BAYESIAN STATISTICAL ANALYSIS OF BACTERIAL DIVERSITY

Jing Tang

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium B123, the Exactum Building (Gustaf Hällströmin katu 2b), on May 15th, 2009, at 12 noon.

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

Supervisors

Prof. Jukka Corander
Department of Mathematics
Åbo Akademi University
Finland

Prof. Elja Arjas
Department of Mathematics and Statistics
University of Helsinki
Finland

Reviewers

Prof. Esa Läärä
Department of Mathematical Sciences
University of Oulu
Finland

Prof. Samuel Kaski
Department of Information and Computer Science
Helsinki University of Technology
Finland

Opponent

Prof. Daniel Thorburn
Department of Statistics
Stockholm University
Sweden

ISBN 978-952-10-5463-1 (paperback)
ISBN 978-952-10-5464-8 (PDF)
http:// ethesis.helsinki.fi
Yliopistopaino
Helsinki 2009

To Jesus.

List of articles

[I] Corander, J. and Tang, J. (2007). Bayesian analysis of population structure based on linked molecular information. *Mathematical biosciences,* 205:19-31.

[II] Tang, J., Hanage, W.P., Fraser C. and Corander, J. (2009) Identifying currents in the gene pool for bacterial populations using an integrative approach, *PLOS Computational biology,* under revision.

[III] Corander, J., Marttinen, P., Sirén, J. and Tang, J. (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics,* 9:539.

[IV] Tang, J., Tao, J.L., Urakawa, H. and Corander,J. (2007) T-BAPS: A Bayesian statistical tool for comparison of microbial communities using terminal-restriction fragment length polymorphism (T-RFLP) data. *Statistical applications in genetics and molecular biology,* 6(1):30.

[V] Marttinen, P., Tang, J., De Baets, B., Dawyndt, P. and Corander, J. Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach. (2009) *IEEE transactions on pattern analysis and machine intelligence,* 31(1):74-85.

[VI] Hanage, W.P., Fraser, C., Tang, J., Connor, T.R. and Corander, J. (2009) Hyper-recombination, diversity and the acquisition of antibiotic resistance in the pneumococcus. *Science,* under revision.

Author's contributions

[I] JT and JC contributed equally to constructing the model and conducting the analysis. JC was the principal author of the paper.

[II] JT and JC formulated the model, conducted the analysis and wrote the paper. WPH and CF contributed to the planning of the study and provided the data set.

[III] JT and JC were jointly responsible for all aspects of the paper. All authors participated in the development of the methodology, the modeling approaches and the data analysis. JC and JT wrote the manuscript.

[IV] JT and JC were jointly responsible for all aspects of the paper.

[V] JT contributed in the model formulation and data analysis. PM was the principal author of the paper.

[VI] JT contributed to the data analysis and writing of the paper. WPH was responsible for writing the paper.

Abstract

Bacteria play an important role in many ecological systems. The molecular characterization of bacteria using either cultivation-dependent or cultivation-independent methods reveals the large scale of bacterial diversity in natural communities, and the vastness of populations within a species or genus. Understanding how bacterial diversity varies across different environments and also within populations should provide insights into many important questions of bacterial evolution and population dynamics.

This thesis presents novel statistical methods for analyzing bacterial diversity using widely employed molecular fingerprinting techniques. The first objective of this thesis was to develop Bayesian clustering models to identify bacterial population structures. Bacterial isolates were identified using multilous sequence typing (MLST), and Bayesian clustering models were used to explore the evolutionary relationships among isolates. Our method involves the inference of genetic population structures via an unsupervised clustering framework where the dependence between loci is represented using graphical models. The population dynamics that generate such a population stratification were investigated using a stochastic model, in which homologous recombination between populations can be quantified within a gene flow network.

The second part of the thesis focuses on cluster analysis of community compositional data produced by two different cultivation-independent analyses: terminal restriction fragment length polymorphism (T-RFLP) analysis, and fatty acid methyl ester (FAME) analysis. The cluster analysis aims to group bacterial communities that are similar in composition, which is an important step for understanding the overall influences of environmental and ecological perturbations on bacterial diversity. A common feature of T-RFLP and FAME data is zero-inflation, which indicates that the observation of a zero value is much more frequent than would be expected, for example, from a Poisson distribution in the discrete case, or a Gaussian distribution in the continuous case. We provided two strategies for modeling zero-inflation in the clustering framework, which were validated by both synthetic and empirical complex data sets.

We show in the thesis that our model that takes into account dependencies between loci in MLST data can produce better clustering results than those methods which assume independent loci. Furthermore, computer algorithms that are efficient in analyzing large scale data were adopted for meeting the increasing computational need. Our method that detects homologous recombination in populations may provide a theoretical criterion for defining bacterial species. The clustering of bacterial community data include T-RFLP and FAME provides an initial effort for discovering the evolutionary dynamics that structure and maintain bacterial diversity in the natural environment.

Keywords

Bacterial diversity, Statistical analysis, Bayesian model-based clustering, MLST, T-RFLP, FAME

Contents

1 Introduction **1**
1.1 Bacterial diversity 1
1.2 Analysis of population structure 2
1.3 Analysis of community structure 3
1.4 The thesis structure 4

2 Discovering bacterial population structures **5**
2.1 Clustering MLST data 5
2.2 Model-based clustering methods 6
2.3 Bayesian unsupervised clustering framework 7
2.4 Modeling nucleotide dependency 7
2.5 Statistical inference of the clustering 8
2.6 Modeling recombination between subpopulations 8
2.7 Software development 9
2.8 Discussion 9

3 Comparing bacterial communities **13**
3.1 T-RFLP and FAME data overview 13
3.2 The zero-inflated data 13
3.3 Bayesian latent mixture modeling 14
3.4 Bayesian fuzzy likelihood clustering 15
3.5 Software development 16
3.6 Discussion 16

4 Conclusion **17**

Acknowledgement **20**

References **21**

1 Introduction

1.1 Bacterial diversity

Bacteria are microscopic organisms which are extremely abundant. They can be found in diverse environments ranging from commonly seen places such as soil and water, to the most extreme conditions such as volcano vents and hot springs. Unlike multicellular organisms such as animals and plants, most bacteria are made of a single cell. Despite their simple structure, bacteria have developed into a wide array of different species. It is estimated that there are more than one million bacterial species living on Earth. The number of bacterial species that exist in a single environment, such as a tiny soil sample, could easily exceed the order of thousands (Kirk et al., 2004).

Bacteria are typically the first organisms to sense and respond to chemical and physical changes in their environment. The study of bacterial diversity is therefore crucial for understanding the dynamics of many ecological processes. For example, by characterizing the composition of soil bacterial communities in transgenic plant fields, one can assess the environmental impacts brought about by genetically modified plants (Park et al., 2006). Evidence in recent medical research indicates that metabolic diseases such as diabetes are associated with fundamental changes in bacterial diversity in the human intestine. Therefore, by identifying these critical changes we might be able to predict whether a person will develop a certain disease (Jia et al., 2008).

Knowledge about bacterial diversity is traditionally based on the cellular shapes, metabolic functions and other phenotypic traits of bacteria. For example, most bacteria adopt one of the three basic cell shapes: sphere, rod-shaped or spiral. Spherical bacteria are referred to as *cocci*, for example *streptococci pneumoniae* that causes lung pneumonia. Rod-shaped bacteria are termed *bacilli*. Examples from this group include *E.coli*, which is found in the lower intestine of humans. Spiral-shaped bacteria, *spirilla*, are the largest of these three types. Based on the cellular structure and other phenotypic properties, bacteriologists can classify similar bacteria into a single species.

However, the conventional phenotypic characterization is applicable only for those bacteria that can be isolated from environmental samples using culture medium. Even with the many advances that have been made in microbiological culture techniques, strain isolation is still impossible for the majority of bacterial species. Furthermore, the information on phenotypic diversity tells us little about the way in which different species are related and thus provides limited understanding of their evolutionary history. In fact, for much of the last century, the nature and identity of most bacteria remained unknown and our knowledge of bacterial diversity was minimal (Kapur and Jain, 2004).

During the past few years, our knowledge of bacterial diversity has shifted dramatically from the phenotypic scale towards the genetic scale. Following the development of widely available molecular fingerprinting techniques, detecting the genetic diversity for both cultivable and uncultivable species has now become possible. The analysis of genetic information found in these molecular surveys has shown a range of bacterial diversity that broadens, and even revolutionizes, our previous views on the complexity of bacterial evolution (Achtman, 2004; Whitaker and Backfield, 2006).

A detailed analysis of genetic diversity consists of two research categories: population

structure analysis and community structure analysis. Population structure analysis concerns the individual-level variation within a species or a genus population. Such a variation is deemed structural if it divides the whole population into multiple genetically distinct subpopulations. The degree of such subdivision is commonly unknown and may be affected by geographical, temporal and ecological factors. The main objective is then to detect the population structure and infer the generative mechanisms (Spratt and Maiden, 1999; Feil and Spratt, 2001). In contrast, community structure analysis focuses on the species-level diversity within a bacterial community. A bacterial community refers to an environmental sample in which different bacterial species cohabit. The primary task in analyzing bacterial communities is to determine the identity and richness of these cohabiting species, *i.e.* the community composition. The subsequent analyses include comparing community compositions across different samples and assessing the effects of environmental disturbances (Kirk et al., 2004). Details of both analysis types and their statistical aspects are given in the next two sections.

1.2 Analysis of population structure

Population structure analysis aims to describe the genetic variation within natural populations. A sufficient description of the genetic variation requires the isolation of a large number of strains that are representative of the whole population. Two fingerprinting methods for bacterial strains are commonly in use: multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST). MLEE indirectly identifies genetic variation through differences in the electrophoretic mobility of a selection of metabolic enzymes, while MLST directly pinpoints the DNA sequence variation on chromosomes. These techniques allow rapid strain characterization on a large scale and have been applied to study the population structure of many bacterial species (Smith et al., 2000).

With the immense amounts of genetic variation discovered, the real challenge is to determine the population dynamics that generate such variation. The population dynamics are, in principle, a complex interplay of evolutionary processes such as mutation and recombination. In bacteria, recombination involves the genetic exchange of chromosome regions between isolates within a species, or, in some cases, from closely related species. The genetic exchange through recombination can be regarded as a 'horizontal' component of the population dynamics, as opposed to a 'vertical' movement of genetic variation which occurs through mutations (Spratt and Maiden, 1999). The relative contributions of these two components are regulated by many environmental factors, such as genetic drift, migration and selection. The analysis of bacterial population dynamics often reveals that instead of there being a single homogenous population, the population is further subdivided into a set of subpopulations. A subpopulation is manifested by a group of individuals which interact more frequently with each other than with more distant individuals. The genetic kinship shared by individuals within a subpopulation is characterized by a higher-than-expected similarity and should be detectable from genetic variation data. Discovering the population structure and its underlying population dynamics is a major goal of population structure analysis (Spratt and Maiden, 1999; Feil and Spratt, 2001).

From a statistical viewpoint, the questions arising from the study of population structure analysis can be formulated as a series of statistical inference tasks. First of all, iden-

tifying a population structure corresponds to assigning individual strains into genetically distinct subpopulations. This is essentially a data clustering problem in the statistical domain (Pritchard et al., 2000; Corander et al., 2004). The second task is then to infer the underlying population dynamics that lead to the population structure. Note that a correlation between subpopulations is often a result of recombination being present as one of the population dynamics. By analyzing the correlation using statistical methods one can find clues to the role of recombination for the evolution of bacterial species (Feil et al., 2004; Corander and Marttinen, 2006).

1.3 Analysis of community structure

A typical microbial community may comprise hundreds, or even thousands, of bacterial species interacting within a particular environmental site. Understanding the bacterial composition of a community is vital for assessing the functions of the associated ecosystem processes. Due to the fact that only a small fraction of bacteria in natural communities can be isolated, analysis of community structure commonly utilizes culture independent methods which overcome the reliance on isolation. The culture independent methods can be either biochemical-based or molecular-based. An example of a biochemical-based method is fatty acid methyl ester (FAME) analysis, which extracts and identifies distinctive fatty acids as biomarkers for species identification. It is known that some unique fatty acids are species-specific and therefore a change in the fatty acid profile implies a change in the bacterial community composition (Ritchie et al., 2000). Molecular-based methods include the commonly used terminal restriction fragment length polymorphism (T-RFLP) analysis. T-RFLP extracts genomic sequences directly from environmental samples. The extracted sequences are digested with polymerase chain reaction (PCR) amplification which targets a genetic marker (for example, the 16s ribosomal RNA gene) which can be used to discriminate between species. An estimation of the community composition is then made through a comparative analysis of the digestion products given by that specific marker (Marsh, 1999).

The community compositions, irrespective of the fingerprinting method used, are in general represented as a data set that consists of a set of variables measured for each of the n samples. Note that each variable is either a phenotypic or genetic feature that indicates the abundance of a certain species. We may consider the data set as a matrix in which the individual samples are represented by rows and the abundances of species are reflected by the columns.

Statistical analysis of the community structure, using such a single data matrix, usually starts with a description of the similarity of species abundance between samples. As in the analysis of population structure, the community structure is often unknown *a priori*. We seek to partition the data into groups of samples which are similar, which is again a data clustering problem. Once communities have been clustered into groups, the complex interactions between the community structure and its environment can be investigated (Blackwood et al., 2003; Abdo et al., 2006).

1.4 The thesis structure

The focus of this thesis is the study of the bacterial diversity which is revealed via some widely used fingerprinting methods of population and community structure analyses. Our ultimate aims include: (i) detecting the population structure using MLST data; (ii) understanding the population dynamics, especially the role of recombination in shaping the bacterial population structures; (iii) detecting the community structure using T-RFLP data; and (iv) detecting the community structure using FAME data. While these fingerprinting methods provide high-throughput data sets that contain much information on the bacterial diversity, they do not provide direct answers to these issues. For this reason, the thesis develops computational tools for clustering these data sets, and explains the underlying data generation mechanisms in a statistically extensive and cost-effective manner.

In this thesis, we developed several model-based clustering approaches for this purpose. A model is a mathematical simplification of the biological process that has generated the observed sample data. Model-based clustering approaches start with formulating a clustering model in which its parameters, *e.g.* the number of clusters (K), are directly estimated from the data. Note that mathematical modeling in the context of bacterial diversity is inevitably a simplified approximation to the reality, so the model parameters should preferably be estimated using a statistical approach. Furthermore, additional uncertainty can emanate from the non-deterministic measurement noise in the fingerprinting techniques. Recent developments in Bayesian unsupervised clustering have provided a promising modeling framework that can handle and represent such uncertainty (Corander et al., 2003, 2004). Our model-based clustering approaches are based on this framework, and adapted to suit our data sets and the biological questions which we seek to answer.

Six research papers are presented in this thesis. Paper [I] focuses on inferring hidden population structures using MLST data, for which potential dependency between nucleotide sites is explicitly modeled. Paper [II] delves into the evolutionary mechanisms that generate the population structures identified in paper [I], by developing a stochastic gene flow model that quantifies the extent and the directionality of recombination between closely related subpopulations. The third paper, [III], describes the software package BAPS (Bayesian Analysis of Population Structures). This software contains computational tools for fitting models and visualizing the output of the methods in [I] and [II]. Paper [VI] is an application of our methods to a real bacterial pathogen data set, for which a significant association between elevated recombination and antibiotic resistance is further explored quantitatively. Paper [IV] and [V] tackle the clustering of zero-inflated data, such as that obtained from T-RFLP and FAME analysis, using two different approaches. In [IV] the zero-inflation is modeled as a mixture of discrete and continuous distributions, while in [V] the zero-inflation is modeled by fuzzy vectors. In each case, the appropriate statistical learning algorithm was used. In addition, the T-BAPS software for clustering T-RFLP data is introduced in [IV]. In summary, the six papers can be grouped into two modules: papers [I], [II], [III] and [VI] concern the analysis of population structure, while papers [IV] and [V] focus on the community structure analysis. The schematic diagram describing the connection between these papers is shown in Figure 1.

