UNIVERSITY OF TECHNOLOGY SYDNEY

STATISTICAL METHODS FOR INFERRING RECOMBINATION IN BACTERIAL GENOMES

FATEMEH-NEHLEH KARGARFARD

PhD

September 2022

UNIVERSITY OF TECHNOLOGY SYDNEY

STATISTICAL METHODS FOR INFERRING RECOMBINATION IN BACTERIAL GENOMES

FATEMEH-NEHLEH KARGARFARD

Thesis submitted in fullfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Science

September 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Fatemeh-Nehleh Kargarfard, declare that this thesis is submitted to fulfil the requirements for the award of Doctor of Philosophy in the Faculty of Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with the NSW Department of Primary Industries and the University of Technology Sydney. The Australian Research Council supported this work; linkage grants LP180100593—the Australian Centre funded this project for Genomic Epidemiological Microbiology (AusGEM).

This research is supported by the Australian Government Research Training Program.

(Fatemeh Nehleh Kargarfard)

Date: 08/09/2022

Production Note:
Signature removed prior to publication.

ABSTRACT

Homologous recombination events in bacterial genomes have wide-ranging effects on public health in society. This phenomenon is a significant factor explaining the prevalence of antimicrobial resistance. When recombination occurs in bacteria, a segment of foreign DNA is introduced into its chromosome. This evolutionary mechanism can give rise to antibiotic resistance. On the other hand, reconstructing bacteria's evolutionary history in the presence of recombination is notoriously tricky. Phylogenetic trees investigate evolutionary history and relationships between organisms, which is essential for understanding and analyzing natural processes. These trees are required tools for numerous fundamental and practical research. Recombination detection in a bacterial genome is one application of phylogenetic trees. Advanced phylogenetic inference methods, such as maximum likelihood and Bayesian inference, use probabilistic models that are computationally expensive, especially for large datasets. Detecting the boundary of recombination events and reconstructing a global phylogenetic tree illustrating the underlying evolutionary pattern of biological sequences has never been a straightforward problem. At the same time, the rapidly growing number of bacterial whole-genomes has produced an extra challenge for the computational approaches to reconstructing fast and accurate phylogenetic trees with the presence of recombination.

In this thesis, we introduce PhiloBacter, a maximum likelihood-based tool to detect recombination in bacterial genomes and account for it during phylogenetic reconstruction. Specifically, it estimates the probability of each site in an alignment to be recombinant. We then presented two approaches to incorporate these probabilities to infer the clonal history of these genomes. The first borrows ideas from sequencing error estimation, and the other uses mixtures of matrices to account for uncertainty introduced through recombination.

We also present a new simulation tool, BaciSim, for bacterial genomes that undergo recombination.

Finally, we developed a software pipeline for the semi-automatic identification of recombination and reconstruction of a phylogenetic tree from an alignment bacterial genome. Using simulated datasets, we investigated the accuracy and reliability of our approach to detect recombination events and to get better estimates of the clonal history of a collection of genomes that underwent recombination. We benchmarked our methods with other widely used methods (Gubbins and ClonalFrameML). Our simulations show that PhiloBacter tends to outperform these two methods.

ACKNOWLEDGEMENT

First, I would like to express my sincere thankfulness to both of my supervisors, Professor Aaron Darling and Dr. Mathieu Fourment, for the scientific and non-scientific lessons they taught me. I'm proud of myself for working under their supervision. Their valuable advice, continuous encouragement, and constant and kind support were critical throughout the development of my research work. Without their help and support, this project would not have been possible.

I am also grateful to all team members for their technical and emotional support: Sid Krishnan, Frederick Jaya, Dr. Daniela Gaio, Dr. Barbara Brito-Rodriguez, Dr. Kay Anantanawat, Dr. Leigh Monahan, and Dr. Matthew Macaulay.

I am sincerely thankful to Dr. Mehrad Hamidian for all his valuable advice and support.

I also thank all the Faculty and institute staff for administratively helping me during these years of my study.

Beyond work, I would like to thank my parents and sister for their consistent support and encouragement during all moments of my life. Their love was always close to me, even if they were geographically thousands of kilometers away.

I am sincerely thankful to all my friends for being in my life since I came to Australia. They always encouraged me during difficult times; my life wasn't so colorful without their presence.

I must also sincerely and deeply thank my partner. His companionship and love were the best encouragement during the last months of my studies.

Lastly, I genuinely acknowledge the organizations that provided financial support for this research, including the Australia Research Council and the Australian Centre for Genomic Epidemiological Microbiology (AusGEM).

THESIS FORMAT

This is a conventional thesis consisting of 6 chapters. This thesis aims to develop statistical methods for inferring recombination in bacterial genomes.

Chapter 1 is an introduction chapter, including the research background, motivation, aim, and objectives.

Chapter 2 is a literature review of conventional methods for inferring phylogenetic trees and recombination detection methods.

Chapter 3 is about general methods in this research work which introduce PhiloBacter.

Chapter 4 presents a new simulation tool for bacterial genomes that undergo recombination.

Chapter 5 is Result and Discussion.

Chapter 6 summarizes the main research outcomes of this thesis and possible future research directions that may arise based on the advances made.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES