BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–2

MetAmp: combining amplicon data from multiple markers for OTU analysis

Ilya Y. Zhbannikov [1,*], James A. Foster [1,2 *]

[1]Graduate Program in Bioinformatics and Computational Biology, University of Idaho.
[2]Institute of Bioinformatics and Evolutionary Studies (IBEST), University of Idaho.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: We present a novel method and corresponding application, MetAmp, to combine amplicon data from multiple genomic markers into Operational Taxonomic Units (OTUs) for microbial community analysis, calibrating the markers using data from known microbial genomes. When amplicons for multiple markers such as the 16S rRNA gene hypervariable regions are available, MetAmp improves the accuracy of OTU-based methods for characterizing bacterial composition and community structure. MetAmp works best with at least three markers, and is applicable to non-bacterial analyses and to non 16S markers. Our application and testing have been limited to 16S analysis of microbial communities.

Results: We clustered standard test sequences derived from the Human Microbiome Mock Community (HMMC) test sets and compared MetAmp and other tools with respect to their ability to recover Operational Taxonomic Units (OTUs) for these benchmark bacterial communities. MetAmp compared favorably to QIIME, UPARSE and Mothur using amplicons from one, two, and three markers.

Availability: MetAmp is available at http://izhbannikov.github.io/MetAmp/

Contact: ilyaz@uidaho.edu, foster@uidaho.edu

Supplementary information: available at Bioinformatics online.

1 INTRODUCTION

High-throughput sequencing technologies allow researchers to characterize microbial community composition and structure without first cultivating the microbes. There are two current techniques for analyzing microbial communities: metagenomic and genomic marker sequencing. Metagenomic analysis fragments and sequences all DNA in a sample, and then optionally assembles and maps genes into annotated genomes. One often uses metagenomic analysis to characterize metabolic potential.

Marker sequencing amplifies and sequences conserved but variable genomic regions, usually hypervariable regions of the essential 16S rRNA gene. One then clusters these sequences, known as amplicons, into Operational Taxonomic Units (OTUs). This approach can handle unknown and highly diverged populations, provided that the OTUs correspond with sufficient accuracy to relevant ecological units. Metagenomic and amplicon analysis therefore answer different research questions. MetAmp targets only amplicon analysis.

Most marker based studies use amplicons of just one marker, and it is difficult to know which marker to select *a priori*. Amplicons from different markers in a given sample, or even from a single species, can cluster differently, giving very different pictures of the underlying microbial community. Different markers may also lead to very different phylogenies and taxonomies, which can differ substantially from those deduced from known genomes. Unfortunately, current bioinformatics techniques assume a single marker. This is true of the best currently available software for amplicon OTU analysis, such as UPARSE (Edgar, 2013), QIIME (Kuczynski *et al.*, 2011), and Mothur (Schloss *et al.*, 2011). We introduce the "Meta-amplicon analysis" technique, MetAmp, which makes it possible to clusters and analyze amplicons from multiple markers.

2 METHODS AND ALGORITHMS

MetAmp borrows from image registration algorithms in image processing. These identify known reference points, called "registration marks", in several images and map them onto known features in a reference image, transforming other pixels accordingly (Zitov and Flusser, 2003). In MetAmp, marker sequences play the role of registration marks, and amplicon sequences that of pixels. MetAmp works as follows:

1. MetAmp builds a 2D *reference topology* of microbial populations (Figure 1(a), top plane) in which points correspond to known, full 16S gene sequences and the distances between the points approximates the distances between the corresponding full 16S sequences region (*reference points: green on top plane*). Later steps will use this topology as a "gold standard" onto which MetAmp will map similar amplicon-induced topologies for each marker. To do this, MetAmp:
   a. Computes the pairwise distances of a set of known, full-length 16S gene sequences in a global alignment; and
   b. Maps the sequences onto a 2D plane (*reference points, green on top plane*) using Sammon Nonlinear Multi-Dimensional Scaling (SNMDS) (Sammon, 1969).
2. MetAmp then builds a separate *guided empirical topology* for each marker sequence. These 2D topologies contain both *anchor points* ("registration marks" in the image processing literature) and the user's empirical amplicon sequences, with distances in the plane approximating distances between the underlying sequences. To do this, *for each marker*, MetAmp:
   a. Extracts the marker sequences from the full 16S sequences used in the reference topology, to use as anchor points (Figure 1(a, b, c), hollow green circles). MetAmp extracts the marker sequences

*to whom correspondence should be addressed

bioinformatically using same forward and reverse primers that define the marker regions. In essence, this is a digital PCR of marker regions from known full length 16S genes, with perfect primer matching.

b. Adds the empirical amplicon sequences (Figure 1(b, c), filled blue circles) which the user acquired by from high throughput sequencing.

c. Builds a distance matrix from global pairwise alignment of both anchor and empirical sequences combined, using the same methodology as in building the reference topology above.

d. Maps both anchor and empirical amplicon sequences onto a 2D plane in a distance-approximating way using SNMDS. This guided empirical topology now contains user sequence data embedded into a topology with anchor points, which will guide the next step (Figure 1c, smaller planes).

