

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 fulfillment
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

	Page
CERTIFICATE OF ORIGINAL AUTHORSHIP	i
ABSTRACT	ii
ACKNOWLEDGEMENT	iii
Thesis Format	iv
TABLE OF CONTENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER ONE: INTRODUCTION	1
1.1 Research Motivation	1
1.2 Recombination	3
1.2.1 Homologous recombination in bacteria	4
1.3 What is a phylogenetic tree?	8
1.3.1 Mathematical definition of phylogenetic tree	9
1.3.2 Computer format of phylogenetic tree	10
1.4 Impact of recombination on phylogenetic analyses	12

1.5	Research Objectives and Contributions	13
1.6	Thesis Structure	14
CHAPTER TWO: LITERATURE REVIEW		15
2.1	Methods for inferring phylogenetic trees	15
2.1.1	Distance-based Approaches:	15
2.1.1.1	UPGMA and WPGMA	16
2.1.1.2	Neighbor-Joining	16
2.1.1.3	Maximum parsimony (MP)	16
2.1.2	Modeling Nucleotide Evolution	17
2.1.2.1	JC69 model	19
2.1.2.2	Some other models	20
2.1.2.3	GTR model	20
2.1.3	Statistical Approaches	21
2.1.3.1	Maximum Likelihood Method	21
2.1.3.2	Bayesian analysis using MCMC	24
2.2	Detection of recombination events in bacterial genomes	26
2.2.1	Phylogenetic-based recombination detection methods	28
2.2.1.1	ClonalFrame and ClonalFrameML	29
2.2.1.2	ClonalOrigin	30
2.2.1.3	Gubbins	30
2.2.1.4	Bacter	31
2.3	Summary	31
CHAPTER THREE: RESEARCH METHODOLOGY: PhiloBacter		33

3.1	Introduction to HMM	33
3.1.1	Basic Understanding of Markov Model	34
3.1.2	Hidden Markov Model (HMM)	36
3.1.3	What kind of problems can be resolved through the Hidden Markov Model?	38
3.2	Incorporating Sequence Uncertainty in Phylogenetic Inference	40
3.2.1	Uncertainty models in DNA sequences	42
3.3	PhiloBacter: A new tool to infer phylogenetic trees from recombinant bacterial genomes	43
3.3.1	Overview	43
3.3.2	Step one: Recombination Estimation	44
3.3.2.1	Description of algorithm	44
3.3.2.2	Lesson learned	52
3.3.3	Step two: Clonal tree Inference using whole genomes	55
3.3.3.1	PhiloBacter: Maximum Likelihood Calculation	56
3.3.3.2	PhiloBacter: Uncertainty in clonal tree Inference	58
3.4	Summary	62
CHAPTER FOUR: RESEARCH METHODOLOGY: BaciSim		64
4.1	Introduction	64
4.2	Background	65
4.3	Methods	67
4.3.1	Overview of BaciSim Simulator	68
4.4	Discussion	70
CHAPTER FIVE: RESULTS AND DISCUSSIONS		73

5.1	Pipeline introduction	73
5.1.1	Installing the pipeline	73
5.1.1.1	Prerequisites:	73
5.1.1.2	Installation Steps:	74
5.1.2	Analysis mode	75
5.1.3	Simulation mode	76
5.2	Evaluation Metrics	78
5.2.1	Evaluating Recombination Detection	78
5.2.2	PhiloBacter as a regression model	79
5.2.3	PhiloBacter as a classification model	79
5.2.4	Evaluating Phylogenetic reconstruction	81
5.3	Performance Evaluation	82
5.3.1	BaciSim Simulator	82
5.3.1.1	Investigating different values of v	82
5.3.1.2	Investigating different recombination lengths	86
5.3.1.3	Investigating different recombination rate	88
5.3.1.4	Investigating different tMRCA	89
5.3.2	SimBac Simulator	91
5.3.3	FastSimBac Simulator	93
5.4	Comparison with fastGEAR	95
5.4.1	fastGEAR Simulated Data	96
5.4.2	BaciSim Simulator	98
5.4.2.1	Detecting Recent Recombinations	98
5.4.2.2	Detecting Ancestral Recombinations	99

5.4.2.3	Detecting All Recombinations	101
5.5	Application to empirical data	101
5.5.1	Application to <i>Streptococcus pneumoniae</i>	101
5.5.2	Application to <i>Staphylococcus aureus</i>	106
5.5.3	Application to <i>Bacillus Cereus</i>	108
5.6	Resource Usage	111
5.7	Discussion	116
CHAPTER SIX: Conclusion and future directions		119
6.1	Conclusion	119
6.2	Perspectives and Future Research	121
6.2.1	PhiloBacter as a fast and more efficient tool	121
6.2.2	Generalised uncertainty approach	122
6.2.3	Internal and external nodes	123
6.2.4	A dynamic and general simulator	124
6.2.5	PhiloBacter and quantum computing	125
REFERENCES		126
APPENDICES		145