The outline of the thesis summary is as follows. In Section 2, we present the Bayesian unsupervised clustering framework that underlies [I],[II], [III] and [VI], for the task of

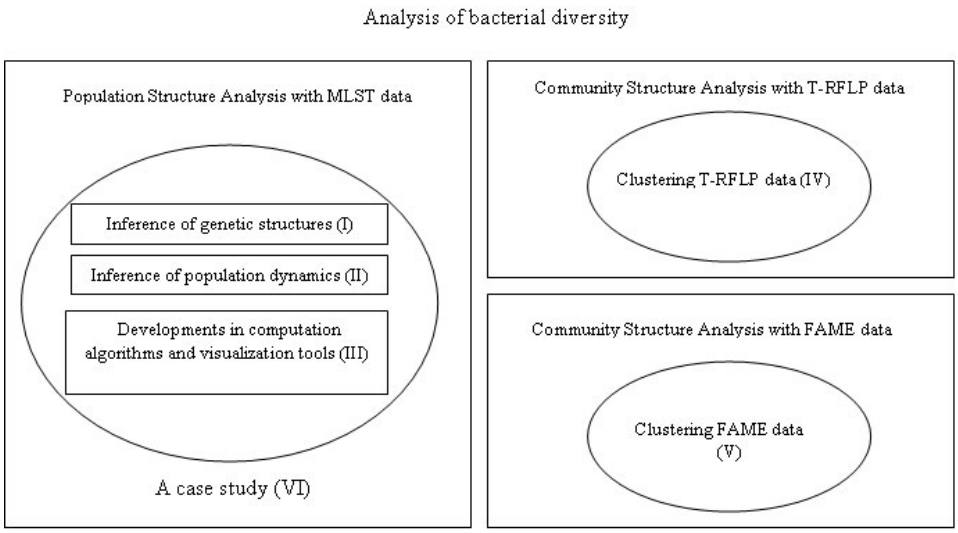

Figure 1: Schematic representation of the thesis structure.

analyzing population structures using MLST data. In section 3, the two clustering methods in [IV] and [V] are summarized, with applications to analyzing community structure T-RFLP and FAME data. Section 4 contains concluding remarks, as well as a discussion of some of the aspects that are missing in our current models, and presents the future challenges that need to be addressed.

2 Discovering bacterial population structures

2.1 Clustering MLST data

The recently developed MLST method detects genetic diversity in bacterial populations by sequencing chromosomal fragments of several housekeeping genes. Housekeeping genes refer to the genes that code for fundamental proteins in bacteria and are known to be evolutionarily conserved. Genetic diversity observed at housekeeping genes therefore provides reliable information concerning the evolutionary mechanisms underlying a population structure. In MLST, typically seven housekeeping genes with a length of approximately 500 nucleotides each are sequenced for hundreds or thousands of isolates. Identification of the population structure corresponds to a clustering of these isolates based on their nucleotide variation in four letters A, T, G and C (Cooper and Feil, 2004).

Statistical analysis of population structures for such high-throughput data is challenging. Generic distance-based clustering techniques, such as dendrograms, are used to graphically display the genetic relatedness of isolates using distance measures (Feil and Spratt, 2001; Achtman, 2004). However, the choice of distance measures is somewhat arbitrary and usually reflects only vaguely the biological reality. A popular clustering approach tailored for MLST data is EBURST (Feil et al., 2004). The EBURST method assigns different sequences at each gene with unique allelic numbers. The allelic profile for each isolate is then defined as a string of allelic numbers over multiple (typically seven) genes. However,

clustering of the isolates is based on the allelic profile, rather than the underlying sequence itself, and thus incurs huge information loss about the nucleotide polymorphisms.

2.2 Model-based clustering methods

In order to overcome the difficulties that exist in distance-based methods, model-based clustering methods within a Bayesian framework have been suggested as a more satisfactory approach to learning population structures. Bayesian model-based clustering provides an effective way of characterizing the uncertainty of data clustering arising from latent population structures.

STRUCTURE is a popular Bayesian model-based clustering method using the standard Markov chain Monte Carlo (MCMC) algorithm (Pritchard et al., 2000; Falush et al., 2003). However, when applying STRUCTURE to MLST data, some limitations should be addressed carefully. Firstly, STRUCTURE estimates the number of clusters (K) by iteratively running the algorithm with a range of candidate cluster numbers, and choosing the one that gives the maximal log likelihood of data. For analyzing bacterial populations, STRUCTURE might underestimate the real number of clusters simply because the range of candidate cluster numbers is too small. On the other hand, increasing the range of cluster numbers will inevitably increase the computational load since each of the numbers in the range must be examined. In particular, with the increasing size and complexity of MLST data (usually at least hundreds of isolates and thousands of nucleotide bases), the performance of the standard MCMC algorithm deteriorates rapidly, which is also a major factor for the unreliable estimation of population structures (Gao and Starmer, 2007).

Recent advances in the Bayesian model-based clustering have been made in a series of publications by Corander et al. (2003, 2004, 2006b, 2007), where the functionality for conducting the generic population structure analysis has been readily applicable in a program BAPS. In contrast to the some what *ad hoc* way of inferring the cluster number K in STRUCTURE, BAPS treats K as a model parameter which is unknown *a priori*, and returns a posterior distribution of $P(K|data)$ where its mode is considered to be the optimum value. In the recent versions of BAPS (version 3 onwards) a greedy stochastic optimization algorithm has been utilized to improve the computational efficiency. However, the BAPS model was formulated initially for clustering sparse molecular markers where the dependency within the loci remains to be incorporated when dealing with sequence data.

In paper [I], we continue along the lines of the Bayesian modeling presented in Corander et al. (2003, 2004), by providing a more biologically reasonable model for characterizing MLST data. A major novelty is the mathematical modeling of the dependency between neighboring nucleotide sites using a decomposable graph. Under the assumption of unrestricted exchangeability, the decomposability enables an explicit characterization of the data marginal likelihood for a partition. For analyzing high-throughput sequence data from bacterial populations, we avoid the use of standard MCMC algorithms for learning the clustering parameters, but apply instead a much faster stochastic greedy algorithm to pinpoint the most likely population structure ([II] and [III]). The proposed method is compared with the STRUCTURE program against the same real bacterial data in [VI].

2.3 Bayesian unsupervised clustering framework

Here we introduce the general Bayesian unsupervised clustering framework for papers [I], [II], [III] and [VI]. This clustering framework was detailed in Corander et al. (2007).

A Bayesian approach to the problem of unsupervised classification can be neatly formulated in terms of a generic model-learning task. If we consider a finite collection of partition solutions to be a partition space $\mathcal{S}$, finding an optimal partition is then equivalent to choosing a model that gives the best probabilistic characterization of the data. The Bayesian approach to learning the plausibility of a model, *i.e.,* a partition S in the partition space $\mathcal{S}$, is defined by the marginal data likelihood as the integral over the conditional likelihood in the parameter space times the prior density

$$\begin{aligned} p(data) &= \sum_{S \in \mathcal{S}} p(S) p(data \mid S) \\ &= \sum_{S \in \mathcal{S}} p(S) \int p(data \mid \theta_S) p(\theta_S) d\theta_S, \end{aligned} \tag{1}$$

where $p(S)$ is the prior probability of model S, and $p(data|\theta_S)$ is the likelihood of the data under model S given its associated parameters θ_S.

Using Bayes rule, the model plausibility to explain the data can be defined as posterior probability

$$p(S \mid data) = \frac{p(S)p(data \mid S)}{\sum_{S \in \mathcal{S}} p(S)p(data \mid S)}. \tag{2}$$

If each partition is considered *a priori* to be equally likely, *i.e.,* $p(S)$ is distributed uniformly over the partition space $\mathcal{S}$, then the posterior distribution simplifies to

$$p(S \mid data) \propto p(data \mid S) \tag{3}$$

The optimal partition $\hat{S}$ is then obtained by maximizing the posterior probability

$$\hat{S} = \arg\max_{S \in \mathcal{S}} p(S \mid data) = \arg\max_{S \in \mathcal{S}} p(data \mid S). \tag{4}$$

2.4 Modeling nucleotide dependency

An important feature of MLST data is the concatenation of nucleotide sequences of several housekeeping genes. The housekeeping genes are typically positioned sparsely around the bacterial genome, and can be treated as mutually independent when formulating a clustering model. However, the nucleotides within a gene are tightly linked as they are aligned in a sequence and thus should not be considered as independent variables. In paper [I], we explicitly represent the physical linkage of the nucleotides within each gene by a second order Markov dependence graphical model. As shown in Fig. 1 (b) in [I], each node of the model represents a nucleotide site and the edges between nodes indicate physical linkage. Each node is fully connected with its two neighboring nodes to represent a stronger linkage than those at distance. The biological rationale behind the triplet model is that, every triplet of nucleotides in a DNA sequence, commonly referred to as a *codon*, is a single unit that codes for a particular amino acid, and thus forms a group of variables with strong correlation.

