control
2	IF	*Bacteroides* (3) low AND *Bacteroides* (4) low AND *Bacteroides* (6) low AND *Anaerostipes* low AND *Ruminococcaceae* low	THEN	control
3	IF	*Bacteroides* low AND *Bacteroides* (1) low AND *Bacteroides* (4) low AND *Bacteroides* (6) low AND *Ruminococcaceae* low AND *Odoribacter* low AND *Anaerostipes* low	THEN	control
1	IF	*Bacteroides* (1) low AND *Bacteroides* (6) high AND *Phascolarctobacterium* low AND *Ruminococcaceae* low	THEN	depressed
2	IF	*Bacteroides* (3) low AND *Ruminococcaceae* low AND *Bacteroides* (6) high	THEN	depressed
3	IF	*Bacteroides* (1) low AND *Bacteroides* (4) low AND *Bacteroides* (6) high AND *Ruminococcaceae* low	THEN	depressed
4	IF	*Alistipes* low AND *Odoribacter* high	THEN	depressed

Table 6.5: Rules that characterise the gut microbiome for the remission cohort

Rule		Antecedent		Consequent
1	IF	*Bacteroides* (3) medium AND *Bacteroides* (6) low AND *Odoribacter* low AND *Oscillospira* low AND *Anaerostipes* low	THEN	remission
2	IF	*Bacteroides* (3) medium AND *Bacteroides* (4) low AND *Phascolarctobacterium* low AND *Oscillospira* low	THEN	remission
3	IF	*Bacteroides* low AND *Bacteroides* (3) medium AND *Bacteroides* (6) low AND *Odoribacter* low	THEN	remission
4	IF	*Bacteroides* low AND *Bacteroides* (1) high AND *Odoribacter* low	THEN	remission
5	IF	*Bacteroides* low AND *Bacteroides* (1) high AND *Bacteroides* (3) medium	THEN	remission
6	IF	*Bacteroides* (3) medium AND *Phascolarctobacterium* low AND *Odoribacter* low AND *Oscillospira* low	THEN	remission
7	IF	*Bacteroides* (3) medium AND *Bacteroides* (6) low AND *Ruminococcaceae* low AND *Odoribacter* low	THEN	remission
8	IF	*Bacteroides* (1) high AND *Bacteroides* (3) low AND *Odoribacter* high AND *Oscillospira* medium	THEN	remission

Table 6.6: Rules that characterise the oral microbiome for depressed and control cohorts

Rule		Antecedent		Consequent
1	IF	*Selenomonas* high AND *Granulicatella elegans* low AND *Prevotella* low	THEN	control
2	IF	*Streptococcus* low AND *Prevotella melaninogenica* low	THEN	control
3	IF	*Haemophilus parainfluenzae* high AND *Granulicatella elegans* low AND *Selemonas* low	THEN	control
4	IF	*Selenomonas* low AND *Streptococcus* low	THEN	control
5	IF	*Haemophilus parainfluenza* high AND *Granulicatella elegans* high	THEN	control
1	IF	*Selomonas* high AND *Streptococcus* high	THEN	depression
2	IF	*Streptococcus* high AND *Granulicatella elegans* low	THEN	depression

of high and low abundant the bacterial genera subset (see Figure 6.5). There are significant biological justifications for the four rules that characterise the depressed gut microbiome. *Phascolarctobacterium* is a bacterial genus that is abundant in the human gut and produces short chain fatty acids, which are associated with modifying host metabolism and mood (Cryan and Dinan, 2012). Additionally, *Phascolarctobacterium* has been previously positively correlated with positive mood in healthy adults (Li et al., 2016). The second and third rules for depression contain the multiple bacteria in the *Bacteroides* genus; *Bacteroides* are a major mutualistic member of the normal human intestinal microbiome, the described abundance patterns indicate a type of gut dysbiosis has occured, which has been frequently associated with various diseases. It is useful to compare the rough results for the gut microbiome with the original analysis that used traditional (i.e. non-RST) methodology (Jiang et al., 2015). The low levels of *Ruminococcaceae* in rules 2 and 3 are concordant with the traditional analysis. The low abundance of *Alistipes* in rule 4 is not consistent with the original analysis. However, the low abundance is combined with a high abundance of *Odoribacter*, which was not mentioned in

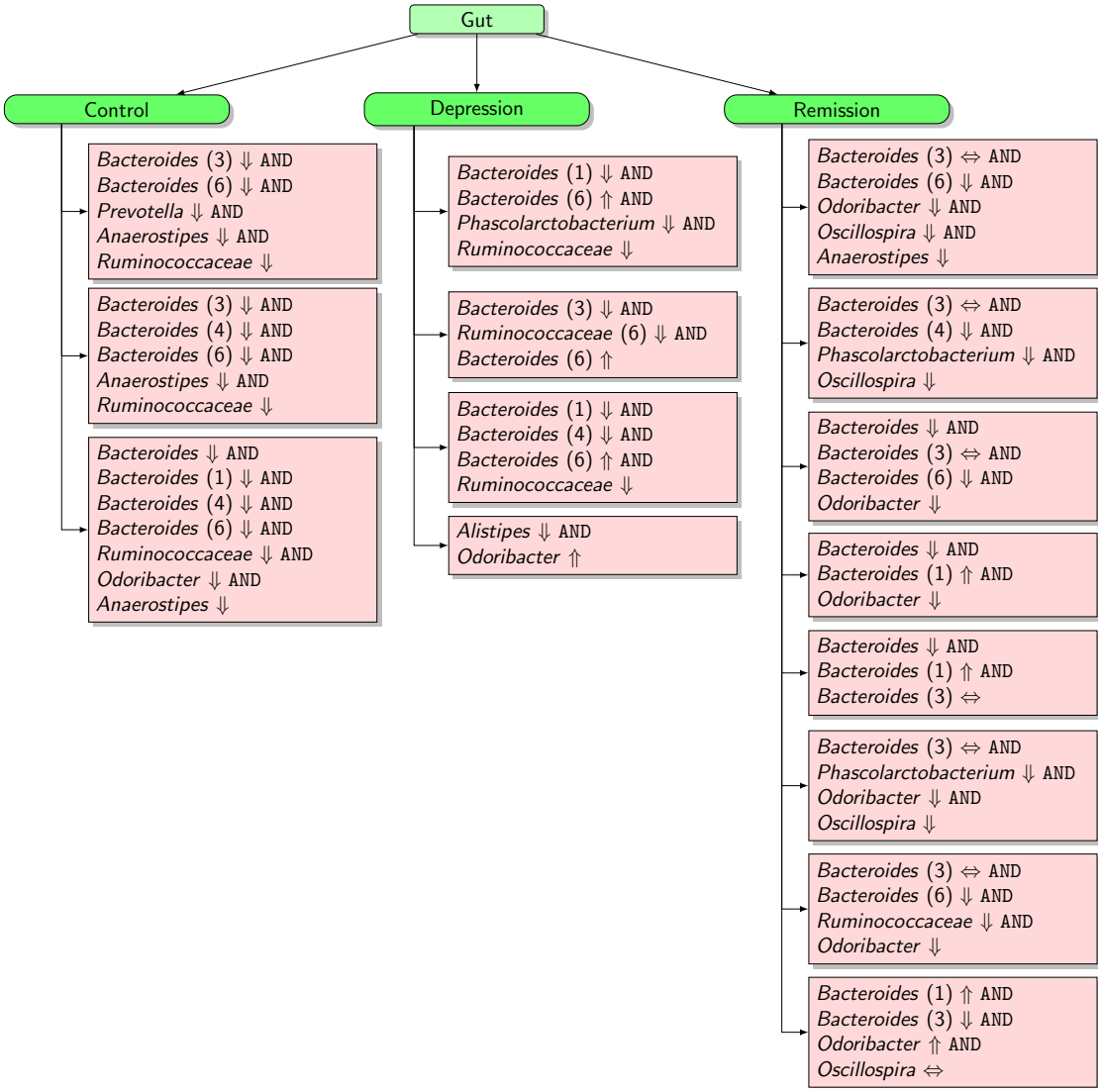


Figure 6.5: Rules that characterise the gut microbiome. ⇑ indicates high abundance, ⇔ indicates medium abundance, and ⇓ indicates low abundance

the original analysis. *Odoribacter* are typically opportunistic pathogens, which can activate inflammatory pathways associated with the microbiome-gut-brain axis (Hardham et al., 2008). In the oral microbiome five rules were induced to characterise control samples and two to characterise depressed samples (see Figure 6.6). There is again compelling biological evidence that supports the induced rules: many oral streptococci and selemonads are opportunistic pathogens (Kreth et al., 2009; Gonçalves et al., 2012). The original analysis of the oral

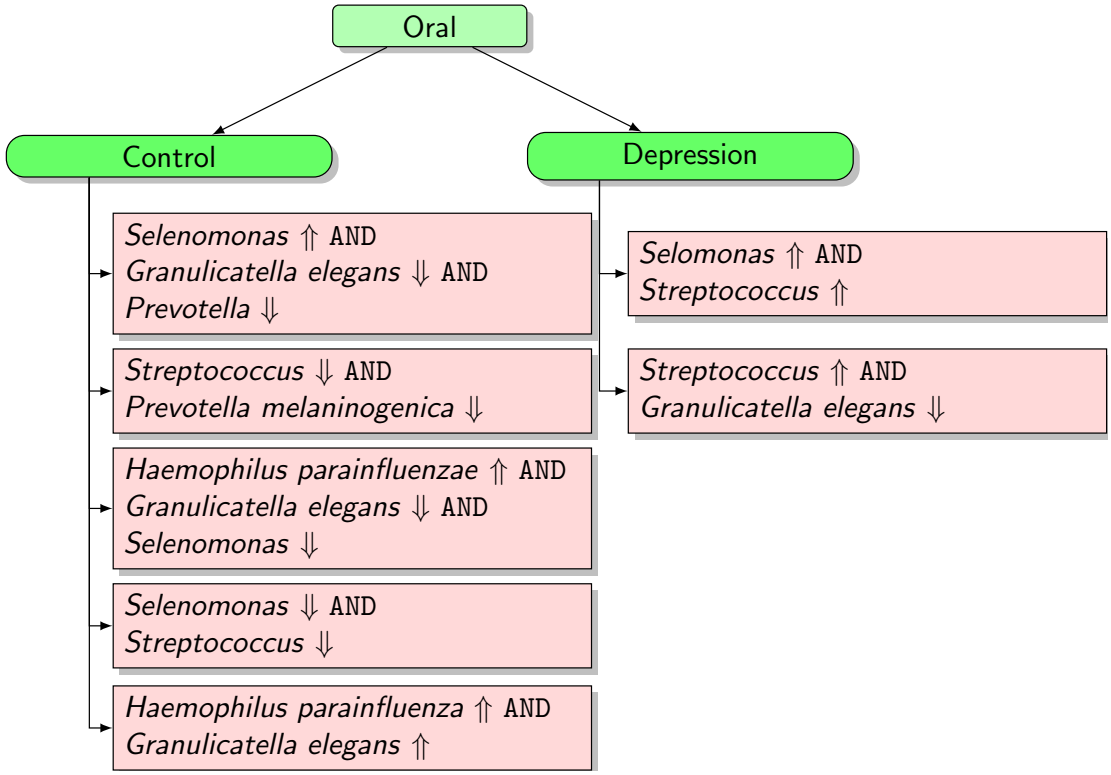


Figure 6.6: Rules that characterise the oral microbiome. ⇑ indicates high abundance, and ⇓ indicates low abundance

microbiome showed a similar pattern of abundance for opportunistic pathogens that is demonstrated in both the control and depressed rules. The rules that describe subjects in remission are unique in that they describe patterns of bacterial abundance at a level other than high or low. This could indicate that the gut microbiome is in a state of recovery from dysbiosis.

In the gut microbiome the support for control rules (83.3% average) was greatly higher compared with the support for depressed rules (28.4%). The lower support for depressed rules is in line with current theories regarding the microbiome-gut-brain axis: it is thought the gut in depressed subjects is in a state of dysbiosis. Dysbiosis describes microbial imbalance, which can vary significantly across different subjects. Additionally, microbiome composition can differ significantly across individuals with dysbiosis whilst the overall gene content is the same (i.e. the functions of the bacteria) (Dash et al., 2015). However, this pattern of support does not hold for the oral microbiome, as the average support for rules is similar for both control (38.94%) and depressed (45%) subjects. This may be related to the method of

Gut microbiome		
Decision	Rule #	Support
Control	1	86.7%
	2	83.3%
	3	80.0%
Depressed	1	31.0%
	2	27.5%
	3	27.5%
	4	27.5%

Oral microbiome		
Decision	Rule #	Support
Control	1	79.0%
	2	26.3%
	3	42.1%
	4	18.4%
	5	28.9%
Depressed	1	55.0%
	2	35.0%

Table 6.7: Quality of microbiome characterisation.

sample collection. The oral microbiome dataset was gathered via saliva samples, which can vary significantly. The gut microbiome dataset was gathered via faeces, which will be more consistent across samples. Therefore defining a control subject from saliva may be a more difficult task for the oral microbiome dataset.

In Chapter 5 a standard microbial analysis was performed of oral microbiome census data, and its possible links with depression. There are some similarities across both the rough set characterisation and the standard analysis. For example, low *Prevotella* is a rule for control subjects (see Table 6.6), and a differential abundance analysis found that some *Prevotella* ASVs are significantly more abundant in depressed subjects (see Figure 5.7). Additionally, *Haemophilus parainfluenzae* is found to be high in control samples, and the differential abundance analysis found that an ASV in the *Haemophilus* genus is significantly less abundant in depressed subjects (see Figure 5.7). Indeed, the *Haemophilus* ASV shows the highest abundance change of all differentially abundant ASVs. No other bacterial genera from the differential abundance analysis intersect with the results of the rough set characterisation. However, high *Streptococcus* abundance was repeatedly part of depression characterisation. *Streptococcus sanguinis* was found to have statistically significant microbial interactions with *Neisseria flavescens* and *Neisseria elongata* in the network analysis (see Figure 5.8).

The gut microbiome dataset was previously analysed with a standard microbial ecology methodology (Jiang et al., 2015). It is difficult to directly compare results as the original work used a clustering based operational taxonomic unit (OTU) method of processing the 16S sequences, and the rough set characterisation detailed in this work uses the superior ASV paradigm. A differential abundance analysis showed that *Bacteroides* was significantly less abundant in the control group, which is supported by all of the generated control rules. However, some differences

include *Phascolarctobacterium* and *Alistipes* being reported as more abundant in the depressed cohort, while the generated depressed rules identify decreased abundance for both genera. However, the significant methodological differences between the two analysis limit the conclusions that can be drawn from any comparisons.

The datasets analysed in this section stratified samples with any confounding conditions into a separate group that was not considered for analysis. Confounding conditions could impact the composition of the microbiome, and any changes should be reflected in the characterisation process also. Including subjects with confounding conditions could significantly reduce the quality of the generated rules by introducing extra variability (i.e. the generated rules would have less support). The rough set characterisation procedure has used simple classes (e.g. depressed or healthy) so far to characterise subjects. To incorporate confounding conditions additional classes could be added. For example, a subject that has recently taken antibiotics will have a significantly different microbiome compared with a subject that has not, even if both have a similar underlying condition. In the case of this dataset, four classes could be chosen to specify control and depressed subjects with and without antibiotic use. This approach will not scale to a large amount of confounders, although including too many confounders may introduce too much variation to the data and limit the effectiveness of any analysis.

