Bioinformatics, 32(4), 2016, 605–607
doi: 10.1093/bioinformatics/btv638
Advance Access Publication Date: 29 October 2015
Applications Note

Sequence analysis

MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets

Yu-Wei Wu[1,2,*], Blake A. Simmons[1,2,3] and Steven W. Singer[1,2]

[1]Joint BioEnergy Institute, Emeryville, CA 94608, USA, [2]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [3]Biological and Engineering Sciences Center, Sandia National Laboratories, Livermore, CA 94551, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 4, 2015; revised on October 1, 2015; accepted on October 26, 2015

Abstract

Summary: The recovery of genomes from metagenomic datasets is a critical step to defining the functional roles of the underlying uncultivated populations. We previously developed MaxBin, an automated binning approach for high-throughput recovery of microbial genomes from metagenomes. Here we present an expanded binning algorithm, MaxBin 2.0, which recovers genomes from co-assembly of a collection of metagenomic datasets. Tests on simulated datasets revealed that MaxBin 2.0 is highly accurate in recovering individual genomes, and the application of MaxBin 2.0 to several metagenomes from environmental samples demonstrated that it could achieve two complementary goals: recovering more bacterial genomes compared to binning a single sample as well as comparing the microbial community composition between different sampling environments.
Availability and implementation: MaxBin 2.0 is freely available at http://sourceforge.net/projects/maxbin/ under BSD license.
Contact: ywwei@lbl.gov
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recovering individual genomes from metagenomic samples remains a difficult task since sequences of all microbial populations are sampled simultaneously; however, it is a critical procedure to understand the functional potential of the uncultivated microbes in natural and engineered ecosystems. Binning is a process to classify unassembled reads or assembled contigs into discrete clusters, in which one bin represents one genome. Predominant binning approaches include clustering by nucleotide frequencies (Dick *et al.*, 2009; Iverson *et al.*, 2009; Mackelprang *et al.*, 2011; Wang *et al.*, 2012; Wrighton *et al.*, 2012) or sequence coverages (Albertsen *et al.*, 2013; Sharon *et al.*, 2013; Wu and Ye, 2011). Automating the binning process provides a high-throughput method to recover microbial genomes from metagenomes. We previously developed MaxBin (Wu *et al.*, 2014), an automated binnning algorithm that classifies assembled genomic sequences from metagenomic datasets. Based on both tetranucleotide frequencies and sequence coverages, MaxBin automatically estimates the bin number from the target metagenome, classifies the sequences into genome bins, and measures the coverage levels for the binned genomes in the metagenome. In total, 19 and 26 draft genomes were recovered from two cellulolytic bacterial consortia enriched from compost using MaxBin (Wu *et al.*, 2014), demonstrating its utility in recovering the genomes of uncultivated microbes.

Recovering genomes from multiple samples may improve the performance of binning algorithms. A differential coverage binning approach was developed to use the coverage information of the contigs from two metagenomic samples to bin genomes by plotting them on a two-dimensional map (Albertsen *et al.*, 2013). Binning tools that employ multiple samples also demonstrated better performances when

two or more samples are co-assembled and binned (Alneberg *et al.*, 2014; Imelfort *et al.*, 2014; Kang *et al.*, 2015).

Here we describe MaxBin 2.0, the next generation of the MaxBin algorithm that recovers genomes from co-assembly of multiple metagenomic samples. By exploiting contig coverage levels across multiple metagenomic datasets, MaxBin 2.0 achieves better binning results than binning individual metagenomic samples. In comparison to other binning algorithms that utilize multiple metagenomic datasets, MaxBin 2.0 is highly accurate in recovering genomes from simulated metagenomes. The ability of MaxBin 2.0 to measure the coverage levels of the genome bins also allows comparisons of the genome-resolved microbial community composition across multiple samples.

2 Methods

MaxBin 2.0 employs an Expectation–Maximization (EM) algorithm to recover draft genomes from metagenomes. Briefly, after co-assembling sequencing reads of multiple metagenomic datasets, MaxBin 2.0 measures the tetranuncleotide frequencies of the contigs and their coverages for all involved metagenomes and classifies the contigs into individual bins. The abundances of all genome bins, which were unknown before the binning process, are estimated by the EM algorithm.

