Systemic Testing of Improved Bacteria Speciation Algorithms

Diego Calderon

April 2013

Acknowledgements

Parents
Cohan and Krizanc
TheLab
Thesis committee
Buds

First I would like to thank my parents for always providing their warmth and support when I needed it most.

Abstract

Understanding Bacteria Speciation is difficult
Diversity
Quirkiness
Molecular Models for Demarcation
ES2 is fast
One line of comparison results

Contents

1 Introduction **3**
- 1.1 Description of Problem . . . 3
  - 1.1.1 Diversity of Bacterial Species . . . 4
  - 1.1.2 Differences of Bacterial Population Dynamics . . . 4
- 1.2 Benefits . . . 4
- 1.3 My Thesis Purpose . . . 4

2 Prior Work **5**
- 2.1 Background and Theory . . . 5
- 2.2 Algorithms . . . 5
- 2.3 Ecotype Simulation . . . 5

3 Improvements of Ecotype Simulation **6**
- 3.1 Ecotype Simulation 2.0 . . . 6
- 3.2 Binning . . . 6
- 3.3 Parallelization . . . 6

4 Systemic Testing of Available Bacteria Speciation Models **7**
- 4.1 Methods . . . 7
  - 4.1.1 Generation of sequences for analysis . . . 7
  - 4.1.2 Variation of Information Metric . . . 7
  - 4.1.3 Running Time Tests . . . 7
- 4.2 Results and Discussion . . . 7
  - 4.2.1 Analysis of in silico-generated sequences . . . 7
  - 4.2.2 *Bacillus* sequences . . . 7
  - 4.2.3 Running time . . . 7

5 Conclusion **8**

Bibliography **9**

Index **9**

Chapter 1

Introduction

> “The sure and definite determination (of species of bacteria) requires so much time, so much acumen of eye and judgement, so much of perseverance and patience that there is hardly anything else so difficult.”
>
> *Otto F. Müller*

1.1 Description of Problem

The quote above, by no mistake, graced the cover of the International Journal of Systematic and Evolutionary Microbiology for decades. Whereas plant and animal systematists are guided by a theory based approach to species, microbiologists have yet to agree on a set of ecological and evolutionary properties that could serve to identify bacterial species [2]. They are naturally handicapped by the paucity of morphological differences that could aide in differentiation of closely related bacterial species. Another factor is that microbiologists cannot predict which traits will cause a speciation event since bacteria are capable of receiving genes from distant relatives through a process known as horizontal gene transfer (HGT) [2]. Thus, in order to effectively understand the microbiome we must strive towards developing a method for consistently demarcating groups, from bacterial diversity, that play distinct ecological roles [4].

Initially, closely related bacterial species were identified based on metabolic phenotype. Systematists now rely on molecular approaches that utilize the decreasing cost of DNA sequencing to compare genetic information. A 70% cutoff was established for whole genome hybridization studies (comparing loss and gain of large chunks of DNA), replaced by varying degrees of sequence identities in

homologous genes [2, 3]. From these technological breakthroughs scientists have taken great steps towards understanding bacteria speciation, at the same time they have brought into focus new difficulties.

1.1.1 Diversity of Bacterial Species

Estimates of eukaryotic diversity fall within the range of 10 to 50 million species. Even though we have only observed approximately 9000 prokaryotic species, indirect approaches that do not rely on cultivation hint at the existence of a billion or more prokaryotic species worldwide and 10 million within a given habitat [1]. The only rational approach to grouping such a large group of diverse organisms effectively is with a theory based molecular method.

Current methods for deciding bacterial lineage are not complete. Recent ecological studies show that a named bacterial species is typically an assemblage of closely related but ecologically distinct populations [2]. Our envisioned demarcation algorithm would be capable of identifying putative clusters of ecologically distinct organisms within named bacterial clades.

1.1.2 Differences of Bacterial Population Dynamics

As briefly mentioned earlier their exist peculiarities of bacterial population dynamics that complicate demarcation.

1.2 Benefits

Bioremediation Antibiotic resistance countless health benefits others

1.3 My Thesis Purpose

My aims for this final project are several.
Should I put this section at the beginning?

Chapter 2

Prior Work

2.1 Background and Theory

2.2 Algorithms

2.3 Ecotype Simulation

Chapter 3

Improvements of Ecotype Simulation

3.1 Ecotype Simulation 2.0

3.2 Binning

3.3 Parallelization

Chapter 4

Systemic Testing of Available Bacteria Speciation Models

4.1 Methods

4.1.1 Generation of sequences for analysis

Preparing the input

Bacillus sequences

4.1.2 Variation of Information Metric

4.1.3 Running Time Tests

4.2 Results and Discussion

4.2.1 Analysis of in silico-generated sequences

4.2.2 *Bacillus* sequences

4.2.3 Running time

Chapter 5

Conclusion

Bibliography

[1] Frederick M Cohan and Alexander F Koeppel. The origins of ecological diversity in prokaryotes. *Current Biology*, 18(21):R1024–R1034, 2008.

[2] Frederick M Cohan, Elizabeth B Perry, et al. A systematics for discovering the fundamental units of bacterial diversity. *Current Biology*, 17(10):373, 2007.

[3] Juan Carlos Francisco, Frederick M. Cohan, and Danny Krizanc. Demarcation of bacterial ecotypes from dna sequence data: A comparative analysis of four algorithms. *Computational Advances in Bio and Medical Sciences, IEEE International Conference on*, 0:1–6, 2012.

[4] Alexander Koeppel, Elizabeth B Perry, Johannes Sikorski, Danny Krizanc, Andrew Warner, David M Ward, Alejandro P Rooney, Evelyne Brambilla, Nora Connor, Rodney M Ratcliff, et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences*, 105(7):2504–2509, 2008.

Index

demarcating, 3

testing, 3