6.5 Measuring the robustness of rough set characterisation

In $p >> n$ data there is a reasonable chance that random correlations between feature vectors and a class label will be present, due to the sheer size of the feature space (see Figure 6.7; Smith and Ebrahim, 2002). A closely related phenomenon in statistics is called the multiple comparisons problem (Noble, 2009). Briefly, in the multiple comparisons problem the more hypotheses you wish to test simultaneously, the more likely it is that an erroneous significant result will be identified. For example, testing if the relative abundance of 10 bacterial genera are associated with a particular disease at the same time, with a significance level of 0.05:

$$\begin{aligned} \mathbb{P}\text{ at least one significant result} &= 1 - \mathbb{P}\text{ no significant results} && (6.30)\\ &= 1 - (1 - 0.05)^{10} && (6.31)\\ &\approx 40\% && (6.32) \end{aligned}$$

Results in an approximately 40% chance of a significant test result, even if none actually are. In computational biology it is common to do hundreds or thousands of simultaneous hypothesis tests, which rapidly increases the probability of a false

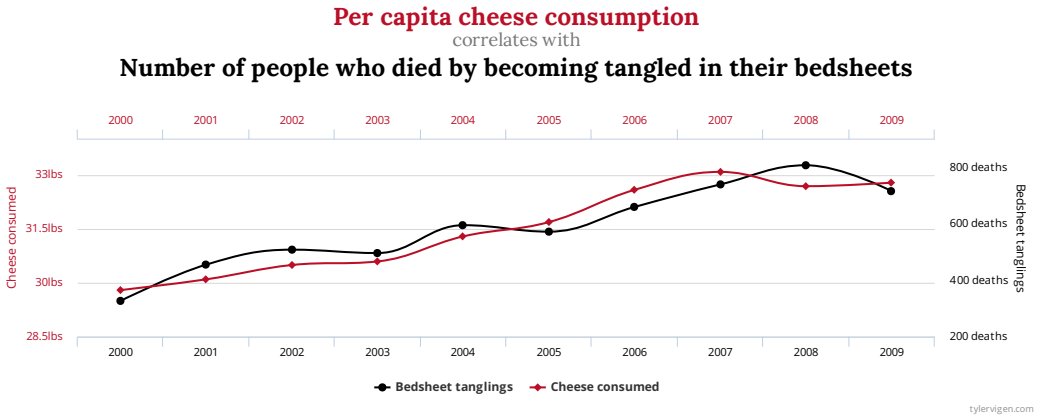


Figure 6.7: Searching through large amounts of data can identify correlated features by chance. The number of people that die by becoming tangled in their bedsheets correlates almost perfectly with U.S. per-capita cheese consumption ($R \approx 0.95$). Graph by Tyler Viglen, available under a Creative Commons license.

positive (Noble, 2009). Several methods exist to correct for multiple comparison testing, including Bonferroni correction (Bland and Altman, 1995) or the Benjamini-Hochberg procedure (Hochberg and Benjamini, 1990).

The problem of ensuring that the output of a feature selector is robust has been covered extensively in Chapter 4, in which robust microbial markers of Inflammatory Bowel Disease (IBD) were identified using aggregating EFS. The same strategy of repeatedly resampling a subset of data can be used to measure the robustness of rough set reducts, and therefore the robustness of the rough set characterisation. In this section the robustness of the oral and gut rough set characterisations (discussed in Section 6.4) is investigated.

6.5.1 Implementation and results

To assess the robustness of the oral and gut characterisations 80% of the microbiome census data for each characterisation task was repeatedly resampled with replacement (bootstrapped; see Figure 6.8). A bootstrap procedure randomly draws samples with replacements from a dataset (i.e. the drawn sample is added back to the dataset and can be redrawn). Bootstrapping is commonly used to estimate the precision of sample statistics, significance tests, and in model validation (Varian, 2005). The rough set characterisation procedure was then performed on each of the data resamples generated by the bootstrapping procedure. To measure the robustness of the rough set characterisation procedure, the similarity of the bacterial species present in the generated rules across different bootstraps was combined into

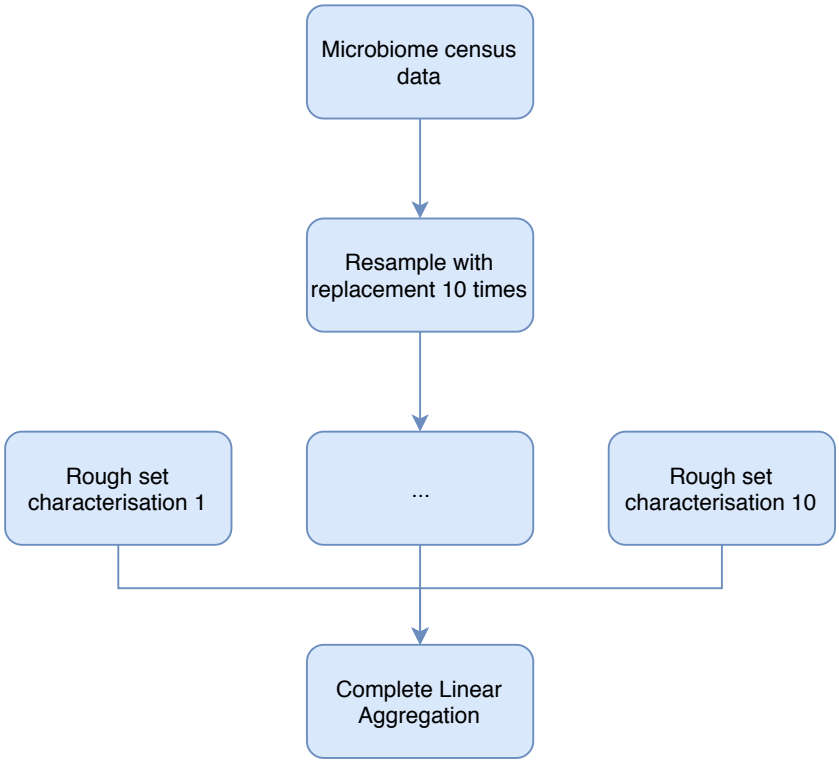


Figure 6.8: The robustness of a rough set characterisation can be measured by bootstrapping microbiome census data and combining the bacterial species present in the reducts. Complete linear aggregation (Abeel et al., 2010) was used to combine the lists of bacterial species into a ranked list of robust characterisations.

a single ranked list of bacterial genera using complete linear aggregation (Abeel et al., 2010). A strong characterisation would be a rule that has been generated consistently across multiple data resamples, and a weak characterisation would be present in few resamples (or absent entirely).

The best gut characterisations included *Bacteroides*, *Alipstipes*, and *Phascolarctobacterium*, which were identified in at least half of the resampled rough sets (see Table 6.8). Overall, the most robust characterisation was *Bacteroides*, which appeared in 80% of gut resamples, and was present in nearly every single generated rule for control, depressed, and remission subjects. Two new bacterial genera were identified in the resampled characterisations that were absent from the original: *Parabacteroides* and *Ersyipelotrichaceae*. One genera present in the original characterisation was absent from every resample (*Ruminococcaceae*).

The best oral characterisations included *Streptococcus*, *Prevotella*, *Haemophilus*, and *Selenomonas*, which were identified in at least half of the resampled rough

Table 6.8: Robustness of gut microbiome rough characterisation.

Genera	Consistency	Frequency	Rule no.
Bacteroides	8	0.8	1 – 3 (control), 1 – 3 (depressed), 1 – 8 (remission)
Alistipes	7	0.7	4 (depressed)
Parabacteroides	6	0.6	None
Phascolarctobacterium	5	0.5	1 (depressed), 2 & 6 (remission)
Erysipelotrichaceae UCG.003	4	0.4	None
Odoribacter	3	0.3	3 (control), 3 (depressed), 1, 3, 4, 6 – 8 (remission)
Prevotella 9	2	0.2	1 (control)
Anaerostipes	1	0.1	1 – 3 (control), 1 (remission)

Table 6.9: Robustness of oral microbiome rough characterisation.

Genera	Consistency	Frequency	Rule no.
Streptococcus	10	1	2 & 4 (control), 1 – 2 (depression)
Prevotella 7	9	0.9	1 – 2 (control)
Haemophilus	8	0.8	3 & 5 (control)
Veillonella	7	0.7	None
Selenomonas 3	7	0.7	1, 3, 4 (control), 1 (depression)
Capnocytophaga	5	0.5	None
Megasphaera	4	0.4	None
Neisseria	2	0.2	None
Gemella	2	0.2	None
Prevotella 6	2	0.2	1 – 2 (control)

sets (see Table 6.9). The most robust characterisation was *Streptococcus*, which was identified in every resample, and was present in over half the generated rules from the original characterisation for both depressed and control subjects. Many new bacterial genera appear that were not present in the original rules that characterised the oral microbiome, including *Veillonella*, *Capnocytophaga*, *Megasphaera*, *Neisseria*, and *Gemella*. The genera *Granulicatella* was present in the original characterisation, but absent from every resample.

6.5.2 Discussion

The rough set resampling approach demonstrates that both gut and oral characterisations of depressed cohorts were on the whole fairly robust. However, the analysis did reveal that certain rules were significantly more robust than others. For example, *Bacteroides* (gut) and *Streptococcus* (oral) were present in nearly every resample, and were both used in a large number of rules. On the other hand, some bacterial genera were only present in the original characterisation and did not appear (e.g. *Ruminococcaceae* in the gut and *Granulicatella* in the oral characterisations). In addition, several new bacterial genera appeared many of the resamples that were absent in the original characterisation. The most robust of which include *Veillonella* and *Capnocytophaga* in the majority of oral resamples, and *Parabacteroides* which appears in the majority of gut resamples. The robustness of characterisations should be taken into account when attempting to investigate potential mechanisms of actions related to biological species.

In future work, incorporating the resampling procedure when identifying reducts from high dimensional data would be an invaluable method of ensuring that outputs (in the form of *IF-THEN* rules) are robust, and not the result of random chance. Under normal conditions this would normally be assessed by testing generated *IF-THEN* rules on unseen data. However, because the characterisation process produces descriptive rules and not predictive rules, this was not possible. Therefore the resampling process was required to further assess the quality of characterisations.

6.6 Summary

Modelling microbiome census data is a difficult task due to the problematic properties associated with high-throughput sequencing data. Recently the widespread application of models that are inappropriate for microbiome census data has rendered the analysis of many microbiome experiments deeply flawed at best (McMurdie and Holmes, 2014). Even analysing microbiome census data with models that are thought to be appropriate is associated with caveats and pitfalls that are unlikely to be noticed by many (Weiss et al., 2017). By applying data-driven CI approaches with weak prior assumptions this metaphorical minefield can be avoided entirely. In addition, the first goal of many microbiome experiments is to characterise an environment. The description of patterns in data is also a significant part of data mining and knowledge discovery — despite the focus on classification throughout this thesis and the popularity of classification for microbiome research generally — and the application of rough set theory in this chapter has enabled the thorough description of problematic data.

The experiments in this chapter found that RST was capable of characterising

microbiomes well. The quality of characterisation was measured with three metrics (fully defined in Section 6.3): the accuracy of characterisation, the quality of characterisation, and the support each induced rule received (i.e. the number of samples that agreed with a generated rule). Furthermore, an additional bootstrapping procedure found that the rough characterisations were fairly robust. The rough sets could perfectly discern samples collected from different locations. Simple rules were induced to generate knowledge about key characteristics of the microbiomes and what distinguishes them from similar environments. The rough sets were also able to perfectly characterise the depression datasets, and more complex rules were induced to generate knowledge from the data. The generated rules included a compelling mix of old and new insights: comparison of the rules to existing analysis using standard approaches found many identified abundance patterns were also present in the RST characterisation. In addition, novel abundance patterns have also been identified, implicating new bacterial genera in the aetiology of the microbiome-gut-brain axis for the first time such as *Odoribacter*. However, it is important to note that further work will be required to generate new data (e.g. the recruitment and analysis of a new cohort to analyse the gut and oral microbiome in depressed subjects) to confirm the novel results independently of the rough set characterisation procedure. Work on subjects with depression in remission revealed that the microbiome could be showing preliminary signs of a recovery process, which provides support for the leaky gut hypothesis of depression.

Preliminary evidence of disease prediction from microbiome census data has been presented in this thesis and across microbiome research more generally. However, highly popular high performance supervised learning algorithms (Statnikov et al., 2013) are almost entirely black box models. Although black box models are acceptable for applications such as image processing and speech recognition, it is important to trust the output of models for clinical use of medical recommendation systems. To enable trust it is critical for models to be transparent and understood by human experts. The RST approach described throughout this chapter can be easily extended from description to transparent disease prediction if sufficient data are available, and disease prediction would be necessary to develop any future clinical applications of RST characterisation.

Rule-based systems suffer from the combinatorial rule explosion problem. As the number of features being considered increases, the number of rules increases exponentially (Combs and Andrews, 1998). This drastically reduces the performance and transparency of rule-based systems. The rough set theory applications to the microbiome census data in this chapter have generated small reducts, with less than a dozen features, which avoids this problem. However, it is important to note some applications of the rough set characterisation may result in a rule explosion if many bacterial species are relevant to the characterisation. Rule optimisation

would be an important method of tackling this problem while maintaining the transparency of the system.

Publications arising from this work

The basis of this work is under preparation for submission:

Wingfield, B., S. Coleman, T. McGinnity and A. J. Bjourson (2019). 'Rough Set Microbiome Characterisation'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Note: manuscript under preparation.

> My methods of navigation have their advantages. I may not have gone where I intended to go, but I think I have ended up where I needed to be.
>
> Douglas Adams

7.1 Introduction

Since germ theory was popularised by John Snow and Louis Pasteur in the 19th century, microorganisms have been viewed by healthcare professionals and the general public, as a pest that must be destroyed. However, the vast majority of microorganisms that inhabit the human body are not pathogenic, and many are responsible for maintaining health. Subtle imbalances in the microbiome have been linked to a large number of diseases with a complex and uncertain aetiology, including Inflammatory Bowel Disease (IBD) and depression. IBD caused 53,000 deaths worldwide in 2013 and its prevalence is increasing, particularly in western countries (Molodecky et al., 2012). Each week 3.3% of the adult population in England report having a depressive episode (McManus et al., 2016) and depression is one of the leading sources of disability globally. Both diseases are currently difficult to diagnose: there are no empirical tests for depression in clinical use, and IBD requires invasive colonoscopy.

The aim of this thesis is to develop computational models of microbiomes across the entire gastrointestinal tract in order to investigate the mechanisms that are involved in the aetiology of disease, with a focus on depression. To achieve the aim of this thesis six objectives were determined:

1. Review computational approaches including Computational Intelligence (CI) and machine learning that have been applied for knowledge discovery from biological data;

2. Review the microbiome literature to identify how the microbiome is thought to be linked with diseases (with a focus on depression), how microbiome census data are created, and CI applications to microbiome census data;

3. Identify methodologies that overcome current limitations in the application of computational models to microbiome census data;

4. Develop computational models that accurately predict IBD from microbiome census datam, and identify a robust subset of bacterial species that can enable knowledge discovery;

5. Using Artificial Intelligence (AI) and CI techniques, identify associations between the oral microbiome and depression in a cohort of young adults;

6. Develop an approach that could characterise microbial environments while preserving data semantics that are destroyed by standard normalisation procedures.

Chapter 2 presented a review of CI applied to knowledge discovery from biological data, while Chapter 3 presented a review of the microbiome-gut-brain axis, its role in disease, and applications of CI to microbiome census data. These reviews identified key research challenges related to microbiome census data analysis and the microbiome-gut-brain axis that have not been considered to date:

1. The impact of microbiome variability on feature selection algorithm output;

2. The role the oral microbiome plays in the microbiome-gut-brain axis in a depressed cohort;

3. The role microbiome census data normalisation algorithms play in destroying data semantics and impairing data interpretability.

Due to the variability of taxonomic profiles (i.e. the count of different bacterial species that represent the microbiome) across individuals, the output of feature selection algorithms to microbiome census data was inconsistent after small changes were made to the input data. This can additionally impact classification performance. This is a common problem when applying feature selection algorithms to complex biological data, as biological data can be variable and highly dimensional, which impacts the performance of standard feature selectors (Abeel et al., 2010). Typical feature selection approaches have traditionally focused on metrics such as execution time and classification accuracy, rather than output stability. However, to enable knowledge discovery, and to to gain the confidence of non-computer scientist domain experts, the output of feature selection algorithms should be robust.

A growing body of evidence suggests that the gastrointestinal microbiome plays an important role in the aetiology of depression. However, the limited work done in a human cohort has focused exclusively on the lower half of the gastrointestinal tract. The role of the oral microbiome has not been investigated in a depressed cohort to date. Saliva can absorb blood-based biomarkers and can represent an important source of disease information. Saliva can be collected non-invasively, which offers significant sampling and handling advantages compared with the collection of faecal

or biopsy samples, which are required to analyse the gut microbiome. Therefore the oral microbiome presents a compelling target for identifying new links between the gastrointestinal microbiome and depression.

Despite a huge variety of normalisation approaches being developed and benchmarked for microbiome census data, there is no universal optimal approach. A chosen approach must carefully balance the properties of the data and the objectives of the experiment. For example, certain microbial environments can violate the sparsity assumptions of popular transformations, and some transformations can produce negative counts which are incompatible with traditional ecological analysis approaches. Additionally, the transformations can destroy the semantics of the original data, and create data that are difficult for scientists with no background in data analytics to understand.

Two models were developed and presented in Chapter 4 that use microbiome census data gathered from subjects with IBD to enable and enhance the non-invasive prediction of IBD. Firstly, a hybrid model was developed that decomposes full IBD diagnosis (including presence, subtype, and severity) into a series of simpler classification problems. The importance of functional profile data (which is less variable than standard taxonomic profiles) was assessed for each stage of the hybrid model to determine the effect of variability on feature selector output. The less variable functional data was found to be the most important type of data for all stages of the hybrid classifier. Secondly, an aggregating ensemble feature selection (EFS) procedure was applied to taxonomic profiles to identify robust microbial markers, enable knowledge discovery, and to mitigate the impact of taxonomic profile variability across subjects on feature selection algoritm output.