According to the general theory of graphical models, such an undirected graph can be decomposed into a perfect ordering using its subsets of cliques and separators, such that the predictive probability of the data $p(\mathrm{x}|S)$ for a partition S can be factorized over these clique and separator components, as shown in formula (5) of [I]. Under the assumption of unrestricted exchangeability which implies that the predictive probability is invariant under the permutation of the feature vectors over the clusters and the linkage groups, the explicit form of the marginal likelihood can be obtained as a factorization of multiple Dirchlet-Multinomial likelihoods, as shown in formula (6) in [I].

2.5 Statistical inference of the clustering

Given the explicit form of the marginal likelihood (formula (6) in [I]), finding an optimal partition $\hat{S}$ defined in formula (4) in section 2.3, requires efficient searching in the partition space $\mathcal{S}$. Due to the large size of $\mathcal{S}$, an exhaustive search for $\hat{S}$ through all the members of $\mathcal{S}$ is numerically impractical. Standard MCMC algorithms, such as the Gibbs and Metropolis-Hastings samplers, have gained considerable popularity for such learning tasks. However, the computational burden associated with these algorithms increases considerably as the model complexity and the size of the investigated data set increase. In our model, the number of parameters for the second order Markov structure is typically beyond what the standard MCMC algorithms can reliably handle. Therefore, two alternative algorithms that require much less computational effort have been suggested in [I]: the non-reversible parallel Metropolis-Hastings algorithm (Corander et al., 2006a) and the stochastic greedy optimization algorithm (Corander et al., 2006b). The non-reversible parallel Metropolis-Hastings algorithm allows a mixing of multiple Markov chains in the learning process with simplified non-reversible transition kernels. Such a strategy enables distinct search processes to learn from one other and thus improves the algorithm convergence and enables a better predictive ability. To further decrease the computational load, the stochastic greedy optimization algorithm aims to find the posterior mode of $p(S|data)$, rather than an estimate from the whole posterior distribution. This is in practice sufficient for the majority of population structure inference problems (Latch et al., 2006).

2.6 Modeling recombination between subpopulations

The transmission of genetic information through homologous recombination between genetically distinct subpopulations, often referred to as gene flow, is one of the important evolutionary processes that shapes bacterial population structure. Unlike eukaryotes, the rates of recombination in bacteria can vary widely between different subpopulations even within a single species. Such a variety of recombination rates can raise additional challenges when attempting to infer the phylogenetic relationships in the population structure. Under the process of lateral gene transfer, an individual strain of a subpopulation may have genetic material acquired from a variety of other subpopulations. The complex interactions between the recipient and donor subpopulations are therefore more reasonably represented as a network rather than a bifurcating tree, as provided by most of the currently available phylogenetic methods (Spratt and Maiden, 1999; Smith et al., 2000).

Based on a population structure resolved by paper [I], paper [II] continues to develop an integrative model for exploring the influence of gene flow upon the evolution of bacterial populations. The model focuses on a gene flow rate **W** that represents, on average, the frequency of gene flow from a donor subpopulation to a recipient subpopulation. Such a model parameter can be decomposed as a product of two hyperparameters: the gene flow propensity **P** and the recombination intensity **Λ**. The gene flow propensity **P** denotes the probability of an individual in the recipient subpopulation acquiring DNA through homologous recombination from a donor subpopulation, and the recombination intensity **Λ** corresponds to the average rate of such recombinational events over the analysed sequence. Statistical learning of these parameters can be done by estimating the admixture coefficients in the population structure using a generalization of the method by Corander and Marttinen (2006), which was introduced in [III]. The gene flow dynamics between subpopulations captured by the rate parameter W is further visualized as a gene flow network, as illustrated in Figure 7 in [II].

2.7 Software development

The non-reversible parallel Metropolis-Hastings algorithm used in [I] was made available in the BAPS 2 software. The stochastic greedy optimization algorithm for [I] and [II] is implemented in the versions 3-5 of BAPS. In addition, we introduce an array of updated Bayesian tools for the analysis of genetic population structures ([III]). With these methods it is possible to fit genetic mixture models using user-specified numbers of clusters and to estimate levels of admixture under a genetic linkage model. Also, mutations representing a different ancestry can be tracked for the sampled individuals, and *a priori* specified hypotheses about genetic population structure can be directly compared. We have improved further the computational characteristics of the algorithms behind the methods implemented in BAPS, thus facilitating the analysis of MLST data sets. In particular, analysis of a single data set can now be distributed over multiple computers using a script interface to the software. These enhanced functions on learning the genetic structure of bacterial populations are expected to meet the increasing need for analyzing large and complex data sets. The software development work is summarized in paper [III].

2.8 Discussion

The analysis of population structure using genetic data is key to population genetics. Bacterial population genetics is the branch of population genetics that deals with the analysis of variation in bacterial populations. Until rather recently, bacterial population genetics remained outside of the focus of the majority of evolutionary biologists. It was generally assumed that, since bacteria reproduce asexually by binary fission, recombination would not play a significant role in shaping genetic diversity. Mutation was therefore expected to be the major evolutionary force to leads to chromosomal variations. If this were the case, bacterial populations would be characterized by clones of nearly identical individuals, with the genetic diversity brought only by accumulative mutations (Spratt and Maiden, 1999).

Following the development of culture-independent PCR methods, DNA-based fingerprinting has been widely adopted for detecting genetic diversity on a finer scale. The ge-

netic data have unveiled much higher heterogeneity in bacterial populations than would be expected in the absence of recombination. Many bacterial population biologists started questioning the validity of clonal structures and are seeking evidence of recombination. However, quantifying the contribution of recombination to population structures remains unclear for most of the bacterial species (Holmes et al., 1999; Hanage et al., 2005).

The necessity to account for recombination in population structure analysis has led to the increasing analysis of multilocus sequence typing (MLST), in which the genetic variation is indexed directly by the nucleotide polymorphisms over a selection of housekeeping genes. The housekeeping genes are involved in essential functions of cell metabolism and are therefore sequenced for each of the sampled strains. Compared to other genes that are hyper-variable and sensitive to environmental pressures, the nucleotide changes within the housekeeping genes are more likely to be evolutionarily conserved. Analysis of these genes therefore provides a more reliable indicator of the effect of recombination on the genome as a whole, and thus an ideal strategy for analyzing the long-term evolutionary relationships in bacterial populations (Cooper and Feil, 2004).

As a standard approach for analyzing MLST data, EBURST involves assigning an allele number to each unique sequence and identifies closely related strains according to the allelic profiles rather than the sequences themselves. Although EBURST captures the overall structure of bacterial populations, it ignores most of the variational information on the sequence level which might be essential in recovering recombination.

In [I], [II], [III] and [VI], we tackle this challenge by developing a statistical modeling framework for inferring population structures using DNA sequence data. Furthermore, recombination can be represented as a mechanism which generates mutual correlations between subpopulations.

The clustering method proposed in [I] has been shown to yield a more sensible partition for a data set where the genetic markers are densely linked, than a framework that assumes independent markers. As shown in Fig. 2 in [I], a consequence of using the independence assumption on data which is densely linked is the overestimation of the number of clusters. The excess number of clusters stems from the fact that the data heterogeneity is inappropriately characterized due to the independence assumption. From a model selection perspective, the empirical data set can, under the assumption of independence, favor a model with a larger number of clusters than is actually feasible. In other words, observing individuals that come from the same cluster is considered *surprising* given that their molecular markers are independent, and thus the model which overweighs the dissimilarity between individuals is favored. By taking into account the existing dependence in DNA sequence markers, our method provides a realistic clustering model for estimating the population structure without a tendency to produce spurious clusters.

