CSCI 5314

Algorithms in Molecular Biology

Spring 2016 Final Project Proposals

Group Members - QIIME	Group Members - mothur
Michael Oberg (**Grad**)	Mike Watson (**Grad**)
Johnathan Thompson	Ian Courtney
Chris Glass	Zane O'Brien
Matt Hong	Joseph Jackson
Charles Luckhardt	Jeffrey Stinemetze

Problem Domain

The Unifrac[1] algorithm is a distance metric for comparing β-diversity in separate biological samples. β-diversity as defined by Whittaker, et al (1972)[2] is "the extent of species replacement or biotic change along environmental gradients." This metric measures the distance between two sets of samples by determining the amount of evolutionary history that is shared. This is accomplished by comparing the phylogenetic tree branch lengths and uniqueness between the two sample sets. The Unifrac algorithm is implemented in both the QIIME[3] and mothur[4] analysis frameworks, in python and C++, respectively. There is also a weighted version of the Unifrac algorithm that takes into account relative abundances. The two variants of the Unifrac algorithm are typically referenced as *Qualitative* (unweighted) or *Quantitative* (weighted). In most analysis, both methods would be deployed. Unweighted Unifrac is the dominant metric when the difference between the samples can be characterized by one or more environmental elements (such as temperature), where abundance information obscures significant patterns of variation in a given taxonomy. Weighted Unifrac (a *Quantitative* measure) helps to reveal differences when there is relative abundance due to a factor such as the availability of a rate-limiting nutrient.

---

[1] Lozupone, Catherine, and Rob Knight. "UniFrac: a new phylogenetic method for comparing microbial communities." *Applied and environmental microbiology* 71.12 (2005): 8228-8235.
[2] Whittaker, Robert H. "Evolution and measurement of species diversity." *Taxon* (1972): 213-251.
[3] Caporaso, J Gregory et al. "QIIME allows analysis of high-throughput community sequencing data." *Nature methods* 7.5 (2010): 335-336.
[4] Schloss, Patrick D et al. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." *Applied and environmental microbiology* 75.23 (2009): 7537-7541.

Algorithm Benchmarking

Methods to Generate Test Data

The Grinder[5] sample simulator will be used to generate simulated samples from the GreenGenes[6,7] 16S rRNA gene database. Developing simulated sample sets for this project from a known, canonical set will allow us to process these samples and benchmark the target Unifrac algorithm, while retaining the ability to compare our analysis pipeline results back to a reference sequence database.

Test Cases to Test Algorithm Limits

The development of the simulated samples using Grinder, which will then be run and benchmarked through our two parallel pipelines included the development of test cases that highlight the strengths and weaknesses of the Unifrac algorithm. These test cases will be developed against the QIIME and mothur implementations of the Unifrac algorithm, and will be used to test the efficacy and suitability of the re-implemented versions.

Algorithm Re-implementation

The Unifrac implementation included in the QIIME suite (implemented as *beta_diversity.py*) is written completely in Python, and the implementation included in mothur is in C++. For this project we will re-implement both of these in optimized C, and compare to the original versions.

This development process will include an analysis of the original performance of the implementations (in Python and C++), and comparison to the re-written versions for each analysis framework.

---

[5] Angly, Florent E et al. "Grinder: a versatile amplicon and shotgun sequence simulator." *Nucleic acids research* 40.12 (2012): e94-e94.
[6] DeSantis, Todd Z et al. "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." *Applied and environmental microbiology* 72.7 (2006): 5069-5072.
[7] McDonald, Daniel et al. "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea." *The ISME journal* 6.3 (2012): 610-618.

Benchmarking Environment

The benchmarking environment will consist of a mix of dedicated systems and portions of the Janus supercomputer[8]. These systems will utilize a combination of runtime, OS and hardware level metrics, in the following hierarchy:

1. Runtime data collection framework
    a. Python/C++ internal timing
        i. Identify Python/C++ best practices, compare to current implementations
    b. UNIX-level timing scaffold (timing methodology, output formatting using a combination of matplotlib, Matlab, R, etc)
    c. Statistical analysis and plotting framework (runtime confidence intervals, plotting)
2. Data collection improvements
    a. Python Profiling
    b. Valgrind, etc for improved runtime data collection
    c. C profiling
3. UNIX and CPU-level performance counters (Linux kernel CPU perfctr patch, TAU, etc)

Preliminary Division of Work and Milestones

Principal means this person has primary responsibility. ***Partner***(s) will support Principal as needed. These are preliminary assignments that will need to be revised by the respective teams and team members.

QIIME:

Project Deliverable	QIIME Principal	QIIME Partner(s)	Date
Grinder Simulated Sample Creation	Matt Hong	Charles Luckhardt	04/12/16
Benchmark Environment	Chris Glass	Michael Oberg / Johnathan Thompson	04/12/16
Unifrac Re-implementation	Johnathan Thompson	Chris Glass / Michael Oberg	04/21/16
Project Writeup	Michael Oberg	Matt Hong / Charles Luckhardt	04/28/16

[8] Oberg, Michael et al. "A system architecture supporting high-performance and cloud computing in an academic consortium environment." *Computer Science-Research and Development* 26.3-4 (2011): 317-324.

mothur:

Project Deliverable	mothur Principal	mothur Partner(s)	Date
Grinder Simulated Sample Creation	Ian Courtney	Zane O'Brien	04/12/16
Benchmark Environment	Jeffrey Stinemetze	Mike Watson / Joseph Jackson	04/12/16
Unifrac Re-implementation	Joseph Jackson	Jeffrey Stinemetze / Mike Watson	04/21/16
Project Writeup	Mike Watson	Ian Courtney / Zane O'Brien	04/28/16

Application Sources

https://github.com/biocore/qiime/blob/master/qiime
https://github.com/mothur/mothur

References

1. Whittaker, Robert H. "Evolution and measurement of species diversity." *Taxon* (1972): 213-251.
2. Lozupone, Catherine, and Rob Knight. "UniFrac: a new phylogenetic method for comparing microbial communities." *Applied and environmental microbiology* 71.12 (2005): 8228-8235.
3. Caporaso, J Gregory et al. "QIIME allows analysis of high-throughput community sequencing data." *Nature methods* 7.5 (2010): 335-336.
4. "beta_diversity.py – Calculate beta diversity (pairwise ... - Qiime." 2011. 24 Mar. 2016 <http://qiime.org/scripts/beta_diversity.html>
5. Schloss, Patrick D et al. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." *Applied and environmental microbiology* 75.23 (2009): 7537-7541.
6. "mothur." 2009. 26 Mar. 2016 <http://www.mothur.org/>
7. https://www.researchgate.net/post/What_is_the_difference_between_alpha-diversity_and_beta-diversity_in_microbial_ecology
8. Oberg, Michael et al. "A system architecture supporting high-performance and cloud computing in an academic consortium environment." *Computer Science-Research and Development* 26.3-4 (2011): 317-324.