Chapter 5 explored the oral microbiome and its role in the microbiome-gut-brain axis in a depressed cohort. The results found alterations present in the composition and structure of the oral microbiome in depressed subjects for the first time. The differences were large enough to enable accurate prediction (83.3% balanced accuracy) of depression from a saliva sample. This novel result has significant implications for the microbiome-gut-brain axis theory of depression, which to date has focused on the lower gastrointestinal tract, and for the current understanding of depression pathophysiology. The predictive performance of the methods developed in Chapters 4 and 5 exceed current clinical best practice. The Hamilton depression scale has a sensitivity of 78.1% and a specificity of 74.6% for screening purposes (Aben et al., 2002). Current non-invasive methods of diagnosing IBD have a maximum accuracy of 81.4% for crohn's disease (CD) (calprotectin), and 83.3% for ulcerative colitis (UC) (faecal lactoferrin) (Langhorst et al., 2008). The robust microbial markers identified in Chapter 4 have a sensitivity of 100% for CD and 87.5% for UC, and a specificity of 94.4% for CD and 100% for UC.

Chapter 6 outlined the development of the rough microbiome characterisation

approach, which applied rough set theory to describe microbial environments. A variety of normalisation techniques were applied to microbiome census data throughout Chapters 4 and 5. Different normalisation approaches were used as there is no single optimal normalisation technique, and the chosen approach must carefully consider both the data and analysis task. By using a data driven approach with minimal prior assumptions this problem can be avoided. Additionally, rough set theory offers an attractive suite of tools for extracting knowledge from data. By applying the rough microbiome characterisation approach to the oral microbiome census data from Chapter 5 and a publicly available dataset gathered from the gut of depressed subjects, new insights were identified and existing results confirmed regarding the microbiome-gut-brain axis in depressed subjects. Four rules were generated to characterise the depressed gut microbiome. Of particular note is that low abundance of *Phascolarctobacterium* is associated with depression. Significantly, this observation has both theoretical and empirical justification. *Phascolarctobacterium* produces Short Chain Fatty Acids (SCFAs), which are thought to be associated with modifying host metabolism and mood (Cryan and Dinan, 2012). Additionally, *Phascolarctobacterium* has been directly observed to be positively correlated with positive mood in a human cohort (Li et al., 2016). The pattern of decreased abundance of some opportunistic pathogens (e.g. *Granulicatella elegans*) that was observed in Chapter 5 was also found by the rough characterisation of the oral microbiome. It is important to note that it was not possible to determine relationships between the oral and gut microbiome in depressed or healthy subjects. The gut microbiome data was gathered from a Chinese cohort and the oral microbiome data was gathered from a European cohort. Any variation or correlation between the oral and gut microbiome and depression could be caused by factors such as differences in diet or ethnicity, which are known to significantly impact the composition and structure of microbiomes (Prideaux et al., 2013).

7.2 Summary of original contributions

The primary aim of this thesis was to develop and apply analytical models to investigate the oral and gut microbiomes for association with disease and enable knowledge discovery, with a focus on depression and the microbiome-gut-brain axis. This aim was achieved by the work presented throughout Chapters 4–6. The following sections below summarise the novel contributions of this thesis (Sections 7.2.1 – 7.2.3).

7.2.1 Non-invasive prediction of IBD and identification of robust microbial markers

The hybrid model and aggregating EFS approaches, developed throughout Chapter 4, enables the non-invasive identification of IBD in paediatric subjects from faecal samples. Additionally, the aggregating EFS enabled knowledge discovery via the generation of a consensus ranked feature list. The following original contributions were delivered during the development of these approaches:

Comprehensive IBD prediction: IBD is a complex disease, and during standard diagnosis its presence, subtype, and severity is assessed by clinicians via invasive colonoscopy. The hybrid model decomposed this complex classification problem into a series of simpler classification tasks to allow non-invasive diagnosis.

Assessing functional feature relevance: Taxonomic profiles are extremely variable across different subjects. It was unclear if functional profiles, which are less variable, could be useful for disease prediction. The relevance of functional features was assessed with the Boruta algorithm. The majority of relevant features were found to be functional features for all stages of the hybrid classifier.

Aggregating ensemble feature selection: Due to the variability of microbiomes across individuals, the output of feature selection algorithms to microbiome census data was inconsistent after small changes were made to the input data. An aggregating EFS approach was applied to microbiome census data to generate a consensus feature ranking that could non-invasively predict IBD in a treatment-naïve paediatric cohort from taxonomic data.

IBD knowledge discovery: By combining state-of-the-art bioinformatics algorithms — the amplicon sequence variant (ASV) approach — to an existing dataset with aggregating EFS, biologically plausible species of bacteria were implicated in the pathogenesis of IBD for the first time.

7.2.2 Analysing oral microbiome in a depressed cohort

Recent work has shown that the microbiome plays an important role in the aetiology of depression (Foster et al., 2017). Much interest has been focused on the role of the lower gastrointestinal tract in the microbiome-gut-brain axis. However, the importance of the oral microbiome has received little attention to date. To

characterise the oral microbiome in a depressed cohort saliva samples were selected from the Ulster University Student Wellbeing Study, and bacterial identification performed using 16S high-throughput sequencing. Through this the following contributions were achieved:

> Structural changes in the depressed oral microbiome: Novel changes to the structure of the oral microbiome in depressed subjects were observed via analysis of beta diversity with statistical tests and constrained ordination. The depressed cohort was found to have significant microbial co-occurrence relationships that were absent in control subjects.
>
> Compositional changes in the depressed oral microbiome: Novel changes to the composition of the oral microbiome in depressed subjects were observed via differential abundance analysis of ASVs and inferred functional pathways.
>
> Prediction of depression from a saliva sample: Multimodal classification was applied with a Self-Organising Map (SOM) to the oral microbiome data, enabling accurate prediction of depression from microbiome census data collected from a saliva sample for the first time. The predictive power of multimodal classification can exceed diagnostic criteria in current clinical use (e.g. the Hamilton depression scale, Aben et al., 2002).

7.2.3 Rough characterisation of oral and gut microbiomes in depressed cohorts

The contributions summarised in sections 7.2.1 and 7.2.2 implicate novel bacterial species in the pathophysiology of IBD depression for the first time. However, the applied models were black box models and cannot be interpreted. Additionally, a variety of normalisation approaches were applied as there is no optimal normalisation technique for all microbiome census data. This can impact the interpretability of the model output by destroying the semantics of the original data. Models based on rough set theory are transparent, data driven, and can measure the relevance and significance of features. The application of rough set theory to microbiome census data enabled the following contributions:

> Rough characterisation: A key goal of microbiome experiments is to characterise (i.e. describe) the microbial community. High-throughput sequencing is used to generate microbiota profiles, but data gathered via this method are extremely challenging to analyse as the data violate

multiple strong assumptions of standard models. Rough set theory has weak assumptions and offers a range of attractive tools for extracting knowledge from complex data. The application of RST simultaneously provides a solution to an open research question regarding identifying an optimal normalisation approach for microbiome census data while providing a clear interpretation of microbial communities. No normalisation approach is optimal for all microbiome census data, and misapplying a normalisation algorithm can significantly impact downstream analysis. As RST is a data-driven approach with minimal prior assumptions, this problem does not occur in the rough characterisation approach demonstrated on a benchmark dataset.

Depression knowledge discovery: Testing the rough characterisation approach on oral and gut microbiome datasets gathered from depressed subjects showed that RST is capable of perfectly characterising the gut and oral microbiomes and identifies previously undescribed alterations to the microbiome-gut-brain axis including increased abundance of the opportunistic pathogen *Odoribacter* in the gut of depressed subjects. The observed alterations have clear biological justifications: bacterial genera that are known to alter human mood and behaviour and have been independently shown to be positively correlated with positive mood were observed to have an decreased abundance in RST analysis of the depressed cohort.

7.3 Future work

The research within this thesis provides new contributions to microbiome research, CI, and the microbiome-gut-brain axis theory. However, this work could be extended in a variety of directions. Some potential avenues are outlined throughout sections 7.3.1 – 7.3.4. The work in this thesis has identified for the first time that the oral microbiome in depressed subjects is significantly different. The first stage of further work should be to confirm the alterations in larger cohorts, and to potentially identify new alterations, by directly measuring taxonomic and functional data with shotgun sequencing. The next stage of further work should be to study the depressed microbiome in longitudinal experiments to identify key changes associated with the onset of depression and remission. Once longitudinal changes have been identified, the final stage of further work would be to determine if the alterations to the microbiome can be reversed or ameliorated by the administration of psychobiotics, and if such approaches are capable of effectively treating depression.

7.3.1 Fuzzy-rough microbiome characterisation

One of the most significant disadvantages of rough set theory is its requirement for data to be discrete. Although raw microbiome count data consists of discrete sequence reads, total sum scaling normalisation (i.e proportions) mitigates uneven library size across samples while maintaining data interpretability. Continuous data must be discretised before it can be analysed, and information loss is guaranteed by the discretisation process. Fuzzy-rough sets combine the concepts of vagueness (fuzzy sets) and uncertainty (rough set theory), which both arise from uncertainty in knowledge. The application of fuzzy-rough sets would enable the use of real-valued data, mitigating one of the most significant drawbacks associated with the rough microbiome characterisation approach.

7.3.2 Microbiome characterisation with shotgun sequenced data

The microbiome census data analysed throughout this thesis was gathered via a marker gene survey. Although marker gene surveys are one of the most popular methods of estimating the structure and composition of microbial communities, there is a number of disadvantages associated with the protocol that could impact results. Firstly, significant bias is introduced by differential amplification and variable 16S copy number across different species. The marker gene survey protocol uses universal primers that can identify a broad spectrum of bacterial and archaeal species. However, it is well known that the amplification efficiency of universal primers varies across different organisms (as the 16S sequences vary slightly). This will impact estimates of bacterial abundance. 16S copy number bias will also impact estimates of bacterial abundance, but this can be somewhat mitigated via the use of copy number databases and ancestral state reconstruction algorithms. Additionally, 16S marker gene surveys can only directly measure taxonomic content of a microbial community. Functional content must be inferred *in silico* from characterised genomes using bioinformatics software packages such as Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt). Compared with direct measurement of functions, it is certain that some information loss will occur because of bias introduced by the marker gene survey or problems with the inference procedure.

Shotgun sequencing sequences long DNA strands instead of small sections of marker genes. This means that the bias described above would be mitigated and functional content can be directly measured. The hybrid model presented in Chapter 4 investigated the relevance of functional profiles for disease prediction. Chapter 6 presented the rough characterisation of oral and gut microbiomes in depressed subjects. Although rough set theory was able to model the microbiomes

perfectly, the generated rules that characterised the environments had somewhat low support. As functional profiles have less variance compared with taxonomic profiles, and it has been shown that alterations in both depressed microbiomes do exist, it would be valuable to perform rough characterisation on directly measured shotgun sequenced data (both taxonomic and functional) to determine if functional profiles generate rules with more support. However, it is important to note that shotgun sequenced sequence data are significantly more expensive to collect and process, and are extremely time-consuming to analyse compared with 16S marker gene survey data. Characterising the microbiome of depressed subjects from shotgun sequenced data would require significant investment and resources.

7.3.3 Longitudinal analysis of depressed cohort

The conclusions that could be drawn from the microbial co-occurrence networks were limited because the analysis provides more compelling results for time series data. Identifying how microbial relationships change over time in response to changes in state can significantly improve our understanding of the underlying biological phenomenon. However, time-series microbiome datasets are often collected from non-medical datasets (e.g. microbial communities present in the ocean), and no longitudinal datasets exist for a depression cohort. Repeated sampling over time could reveal changes correlated to the onset of depression, antidepressant prescription, and depression entering remission. Time series analysis is not limited only to microbial co-occurrence analysis: monitoring the structure and composition of the microbiome over time would also be invaluable. Time series data would provide additional classification or regression problems that could be modelled. For example, subject prognosis could be predicted by modelling non-response to antidepressant medication, which is thought to be linked to inflammation, from time-series data.

7.3.4 Psychobiotics

Once the onset of depression or remission has been associated with alterations to microbiomes across the human body from time-series data, the next logical step would be to identify if reversing such changes can be used to treat depression. Restoring altered microbiomes to their original state could provide a novel method of treating depression, possibly in combination with traditional methods including pharmacological intervention and cognitive behavioural therapy. In addition, novel psychobiotics — live bacteria that when ingested confer mental health benefits through interactions with commensal gut bacteria — could be identified by understanding the overall effect that the microbiome alterations have on the host. For example, the alterations associated with the onset of depression could

lead to an increased amount of SCFAs being produced by the microbiome (and hence absorbed by the human host). SCFAs can engage in epigenetic regulation of inflammation, which is linked to the aetiology of depression. Identifying and administering a bacterial species that metabolises SCFAs could theoretically prevent the development of depression, or treat it.

7.4 Conclusion

The aim of this thesis was to develop and apply analytical CI models to investigate the oral and gut microbiomes for association with disease and to enable knowledge discovery, with a focus on depression and the microbiome-gut-brain axis. The work contained in this thesis addresses this objective by developing and applying CI techniques on publicly available IBD microbiome census data, and applying the same approaches to oral microbiome samples collected in collaboration with the Northern Ireland Centre for Stratified Medicine. The analysis of the oral microbiome samples identified key alterations correlated with depression. The alterations were significant enough to allow the non-invasive prediction of depression from a saliva sample for the first time, with a sensitivity and specificity that exceeds some diagnostic criteria in current clinical use. This discovery has extended the microbiome-gut-brain axis theory to include oral microbiome for the first time, which has to date focused on the lower gastrointestinal tract. In addition, this work developed a rough characterisation procedure that provides a potential solution to an open research question regarding identifying an optimal normalisation technique for microbiome census data. The rough characterisation approach was applied to the oral microbiome data described in Chapter 5 and publicly available gut microbiome data gathered from a depressed cohort and generated new insights into the microbiome-gut-brain axis.

BIBLIOGRAPHY

Aagaard, K., J. Petrosino, W. Keitel, M. Watson, J. Katancik, N. Garcia, S. Patel, M. Cutting, T. Madden, H. Hamilton et al. (2013). 'The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters'. In: *The FASEB Journal* 27.3, pp. 1012–1022.

Abdi, H. and L. J. Williams (2010). 'Principal component analysis'. In: *Wiley interdisciplinary reviews: computational statistics* 2.4, pp. 433–459.

Abeel, T., T. Helleputte, Y. Van de Peer, P. Dupont and Y. Saeys (2010). 'Robust biomarker identification for cancer diagnosis with ensemble feature selection methods'. In: *Bioinformatics* 26.3, pp. 392–398.

Aben, I., F. Verhey, R. Lousberg, J. Lodder and A. Honig (2002). 'Validity of the beck depression inventory, hospital anxiety and depression scale, SCL-90, and hamilton depression rating scale as screening instruments for depression in stroke patients'. In: *Psychosomatics* 43.5, pp. 386–393.

Achtman, M. and M. Wagner (2008). 'Microbial diversity and the genetic nature of microbial species'. In: *Nature Reviews Microbiology* 6.6, pp. 431–440.

Aha, D. W. and R. L. Bankert (1996). 'A comparative evaluation of sequential feature selection algorithms'. In: *Learning from data*. Springer, pp. 199–206.

Aitchison, J. (1986). 'The statistical analysis of compositional data'. In:

Aitchison, J. et al. (1994). 'Principles of compositional data analysis'. In: *Multivariate analysis and its applications* 24, pp. 73–81.

Aitchison, J. and J. J. Egozcue (2005). 'Compositional data analysis: where are we and where should we be heading?' In: *Mathematical Geology* 37.7, pp. 829–850.

Allen, A. P., W. Hutch, Y. E. Borre, P. J. Kennedy, A. Temko, G. Boylan, E. Murphy, J. F. Cryan, T. G. Dinan and G. Clarke (Nov. 2016). 'Bifidobacterium longum 1714 as a translational psychobiotic: modulation of stress, electrophysiology and neurocognition in healthy volunteers'. In: *Translational Psychiatry* 6. Original Article, e939 EP -. URL: http://dx.doi.org/10.1038/tp.2016.191.

American Psychiatric Association et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Amir, A., D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Z. Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez et al. (2017). 'Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns'. In: *mSystems* 2.2, e00191–16.

Amsel, R., P. A. Totten, C. A. Spiegel, K. C. Chen, D. Eschenbach and K. K. Holmes (1983). 'Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations'. In: *The American journal of medicine* 74.1, pp. 14–22.