The number of subpopulations, K, is usually treated as an unknown parameter when considering model-based clustering approaches. The approach of Pritchard et al. (2000), for example, tests different K values in the clustering algorithm and determines the most likely K by comparing the corresponding approximate marginal likelihoods of the data. Specifically, the algorithm itself requires K to be specified and fixed in advance. Finding the correct number requires repeated execution of the algorithm with different initiated values of K. In contrast, our method takes a different approach by allowing K to vary within a range in the stochastic search algorithm. Under this approach, the partition and estimation

of K are conducted simultaneously. When no auxiliary information is available, the maximal value for K is equal to the number of individuals in the data. In practice, the upper bound for K is usually selected *a priori* based on the relevant background information. The marginal probability that K is equal to k, given the data, is obtained by summing the posterior probabilities for all the partitions that have k clusters, *e.g.*,

$$p(|S| = k \mid data) = \sum_{S \in \mathcal{S}:|S|=k} p(S \mid data). \tag{5}$$

Figure 7 in [II] shows the gene flow network inferred from the *Neisseria* data set which consists of 5086 strains with a total length of 3284 nucleotides concatenated from seven housekeeping genes. This gene flow network reveals evidence of highly complex genetic exchanges brought about by homologous recombination between the 32 identified subpopulations. Clearly, homologous recombination is an important population dynamics which has shaped the genetic makeup of each sampled individual strain to various degrees. We presented in Figure 2 the profile of gene *recA* for each of the identified subpopulations. The similarities and dissimilarities in the DNA sequences when comparing different subpopulations could suggest the local influences of homologous recombination. This would provide a nice illustration of the level of complexity that can exist. BAPS is currently the only Bayesian population genetics software which is able to deal with such levels of complexity.

In microbiology, a highly relevant question is: what is a bacterial species? The continuous exploration of genetic variation during the genomic era has intensified the debate on the definition of a bacterial species (Fraser et al., 2007). A primary criterion for differentiating species is a certain level of similarity based on DNA hybridization. Individual strains are currently assigned to a common species if their pairwise DNA-DNA associations are 70% or higher. However, such a threshold level was determined as a tradition for purely pragmatic reasons, since it matched the pre-existing species definition based on phenotypic consistency. Even though this criterion is currently widely accepted, its theoretical justification is still missing. Various theoretical concepts for defining bacterial species have been suggested, but none of them have been generally accepted. A recent paper by Achtman and Wagner (2008) proposes a theory-based species concept based on cohesive evolutionary forces. They suggest that species should be defined as a metapopulation lineage that consists of a set of related subpopulations that evolved separately from other lineages. It is those evolutionary dynamics that constantly affect all of the subpopulations which make the species distinguished from the others.

The gene flow network provided by our method (II and VI) might serve as a plausible evolutionary criterion for the theory-based species definition suggested in Achtman and Wagner (2008). Subpopulations that are identified with incoming and outgoing gene flows are indicative of cohesive evolutionary dynamics and therefore should be classified as a common species. A lack of gene flow between subpopulations indicates the existence of genetic barriers to homologous recombination. Analysis of an *N. meningitidis* and *N. lactamica* MLST data set identifies three subpopulations in *N. lactamica*, two of which are distinctly separated, while the third one is linked to subpopulations in *N. meningitidis* (Figure 7 in [II]). By adopting the theory-based species concept by Achtman and Wagner (2008), we may infer that the two discrete subpopulations are evolving separately and

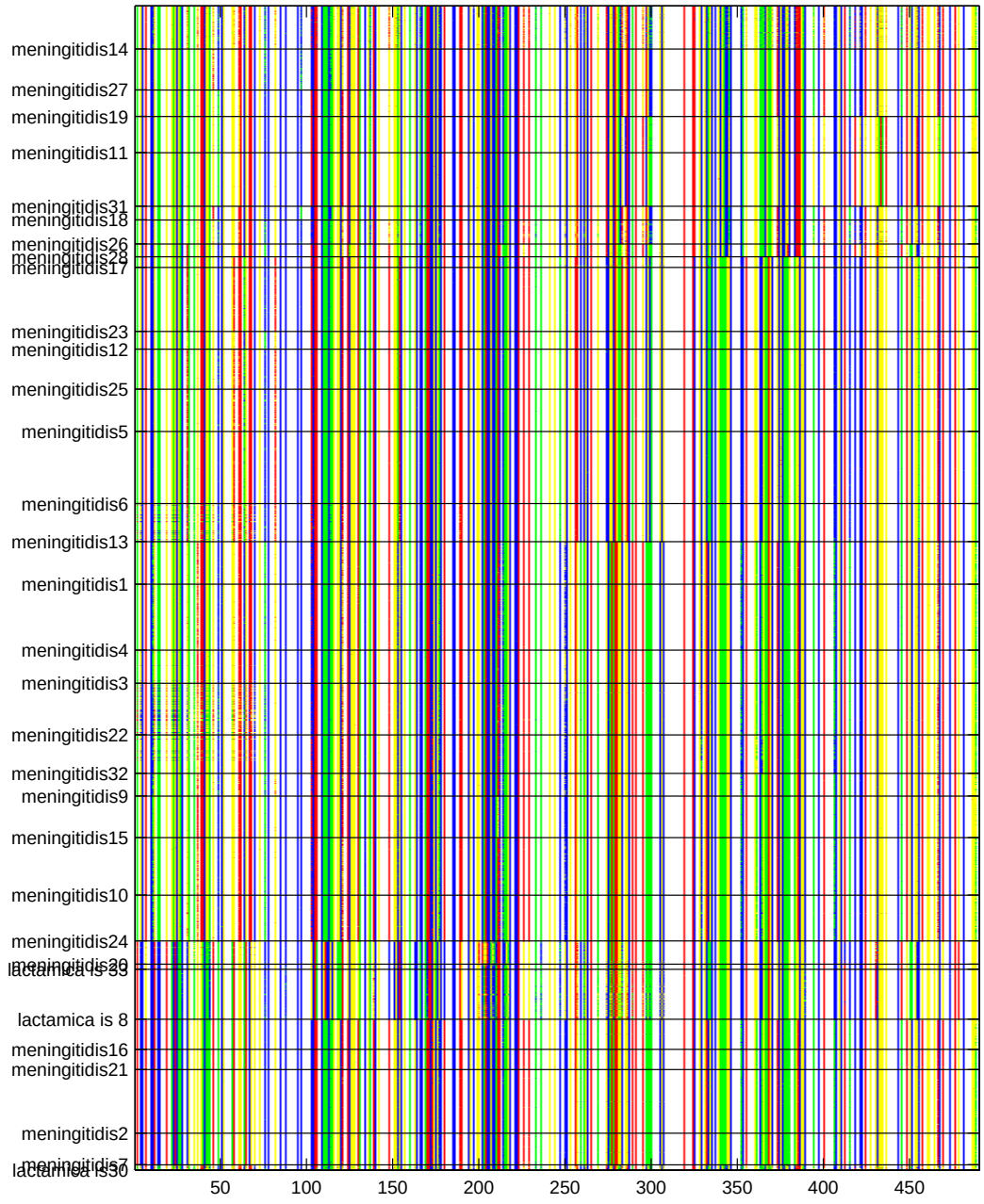

Figure 2: The color-coded image of the *Neisseria* data matrix based on the DNA sequence at gene *recA*. Rows correspond to the individual strains and columns display the nucleotide bases. The data matrix was ordered according to the estimated partition of 32 clusters, which are separated by black lines and labeled according to their named species and cluster numbers.

thus potentially belong to distinct species, while the third subpopulation is influenced constantly by *N. meningitidis*, and should therefore eventually be classified as a *N. meningitidis* species rather than *N. lactamica*. In addition, our result confirms that *N. meningitidis* is a metapopulation that consists of numerous, semi-discrete populations that are influenced by frequent gene flow, and thus forms a coherent group which may be recognized as a fundamental unit of diversity in bacterial species.

3 Comparing bacterial communities

3.1 T-RFLP and FAME data overview

T-RFLP and FAME are two culture-independent methods for characterizing bacterial communities (Danovaro et al., 2006; Ritchie et al., 2000). As a genetic fingerprint method, T-RFLP analysis provides a quantitative measurement of the relative abundance of bacterial species present in a sample (Figure 1 in [IV]). In contrast, FAME is a phenotypic-based method that extracts fatty acid methyl ester profiles which are known as signature markers for assessing community compositions. Despite the fundamental differences in the procedures for these methods, the output profiles produced share similar patterns, *i.e.,* a series of peaks in a spectrum. In T-RFLP analysis, terminal restriction fragments (T-RFs), that represent different bacterial species, are separated by electrophoresis and the output is visualized as an electropherogram. Each peak position corresponds to the size of a particular terminal-restriction fragment (T-RF) and the height of the peak tells its intensity. In FAME analysis each individual FAME is quantified by gas chromatography with its retention time displayed on the horizontal axis and the associated peak area on the vertical axis. For both of the analyses, it is assumed that the number of peaks corresponds to the species number and that the peak height represents the species abundance in the sample.