LIST OF TABLES

Tables	Title	Page
Table 5.1	Summary of recent recombination (leaves) detection for three selected lengths (500, 1000, and 1500) using BaciSim, fastGEAR, Gubbins, CFML, and PhiloBacter. The simulated data were generated using the following settings: v : 0.03, Number of genomes: 10, Alignment length: 100K, and tMRCA: 0.01. The left column indicates the number of simulated recombination events, with the 'Anc' columns set to zero to represent the absence of ancestral recombination in the simulated data. Column 'Rec' shows the number of recent recombination events.	99
Table 5.2	Summary of ancestral recombination (internal nodes) detection for three selected lengths (500, 1000, and 1500) using BaciSim, fastGEAR, Gubbins, CFML, and PhiloBacter. The simulated data were generated using the following settings: v : 0.03, Number of genomes: 10, Alignment length: 100K, and tMRCA: 0.01. The left column indicates the number of simulated recombination events, with the 'Rec' columns set to zero to represent the absence of recent recombination in the simulated data. Column 'Anc' shows the number of ancestral recombination events.	100

LIST OF FIGURES

Figures	Title	Page
Figure 1.1	A schematic view of transduction in which external DNA is entered into a bacterial cell by a virus (Image modified from “Conjugation,” by Adenosine (CC BY-SA 3.0). The modified image is licensed under a CC BY-SA 3.0 license.)	5
Figure 1.2	A schematic view of transformation in bacteria in which some bacteria take up naked DNA from the environment (Image modified from “Conjugation,” by Adenosine (CC BY-SA 3.0). The modified image is licensed under a CC BY-SA 3.0 license.)	7
Figure 1.3	A schematic view of transformation in bacteria in which one bacterial cell transfers genetic material to the other cell through straight contact (Image modified from “Conjugation,” by Adenosine (CC BY-SA 3.0). The modified image is licensed under a CC BY-SA 3.0 license.)	8
Figure 1.4	The topology of all trees is the same. (a) Rooted phylogenetic tree - Cladogram (b) Rooted phylogenetic tree - Phylogram (c) Unrooted phylogenetic tree.	10
Figure 2.1	An example of unrooted tree topology for four taxa with two internal nodes [71].	22
Figure 3.1	Representation of a Markov Chain with two states: This shows a Markov Chain with states S_0 and S_1 , and the arrows indicate the probability of departing from one state to the other state or remaining the same [114]	35
Figure 3.2	A graphical representation of an HMM	38

Figure 3.3	Overview of the PhiloBacter process, illustrating (Step 1) estimation of recombination events and (Step 2) phylogenetic inference accounting for the presence of recombination.	44
Figure 3.4	Schematic representation of the PhiloBacter HMM framework. This model encompasses five core components: (1) Hidden State Sequence, (2) Start Probability, (3) Transition Probability, (4) Emission Probability, and (5) Observation Sequence,	45
Figure 3.5	Illustration of the bi-state HMM approach utilized in this study for recombination state identification. The red node exemplifies a target node, highlighting the hypothesis that recombination amplifies polymorphism within each branch, subsequently elongating it.	46
Figure 3.6	Example of the HMM model observation for three different nodes, including external and internal nodes.	47
Figure 3.7	Depiction of the Four-State HMM framework. The red node serves as a representative example of the target node within the model. State 1 shows ClonalFrame, which didn't undergo any recombination.	53
Figure 3.8	Eight trees indicate the hidden states of HMM model. The red node serves as a representative example of the target node within the model. State 1 shows ClonalFrame, which didn't undergo any recombination.	55
Figure 4.1	A collection of clonal and local trees which simulates various evolutionary schemes for every segment of the genomes might represent mosaic evolutionary histories.	68
Figure 4.2	Overview of simulation studies. A flow diagram of the steps in a simulation study [130].	69