Ananthakrishnan, A. N. (2015). 'Epidemiology and risk factors for IBD'. In: *Nature Reviews Gastroenterology & Hepatology* 12.4, pp. 205–217.

Anders, S. and W. Huber (2010). 'Differential expression analysis for sequence count data'. In: *Genome biology* 11.10, R106.

Anderson, E. (1935). 'The irises of the Gaspe Peninsula'. In: *Bulletin of American Iris Society* 59, pp. 2–5.

Ang, J. C., A. Mirzal, H. Haron and H. N. A. Hamed (2016). 'Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection'. In: *IEEE/ACM transactions on computational biology and bioinformatics* 13.5, pp. 971–989.

Apté, C. and S. Weiss (1997). 'Data mining with decision trees and decision rules'. In: *Future generation computer systems* 13.2-3, pp. 197–210.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto et al. (2011). 'Enterotypes of the human gut microbiome'. In: *nature* 473.7346, pp. 174–180.

Azar, A. T., N. Bouaynaya and R. Polikar (2015). 'Inductive learning based on rough set theory for medical decision making'. In: *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. IEEE, pp. 1–8.

Bäckhed, F., R. E. Ley, J. L. Sonnenburg, D. A. Peterson and J. I. Gordon (2005). 'Host-bacterial mutualism in the human intestine'. In: *science* 307.5717, pp. 1915–1920.

Bag, S., B. Saha, D. A. Ojasvi Mehta, N. Kumar, M. Dayal, A. Pant, P. Kumar, S. Saxena, K. H. Allin, T. Hansen et al. (2016). 'An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples'. In: *Scientific reports* 6.

Bagby, R. M., A. G. Ryder, D. R. Schuller and M. B. Marshall (2004). 'The Hamilton Depression Rating Scale: has the gold standard become a lead weight?' In: *American Journal of Psychiatry* 161.12, pp. 2163–2177.

Balakrishnama, S. and A. Ganapathiraju (1998). 'Linear discriminant analysis-a brief tutorial'. In: *Institute for Signal and information Processing* 18, pp. 1–8.

Balasubramanian, M. and E. L. Schwartz (2002). 'The isomap algorithm and topological stability'. In: *Science* 295.5552, pp. 7–7.

Bálint, M., M. Bahram, A. M. Eren, K. Faust, J. A. Fuhrman, B. Lindahl, R. B. O'Hara, M. Öpik, M. L. Sogin, M. Unterseher et al. (2016). 'Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes'. In: *FEMS microbiology reviews* 40.5, pp. 686–700.

Banerjee, M., S. Mitra and H. Banka (2007). 'Evolutionary rough feature selection in gene expression data'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.4, pp. 622–632.

Bartram, A. K., M. D. Lynch, J. C. Stearns, G. Moreno-Hagelsieb and J. D. Neufeld (2011). 'Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads'. In: *Applied and environmental microbiology*.

Basheer, I. A. and M. Hajmeer (2000). 'Artificial neural networks: fundamentals, computing, design, and application'. In: *Journal of microbiological methods* 43.1, pp. 3–31.

Baumeister, A. A., M. F. Hawkins and S. M. Uzelac (2003). 'The myth of reserpine-induced depression: role in the historical development of the monoamine hypothesis'. In: *Journal of the History of the Neurosciences* 12.2, pp. 207–220.

Bazan, J. G., H. S. Nguyen, S. H. Nguyen, P. Synak and J. Wróblewski (2000). 'Rough set algorithms in classification problem'. In: *Rough set methods and applications*. Springer, pp. 49–88.

Bazan, J. G. and M. Szczuka (2000). 'RSES and RSESlib-a collection of tools for rough set computations'. In: *International Conference on Rough Sets and Current Trends in Computing*. Springer, pp. 106–113.

Belstrøm, D., P. Holmstrup, A. Bardow, A. Kokaras, N.-E. Fiehn and B. J. Paster (2016). 'Comparative analysis of bacterial profiles in unstimulated and stimulated saliva samples'. In: *Journal of oral microbiology* 8.1, p. 30112.

Benjamini, Y. and Y. Hochberg (1995). 'Controlling the false discovery rate: a practical and powerful approach to multiple testing'. In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Bennett, C. M., A. A. Baird, M. B. Miller and G. L. Wolford (2011). 'Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction'. In: *Journal of Serendipitous and Unexpected Results* 1, pp. 1–5.

Berger, S. L., T. Kouzarides, R. Shiekhattar and A. Shilatifard (2009). 'An operational definition of epigenetics'. In: *Genes & development* 23.7, pp. 781–783.

Bezabeh, T., R. L. Somorjai and I. C. Smith (2009). 'MR metabolomics of fecal extracts: applications in the study of bowel diseases'. In: *Magnetic Resonance in Chemistry* 47.S1.

Bland, J. M. and D. G. Altman (1995). 'Multiple significance tests: the Bonferroni method'. In: *Bmj* 310.6973, p. 170.

Blaxter, M. (2003). 'Molecular systematics: counting angels with DNA'. In: *Nature* 421.6919, p. 122.

Boettiger, C. (2015). 'An introduction to Docker for reproducible research'. In: *ACM SIGOPS Operating Systems Review* 49.1, pp. 71–79.

Bokulich, N. A., J. R. Rideout, W. G. Mercurio, A. Shiffer, B. Wolfe, C. F. Maurice, R. J. Dutton, P. J. Turnbaugh, R. Knight and J. G. Caporaso (2016).

'mockrobiota: a public resource for microbiome bioinformatics benchmarking'. In: *mSystems* 1.5, e00062–16.

Borg, I. and P. J. Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bravo, J. A., P. Forsythe, M. V. Chew, E. Escaravage, H. M. Savignac, T. G. Dinan, J. Bienenstock and J. F. Cryan (2011). 'Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve'. In: *Proceedings of the National Academy of Sciences* 108.38, pp. 16050–16055.

Bray, J. R. and J. T. Curtis (1957). 'An ordination of the upland forest communities of southern Wisconsin'. In: *Ecological monographs* 27.4, pp. 325–349.

Breiman, L. (2001). 'Random forests'. In: *Machine learning* 45.1, pp. 5–32.

Breznak, J. A. (2002). 'A need to retrieve the not-yet-cultured majority'. In: *Environmental microbiology* 4.1, pp. 4–5.

Brodin, J., M. Mild, C. Hedskog, E. Sherwood, T. Leitner, B. Andersson and J. Albert (2013). 'PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data'. In: *PloS one* 8.7, e70388.

Brown, E. M., M. Sadarangani and B. B. Finlay (2013). 'The role of the immune system in governing host-microbe interactions in the intestine'. In: *Nature immunology* 14.7, p. 660.

Brown, T. A., L. A. Campbell, C. L. Lehman, J. R. Grisham and R. B. Mancill (2001). 'Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample.' In: *Journal of abnormal psychology* 110.4, p. 585.

Brüls, T. and J. Weissenbach (2011). 'The human metagenome: our other genome?' In: *Human molecular genetics*, ddr353.

Brunoni, A. R., M. Lopes and F. Fregni (2008). 'A systematic review and meta-analysis of clinical studies on major depression and BDNF levels: implications for the role of neuroplasticity in depression'. In: *International Journal of Neuropsychopharmacology* 11.8, pp. 1169–1180.

Bühlmann, P. (2012). 'Bagging, boosting and ensemble methods'. In: *Handbook of Computational Statistics*. Springer, pp. 985–1022.

Burke, H. M., M. C. Davis, C. Otte and D. C. Mohr (2005). 'Depression and cortisol responses to psychological stress: a meta-analysis'. In: *Psychoneuroendocrinology* 30.9, pp. 846–856.

Bushman, B. J., C. N. DeWall, R. S. Pond and M. D. Hanus (2014). 'Low glucose relates to greater aggression in married couples'. In: *Proceedings of the National Academy of Sciences* 111.17, pp. 6254–6257.

Callahan, B. J., P. J. McMurdie and S. P. Holmes (2017). 'Exact sequence variants should replace operational taxonomic units in marker-gene data analysis'. In: *The ISME journal* 11.12, p. 2639.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson and S. P. Holmes (2016a). 'DADA2: high-resolution sample inference from Illumina amplicon data'. In: *Nature methods*.

Callahan, B. J., K. Sankaran, J. A. Fukuyama, P. J. McMurdie and S. P. Holmes (2016b). 'Bioconductor workflow for microbiome data analysis: from raw reads to community analyses'. In: *F1000Research* 5.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon et al. (2010). 'QIIME allows analysis of high-throughput community sequencing data'. In: *Nature methods* 7.5, pp. 335–336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer and R. Knight (2011). 'Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample'. In: *Proceedings of the National Academy of Sciences* 108.Supplement 1, pp. 4516–4522.

Carrigg, C., O. Rice, S. Kavanagh, G. Collins and V. O'Flaherty (2007). 'DNA extraction method affects microbial community profiles from soils and sediment'. In: *Applied microbiology and biotechnology* 77.4, pp. 955–964.

Carter, J., D. Beck, H. Williams, G. Dozier and J. A. Foster (2014). 'GA-based selection of vaginal microbiome features associated with bacterial vaginosis'. In: *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. ACM, pp. 265–268.

Castrén, E. (2005). 'Is mood chemistry?' In: *Nature Reviews Neuroscience* 6.3, pp. 241–246.

Chao, A. (1984). 'Nonparametric estimation of the number of classes in a population'. In: *Scandinavian Journal of statistics*, pp. 265–270.

Chawla, N. V., K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer (2002). 'SMOTE: synthetic minority over-sampling technique'. In: *Journal of artificial intelligence research*, pp. 321–357.

Chen, C.-S., S. Sullivan, T. Anderson, A. C. Tan, P. J. Alex, S. R. Brant, C. Cuffari, T. M. Bayless, M. V. Talor, C. L. Burek et al. (2009). 'Identification of novel serological biomarkers for inflammatory bowel disease using Escherichia coli proteome chip'. In: *Molecular & Cellular Proteomics* 8.8, pp. 1765–1776.

Chen, L., W. Wang, R. Zhou, S. C. Ng, J. Li, M. Huang, F. Zhou, X. Wang, B. Shen, M. A. Kamm et al. (2014). 'Characteristics of fecal and mucosa-associated microbiota in Chinese patients with inflammatory bowel disease'. In: *Medicine* 93.8.

Chen, T., W.-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan and F. E. Dewhirst (2010). 'The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information'. In: *Database* 2010.

Chor, B., D. Horn, N. Goldman, Y. Levy and T. Massingham (2009). 'Genomic DNA k-mer spectra: models and modalities'. In: *Genome biology* 10.10, R108.

Clarridge, J. E. (2004). 'Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases'. In: *Clinical microbiology reviews* 17.4, pp. 840–862.

Cleland, E. E. (2011). 'Biodiversity and ecosystem stability'. In: *Nature Education Knowledge* 3.10, p. 14.

Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice (2009). 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants'. In: *Nucleic acids research* 38.6, pp. 1767–1771.

Coenye, T. and P. Vandamme (2003). 'Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes'. In: *FEMS Microbiology Letters* 228.1, pp. 45–49.

Cohan, F. M. (2002). 'What are bacterial species?' In: *Annual Reviews in Microbiology* 56.1, pp. 457–487.

Colgan, S. P., V. F. Curtis and E. L. Campbell (2013). 'The inflammatory tissue microenvironment in IBD'. In: *Inflammatory bowel diseases* 19.10, p. 2238.

Colwell, R. K. and J. A. Coddington (1994). 'Estimating terrestrial biodiversity through extrapolation'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 345.1311, pp. 101–118.

Combs, W. E. and J. E. Andrews (1998). 'Combinatorial rule explosion eliminated by a fuzzy rule configuration'. In: *IEEE Transactions on fuzzy Systems* 6.1, pp. 1–11.

Cook, C. E., M. T. Bergman, R. D. Finn, G. Cochrane, E. Birney and R. Apweiler (2015). 'The European Bioinformatics Institute in 2016: data growth and integration'. In: *Nucleic acids research* 44.D1, pp. D20–D26.

Costea, P. I., G. Zeller, S. Sunagawa and P. Bork (2014). 'A fair comparison'. In: *Nature methods* 11.4, pp. 359–359.

Coyte, K. Z., J. Schluter and K. R. Foster (2015). 'The ecology of the microbiome: networks, competition, and stability'. In: *Science* 350.6261, pp. 663–666.

Cryan, J. F. and T. G. Dinan (2012). 'Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour'. In: *Nature reviews neuroscience* 13.10, pp. 701–712.

Csardi, G. and T. Nepusz (2006). 'The igraph software package for complex network research'. In: *InterJournal, Complex Systems* 1695.5, pp. 1–9.

D'Odorico, S., R. Bortolan, R. Cardin, D. D'Inca', A. Martines, G. Ferronato and A. Sturniolo (2001). 'Reduced plasma antioxidant concentrations and increased oxidative DNA damage in inflammatory bowel disease'. In: *Scandinavian journal of gastroenterology* 36.12, pp. 1289–1294.

Dai, J. and Q. Xu (2013). 'Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification'. In: *Applied Soft Computing* 13.1, pp. 211–221.

Dantzer, R., J. C. O'Connor, G. G. Freund, R. W. Johnson and K. W. Kelley (2008). 'From inflammation to sickness and depression: when the immune system subjugates the brain'. In: *Nature reviews neuroscience* 9.1, pp. 46–56.

Darwin, C. (1872). *The expression of the emotions in man and animals*. 1st ed. John Murray.

Dash, S., G. Clarke, M. Berk and F. N. Jacka (2015). 'The gut microbiome and diet in psychiatry: focus on depression'. In: *Current opinion in psychiatry* 28.1, pp. 1–6.

Deacon, B. J. and G. L. Baird (2009). 'The chemical imbalance explanation of depression: reducing blame at what cost?' In: *Journal of Social and Clinical Psychology* 28.4, pp. 415–435.

Deng, L., D. Yu et al. (2014). 'Deep learning: methods and applications'. In: *Foundations and Trends in Signal Processing* 7.3–4, pp. 197–387.

Dennis, J. L., T. R. Hvidsten, E. C. Wit, J. Komorowski, A. K. Bell, I. Downie, J. Mooney, C. Verbeke, C. Bellamy, W. N. Keith et al. (2005). 'Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm'. In: *Clinical Cancer Research* 11.10, pp. 3766–3772.

Determan Jr, C. E. (2015). 'Optimal Algorithm for Metabolomics Classification and Feature Selection varies by Dataset'. In: *International Journal of Biology* 7.1, p. 100.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo and C. Notredame (2017). 'Nextflow enables reproducible computational workflows'. In: *Nature Biotechnology* 35.4, pp. 316–319.

Diamond, J. M. (1975). 'Assembly of species communities'. In: *Ecology and evolution of communities*, pp. 342–444.

Díaz-Uriarte, R. and S. A. De Andres (2006). 'Gene selection and classification of microarray data using random forest'. In: *BMC bioinformatics* 7.1, p. 3.

Dick, G. J., A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton and J. F. Banfield (2009). 'Community-wide analysis of microbial genome sequence signatures'. In: *Genome biology* 10.8, R85.

Dietterich, T. G. et al. (2000). 'Ensemble methods in machine learning'. In: *Multiple classifier systems* 1857, pp. 1–15.

Dinsdale, E. A., R. A. Edwards, B. Bailey, I. Tuba, S. Akhter, K. McNair, R. Schmieder, N. Apkarian, M. Creek, E. Guan et al. (2013). 'Multivariate analysis of functional metagenomes'. In: *Frontiers in genetics* 4, p. 41.

Ditzler, G., R. Polikar and G. Rosen (2015). 'Multi-layer and recursive neural networks for metagenomic classification'. In: *IEEE transactions on nanobioscience* 14.6, pp. 608–616.

Doolittle, W. F. and O. Zhaxybayeva (2009). 'On the origin of prokaryotic species'. In: *Genome research* 19.5, pp. 744–756.

Dougherty, J., R. Kohavi and M. Sahami (1995). 'Supervised and unsupervised discretization of continuous features'. In: *Machine Learning Proceedings 1995*. Elsevier, pp. 194–202.

Downey Jr, T. J., D. J. Meyer, R. K. Price and E. L. Spitznagel (1999). 'Using the receiver operating characteristic to assess the performance of neural classifiers'. In: *Neural Networks, 1999. IJCNN'99. International Joint Conference on*. Vol. 5. IEEE, pp. 3642–3646.

Drachman, D. A. (2005). 'Do we have brain to spare?' In: *Neurology* 64.12, pp. 2004–2005.

Duncan, S. H. and H. J. Flint (2013). 'Probiotics and prebiotics and health in ageing populations'. In: *Maturitas* 75.1, pp. 44–50.