3.2 The zero-inflated data

In order to assess spatial and temporal changes in the community composition, statistical methods are needed for understanding the relationships between communities. This is usually done by cluster analysis that groups together similar communities based on their T-RFLP and FAME data. A striking feature of the T-RFLP and FAME data is the excessive abundance of zeros. Such a data set, called a *zero-inflated* data set, often arises in a biological or ecological experiment in which a large number of features are measured simultaneously. Measurements below a certain threshold are often assigned a zero value, thus implying their total absence. The term *inflation* is used to emphasize that the probability of a feature going undetected is much higher than would be estimated using well-defined distributions, such as the Poisson or Gaussian distributions (Warton, 2005; Lam et al., 2006). In typical T-RFLP and FAME data processing, peaks with heights no larger than a certain threshold will be typically treated as noise. After such a censoring step, the observed values then follow a distribution with a mixture of high-frequency zeros and positive continuous values.

A common aim in the statistical analysis of zero-inflated data is to determine the similarity between observations in terms of their feature patterns. One can derive a partition structure by assigning observations with similar features into the same cluster. When no well-defined clusters exist *a priori*, inferring the correct partition number is also beneficial. Furthermore, a partition structure sets a basic corner stone, upon which many deeper questions can be pursued via statistical inference. These questions might include the identification of explanatory factors that shape a given partition, or predicting the cluster membership of a new sample (McCartney, 2002; Whitaker and Backfield, 2006).

From the Bayesian statistics viewpoint, this data-partitioning problem can be tackled by model-based unsupervised clustering methods. A model-based approach has the advantage of explicitly modeling the data as a mixture of distinguished clusters. The clustering label of an observation is generally treated as an independent sample from a probability distribution. Given the data likelihood, one can update the label distribution via Bayes' rule and thus obtain an uncertainty measure of the partition. In addition, the number of clusters is normally treated as a parameter of the model and thus need not be specified in advance. Learning the optimal number of clusters corresponds to finding the model that fits the data best, which can be done using a variety of model selection criteria (Richardson and Green, 1997; Spiegelhalter et al., 2002).

Despite the increasing popularity of model-based frameworks, there exists very little literature on the application of model-based clustering methods on zero-inflated datasets. A convenient solution is to rescale the real-valued observations into category groups so that a discrete model can be utilized (Rees et al., 2004). Such a categorization procedure is justified by the fact that the proportion of real values is small compared to the proportion of zero values, so that most of the information is preserved after discretization. However, the inevitable loss of information might be essential in differentiating the samples, especially for highly sensitive bacterial data. In addition, determination of the number of category groups is quite arbitrary and a poor partition result might be obtained by an ill-posed categorization. On the other hand, classifiers based on inherently continuous models, such as Gaussian mixture models, might have a better clustering performance, since they tend to preserve all the information within the data. However, ordinary Gaussian mixture models do not reflect zero-inflation either and will thus more than likely fail to produce sensible partitions. Another practical impediment for Gaussian mixture models is the complex construction of the modeling framework and the enormous computation load, especially for high-dimensional data such as T-RFLP and FAME.

3.3 Bayesian latent mixture modeling

In [IV], we develop a hierarchical Bayesian model-based clustering method to tackle the above-mentioned difficulties. The zero-inflation is explicitly modeled as a mixture of Bernoulli and Gaussian distributions. The clustering framework is based on the latent class concept (Duda et al., 2000), and formulated in the OpenBUGS environment (Thomas et al., 2006). To choose the optimal number of clusters, we utilized the recently introduced BICM criterion for model selection (Raftery et al., 2006). Specifically, the data likelihood under

a partition, $p(data \mid S)$, is defined as a mixture of K cluster densities:

$$p(data \mid S) = \sum_{k=1}^{K} p(data \mid Z = k, \theta) p(Z = k \mid \theta), \tag{6}$$

where the likelihood for each component mixture k is given by a linear mixture

$$p(data \mid Z = k, \theta) = \lambda I(data = 0) + (1 - \lambda) f(data; \theta) I(data \neq 0). \tag{7}$$

The posterior of the partition, which is parameterized by Z and θ, can be updated by incorporating its prior and the data likelihood. We use the standard Gibbs sampling algorithm to construct a Markov chain that converges to the target posterior distribution of these parameters. Furthermore, we use a Bayesian model averaging approach to account for the uncertainty for the cluster number K. The optimal cluster number is obtained by maximizing the posterior model probability, where the integrated likelihood of a model with K clusters is given by

$$p(data \mid K) = \int p(data \mid S) p(S \mid K) dS. \tag{8}$$

Such a marginalized model likelihood can be directly estimated from the posterior simulation of $p(data|S)$ using the BICM criterion introduced in Raftery et al. (2006).

3.4 Bayesian fuzzy likelihood clustering

In [V], we provide an alternative method for the unsupervised clustering of zero-inflated data. We propose a quasi-likelihood framework which preserves the continuous nature of the measurements while invoking a discrete model to account for the excessive zeros. The quasi-likelihood framework can be seen as an extension of the basic clustering framework described in section 2.3. By deriving an analytical expression of the data likelihood and employing an efficient stochastic greedy search algorithm, this approach can identify the clusters corresponding to the species level with good accuracy and a low learning complexity.

The essential novelty in the modeling framework is the interpretation of continuous data as fuzzy observations that approximate binary data. The quasi-likelihood for such fuzzy data is analogous to the likelihood for binary data since

$$p(data) = p^{data}(1 - p)^{1 - data}, \tag{9}$$

where for $data = 0$ or $data = 1$, this equals a binary likelihood and for $0 < data < 1$, it can be interpreted as the quasi-likelihood of a fuzzy observation (Wedderburn, 1974). With this transformation, the resulting fuzzy model leads to an analytical expression of data likelihood that is analogous to the ordinary binary clustering model, for which a common learning algorithm, *e.g.,* the stochastic search algorithm, can be applied to find the best partition.

3.5 Software development

For biologists working with microbial communities using T-RFLP, it is essential that appropriate statistical tools are available for doing the analysis. Therefore, we have implemented our method into a user-friendly software package (T-BAPS), utilizing internally the OpenBugs environment for the MCMC simulations. Our package is freely downloadable. It has built-in graphics for exploring the clustering solutions and it enables an intuitive investigation of the estimated posterior clustering uncertainties ([IV]).

3.6 Discussion

There is a wide range of culture-independent methods available for studying the diversity in a bacterial community. FAME analysis has proven to be a simple and fast phenotypic characterization method, while T-RFLP can provide genotypic fingerprints with high reproducibility. Each method has its own limitations and provides only a partial picture of the bacterial composition in different aspects. To overcome this, a combined approach is often necessary. The ability to discriminate between bacterial communities depends not only on the employed markers, but also on the statistical tools that are chosen for making sensible interpretations from the data.

Statistical analysis of bacterial community structures using either T-RFLP or FAME profiles involves several common aspects of data processing and analysis. These include: 1) rules for differentiating true peaks from background noise, 2) methods for aligning peaks which belong to the same categories, and 3) clustering algorithms that assess similarities between profiles (Abdo et al., 2006). The data processing steps in (1) and (2) can be treated as a normalization procedure so that the output data is digitalized for further analysis. Note that the statistical analysis for (1) and (2) is itself a challenging problem and has attracted a significant amount of research. To keep this thesis focused, we ignore the statistical concerns for the data processing steps and aim at providing two novel clustering algorithms by taking into account the particular zero-inflation characteristic. Therefore, to minimize the statistical error, the proposed clustering algorithms in [IV] and [V] should be used on processed data rather than raw data, as long as the zero-inflation property is preserved.

The zero-inflation of the data, previously overlooked in the literature, is explicitly modeled in the thesis using two distinct strategies. The block of zeros that is commonly observed for most of the peak positions indicates that there are species (denoted by T-RFs) that are often jointly absent in some samples, while present in other samples. The abundance of these species is rather varying in the samples. Such a mixture of zeros and diverse positive values does not fit most of the current Gaussian-based models or discrete models. We believe that our methods capture the particular data generation mechanisms for T-RFLP and FAME, and may be extended to other similar data matrices with the zero-inflation property.