The description that follows demonstrates the incorporation of multiple metagenomic datasets into the MaxBin algorithm. Let the number of metagenomes be M. The reads of all metagenomes are combined and co-assembled. Let S be a contig in the co-assembly. The probability that S belongs to a genome G can be measured based on its tetranucleotide frequencies and contig coverages in all metagenomic datasets, which is

$$P(S \in G) = P_{dist}(S \in G) \cdot \prod_{k=1}^{M} P_{cov}\left(S \in G \mid cov(G_k)\right)$$

In which $P_{dist}()$ and $P_{cov}()$ are probability density functions for Euclidean distance of tetranucleotide frequencies and coverages between S and G, respectively. This probability term is then applied in the Expectation-Maximization algorithm for binning genomes from metagenomes. See Supplementary Materials for a more detailed description about the probabilities and the algorithm. Other improvements include adding multi-thread support into the MaxBin algorithm and adding more runtime options to allow users to adapt MaxBin 2.0 to specific applications.

3 Results

The performances of MaxBin 2.0 was compared to three other automated binning software: GroopM (Imelfort *et al.*, 2014), CONCOCT (Alneberg *et al.*, 2014) and MetaBAT (Kang *et al.*, 2015), using metagenomic datasets simulated by MetaSim (Richter *et al.*, 2008) and assembled by Megahit (Li *et al.*, 2015). Benchmarking the tools using different minimum contig length settings (500 and 1000 bps) revealed that MaxBin performed relatively well (in terms of *F*-score, which is the harmonic mean of precision and recall) compared to other binning tools (Fig. 1). It was also ranked first in tests involving 20 or more samples, indicating its accuracy in classifying contigs into distinct genomes.

MaxBin 2.0 was also applied to two sets of real metagenomes. One set consists of two cellulolytic consortia adapted from compost; the other set contains ten Human Microbiome Project (HMP) datasets. Binning the co-assembly of the two compost metagenomes yielded 84 bins, surpassing the number of bins generated by binning

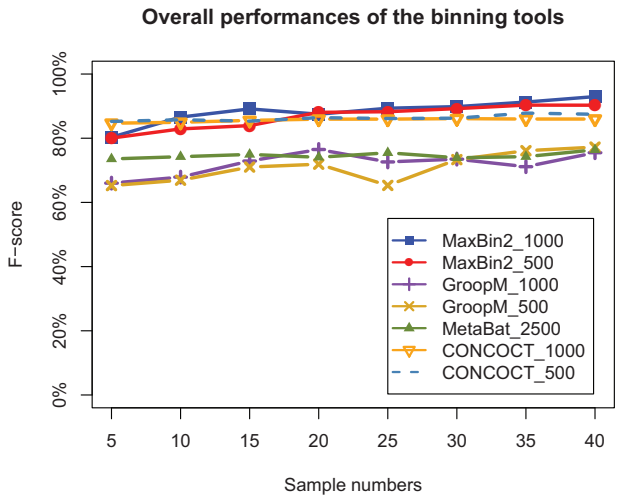

Fig. 1. The overall performances (based on F-score, which is the harmonic mean of precision and recall) estimated for binning 5-40 simulated metagenomes using different binning software. Numbers after each software tools indicate minimum contig lengths

the two metagenomes separately (19 and 26 bins). Bins produced by individual binning of the metagenomes are mostly present in the co-assembly, and more than half of the bins (57) uniquely belong to the co-assembled metagenome.

The co-assembly of the ten HMP metagenomes sampled from five body sites was also binned using MaxBin 2.0. The co-assembled metagenome yielded many more genome bins (96) compared to individual binning of each metagenomic dataset (21.4 in average). Clustering the ten HMP samples based on the coverage levels of the genome bins suggested that similar body sites share similar microbial community composition, demonstrating how MaxBin 2.0 helps users compare microbial content among a number of distinct sampling environments.

Acknowledgements

We thank Christopher S. Miller (University of Colorado at Denver) for valuable discussions that led to the conception of MaxBin 2.0.

Funding

This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Conflict of interest: none declared.

References

Albertsen,M. *et al* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.

Alneberg,J. *et al* (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.

Dick,G.J. *et al*. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*, **10**, R85.

Imelfort,M. *et al*. (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, **2**, e603.

Iverson,V. *et al*. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, **335**, 587–590.

Li,D. *et al*. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.

Kang,D.D. *et al*. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165.

Mackelprang,R. *et al*. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, **480**, 368–371.

Richter,Y. *et al*. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.

Sharon,I. *et al*. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, **23**, 111–120.

Wang,Y. *et al*. (2012) MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics*, **28**, i356–i362.

Wrighton,K.C. *et al*. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, **337**, 1661–1665.

Wu,Y.W. and Ye,Y. (2011) A novel abundance-based algorithm for binning metagenomic sequences using *l*-tuples. *J. Comput. Biol.*, **18**, 523–534.

Wu,Y.W. *et al*. (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**, 26.