Duran-Pinedo, A. E. and J. Frias-Lopez (2015). 'Beyond microbial community composition: functional activities of the oral microbiome in health and disease'. In: *Microbes and infection* 17.7, pp. 505–516.

Edgar, R. C. (2010). 'Search and clustering orders of magnitude faster than BLAST'. In: *Bioinformatics* 26.19, pp. 2460–2461.

— (2013). 'UPARSE: highly accurate OTU sequences from microbial amplicon reads'. In: *Nature methods* 10.10, pp. 996–998.

— (2016). 'UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing'. In: *bioRxiv*, p. 081257.

Efraim, T., E. A. Jay, T.-P. Liang and R. McCarthy (2001). 'Decision support systems and intelligent systems'. In: *Upper Saddle River, NK: Prentice Hall*.

El Aidy, S., T. G. Dinan and J. F. Cryan (2015). 'Gut microbiota: the conductor in the orchestra of immune–neuroendocrine communication'. In: *Clinical therapeutics* 37.5, pp. 954–967.

El Aidy, S., T. Dinan and J. Cryan (2014). 'Immune modulation of the brain-gut-microbe axis'. In: *Frontiers in Microbiology* 5, p. 146. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00146. URL: http://journal.frontiersin.org/article/10.3389/fmicb.2014.00146.

Elson, C. O., Y. Cong, V. J. McCracken, R. A. Dimmitt, R. G. Lorenz and C. T. Weaver (2005). 'Experimental models of inflammatory bowel disease

reveal innate, adaptive, and regulatory mechanisms of host dialogue with the microbiota'. In: *Immunological reviews* 206.1, pp. 260–276.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison and M. L. Sogin (2013). 'Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data'. In: *Methods in Ecology and Evolution* 4.12, pp. 1111–1119.

Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis and M. L. Sogin (2015). 'Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences'. In: *The ISME journal* 9.4, pp. 968–979.

Fadrosh, D. W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman and J. Ravel (2014). 'An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform'. In: *Microbiome* 2.1, p. 6.

Faith, D. P. (1992). 'Conservation evaluation and phylogenetic diversity'. In: *Biological conservation* 61.1, pp. 1–10.

Fan, X., A. V. Alekseyenko, J. Wu, B. A. Peters, E. J. Jacobs, S. M. Gapstur, M. P. Purdue, C. C. Abnet, R. Stolzenberg-Solomon, G. Miller et al. (2016). 'Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study'. In: *Gut*, gutjnl–2016.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). 'From data mining to knowledge discovery in databases'. In: *AI magazine* 17.3, p. 37.

Fedoroff, J. P., S. E. Starkstein et al. (1991). 'Are depressive symptoms nonspecific in patients with acute stroke?' In: *The American journal of psychiatry* 148.9, p. 1172.

Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell and G. B. Gloor (2014). 'Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis'. In: *Microbiome* 2.1, p. 15.

Ferrari, A. J., F. J. Charlson, R. E. Norman, S. B. Patten, G. Freedman, C. J. Murray, T. Vos and H. A. Whiteford (2013). 'Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010'. In: *PLoS Med* 10.11, e1001547.

Fisher, R. A. (1936). 'The use of multiple measurements in taxonomic problems'. In: *Annals of human genetics* 7.2, pp. 179–188.

Forney, L. J., J. A. Foster and W. Ledger (2006). 'The vaginal flora of healthy women is not always dominated by Lactobacillus species'. In: *Journal of Infectious Diseases* 194.10, pp. 1468–1469.

Forslund, K., F. Hildebrand, T. Nielsen, G. Falony, E. Le Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H. K. Pedersen et al. (2015).

'Disentangling the effects of type 2 diabetes and metformin on the human gut microbiota'. In: *Nature* 528.7581, p. 262.

Foster, J. A. and K.-A. M. Neufeld (2013). 'Gut–brain axis: how the microbiome influences anxiety and depression'. In: *Trends in neurosciences* 36.5, pp. 305–312.

Foster, J. A., L. Rinaman and J. F. Cryan (2017). 'Stress & the gut-brain axis: Regulation by the microbiome'. In: *Neurobiology of stress*.

Freedland, K. E., P. J. Lustman, R. M. Carney and B. A. Hong (1992). 'Underdiagnosis of depression in patients with coronary artery disease: the role of nonspecific symptoms'. In: *The International Journal of Psychiatry in Medicine* 22.3, pp. 221–229.

Freund, Y. and R. E. Schapire (1997). 'A decision-theoretic generalization of on-line learning and an application to boosting'. In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Friedman, J. and E. J. Alm (2012). 'Inferring correlation networks from genomic survey data'. In: *PLoS Comput Biol* 8.9, e1002687.

Fry, M. and A. V. Ferguson (2007). 'The sensory circumventricular organs: brain targets for circulating signals controlling ingestive behavior'. In: *Physiology & behavior* 91.4, pp. 413–423.

Fulcher, J. (2008). 'Computational intelligence: an introduction'. In: *Computational intelligence: a compendium*. Springer, pp. 3–78.

Gangadhar, B., J. Ancy, N. Janakiranaiah and C. Umapathy (1993). 'P300 amplitude in non-bipolar, melancholic depression'. In: *Journal of affective disorders* 28.1, pp. 57–60.

Gee, J. E., B. K. De, P. N. Levett, A. M. Whitney, R. T. Novak and T. Popovic (2004). 'Use of 16S rRNA gene sequencing for rapid confirmatory identification of Brucella isolates'. In: *Journal of clinical microbiology* 42.8, pp. 3649–3654.

Gevers, D., S. Kugathasan, L. A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour et al. (2014). 'The treatment-naive microbiome in new-onset Crohn's disease'. In: *Cell host & microbe* 15.3, pp. 382–392.

Glenn, T. C. (2011). 'Field guide to next-generation DNA sequencers'. In: *Molecular ecology resources* 11.5, pp. 759–769.

Gloor, G. B. and G. Reid (2016). 'Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data'. In: *Canadian journal of microbiology* 62.8, pp. 692–703.

Glymenaki, M., G. Singh, A. Brass, G. Warhurst, A. J. McBain, K. J. Else and S. M. Cruickshank (2017). 'Compositional changes in the gut mucus microbiota precede the onset of colitis-induced inflammation'. In: *Inflammatory bowel diseases* 23.6, pp. 912–922.

Gonçalves, L. F., D. Fermiano, M. Feres, L. C. Figueiredo, F. R. Teles, M. P. Mayer and M. Faveri (2012). 'Levels of Selenomonas species in generalized aggressive periodontitis'. In: *Journal of periodontal research* 47.6, pp. 711–718.

Goodfellow, M., G. Manfio and J. Chun (1997). 'Towards a practical species concept for cultivable bacteria'. In:

Grosan, C. and A. Abraham (2011). 'Rule-based expert systems'. In: *Intelligent Systems*, pp. 149–185.

Guyon, I. and A. Elisseeff (2003). 'An introduction to variable and feature selection'. In: *Journal of machine learning research* 3, pp. 1157–1182.

Guyon, I. and A. Elisseeff (2006). 'An introduction to feature extraction'. In: *Feature extraction*. Springer, pp. 1–25.

Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). 'Gene selection for cancer classification using support vector machines'. In: *Machine learning* 46.1, pp. 389–422.

Hajjar, E. R., A. C. Cafiero and J. T. Hanlon (2007). 'Polypharmacy in elderly patients'. In: *The American journal of geriatric pharmacotherapy* 5.4, pp. 345–351.

Halfvarson, J., C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez et al. (2017). 'Dynamics of the human gut microbiome in inflammatory bowel disease'. In: *Nature microbiology* 2.5, p. 17004.

Hall, M. A. and L. A. Smith (1998). 'Practical feature subset selection for machine learning'. In:

Hamilton, M. (1960). 'A rating scale for depression'. In: *Journal of Neurology, Neurosurgery & Psychiatry* 23.1, pp. 56–62.

Hanauer, S. B. (2006). 'Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities'. In: *Inflammatory bowel diseases* 12.5, S3–S9.

Hardham, J. M., K. W. King, K. Dreier, J. Wong, C. Strietzel, R. R. Eversole, C. Sfintescu and R. T. Evans (2008). 'Transfer of Bacteroides splanchnicus to Odoribacter gen. nov. as Odoribacter splanchnicus comb. nov., and description of Odoribacter denticanis sp. nov., isolated from the crevicular spaces of canine periodontitis patients'. In: *International journal of systematic and evolutionary microbiology* 58.1, pp. 103–109.

Hassanien, A.-E., M. G. Milanova, T. G. Smolinski and A. Abraham (2008). 'Computational intelligence in solving bioinformatics problems: Reviews, perspectives, and challenges'. In: *Computational Intelligence in Biomedicine and Bioinformatics*. Springer, pp. 3–47.

Hastie, T., R. Tibshirani and J. Friedman (2009). 'Overview of Supervised Learning'. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, pp. 9–41. ISBN: 978-0-387-84858-7. DOI:

10.1007/978-0-387-84858-7_2. URL: http://dx.doi.org/10.1007/978-0-387-84858-7_2.

He, H. and E. A. Garcia (2009). 'Learning from imbalanced data'. In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

He, Y., J. G. Caporaso, X.-T. Jiang, H.-F. Sheng, S. M. Huse, J. R. Rideout, R. C. Edgar, E. Kopylova, W. A. Walters, R. Knight et al. (2015). 'Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity'. In: *Microbiome* 3.1, p. 20.

Herculano-Houzel, S. (2009). 'The human brain in numbers: a linearly scaled-up primate brain'. In: *Frontiers in human neuroscience* 3.

Heuser, I., A. Yassouridis and F. Holsboer (1994). 'The combined dexamethasone/CRH test: a refined laboratory test for psychiatric disorders'. In: *Journal of psychiatric research* 28.4, pp. 341–356.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al. (2012). 'Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups'. In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97.

Hira, Z. M. and D. F. Gillies (2015). 'A review of feature selection and feature extraction methods applied on microarray data'. In: *Advances in bioinformatics* 2015.

Hirschfeld, R. (2000). 'History and evolution of the monoamine hypothesis of depression.' In: *The Journal of clinical psychiatry*.

Hochberg, Y. and Y. Benjamini (1990). 'More powerful procedures for multiple significance testing'. In: *Statistics in medicine* 9.7, pp. 811–818.

Hodes, G. E., V. Kana, C. Menard, M. Merad and S. J. Russo (2015). 'Neuroimmune mechanisms of depression'. In: *Nature neuroscience* 18.10, p. 1386.

Holte, R. C. (1993). 'Very simple classification rules perform well on most commonly used datasets'. In: *Machine learning* 11.1, pp. 63–90.

Holzinger, A., M. Dehmer and I. Jurisica (2014). 'Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions'. In: *BMC bioinformatics* 15.6, p. I1.

Hooper, L. V., D. R. Littman and A. J. Macpherson (2012). 'Interactions between the microbiota and the immune system'. In: *Science* 336.6086, pp. 1268–1273.

Horner-Devine, M. C., J. M. Silver, M. A. Leibold, B. J. Bohannan, R. K. Colwell, J. A. Fuhrman, J. L. Green, C. R. Kuske, J. B. Martiny, G. Muyzer et al. (2007). 'A comparison of taxon co-occurrence patterns for macro-and microorganisms'. In: *Ecology* 88.6, pp. 1345–1353.

Hornik, K., M. Stinchcombe and H. White (1989). 'Multilayer feedforward networks are universal approximators'. In: *Neural networks* 2.5, pp. 359–366.

Hourigan, S., L. Chen, Z. Grigoryan, G. Laroche, M. Weidner, C. Sears and M. Oliva-Hemker (2015). 'Microbiome changes associated with sustained eradication of Clostridium difficile after single faecal microbiota transplantation in children with and without inflammatory bowel disease'. In: *Alimentary pharmacology & therapeutics* 42.6, pp. 741–752.

Hoyer, B. H., N. Van de Velde, M. Goodman and R. Roberts (1972). 'Examination of hominid evolution by DNA sequence homology'. In: *Journal of Human Evolution* 1.6, pp. 645–649.

Huang, Y., B. Niu, Y. Gao, L. Fu and W. Li (2010). 'CD-HIT Suite: a web server for clustering and comparing biological sequences'. In: *Bioinformatics* 26.5, pp. 680–682.

Hugenholtz, P., B. M. Goebel and N. R. Pace (1998). 'Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity'. In: *Journal of bacteriology* 180.18, pp. 4765–4774.

Human Microbiome Project Consortium (2012). 'A framework for human microbiome research'. In: *Nature* 486.7402, pp. 215–221.

Humphrey, S. P. and R. T. Williamson (2001). 'A review of saliva: normal composition, flow, and function'. In: *Journal of Prosthetic Dentistry* 85.2, pp. 162–169.

Hunter, L. (1993). 'Molecular biology for computer scientists'. In: *Artificial intelligence and molecular biology*, pp. 1–46.

Huse, S. M., D. M. Welch, H. G. Morrison and M. L. Sogin (2010). 'Ironing out the wrinkles in the rare biosphere through improved OTU clustering'. In: *Environmental microbiology* 12.7, pp. 1889–1898.

IHMS Consortium (2015a). *IHMS_SOP 01 V1: Standard Operating Procedure for sample identification*. Accessed: 2017-05-12. URL: http://microbiome-standards.org/fileadmin/SOPs/IHMS_SOP_01_V2.pdf.

— (2015b). *IHMS_SOP 06 V1: Standard Operating Procedure for faecal samples DNA extraction, Protocol Q*. Accessed: 2017-05-12. URL: http://microbiome-standards.org/fileadmin/SOPs/IHMS_SOP_06_V2.pdf.

Jain, A. K., J. Mao and K. M. Mohiuddin (1996). 'Artificial neural networks: A tutorial'. In: *Computer* 29.3, pp. 31–44.

Jain, R., M. C. Rivera and J. A. Lake (1999). 'Horizontal gene transfer among genomes: the complexity hypothesis'. In: *Proceedings of the National Academy of Sciences* 96.7, pp. 3801–3806.

James, G., D. Witten and T. Hastie (2014). *An Introduction to Statistical Learning: With Applications in R*.

James, G., D. Witten, T. Hastie and R. Tibshirani (2013). 'Linear Model Selection and Regularization'. In: *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer New York, pp. 203–264. DOI: 10.1007/978-

1-4614-7138-7_6. URL: http://dx.doi.org/10.1007/978-1-4614-7138-7_6.

Japkowicz, N. and S. Stephen (2002). 'The class imbalance problem: A systematic study'. In: *Intelligent data analysis* 6.5, pp. 429–449.

Jensen, R. and Q. Shen (2008). *Computational intelligence and feature selection: rough and fuzzy approaches*. Vol. 8. John Wiley & Sons.

Jian, X., L. Fu, H. H. Tao and W. Haiwei (2016). 'Rough reduction algorithm for reduction of metagenomic DNA digital signature'. In: *Control and Decision Conference (CCDC), 2016 Chinese*. IEEE, pp. 5545–5550.

Jian, X., L. Fu and H. Tao (2015). 'The reduction and classification research on DNA fragment species attributes in meta genome'. In: *Control and Decision Conference (CCDC), 2015 27th Chinese*. IEEE, pp. 4766–4771.

Jiang, H., Z. Ling, Y. Zhang, H. Mao, Z. Ma, Y. Yin, W. Wang, W. Tang, Z. Tan, J. Shi et al. (2015). 'Altered fecal microbiota composition in patients with major depressive disorder'. In: *Brain, behavior, and immunity* 48, pp. 186–194.

Jiang, Y., W. T. Clark, I. Friedberg and P. Radivojac (2014). 'The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective'. In: *Bioinformatics* 30.17, pp. i609–i616.

Johansson, M. E., M. Phillipson, J. Petersson, A. Velcich, L. Holm and G. C. Hansson (2008). 'The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria'. In: *Proceedings of the national academy of sciences* 105.39, pp. 15064–15069.

John, G. H., R. Kohavi, K. Pfleger et al. (1994). 'Irrelevant features and the subset selection problem'. In: *Machine learning: proceedings of the eleventh international conference*, pp. 121–129.

Johnson, A. (1983). 'The pathogenic potential of commensal species of Neisseria'. In: *Journal of clinical pathology* 36.2, pp. 213–223.

Kalousis, A., J. Prados and M. Hilario (2007). 'Stability of feature selection algorithms: a study on high-dimensional spaces'. In: *Knowledge and information systems* 12.1, pp. 95–116.

Kamada, N., S.-U. Seo, G. Y. Chen and G. Núñez (2013). 'Role of the gut microbiota in immunity and inflammatory disease'. In: *Nature Reviews Immunology* 13.5, p. 321.