Typically, the number of different types of bacterial communities, i.e. the cluster number, is subjectively determined. A common procedure is to apply ordination methods such as PCA(principal component analysis) or MDS(multidimensional scaling) to visualize the distance between samples and then determine the cluster number by drawing cluster boundaries manually (Rees et al., 2004). Determination of cluster numbers in such an *ad-hoc* manner clearly lacks objectivity and statistical rigor. In contrast, [IV] and [V] provide two

model-based clustering algorithms that can objectively determine the optimal number of clusters. [IV] treats the learning of cluster numbers as a model selection problem, which is similar to the one used in the hierarchical setting. Furthermore, the integrated likelihood of data can be approximated directly from the sequence of likelihoods in the posterior simulation. Such a strategy will enhance the model performance. The algorithm used in [V] is essentially the same as that in [I] and [II] and is associated with lower computational complexity.

Other clustering algorithms, such as UPGMA (Unweighted Pair Group Method with Arithmetic mean) or hierarchical clustering, typically arrange community samples as a dendrogram. In contrast, our methods do not attempt to create such a tree structure, since it is generally not clear which levels of the tree should correspond to ecologically distinct communities. Furthermore, it is difficult to interpret the similarity between samples in a dendrogram in an evolutionary context (Blackwood et al., 2003). We recommend that the evolutionary relationships between samples should be inferred by statistical modeling, rather than by a simple description using somewhat arbitrary distance measures. By formulating a realistic model of the way in which bacterial communities emerge and diversify, it is possible for bacterial population biologists to start exploring underlying mechanisms that may be responsible for the geographical and temporal changes in community structures. The identification of communities clusters that are similar in bacterial compositions, as described in [IV] and [V], will be a critical step towards understanding the complex ecological interactions in natural microbial communities.

4 Conclusion

In this thesis, we have presented several Bayesian model-based clustering methods that can be used for studying different aspects of bacterial diversity. We showed that these methods can be applied to the analysis of data generated by both culture-dependent and culture-independent fingerprinting methods, in an effort to discover the bacterial population or community structures.

Firstly, we presented a Bayesian unsupervised clustering model for MLST data, aimed at providing a quantitative explanation of the hidden genetic population structures. We presented the results of a simulated data analysis, demonstrating that our method that considers the nucleotide linkage works better for identifying subpopulations than the previous method in Corander et al. (2003, 2004) which lacked a linkage model. Furthermore, we combined our clustering results with an admixture analysis to formulate a stochastic model for inferring homologous recombination between subpopulations. The analysis of the *Neisseria* MLST data showed that our method is able to identify the gene flow between subpopulations. In particular, it captures the tiny, but significant recombination between the *N.meningitids* and *N.lactamica* species. As discussed in [II], many of the inferred gene flows have been validated in the literature, allowing us to derive global insights regarding the macroevolution of bacterial populations ([I-II]). A case study on *Streptococcus pneumoniae* MLST data identified a subpopulation characterized by a high rate of recombination and an increased resistance to antibiotics, suggesting that homologous recombination might result in important genetic changes that are crucial for bacteria to develop antibiotic

resistance ([VI]).

Secondly, we presented two modeling frameworks for clustering bacterial fingerprinting data featuring a mixture of discrete and continuous observations. The so-called zero-inflation of the data is explicitly modeled via two distinct strategies. In [IV], the data is assumed to be generated from a mixture of Bernoulli and Gaussian distributions. Under this model assumption, the data likelihood can be derived analytically. In [V], we treated a positive value as a fuzzy observation, reflecting the uncertainty about its actual presence. Such a model formulation enables us to derive a quasi marginal likelihood of a partition in an analytical form. These two methods can be applied to clustering T-RFLP and FAME data, which carry the zero-inflation feature, so that the community structures can be compared between samples.

For all of the above models, we presented algorithms that search the correct partitions in a stochastic way. This learning task is computationally challenging due to the high throughput and complexity nature of the bacterial data. As discussed earlier, MLST data usually comprises, at the very least, hundreds of individuals and thousands of sequence bases. T-RFLP and FAME data usually contain less samples and a smaller number of variables, but the inherent continuous measurements increase the model complexity. To address these challenges, we adapted existing algorithms that exploit problem specific structures and formulated efficient learning algorithms.

Finally, for computational tools to have a broad impact, they must be accompanied by visualization and browsing tools that are easily accessible to biologists. To support this effort, we developed the BAPS software for the visualization and statistical analysis of MLST sequence data, and the T-BAPS software for the analysis of T-RFLP data. These software packages are both freely available online and contain detailed instructions for implementation [III-IV].

We hope that the use of model-based statistical methods in the thesis, coupled with high-throughput fingerprinting techniques, would serve as a starting point for addressing further problems in understanding bacterial diversity. Obviously, there are limiting factors to our model formulations and to the computational algorithms, which need to be addressed in future work. One drawback of our clustering model in [I] is that the dependency structure needs to be specified *a priori*. For analyzing MLST data such a limitation is justifiable since the nucleotide sequences are constrained to be linked linearly as codons. However, for other types of genetic markers, such as AFLP loci (Vatcher et al., 2002), the dependency between the markers, commonly referred to as the *linkage map*, is generally unknown. Modeling of population structures using such data requires additional parameterizations of the unknown dependency structure. This would lead to a more complex model, due to the increasing size of the parameter space. Learning such models from data is computationally very challenging, even when an efficient stochastic search algorithm is applied.

The genetic shape concept introduced in [II] for describing the genetic affinity between subpopulations would be a potential target for further theoretical investigation. Namely, the average change in log likelihoods of moving a randomly chosen individual in a subpopulation to another subpopulation can be interpreted as an indicator of the probability of gene flow. A parallel can be drawn with chemical reactions, where a change in the free energy of a system can potentially trigger a chemical reaction (Weast, 1978). The 'reaction' between subpopulations in the context, characterized by the gene flow events, is thus

regulated by the genetic affinity. Finding out the inherent relationships between the genetic affinity and the gene flow propensity is a key challenge for understanding the systematic picture of bacterial population evolution (Fraser et al., 2007).

As an important component in the analysis of bacterial population structures, the gene flow pathway identified from MLST data can be seen as an explanation of the evolutionary mechanisms that introduce homologous recombination across the subpopulations. Due to the discrete nature of MLST data, recombination is readily detectable as it can be inferred statistically from a deviation of the expected allele frequencies when considering mutation as the only evolutionary mechanism ([II]). However, in the case of community structure analysis, this thesis does not provide the equivalent modeling for detecting evolutionary relationships between different community clusters. This is due to the difficulties of decomposing a continuous observation as a mixture of components from different sources, e.g. community clusters. Independent component analysis (ICA) has been applied to continuous data for discovering mutually independent hidden factors (Shohei and Hoyer, 2005). We are currently investigating the possibilities of applying ICA to T-RFLP and FAME data, in an effort to detect independent evolutionary forces that are influencing the community structures of bacteria.

Acknowledgements

As a statistician I have been trained to choose the action that always gives rise to the highest expected value. Most of time I would admit that I just miss the target. I have made too many choices that usually mean nothing, and sometimes even could incur unpleasant consequences that I surely feel regret for a while. At the final stage of my Phd thesis I am always pondering: what is good about this at the cost of the most previous time in my age? I still remember five years ago when I was struggling on writing my Master thesis, I kept asking a same question over and over: Are you sure you want to continue?

Yes, I am, and I knew that it was going to be an important, and perhaps a great decision. A great decision by nature is never easy, as it requires the best of my energy and patience. I chose to pursue the Phd, not because it is easy, but because it is difficult, and I knew that I will not be alone when doing this. In spite of all the difficulties, the biggest comfort that I have been constantly received is the fact that, from the beginning to the end, there have been the best people accompanying me all the time. It is their countless support and love that makes the whole thing meaningful.

First and foremost I wish to thank my supervisor, Jukka Corander, for his guidance and support throughout the entire work. I have learned tremendously from his enthusiasm and brilliance for choosing and pursuing interesting research topics, from the basic ideas of Bayesian statistics, to the development of computational biology models and algorithms. Discussions and brainstorming sessions with him have been some of the best moments I have had. I have been deeply impressed by his deep knowledge on both mathematics and biology. His valuable and constructive comments during the process have always been of great help.

I would like to thank Elja Arjas, my co-supervisor, for creating a nice atmosphere to work with and for helpful suggestions in conducting the thesis plan and all the practical issues. This project would not have been possible without Elja's administrative work. He is always smiling and kind to me, like a father to his son. I greatly enjoyed working with him and am indebted to him for his dedication and for the patience and encouragement I learned from him.