3. MetAmp then maps each guided empirical plane, one for each marker, onto the reference plane. In this mapping, the anchor points map onto their corresponding reference points, and the the empirical points map into the reference plane with the same convolution as in the anchor-to-reference mapping (Figure 1(b,c), arrows). Currently, the mapping is an affine transformation, similar to that used to map pixels in images with registration marks onto a reference image.
4. The resulting plane contains reference sequences corresponding to known 16S genes and empirical points corresponding to user amplicon sequences (Figure 1 (c)). The distances between the empirical points correspond to distances between the user amplicon sequences, using the best available (full length 16S) sequence data to correct for distortions caused by the choice of individual markers. The points in the resulting plane are therefore ready for clustering and downstream OTU analysis.

3 VALIDATION

We validated MetAmp by performing OTU analysis on two amplicon datasets from the Human Microbiome Mock Community www.hmpdacc.org/HMMC, namely the ”Even” (SRR072220, SRR072239), and the ”Staggered” communities (SRR072221, SRR072223, SRR072237). These two datasets include amplicons from three marker regions (V13, V35, and V69) from Roche 454 GS FLX Titanium sequencing of two communities with 22 known species. Illumina paired end sequences were not available for the HMMC at the time of this writing.

These experiments test the ability of standard OTU analysis pipeline to recover known OTUs, both with and without multiple marker data that has been pre-processed with MetAmp. We computed the average number of OTUs formed by clustering these data using MetAmp, UPARSE, Mothur, and QIIME and report the ratio of this average to the known number of populations (22). Table 1 reports typical results (see Supplementary Data for full results). A perfect score would be 1.0, with larger values ”recognizing” populations that aren't there, and smaller values failing to recognize some that are. We tested each tool on V13, V35, V69 individually, and additionally tested MetAmp on pairs and triplets of these markers. MetAmp ran for about 8 hours using the whole reference data set (abut 5.1k genes) on Intel Core i5 Macbook Pro laptop.

ACKNOWLEDGEMENT

This work was made possible by NIH Grants P20GM016454, P20GM16448 from the INBRE and COBRE (NCRR), and by NSF DBI0939454.

REFERENCES

Edgar, R. C. (2013) Uparse: highly accurate otu sequences from microbial amplicon reads. *Nat Meth*, **10**, 996–998.

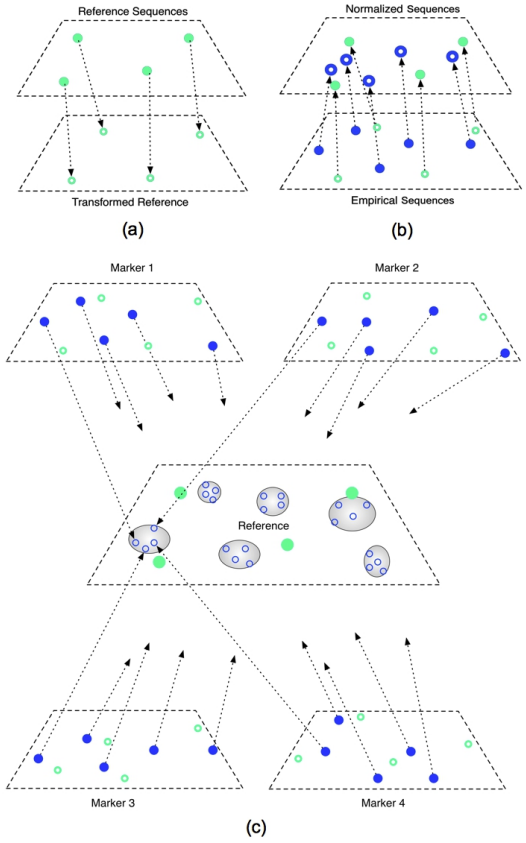

Fig. 1. Illustration of the MetAmp algorithm for combining amplicons from multiple marker sequences. See the user manual and supplementary materials for more on the analysis workflow, algorithmic complexity, parameters, and examples.

Table 1. Ratio of average number of OTUs to actual number of populations detected for MetAmp, UPARSE, Mothur, and QIIME using single and multiple 16S rRNA markers on human microbiome mock communities, using 97% sequence similarity.

Community type	MetAmp [1]	MetAmp [2]	UPARSE [2]	Mothur [2]	QIIME [2]
Even	1.06	0.93	1.62	9.6	42.9
Staggered	1.0	0.85	1.39	10.18	43.3

[1] Combined markers: V13, V35, V69

[2] Average from each of single marker: V13, V35, V69

Kuczynski, J., Stombaugh, J., Walters, W. A., Gonzalez, A., Caporaso, J. G. and Knight, R. (2011) Using qiime to analyze 16s rrna gene sequences from microbial communities. *Curr Protoc Bioinformatics*, **Chapter 10**, Unit10.7.

Sammon, J. (1969) A nonlinear mapping for data structure analysis. *IEEE Trans on Comp*, **18**, 401–409.

Schloss, P. D., Gevers, D. and Westcott, S. L. (2011) Reducing the effects of pcr amplification and sequencing artifacts on 16s rrna-based studies. *PLoS ONE*, **6**.

Zitov, B. and Flusser, J. (2003) Image registration methods: a survey. *Image and Vision Computing*, **21**, 977–1000.