Kans, J. (2013). *Entrez Direct: E-utilities on the UNIX Command Line*. Available from: https://www.ncbi.nlm.nih.gov/books/NBK179288/, accessed: 2017-06-29.

Kazor, C., P. Mitchell, A. Lee, L. Stokes, W. Loesche, F. Dewhirst and B. Paster (2003). 'Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients'. In: *Journal of clinical microbiology* 41.2, pp. 558–563.

Kelly, J. R., A. P. Allen, A. Temko, W. Hutch, P. J. Kennedy, N. Farid, E. Murphy, G. Boylan, J. Bienenstock, J. F. Cryan et al. (2017). 'Lost in translation? The potential psychobiotic Lactobacillus rhamnosus (JB-1) fails to modulate stress or cognitive performance in healthy male subjects'. In: *Brain, behavior, and immunity* 61, pp. 50–59.

Kelly, J. R., G. Clarke, J. F. Cryan and T. G. Dinan (2016). 'Brain-gut-microbiota axis: challenges for translation in psychiatry'. In: *Annals of epidemiology* 26.5, pp. 366–372.

Kelly, J. R., P. J. Kennedy, J. F. Cryan, T. G. Dinan, G. Clarke and N. P. Hyland (2015). 'Breaking down the barriers: the gut microbiome, intestinal permeability and stress-related psychiatric disorders'. In: *Frontiers in cellular neuroscience* 9.

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg and C. O. Webb (2010). 'Picante: R tools for integrating phylogenies and ecology'. In: *Bioinformatics* 26.11, pp. 1463–1464.

Kembel, S. W., M. Wu, J. A. Eisen and J. L. Green (2012). 'Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance'. In: *PLoS Comput Biol* 8.10, e1002743.

Kempton, M. J., Z. Salvador, M. R. Munafò, J. R. Geddes, A. Simmons, S. Frangou and S. C. Williams (2011). 'Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder'. In: *Archives of general psychiatry* 68.7, pp. 675–690.

Keogh, E. and A. Mueen (2011). 'Curse of dimensionality'. In: *Encyclopedia of Machine Learning*. Springer, pp. 257–258.

Kerber, R. (1992). 'Chimerge: Discretization of numeric attributes'. In: *Proceedings of the tenth national conference on Artificial intelligence*. Aaai Press, pp. 123–128.

Kim, M., M. Morrison and Z. Yu (2011). 'Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes'. In: *Journal of microbiological methods* 84.1, pp. 81–87.

Kirsch, I., B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore and B. T. Johnson (2008). 'Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration'. In: *PLoS Med* 5.2, e45.

Klappenbach, J. A., J. M. Dunbar and T. M. Schmidt (2000). 'rRNA operon copy number reflects ecological strategies of bacteria'. In: *Applied and environmental microbiology* 66.4, pp. 1328–1333.

Kohonen, T. (1990). 'The self-organizing map'. In: *Proceedings of the IEEE* 78.9, pp. 1464–1480.

— (1998). 'The self-organizing map'. In: *Neurocomputing* 21.1, pp. 1–6.

Kolenbrander, P. E., R. J. Palmer Jr, S. Periasamy and N. S. Jakubovics (2010). 'Oral multispecies biofilm development and the key role of cell–cell distance'. In: *Nature Reviews Microbiology* 8.7, p. 471.

Kononenko, I. (2001). 'Machine learning for medical diagnosis: history, state of the art and perspective'. In: *Artificial Intelligence in medicine* 23.1, pp. 89–109.

Kopylova, E., J. A. Navas-Molina, C. Mercier, Z. Z. Xu, F. Mahé, Y. He, H.-W. Zhou, T. Rognes, J. G. Caporaso and R. Knight (2016). 'Open-source sequence clustering methods improve the state of the art'. In: *mSystems* 1.1, e00003–15.

Kostic, A. D., D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero et al. (2012). 'Genomic analysis identifies association of Fusobacterium with colorectal carcinoma'. In: *Genome research* 22.2, pp. 292–298.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). 'Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform'. In: *Applied and environmental microbiology* 79.17, pp. 5112–5120.

Krachunov, M., D. Vassilev, M. Nisheva, O. Kulev, V. Simeonova and V. Dimitrov (2015). 'Fuzzy Indication of Reliability in Metagenomics NGS Data Analysis'. In: *Procedia Computer Science* 51, pp. 2859–2863.

Kreth, J., J. Merritt and F. Qi (2009). 'Bacterial and host interactions of oral streptococci'. In: *DNA and cell biology* 28.8, pp. 397–403.

Krizhevsky, A. and G. Hinton (2009). 'Learning multiple layers of features from tiny images'. In:

Krizhevsky, A., I. Sutskever and G. E. Hinton (2012). 'Imagenet classification with deep convolutional neural networks'. In: *Advances in neural information processing systems*, pp. 1097–1105.

Kuczynski, J., Z. Liu, C. Lozupone, D. McDonald, N. Fierer and R. Knight (2010). 'Microbial community resemblance methods differ in their ability to detect biologically relevant patterns'. In: *Nature methods* 7.10, pp. 813–819.

Kumar, M., M. Husian, N. Upreti and D. Gupta (2010). 'Genetic algorithm: Review and application'. In: *International Journal of Information Technology and Knowledge Management* 2.2, pp. 451–454.

Kursa, M. B., W. R. Rudnicki et al. (2010). *Feature selection with the Boruta package*.

Lahat, D., T. Adali and C. Jutten (2015). 'Multimodal data fusion: an overview of methods, challenges, and prospects'. In: *Proceedings of the IEEE* 103.9, pp. 1449–1477.

Langhorst, J., S. Elsenbruch, J. Koelzer, A. Rueffer, A. Michalsen and G. J. Dobos (2008). 'Noninvasive markers in the assessment of intestinal inflammation in inflammatory bowel diseases: performance of fecal lactoferrin, calprotectin,

and PMN-elastase, CRP, and clinical indices'. In: *The American journal of gastroenterology* 103.1, p. 162.

Langille, M. G., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight et al. (2013). 'Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences'. In: *Nature biotechnology* 31.9, pp. 814–821.

Larose, D. T. and C. D. Larose (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Larsen, P. E., D. Field and J. A. Gilbert (2012). 'Predicting bacterial community assemblages using an artificial neural network approach'. In: *Nature methods* 9.6, pp. 621–625.

LeCun, Y., Y. Bengio and G. Hinton (2015). 'Deep learning'. In: *Nature* 521.7553, pp. 436–444.

Ledergerber, C. and C. Dessimoz (2011). 'Base-calling for next-generation sequencing platforms'. In: *Briefings in bioinformatics* 12.5, pp. 489–497.

Lee, G., C. Rodriguez and A. Madabhushi (2007). 'An empirical comparison of dimensionality reduction methods for classifying gene and protein expression datasets'. In: *International Symposium on Bioinformatics Research and Applications*. Springer, pp. 170–181.

Leinonen, R., H. Sugawara, M. Shumway and I. N. S. D. Collaboration (2010). 'The sequence read archive'. In: *Nucleic acids research* 39.suppl_1, pp. D19–D21.

Ley, R. E., J. K. Harris, J. Wilcox, J. R. Spear, S. R. Miller, B. M. Bebout, J. A. Maresca, D. A. Bryant, M. L. Sogin and N. R. Pace (2006). 'Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat'. In: *Applied and Environmental Microbiology* 72.5, pp. 3685–3695.

Li, L., Q. Su, B. Xie, L. Duan, W. Zhao, D. Hu, R. Wu and H. Liu (2016). 'Gut microbes in correlation with mood: case study in a closed experimental human life support system'. In: *Neurogastroenterology & Motility* 28.8, pp. 1233–1240.

Liao, S.-H. (2005). 'Expert system methodologies and applications—a decade review from 1995 to 2004'. In: *Expert systems with applications* 28.1, pp. 93–103.

Liao, S.-H., P.-H. Chu and P.-Y. Hsiao (2012). 'Data mining techniques and applications–A decade review from 2000 to 2011'. In: *Expert systems with applications* 39.12, pp. 11303–11311.

Liu, B., Y. Ma and C. K. Wong (2000). 'Improving an association rule based classifier'. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 504–509.

Liu, J. and Y. Duan (2012). 'Saliva: A potential media for disease diagnostics and monitoring'. In: *Oral oncology* 48.7, pp. 569–577.

Loomba, R., V. Seguritan, W. Li, T. Long, N. Klitgord, A. Bhatt, P. S. Dulai, C. Caussy, R. Bettencourt, S. K. Highlander et al. (2017). 'Gut Microbiome-Based

Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease'. In: *Cell Metabolism* 25.5, pp. 1054–1062.

López-Muñoz, F. and C. Alamo (2009). 'Monoaminergic neurotransmission: the history of the discovery of antidepressants from 1950s until today'. In: *Current pharmaceutical design* 15.14, pp. 1563–1586.

Lorenzetti, V., N. B. Allen, A. Fornito and M. Yücel (2009). 'Structural brain abnormalities in major depressive disorder: a selective review of recent MRI studies'. In: *Journal of affective disorders* 117.1, pp. 1–17.

Love, M. I., W. Huber and S. Anders (2014). 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2'. In: *Genome Biology* 15 (12), p. 550. DOI: 10.1186/s13059-014-0550-8.

Luscher, B., Q. Shen and N. Sahir (2011). *The GABAergic deficit hypothesis of major depressive disorder*.

Lynch, M. D. and J. D. Neufeld (2015). 'Ecology and exploration of the rare biosphere'. In: *Nature Reviews Microbiology* 13.4, p. 217.

Lyte, M. (2013). 'Microbial endocrinology in the microbiome-gut-brain axis: how bacterial production and utilization of neurochemicals influence behavior'. In: *PLoS Pathog* 9.11, e1003726.

— (2014). 'Microbial endocrinology and the microbiota-gut-brain axis'. In: *Microbial Endocrinology: The Microbiota-Gut-Brain Axis in Health and Disease*. Springer, pp. 3–24.

Ma, P. C. and K. C. Chan (2007). 'An effective data mining technique for reconstructing gene regulatory networks from time series expression data'. In: *Journal of Bioinformatics and Computational Biology* 5.03, pp. 651–668.

Maes, M., H. Y. Meltzer, E. Bosmans, R. Bergmans, E. Vandoolaeghe, R. Ranjan and R. Desnyder (1995). 'Increased plasma concentrations of interleukin-6, soluble interleukin-6, soluble interleukin-2 and transferrin receptor in major depression'. In: *Journal of affective disorders* 34.4, pp. 301–309.

Maffei, V. (Apr. 2018). *dada2 to PICRUSt*. Accessed: 2018-04-04. URL: https://github.com/vmaffei/dada2_to_picrust.

Mahé, F., T. Rognes, C. Quince, C. de Vargas and M. Dunthorn (2014). 'Swarm: robust and fast clustering method for amplicon-based studies'. In: *PeerJ* 2, e593.

Mamoshina, P., A. Vieira, E. Putin and A. Zhavoronkov (2016). 'Applications of deep learning in biomedicine'. In: *Molecular pharmaceutics* 13.5, pp. 1445–1454.

Marchesi, J. R. and J. Ravel (2015). 'The vocabulary of microbiome research: a proposal'. In: *Microbiome* 3.1, p. 31.

Mardis, E. R. (2008). 'The impact of next-generation sequencing technology on genetics'. In: *Trends in genetics* 24.3, pp. 133–141.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays'. In: *Genome research* 18.9, pp. 1509–1517.

Masci, J., U. Meier, D. Cireşan and J. Schmidhuber (2011). 'Stacked convolutional auto-encoders for hierarchical feature extraction'. In: *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59.

May, A., S. Abeln, W. Crielaard, J. Heringa and B. W. Brandt (2014). 'Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations'. In: *Bioinformatics* 30.11, pp. 1530–1538.

McGrath, C. L., M. E. Kelley, P. E. Holtzheimer, B. W. Dunlop, W. E. Craighead, A. R. Franco, R. C. Craddock and H. S. Mayberg (2013). 'Toward a neuroimaging treatment selection biomarker for major depressive disorder'. In: *JAMA psychiatry* 70.8, pp. 821–829.

McLafferty, M., C. R. Lapsley, E. Ennis, C. Armour, S. Murphy, B. P. Bunting, A. J. Bjourson, E. K. Murray and S. M. O'Neill (2017). 'Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland'. In: *PloS one* 12.12, e0188785.

McManus, S., P. Bebbington, R. Jenkins and T. Brugha (2016). 'Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014'. In: *Leeds: NHS Digital*, pp. 39–40.

McMurdie, P. J. and S. Holmes (2013). 'phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data'. In: *PloS one* 8.4, e61217.

— (2014). 'Waste not, want not: why rarefying microbiome data is inadmissible'. In: *PLoS Comput Biol* 10.4, e1003531.

Meinshausen, N. and P. Bühlmann (2006). 'High-dimensional graphs and variable selection with the lasso'. In: *The annals of statistics*, pp. 1436–1462.

Melssen, W., R. Wehrens and L. Buydens (2006). 'Supervised Kohonen networks for classification problems'. In: *Chemometrics and Intelligent Laboratory Systems* 83.2, pp. 99–113.

Merriman, B., I. Torrent, J. M. Rothberg, R. Team et al. (2012). 'Progress in ion torrent semiconductor chip based sequencing'. In: *Electrophoresis* 33.23, pp. 3397–3417.

MetaHIT Consortium (2011). 'MetaHIT: The European Union Project on metagenomics of the human intestinal tract'. In: *Metagenomics of the human body*. Springer, pp. 307–316.

Meyer, M. and M. Kircher (2010). 'Illumina sequencing library preparation for highly multiplexed target capture and sequencing'. In: *Cold Spring Harbor Protocols* 2010.6, pdb–prot5448.

Mikheyev, A. S. and M. M. Tin (2014). 'A first look at the Oxford Nanopore MinION sequencer'. In: *Molecular ecology resources* 14.6, pp. 1097–1102.

Mirucki, C. S., M. Abedi, J. Jiang, Q. Zhu, Y.-H. Wang, K. E. Safavi, R. B. Clark and F. C. Nichols (2014). 'Biologic activity of porphyromonas endodontalis complex lipids'. In: *Journal of endodontics* 40.9, pp. 1342–1348.

Molcrani, M.-C., F. Duval, M. Crocq, P. Bailey and J. Macher (1997). '{HPA} axis dysfunction in depression: Correlation with monoamine system abnormalities'. In: *Psychoneuroendocrinology* 22, Supplement 1, S63–S68. ISSN: 0306-4530. DOI: https://doi.org/10.1016/S0306-4530(97)00012-7. URL: http://www.sciencedirect.com/science/article/pii/S0306453097000127.

Molodecky, N. A., S. Soon, D. M. Rabi, W. A. Ghali, M. Ferris, G. Chernoff, E. I. Benchimol, R. Panaccione, S. Ghosh, H. W. Barkema et al. (2012). 'Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review'. In: *Gastroenterology* 142.1, pp. 46–54.

Moreno, E. and I. Moriyón (2002). 'Brucella melitensis: a nasty bug with hidden credentials for virulence'. In: *Proceedings of the National Academy of Sciences* 99.1, pp. 1–3.

Morgan, X. C. and C. Huttenhower (2012). 'Human microbiome analysis'. In: *PLoS Comput Biol* 8.12, e1002808.

Morgan, X. C., T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper et al. (2012). 'Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment'. In: *Genome biology* 13.9, R79.

Mössner, R., O. Mikova, E. Koutsilieri, M. Saoud, A.-C. Ehlis, N. Müller, A. J. Fallgatter and P. Riederer (2007). 'Consensus paper of the WFSBP Task Force on Biological Markers: biological markers in depression'. In: *The world journal of biological psychiatry* 8.3, pp. 141–174.

Mrazek, D. A., J. C. Hornberger, C. A. Altar and I. Degtiar (2014). 'A review of the clinical, economic, and societal burden of treatment-resistant depression: 1996–2013'. In: *Psychiatric services* 65.8, pp. 977–987.

Mullis, K. B. (1990). 'The unusual origin of the polymerase chain reaction'. In: *Scientific American* 262.4, pp. 56–65.

Munson, M., T. Pitt-Ford, B. Chong, A. Weightman and W. Wade (2002). 'Molecular and cultural analysis of the microflora associated with endodontic infections'. In: *Journal of dental research* 81.11, pp. 761–766.

Naseribafrouei, A., K. Hestad, E. Avershina, M. Sekelja, A. Linløkken, R. Wilson and K. Rudi (2014). 'Correlation between the human fecal microbiota and depression'. In: *Neurogastroenterology & Motility* 26.8, pp. 1155–1162.

National Collaborating Centre for Mental Health (2010). 'Depression: the treatment and management of depression in adults (updated edition)'. In: British Psychological Society.