I owe much to Pekka Pamilo and his Center of Excellence in Population Genetics analyses. Pekka is the one who convinced me to come to Helsinki in the first place, with a warm welcome that I had never experienced before. Through a lot of training courses given by the center, Pekka provided me with a great opportunity to learn biology and understand the biologists' way of thinking. Through the interaction with Pekka and the members of his center, I appreciated the essence of using mathematics in deciphering biological puzzles.

I would like to thank Daniel Thorburn for agreeing to be my opponent in the public examination. I also thank Esa Läärä and Sameul Kaski for their critical comments when reviewing my thesis. Also, I thank Mats Gyllenberg for agreeing to be the Custos.

This thesis was jointly funded by the Graduate School in Computational Biology, Bioinformatics and Biometry (ComBI) and the Centre of Population Genetic Analyses supported by the Academy of Finland. Thank you for your financial support.

The work in this thesis is the fruit of a collaboration with a number of researchers from various backgrounds. Special thanks go to William P. Hanage, for his inspiring vision and for his trust in our methods. The collaboration with William has been the most fruitful

one, as the validation of our statistical models for bacterial population genetics led to a significant finding in the identification of antibiotic resistance subpopulations.

I wish to thank every one of the other co-authors in my thesis for advice, comments and support: Pekka Marttinen, Jukka Sirén, Christophe Fraser, Thomas R Connor, Jinglun Tao, Hitetoshi Urakawa, Bernard De Baets and Peter Dawyndt. I would also like to thank my fellow colleagues: Matti Pirinen, Jukka Kohonen, Bob O'hara, Crispin Mwanza, Dario Gasbarra, Mikko Sillanpää, Sarish Talikota, Rashi Gupta, Rossana Moroni, Ping Yan and Martti Nikunen. Thank you for the enjoyable company during seminars, conferences, parties, sports and breaks. It has been a great pleasure to work with you guys, and I look forward to continuing our collaboration in the near future.

Special thanks go to Pasi Tuohino, Riitta Ulmanen, Tarja Hämäläinen, Raili Pauninsalo and all those in the administrative corridor who took care of all the housekeeping matters. Thank you for the excellent work.

A great deal of the writing in the summary was done while I am working at VTT, Technical Research Center of Finland. I thank Catherine Bounsaythip who helped me realize the importance of pushing myself towards the finalization of the thesis. I also wish to thank Matej Oršič, Richard Fagerstöm, Kim Ekroos and Marko Sys-Aho for their understanding and support so that I can make the thesis ready well on schedule.

Many thanks to my friends who brought a lot of joy and fun in the fellowship. The names are too many to be listed here. I thank them for making me realize that there are things more important than a Phd degree.

My deep gratitude goes to my parents, for everything they taught me and for the values they have given me. You might not know what is written here, but please take this as a gift because a big part of this achievement belongs to you.

Finally and above all, I am eternally grateful to my wife Jing for her unconditional love and support, for sticking with me from the beginning and throughout this long and at times very difficult journey, and most recently, for sharing with me the most important and most happy moment of my life, the birth of Gabel.

References

Abdo, Z., Schuette, U. M., J.Bent, S., J.Williams, C., J.Forney, L., and Joyce, P. (2006). Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16s rrna genes. *Environmental microbiology*, 8(5):929–938.

Achtman, M. (2004). Population structure of pathogenic bacteria revisited. *International journal of medical microbiology*, (294):67–73.

Achtman, M. and Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature reviews microbiology*.

Blackwood, C. B., Marsh, T., Kim, S.-H., and Paul, E. A. (2003). Terminal restriction fragment length polymorphism data analysis for quantitative comparsion of microbial communities. *Applied and environmental microbiology*, 69(2):926–932.

Cooper, J. E. and Feil, E. J. (2004). Multilocus sequence typing - what is resolved? *TRENDS in Microbiology*, 12(8):373–377.

Corander, J., Gyllenberg, M., and Koski, T. (2006a). Bayesian model learning based on a parallel MCMC strategy. *Statistics and computing*, 16:355–362.

Corander, J., Gyllenberg, M., and Koski, T. (2007). Random partition models and exchangeability for bayesian identification of population structure. *Bulletin of Mathematical Biology*, 69:797–815.

Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multi-locus molecular markers. *Molecular Ecology*, 15:2833–2843.

Corander, J., Marttinen, P., and Mäntyniemi, S. (2006b). Bayesian identification of stock mixtures from molecular marker data. *Fishery bulletin*, 104:550–558.

Corander, J., Waldmann, P., Marttinen, P., and Sillanpaa, M. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, 20:2363–2369.

Corander, J., Waldmann, P., and Sillanpää, M. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics*, 163:367–374.

Danovaro, R., Luna, G., Anno, A., and Pietrangeli, B. (2006). Comparison of two fingerprinting techniques, terminal restriction fragment length polymorphism and automated ribosomal intergenic spacer analysis, for determination of bacterial diversity in aquatic environments. *Applied and environmental microbiology*, 72(9):5982–5989.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern classification and scene analysis.* New York: Wiley.

Falush, D., Wirth, T., Linz, B., Pritchard, J., and Stephens, M. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science*, 299:1582–1585.

Feil, E. J., Li, B. C., Aanensen, D. M., P.Hanage, W., and Spratt, B. G. (2004). eburst: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of bacteriology*, 186(5):1518–1530.

Feil, E. J. and Spratt, B. (2001). Recombination and the population structures of bacterial pathogens. *Annual review of microbiology*, 55:561–590.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, 315:476–480.

Gao, X. and Starmer, J. (2007). Human population structure detection via multilocus genotype clustering. *BMC Genetics*, 8(34).

Hanage, W. P., Fraser, C., and Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biology*, 3(6).

Holmes, E. C., Urwin, R., and Maiden, M. C. (1999). The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningtidis*. *Molecular Biololgy Evolution*, 16(6):741–749.

Jia, W., Li, H., Zhao, L., and K.Nicholson, J. (2008). Gut microbiota: a potential new territory for drug targeting. *Nature Reviews: Drug Discovery*, 7:123–129.

Kapur, M. and Jain, R. K. (2004). Microbial diversity: Exploring the unexplored. *World Federation of Culture Collection (WFCC) Newsletter*, 39:12–16.

Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H., and Trevors, J. T. (2004). Methods of studying soil microbial diversity. *Journal of Microbiological Methods*, 58:169–188.

Lam, K., Xue, H., and Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated cound data. *Biometrics*, (62):996–1003.

Latch, E., Dharmarajan, G., Glaubitz, J., and Rhodes, O. (2006). Relative performance of bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation genetics*, 7:295–302.

Marsh, T. L. (1999). Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Current opinion in microbiology*, 2:323–327.

McCartney, A. (2002). Application of molecular biological methods for studying probiotics and the gut flora. *British Journal of Nutrition*, 88(1):S29–S37.

Park, S., Ku, Y., Seo, M., Kim, D., and Yeon, J. (2006). Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles of terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities. *Soil Biology and Biochemistry*, 38:2344–2349.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2006). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In Bernardo, J., editor, *Bayesian Statistics 8*, pages 1–45. Oxford University Press.

Rees, G. N., Baldwin, D. S., Watson, G. O., Perryman, S., and Nielsen, D. L. (2004). Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics. *Antonie van Leeuwenhoek*, 86:339–347.

Richardson, S. and Green, P. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society. B*, 59:731–792.

Ritchie, N. J., Schutter, M. E., Dick, R. P., and Myrold, D. D. (2000). Use of length heterogeneity PCR and Fatty acid methyl ester profiles to characterize microbial communities in soil. *Applied and environmental microbiology*, 66(4):1668–1675.

Shohei, S. and Hoyer, P. (2005). Discovery of non-gaussian linear causal models using ICA. *In Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, pages 526–533.

Smith, J. M., Feil, E. J., and Smith, N. H. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays*, 22:1115–1122.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society*, 64:583–640.

Spratt, B. G. and Maiden, M. C. (1999). Bacterial population genetics, evolution and epidemiology. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences.*, (354):701–710.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS open. *R News*, 6:12–17.

Vatcher, G., Smailus, D., and Krzywinski, M. (2002). fRFLP and fAFLP: medium-throughput genotyping by fluorescently post-labeling restriction digestion. *Biotechnqiues*, 33(3):539–546.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, (16):275–289.

Weast, R. C. (1978). *Handbook of chemistry and physics*. CRC press, Inc.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447.

Whitaker, R. J. and Backfield, J. F. (2006). Population genomics in natural microbial communities. *TRENDS in Ecology and Evolution*, 21(9):508–516.