Figure 4.3	An example of the graphic output of BaciSim shows recombination events, their location and length along the alignment.	72
Figure 5.1	A schematic view of the pipeline modes. a) shows the pipeline has two modes; b) shows the steps of the analysis mode.	75
Figure 5.2	A schematic view of the pipeline. Detailing its five primary stages: (1) Simulation, (2) Sequence Generation, (3) Initial Tree Construction, (4) Recombination Detection and Tree Inference, and (5) Analysis.	78
Figure 5.3	RMSE for different value of ν of three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - tMRCA:0.01 - Recombination length:500	83
Figure 5.4	Accuracy for different values of ν for three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - tMRCA:0.01 - Recombination length:500	83
Figure 5.5	F1-Score for different values of ν for three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - tMRCA:0.01 - Recombination length:500	84
Figure 5.6	BaciSim Simulator: Distances between true (clonal) tree and trees of three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - tMRCA:0.01- Recombination rate:0.01 - Recombination length 500	85
Figure 5.7	BaciSim Simulator: Investigating how accurate each method (PhiloBacter, Gubbins and CFML from left to right) can estimate different intervals of recombination length. Number of genome:10 - Alignment length:100K - nu:0.05- tMRCA:0.01	87

- Figure 5.8 **BaciSim Simulator:** Distances between true (clonal) tree and trees of three methods **for different recombination rate:** ClonalFrameML, Gubbins, PhiloBacter- Number of genome:10 - Alignment length:100K - nu:0.05- tM-RCA:0.01 - Recombination length 500 89
- Figure 5.9 **BaciSim Simulator:** Distances between true (clonal) tree and trees of three methods **for different values of tMRCA:** ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - nu:0.05- Recombination rate:0.01 - Recombination length 500 91
- Figure 5.10 **SimBac Simulator:** Distances between true (clonal) tree and trees of three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K- Recombination rate:0.0005 - Recombination length:500 93
- Figure 5.11 **FastSimBac Simulator:** Distances between true (clonal) tree and trees of three methods: ClonalFrameML, Gubbins, PhiloBacter - Number of genome:10 - Alignment length:100K - Recombination rate:0.0005 - Recombination length: 500 95
- Figure 5.12 The figure visually represents the population's genetic structure inferred by fastGEAR. In this illustration, the rows are aligned with the sequences, the columns correspond to specific positions within the alignment, and various colors depict different populations. 97
- Figure 5.13 Analysis of the PMEN1 genome alignment with Gubbins employing different phylogeny construction strategies [87] 103
- Figure 5.14 Analysis of the PMEN1 genome alignment with PhiloBacter. a) PhiloBacter tree, b) CFML tree 104
- Figure 5.15 Analysis of the PMEN1 genome alignment with CFML. Vertical bars indicate recombination events detected by the CFML. 105

Figure 5.16 Analysis of the PMEN1 genome alignment with PhiloBacter. Vertical bars indicate recombination events detected by the analysis.	107
Figure 5.17 The analysis of <i>Bacillus cereus</i> strains isolated in Australia: a) Tree representation using Gubbins, b) CFML-based phylogenetic tree, and c) PhiloBacter tree analysis.	110
Figure 5.18 The analysis of <i>Bacillus cereus</i> strains isolated in Australia with PhiloBacter. Vertical bars indicate recombination events detected by the analysis.	112
Figure 5.19 Resource usage comparison chart showcasing CPU consumption across various tools: PhiloBacter, CFML, Gubbins (Gubbins-result is our custom script for output manipulation), and RAxML.	113
Figure 5.20 Resource usage comparison chart showcasing RAM consumption across various tools: PhiloBacter, CFML, Gubbins (Gubbins-result is our custom script for output manipulation), and RAxML.	114
Figure 5.21 Comparative chart of job duration, illustrating the processing times for PhiloBacter, CFML, Gubbins (enhanced with our tailored script -Gubbins-result- for output refinement), and RAxML.	115
Figure 5.22 I/O resource usage comparison chart, displaying results for PhiloBacter, CFML, Gubbins (modified using our custom script -Gubbins-result- for output optimization), and RAxML.	116
Figure A.1 BaciSim Simulator: This figure is the same as Figure 5.3. We removed Gubbins's data to clarify the differences between CFML and PhiloBacter(s).	147
Figure A.2 BaciSim Simulator: This figure is the same as Figure 5.4. We removed Gubbins's data to clarify the differences between CFML and PhiloBacter(s).	148

Figure A.3	BaciSim Simulator: This figure is the same as Figure 5.5. We removed Gubbins's data to clarify the differences between CFML and PhiloBacter(s).	149
Figure A.4	SimBac Simulator: This figure is the same as Figure 5.10. We removed Gubbins data to clarify the differences between CFML and PhiloBacter.	150
Figure A.5	SimBac Simulator: This figure is the same as Figure 5.11. We removed Gubbins data to clarify the differences between CFML and PhiloBacter.	151
Figure A.6	SimBac Simulator: This figure is the same as Figure 5.12. We removed Gubbins data to clarify the differences between CFML and PhiloBacter.	152
Figure A.7	Details of Australian genomes used in section 5.5.3 Application to <i>Bacillus Cereus</i> .	153