Navas-Molina, J. A., J. M. Peralta-Sánchez, A. González, P. J. McMurdie, Y. Vázquez-Baeza, Z. Xu, L. K. Ursell, C. Lauber, H. Zhou, S. J. Song et al. (2013). 'Advancing our understanding of the human microbiome using QIIME'. In: *Methods in enzymology* 531, p. 371.

Ng, S. C., C. N. Bernstein, M. H. Vatn, P. L. Lakatos, E. V. Loftus, C. Tysk, C. O'morain, B. Moum, J.-F. Colombel et al. (2013). 'Geographical variability and environmental risk factors in inflammatory bowel disease'. In: *Gut* 62.4, pp. 630–649.

Noble, W. S. (2006). 'What is a support vector machine?' In: *Nature biotechnology* 24.12, pp. 1565–1567.

— (2009). 'How does multiple testing correction work?' In: *Nature biotechnology* 27.12, p. 1135.

Noriega, L. (2005). 'Multilayer perceptron tutorial'. In: *School of Computing. Staffordshire University*.

Nørsett, K. G., A. Lægreid, H. Midelfart, F. Yadetie, S. E. Erlandsen, S. Falkmer, J. E. Grønbech, H. L. Waldum, J. Komorowski and A. K. Sandvik (2004). 'Gene expression based classification of gastric carcinoma'. In: *Cancer letters* 210.2, pp. 227–237.

Nørskov-Lauritsen, N. (2014). 'Classification, identification, and clinical significance of Haemophilus and Aggregatibacter species with host specificity for humans'. In: *Clinical microbiology reviews* 27.2, pp. 214–240.

Nugent, R. P., M. A. Krohn and S. L. Hillier (1991). 'Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation.' In: *Journal of clinical microbiology* 29.2, pp. 297–301.

O'Mahony, S., G. Clarke, Y. Borre, T. Dinan and J. Cryan (2015). 'Serotonin, tryptophan metabolism and the brain-gut-microbiome axis'. In: *Behavioural brain research* 277, pp. 32–48.

OED Online (2017). Noise. Accessed: 2017-05-21. Oxford University Press.

Oksanen, J., R. Kindt, P. Legendre, B. O'Hara, M. Stevens, M. Oksanen and M. Suggests (2015). *Vegan community ecology package: ordination methods, diversity analysis and other functions for community and vegetation ecologists. Version 2.3-1*.

Oksanen, J., R. Kindt, P. Legendre, B. O'Hara, M. H. H. Stevens, M. J. Oksanen and M. Suggests (2007). 'The vegan package'. In: *Community ecology package* 10, pp. 631–637.

Oliveros, J. (2015). *VENNY. An interactive tool for comparing lists with Venn Diagrams. 2007*.

Owens, M., J. Herbert, P. B. Jones, B. J. Sahakian, P. O. Wilkinson, V. J. Dunn, T. J. Croudace and I. M. Goodyer (2014). 'Elevated morning cortisol is a stratified population-level biomarker for major depression in boys only with high depressive symptoms'. In: *Proceedings of the National Academy of Sciences* 111.9, pp. 3638–3643.

Papa, E., M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram et al. (2012). 'Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease'. In: *PloS one* 7.6, e39242.

Papakostas, G. I., M. Fava and M. E. Thase (2008). 'Treatment of SSRI-resistant depression: a meta-analysis comparing within-versus across-class switches'. In: *Biological psychiatry* 63.7, pp. 699–704.

Parada, A. E., D. M. Needham and J. A. Fuhrman (2015). 'Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples'. In: *Environmental microbiology*.

Patterson, D. W. (1998). *Artificial neural networks: theory and applications*. Prentice Hall PTR.

Paulson, J. N., O. C. Stine, H. C. Bravo and M. Pop (2013). 'Differential abundance analysis for microbial marker-gene surveys'. In: *Nature methods* 10.12, pp. 1200–1202.

Pawlak, Z. (1996). 'Why rough sets?' In: *Proceedings of IEEE 5th International Fuzzy Systems*. Vol. 2. IEEE, pp. 738–743.

— (1998). 'Rough set theory and its applications to data analysis'. In: *Cybernetics & Systems* 29.7, pp. 661–688.

Pawlak, Z. (2012). *Rough sets: Theoretical aspects of reasoning about data*. Vol. 9. Springer Science & Business Media.

Perry, A. and P. Lambert (2011). 'Propionibacterium acnes: infection beyond the skin'. In: *Expert review of anti-infective therapy* 9.12, pp. 1149–1156.

Petit, J., N. Meurice, J. L. Medina-Franco and G. M Maggiora (2014). 'A rough set theory approach to the analysis of gene expression profiles'. In: *Chemoinformatics for Drug Discovery*, pp. 51–83.

Phipson, B. and G. K. Smyth (2010). 'Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn'. In: *Statistical applications in genetics and molecular biology* 9.1.

Pinheiro, H. P., A. de Souza Pinheiro and P. K. Sen (2005). 'Comparison of genomic sequences using the Hamming distance'. In: *Journal of Statistical Planning and Inference* 130.1, pp. 325–339.

Preston, C. M., K. Y. Wu, T. F. Molinski and E. F. DeLong (1996). 'A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov'. In: *Proceedings of the National Academy of Sciences* 93.13, pp. 6241–6246.

Prideaux, L., S. Kang, J. Wagner, M. Buckley, J. E. Mahar, P. De Cruz, Z. Wen, L. Chen, B. Xia, D. R. van Langenberg et al. (2013). 'Impact of ethnicity, geography, and disease on the microbiota in health and inflammatory bowel disease'. In: *Inflammatory bowel diseases* 19.13, pp. 2906–2918.

Prosberg, M., F. Bendtsen, I. Vind, A. M. Petersen and L. L. Gluud (2016). 'The association between the gut microbiota and the inflammatory bowel disease activity: a systematic review and meta-analysis'. In: *Scandinavian Journal of Gastroenterology* 51.12, pp. 1407–1415.

Qi, Y. (2012). 'Random forest for bioinformatics'. In: *Ensemble machine learning*. Springer, pp. 307–323.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen et al. (2012). 'A metagenome-wide association study of gut microbiota in type 2 diabetes'. In: *Nature* 490.7418, p. 55.

Quail, M. A., I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow and D. J. Turner (2008). 'A large genome center's improvements to the Illumina sequencing system'. In: *Nature methods* 5.12, p. 1005.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012). 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers'. In: *BMC genomics* 13.1, p. 341.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies and F. O. Glöckner (2012). 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools'. In: *Nucleic acids research* 41.D1, pp. D590–D596.

Ralston, A. (2008). 'Operons and prokaryotic gene regulation'. In: *Nature Education* 1.1, p. 216.

Resh, V. H. and R. T. Cardé (2009). *Encyclopedia of insects*. Academic Press.

Ricanek, P., S. M. Lothe, S. A. Frye, A. Rydning, M. H. Vatn and T. Tønjum (2012). 'Gut bacterial profile in patients newly diagnosed with treatment-naive Crohn's disease'. In: *Clinical and experimental gastroenterology* 5, p. 173.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka et al. (2014). 'Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences'. In: *PeerJ* 2, e545.

Riesenfeld, C. S., P. D. Schloss and J. Handelsman (2004). 'Metagenomics: genomic analysis of microbial communities'. In: *Annu. Rev. Genet.* 38, pp. 525–552.

Rifkin, R. and A. Klautau (2004). 'In defense of one-vs-all classification'. In: *Journal of machine learning research* 5.Jan, pp. 101–141.

Ritter, K., J. Schumacher, M. Weygandt, R. Buchert, C. Allefeld, J.-D. Haynes, A. D. N. Initiative et al. (2015). 'Multimodal prediction of conversion to

Alzheimer's disease based on incomplete biomarkers'. In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1.2, pp. 206–215.

Riza, L. S., A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Śle¸zak and J. M. Benítez (2014). 'Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets"'. In: *Information Sciences* 287, pp. 68–89. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2014.07.029. URL: http://www.sciencedirect.com/science/article/pii/S0020025514007294.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data'. In: *Bioinformatics* 26.1, pp. 139–140.

Roesch, L. F., R. R. Fulthorpe, A. Riva, G. Casella, A. K. Hadwin, A. D. Kent, S. H. Daroub, F. A. Camargo, W. G. Farmerie and E. W. Triplett (2007). 'Pyrosequencing enumerates and contrasts soil microbial diversity'. In: *The ISME journal* 1.4, pp. 283–290.

Rognes, T., T. Flouri, B. Nichols, C. Quince and F. Mahé (2016). 'VSEARCH: a versatile open source tool for metagenomics'. In: *PeerJ* 4, e2584.

Rosen, M. J., B. J. Callahan, D. S. Fisher and S. P. Holmes (2012). 'Denoising PCR-amplified metagenome data'. In: *BMC bioinformatics* 13.1, p. 283.

Sacchi, C. T., A. M. Whitney, L. W. Mayer, R. Morey, A. Steigerwalt, A. Boras, R. S. Weyant and T. Popovic (2002). 'Sequencing of 16S rRNA gene: a rapid tool for identification of Bacillus anthracis'. In: *Emerging infectious diseases* 8.10, pp. 1117–23.

Sadava, D. E., D. M. Hillis, H. C. Heller and M. Berenbaum (2009). *Life: the science of biology*. Vol. 2. Macmillan.

Saeys, Y., T. Abeel and Y. Van de Peer (2008). 'Robust feature selection using ensemble feature selection techniques'. In: *Machine learning and knowledge discovery in databases*, pp. 313–325.

Saeys, Y., I. Inza and P. Larrañaga (2007). 'A review of feature selection techniques in bioinformatics'. In: *bioinformatics* 23.19, pp. 2507–2517.

Safavian, S. R. and D. Landgrebe (1991). 'A survey of decision tree classifier methodology'. In: *IEEE transactions on systems, man, and cybernetics* 21.3, pp. 660–674.

Said, H. S., W. Suda, S. Nakagome, H. Chinen, K. Oshima, S. Kim, R. Kimura, A. Iraha, H. Ishida, J. Fujita et al. (2013). 'Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers'. In: *DNA research* 21.1, pp. 15–25.

Sakon, H., F. Nagai, M. Morotomi and R. Tanaka (2008). 'Sutterella parvirubra sp. nov. and Megamonas funiformis sp. nov., isolated from human faeces'. In:

International journal of systematic and evolutionary microbiology 58.4, pp. 970–975.

Sandberg, S., S. Järvenpää, A. Penttinen, J. Paton and D. C. McCann (2004). 'Asthma exacerbations in children immediately following stressful life events: a Cox's hierarchical regression'. In: *Thorax* 59.12, pp. 1046–1051.

Sapolsky, R. M., L. M. Romero and A. U. Munck (2000). 'How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions 1'. In: *Endocrine reviews* 21.1, pp. 55–89.

Sartor, R. B. (2006). 'Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis'. In: *Nature clinical practice Gastroenterology & hepatology* 3.7, pp. 390–407.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan and C. Quince (2015). 'Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform'. In: *Nucleic acids research*, gku1341.

Schleidgen, S., C. Klingler, T. Bertram, W. H. Rogowski and G. Marckmann (2013). 'What is personalized medicine: sharpening a vague term based on a systematic literature review'. In: *BMC medical ethics* 14.1, p. 55.

Schliep, K. P. (2010). 'phangorn: phylogenetic analysis in R'. In: *Bioinformatics*, btq706.

Schloss, P. D., A. M. Schubert, J. P. Zackular, K. D. Iverson, V. B. Young and J. F. Petrosino (2012). 'Stabilization of the murine gut microbiome following weaning'. In: *Gut microbes* 3.4, pp. 383–393.

Schloss, P. D. and S. L. Westcott (2011). 'Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis'. In: *Applied and environmental microbiology* 77.10, pp. 3219–3226.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson et al. (2009). 'Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities'. In: *Applied and environmental microbiology* 75.23, pp. 7537–7541.

Schmidhuber, J. (2015). 'Deep learning in neural networks: An overview'. In: *Neural networks* 61, pp. 85–117.

Schnell, I. B., K. Bohmann and M. T. P. Gilbert (2015). 'Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies'. In: *Molecular ecology resources* 15.6, pp. 1289–1303.

Scholkopf, B. and A. J. Smola (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schüffler, P. J., D. Mahapatra, J. A. Tielbeek, F. M. Vos, J. Makanyanga, D. A. Pendsé, C. Y. Nio, J. Stoker, S. A. Taylor and J. M. Buhmann (2013). 'A model development pipeline for Crohn's disease severity assessment from magnetic

resonance images'. In: *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*. Springer, pp. 1–10.

Segata, N., J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett and C. Huttenhower (2011). 'Metagenomic biomarker discovery and explanation'. In: *Genome biology* 12.6, R60.

Sender, R., S. Fuchs and R. Milo (2016). 'Revised estimates for the number of human and bacteria cells in the body'. In: *PLoS Biol* 14.8, e1002533.

Shade, A. (2017). 'Diversity is the question, not the answer'. In: *The ISME journal* 11.1, p. 1.

Shapiro, B. J. and M. F. Polz (2014). 'Ordering microbial diversity into ecologically and genetically cohesive units'. In: *Trends in microbiology* 22.5, pp. 235–247.

Shen, Q. and A. Chouchoulas (2000). 'A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems'. In: *Engineering Applications of Artificial Intelligence* 13.3, pp. 263–278.

Shimizu, E., K. Hashimoto, N. Okamura, K. Koike, N. Komatsu, C. Kumakiri, M. Nakazato, H. Watanabe, N. Shinoda, S.-i. Okada et al. (2003). 'Alterations of serum levels of brain-derived neurotrophic factor (BDNF) in depressed patients with or without antidepressants'. In: *Biological psychiatry* 54.1, pp. 70–75.

Shoemark, D. K. and S. J. Allen (2015). 'The microbiome and disease: reviewing the links between the oral microbiome, aging, and Alzheimer's disease'. In: *Journal of Alzheimer's Disease* 43.3, pp. 725–738.

Shreiner, A. B., J. Y. Kao and V. B. Young (2015). 'The gut microbiome in health and in disease'. In: *Current opinion in gastroenterology* 31.1, p. 69.

Shyn, S. I. and S. P. Hamilton (2010). 'The genetics of major depression: moving beyond the monoamine hypothesis'. In: *Psychiatric Clinics of North America* 33.1, pp. 125–140.

Silverberg, M. S., J. Satsangi, T. Ahmad, I. D. Arnott, C. N. Bernstein, S. R. Brant, R. Caprilli, J.-F. Colombel, C. Gasche, K. Geboes et al. (2005). 'Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology'. In: *Canadian Journal of Gastroenterology and Hepatology* 19.Suppl A, 5A–36A.

Simpson, E. H. (1949). 'Measurement of diversity.' In: *Nature*.

Sinha, R., J. Chen, A. Amir, E. Vogtmann, J. Shi, K. S. Inman, R. Flores, J. Sampson, R. Knight and N. Chia (2016). 'Collecting fecal samples for microbiome analyses in epidemiology studies'. In: *Cancer Epidemiology and Prevention Biomarkers* 25.2, pp. 407–416.

Skowron, A. and C. Rauszer (1992). 'The discernibility matrices and functions in information systems'. In: *Intelligent Decision Support*. Springer, pp. 331–362.

Smith, G. D., Y. Ben-Shlomo, A. Beswick, J. Yarnell, S. Lightman and P. Elwood (2005). 'Cortisol, testosterone, and coronary heart disease'. In: *Circulation* 112.3, pp. 332–340.

Smith, G. D. and S. Ebrahim (2002). 'Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers'. In: *BMJ: British Medical Journal* 325.7378, p. 1437.

Smyth, R., T. Schlub, A. Grimm, V. Venturi, A. Chopra, S. Mallal, M. Davenport and J. Mak (2010). 'Reducing chimera formation during PCR amplification to ensure accurate genotyping'. In: *Gene* 469.1, pp. 45–51.

Sneed, J. R. and M. E. Culang-Reinlieb (2011). 'The vascular depression hypothesis: an update'. In: *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry* 19.2, p. 99.

Sokal, R. R. and P. H. Sneath (1963). *Principles of numerical taxonomy.* San Francisco and London, WH Freeman & Co.

Sokol, H., B. Pigneur, L. Watterlot, O. Lakhdari, L. G. Bermúdez-Humarán, J.-J. Gratadoux, S. Blugeon, C. Bridonneau, J.-P. Furet, G. Corthier et al. (2008). 'Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients'. In: *Proceedings of the National Academy of Sciences* 105.43, pp. 16731–16736.

Staley, J. T. and A. Konopka (1985). 'Measurement of in situ activities of non-photosynthetic microorganisms in aquatic and terrestrial habitats'. In: *Annual Reviews in Microbiology* 39.1, pp. 321–346.

Statnikov, A. and C. F. Aliferis (2007). 'Are random forests better than support vector machines for microarray-based cancer classification?' In: *AMIA annual symposium proceedings*. Vol. 2007. American Medical Informatics Association, p. 686.

Statnikov, A., M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. J. Blaser, C. F. Aliferis and A. V. Alekseyenko (2013). 'A comprehensive evaluation of multicategory classification methods for microbiomic data'. In: *Microbiome* 1.1, p. 11.

Statnikov, A., L. Wang and C. F. Aliferis (2008). 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification'. In: *BMC bioinformatics* 9.1, p. 319.

Stauffer, R. C. (1957). 'Haeckel, Darwin, and ecology'. In: *The Quarterly Review of Biology* 32.2, pp. 138–144.

Stein, D. J., M. F. Vasconcelos, L. Albrechet-Souza, K. M. M. Ceresér and R. M. M. De Almeida (2017). 'Microglial Over-activation by Social Defeat Stress Contributes to Anxiety-and Depressive-like Behaviours'. In: *Frontiers in behavioral neuroscience* 11, p. 207.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha and G. E. Robinson (2015). 'Big data: astronomical or genomical?' In: *PLoS biology* 13.7, e1002195.

Stewart, E. J. (2012). 'Growing unculturable bacteria'. In: *Journal of bacteriology* 194.16, pp. 4151–4160.

Stilling, R. M., T. G. Dinan and J. F. Cryan (2014). 'Microbial genes, brain & behaviour–epigenetic regulation of the gut–brain axis'. In: *Genes, Brain and Behavior* 13.1, pp. 69–86.

Stingu, C.-S., R. Schaumann, H. Jentsch, K. Eschrich, O. Brosteanu and A. C. Rodloff (2013). 'Association of periodontitis with increased colonization by Prevotella nigrescens'. In: *Journal of investigative and clinical dentistry* 4.1, pp. 20–25.

Stocker, S., R. Snajder, J. Rainer, S. Trajanoski, G. Gorkiewicz, Z. Trajanoski and G. G. Thallinger (2011). 'SnoWMAn: High-throughput phylotyping, analysis and comparison of microbial communities'. In: *Under revision*.

Stoddard, S. F., B. J. Smith, R. Hein, B. R. Roller and T. M. Schmidt (2014). 'rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development'. In: *Nucleic Acids Research* 43.D1, pp. D593–D598.

— (2015). 'rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development'. In: *Nucleic Acids Research* 43.D1, pp. D593–D598.

Strauss, J., G. G. Kaplan, P. L. Beck, K. Rioux, R. Panaccione, R. DeVinney, T. Lynch and E. Allen-Vercoe (2011). 'Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host'. In: *Inflammatory bowel diseases* 17.9, pp. 1971–1978.

Swann, A. C., C. L. Bowden, J. R. Calabrese, S. C. Dilsaver and D. D. Morris (1999). 'Differential effect of number of previous episodes of affective disorder on response to lithium or divalproex in acute mania'. In: *American Journal of Psychiatry* 156.8, pp. 1264–1266.

Swanson, C. A. and M. K. Sliwinski (2013). 'Archaeal assemblages inhabiting temperate mixed forest soil fluctuate in taxon composition and spatial distribution over time'. In: *Archaea* 2013.

Swidsinski, A., J. Weber, V. Loening-Baucke, L. P. Hale and H. Lochs (2005). 'Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease'. In: *Journal of clinical microbiology* 43.7, pp. 3380–3389.

Tamboli, C., C. Neut, P. Desreumaux and J. Colombel (2004). 'Dysbiosis in inflammatory bowel disease'. In: *Gut* 53.1, pp. 1–4.

Tange, O. et al. (2011). 'Gnu parallel-the command-line power tool'. In: *The USENIX Magazine* 36.1, pp. 42–47.

Tenenbaum, J. B., V. De Silva and J. C. Langford (2000). 'A global geometric framework for nonlinear dimensionality reduction'. In: *science* 290.5500, pp. 2319–2323.

Theriot, C. M., M. J. Koenigsknecht, P. E. Carlson Jr, G. E. Hatton, A. M. Nelson, B. Li, G. B. Huffnagle, J. Li and V. B. Young (2014). 'Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection'. In: *Nature communications* 5, p. 3114.

Tietz, A., K. E. Aldridge and J. E. Figueroa (2005). 'Disseminated coinfection with Actinomyces graevenitzii and Mycobacterium tuberculosis: case report and review of the literature'. In: *Journal of clinical microbiology* 43.6, pp. 3017–3022.

Tong, M., X. Li, L. W. Parfrey, B. Roth, A. Ippoliti, B. Wei, J. Borneman, D. P. McGovern, D. N. Frank, E. Li et al. (2013). 'A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease'. In: *PloS one* 8.11, e80702.

Torres, A. and J. J. Nieto (2006). 'Fuzzy logic in medicine and bioinformatics'. In: *BioMed Research International* 2006.

Tran, N. (2011). 'Blood-Brain Barrier'. In: *Encyclopedia of Clinical Neuropsychology*. Springer, pp. 426–426.

Triantafillidis, J. K., G. Nasioulas and P. A. Kosmidis (2009). 'Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies'. In: *Anticancer research* 29.7, pp. 2727–2737.

Tringe, S. G. and P. Hugenholtz (2008). 'A renaissance for the pioneering 16S rRNA gene'. In: *Current opinion in microbiology* 11.5, pp. 442–446.

Tringe, S. G. and E. M. Rubin (2005). 'Metagenomics: DNA sequencing of environmental samples'. In: *Nature reviews genetics* 6.11, p. 805.

Trivedi, M. H., A. J. Rush, S. R. Wisniewski, A. A. Nierenberg, D. Warden, L. Ritz, G. Norquist, R. H. Howland, B. Lebowitz, P. J. McGrath et al. (2006). 'Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: implications for clinical practice'. In: *American journal of Psychiatry* 163.1, pp. 28–40.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight and J. I. Gordon (2007). 'The human microbiome project: exploring the microbial part of ourselves in a changing world'. In: *Nature* 449.7164, p. 804.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis and J. I. Gordon (2006). 'An obesity-associated gut microbiome with increased capacity for energy harvest'. In: *nature* 444.7122, pp. 1027–131.

Vaishnava, S., M. Yamamoto, K. M. Severson, K. A. Ruhn, X. Yu, O. Koren, R. Ley, E. K. Wakeland and L. V. Hooper (2011). 'The antibacterial lectin RegIIIγ promotes the spatial segregation of microbiota and host in the intestine'. In: *Science* 334.6053, pp. 255–258.

Van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths. ISBN: 9780408709293. URL: https://books.google.co.uk/books?id=t-pTAAAAMAAJ.

Vapnik, V. (1998). *Statistical learning theory*. Vol. 1. Wiley New York.

Varian, H. (2005). 'Bootstrap tutorial'. In: *Mathematica Journal* 9.4, pp. 768–775.

Verleysen, M. et al. (2003). 'Learning high-dimensional data'. In: *Nato Science Series Sub Series III Computer And Systems Sciences* 186, pp. 141–162.

Vester-Andersen, M. K., M. V. Prosberg, T. Jess, M. Andersson, B. G. Bengtsson, T. Blixt, P. Munkholm, F. Bendtsen and I. Vind (2014). 'Disease course and surgery rates in inflammatory bowel disease: a population-based, 7-year follow-up study in the era of immunomodulating therapy'. In: *The American journal of gastroenterology* 109.5, pp. 705–714.

Vickery, B. (1979). 'Reviews: van Rijsbergen, CJ Information retrieval. 2nd edn. London, Butterworths, I978. 208pp'. In: *Journal of librarianship* 11.3, pp. 237–237.

Vogel, G. (2015). 'NIH debates human-animal chimeras'. In: *Science* 350.6258, pp. 261–262.

Wade, W. G. (2013). 'The oral microbiome in health and disease'. In: *Pharmacological research* 69.1, pp. 137–143.

Wallace, K. L., L.-B. Zheng, Y. Kanazawa and D. Q. Shih (2014). 'Immunopathology of inflammatory bowel disease'. In: *World J Gastroenterol* 20.1, pp. 6–21.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007a). 'Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy'. In: *Applied and environmental microbiology* 73.16, pp. 5261–5267.

Wang, W., L. Chen, R. Zhou, X. Wang, L. Song, S. Huang, G. Wang and B. Xia (2014). 'Increased proportions of Bifidobacterium and the Lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease'. In: *Journal of clinical microbiology* 52.2, pp. 398–406.

Wang, X., J. Yang, X. Teng, W. Xia and R. Jensen (2007b). 'Feature selection based on rough sets and particle swarm optimization'. In: *Pattern recognition letters* 28.4, pp. 459–471.

Waraich, P., E. M. Goldner, J. M. Somers and L. Hsu (2004). 'Prevalence and incidence studies of mood disorders: a systematic review of the literature'. In: *The Canadian Journal of Psychiatry* 49.2, pp. 124–138.

Warton, D. I., S. T. Wright and Y. Wang (2012). 'Distance-based multivariate analyses confound location and dispersion effects'. In: *Methods in Ecology and Evolution* 3.1, pp. 89–101.

Watson, J. D., F. H. Crick et al. (1953). 'Molecular structure of nucleic acids'. In: *Nature* 171.4356, pp. 737–738.

Wehrens, R., L. M. Buydens et al. (2007). 'Self-and super-organizing maps in R: the Kohonen package'. In: *J Stat Softw* 21.5, pp. 1–19.

Wei, Z., W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. M. Visscher et al. (2013). 'Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease'. In: *The American Journal of Human Genetics* 92.6, pp. 1008–1012.

Weiss, S., Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham et al. (2017). 'Normalization and microbial differential abundance strategies depend upon data characteristics'. In: *Microbiome* 5.1, p. 27.

Westcott, S. L. and P. D. Schloss (2015). 'De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units'. In: *PeerJ* 3, e1487.

— (2017). 'OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units'. In: *mSphere* 2.2, e00073–17.

White, J. R., S. Navlakha, N. Nagarajan, M.-R. Ghodsi, C. Kingsford and M. Pop (2010). 'Alignment and clustering of phylogenetic markers-implications for microbial diversity studies'. In: *BMC bioinformatics* 11.1, p. 152.

Whittaker, R. H. (1972). 'Evolution and measurement of species diversity'. In: *Taxon*, pp. 213–251.

Wickham, H. (2014). *Advanced r*. CRC Press.

Wickham, H. et al. (2014). 'Tidy data'. In: *Journal of Statistical Software* 59.10, pp. 1–23.

Wilbrow, M. (2013). *Draw a Kohonen SOM feature map?* Accessed 2018–06-15. URL: https://tex.stackexchange.com/questions/144366/draw-a-kohonen-som-feature-map.

Willing, B., J. Halfvarson, J. Dicksved, M. Rosenquist, G. Järnerot, L. Engstrand, C. Tysk and J. K. Jansson (2008). 'Twin studies reveal specific imbalances in the mucosaassociated microbiota of patients with ileal Crohn's disease'. In: *Inflammatory bowel diseases* 15.5, pp. 653–660.

Willner, P., A. S. Hale and S. Argyropoulos (2005). 'Dopaminergic mechanism of antidepressant action in depressed patients'. In: *Journal of affective disorders* 86.1, pp. 37–45.

Willner, P., J. Scheel-Krüger and C. Belzung (2013). 'The neurobiology of depression and antidepressant action'. In: *Neuroscience & Biobehavioral Reviews* 37.10, pp. 2331–2371.

Wingfield, B., S. Coleman, T. McGinnity and A. J. Bjourson (2016). 'A metagenomic hybrid classifier for paediatric inflammatory bowel disease'. In: *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, pp. 1083–1089.

— (2018a). 'A journal paper about rough sets and depression'. In: *Somewhere good*. Note: I need to write this first!

— (2018b). 'Depression paper'. In: *Nature Scientific Reports*. Note: manuscript under preparation.

— (2018c). 'Robust Microbial Markers for Non-Invasive Inflammatory Bowel Disease Identification'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Preprint Digital Object Identifier: 10.1109/TCBB.2018.2831212.

Woese, C. R. and G. E. Fox (1977). 'Phylogenetic structure of the prokaryotic domain: the primary kingdoms'. In: *Proceedings of the National Academy of Sciences* 74.11, pp. 5088–5090.

Wolpert, D. H. (2002). 'The supervised learning no-free-lunch theorems'. In: *Soft computing and industry*. Springer, pp. 25–42.

Wolpert, D. H. and W. G. Macready (1997). 'No free lunch theorems for optimization'. In: *Evolutionary Computation, IEEE Transactions on* 1.1, pp. 67–82.

Wong, K. H., Y. Jin and Z. Moqtaderi (2013). 'Multiplex Illumina sequencing using DNA barcoding'. In: *Current protocols in molecular biology*, pp. 7–11.

Woźniak, M., M. Graña and E. Corchado (2014). 'A survey of multiple classifier systems as hybrid systems'. In: *Information Fusion* 16, pp. 3–17.

Wulff, K., S. Gatti, J. G. Wettstein and R. G. Foster (2010). 'Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease'. In: *Nature Reviews Neuroscience* 11.8, pp. 589–599.

Xia, T., J. C. Baumgartner and L. David (2000). 'Isolation and identification of Prevotella tannerae from endodontic infections'. In: *Molecular Oral Microbiology* 15.4, pp. 273–275.

Xu, Z., D. Malmer, M. G. Langille, S. F. Way and R. Knight (2014). 'Which is more important for classifying microbial communities: who's there or what they can do?' In: *The ISME journal* 8.12, p. 2357.

Yang, B., Y. Wang and P.-Y. Qian (2016). 'Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis'. In: *BMC bioinformatics* 17.1, p. 135.

Yang, J. and V. Honavar (1998). 'Feature subset selection using a genetic algorithm'. In: *IEEE Intelligent Systems and their Applications* 13.2, pp. 44–49.

Yirmiya, R., N. Rimmerman and R. Reshef (2015). 'Depression as a microglial disease'. In: *Trends in neurosciences* 38.10, pp. 637–658.

Yoshizawa, J. M., C. A. Schafer, J. J. Schafer, J. J. Farrell, B. J. Paster and D. T. Wong (2013). 'Salivary biomarkers: toward future clinical and diagnostic utilities'. In: *Clinical microbiology reviews* 26.4, pp. 781–791.

Yu, L. and H. Liu (2004). 'Efficient feature selection via analysis of relevance and redundancy'. In: *Journal of machine learning research* 5.Oct, pp. 1205–1224.

Zaidi, S. S. A. and X. Zhang (2016). 'Computational operon prediction in whole-genomes and metagenomes'. In: *Briefings in functional genomics*, elw034.

Zaura, E., B. W. Brandt, M. J. T. de Mattos, M. J. Buijs, M. P. Caspers, M.-U. Rashid, A. Weintraub, C. E. Nord, A. Savell, Y. Hu et al. (2015). 'Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces'. In: *MBio* 6.6, e01693–15.

Zhang, L., H. Lee, M. C. Grimm, S. M. Riordan, A. S. Day and D. A. Lemberg (2014). '*Campylobacter concisus* and inflammatory bowel disease'. In: *World journal of gastroenterology: WJG* 20.5, p. 1259.

Zhang, X., D. Zhang, H. Jia, Q. Feng, D. Wang, D. Liang, X. Wu, J. Li, L. Tang, Y. Li et al. (2015). 'The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment'. In: *Nature medicine* 21.8, p. 895.

Zhao, J., S. C. Ng, Y. Lei, F. Yi, J. Li, L. Yu, K. Zou, Z. Dan, M. Dai, Y. Ding et al. (2013). 'First prospective, population-based inflammatory bowel disease incidence study in mainland of China: the emergence of "western" disease'. In: *Inflammatory bowel diseases* 19.9, pp. 1839–1845.

Zhao, Z. and H. Liu (2007). 'Searching for Interacting Features.' In: *ijcai*. Vol. 7, pp. 1156–1161.

Zheng, P., B. Zeng, C. Zhou, M. Liu, Z. Fang, X. Xu, L. Zeng, J. Chen, S. Fan, X. Du et al. (2016). 'Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism'. In: *Molecular psychiatry* 21.6, pp. 786–796.

Zhong, N., J. Dong and S. Ohsuga (2001). 'Using rough sets with heuristics for feature selection'. In: *Journal of intelligent information systems* 16.3, pp. 199–214.

Zou, H. and T. Hastie (2005). 'Regularization and variable selection via the elastic net'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Zuo, T., M. A. Kamm, J.-F. Colombel and S. C. Ng (2018). 'Urbanization and the gut microbiota in health and inflammatory bowel disease'. In: *Nature Reviews Gastroenterology & Hepatology*, p. 1.