MetaBinG2: a fast and accurate metagenomics sequence classification method for samples with many unknown organisms

Yuyang Qiao[1], Ben Jia[1,2], Chaochun Wei[1,2]

1. Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, 200240

2. Shanghai Center for Bioinformation Technology, Shanghai, China, 200210.

ABSTRACT

Background: Many methods have been developed for metagenomic sequence classification, and most of them depend heavily on the known organisms. In addition, a large portion of reads may be classified as unknown, which greatly impairs our understanding of the whole sample.

Result: Here we present MetaBinG2, a fast method to do metagenomics sequence classification and consequent abundance analysis on complex environments with a large number of unknown organisms. MetaBinG2 is based on sequence composition, and uses GPUs to accelerate its speed. A million 100bp Illumina sequences can be classified within two minutes. We applied MetaBinG2 to MetaSUB Inter-City Challenge and identified microbial community structures for different cities.

Conclusion: Compared to existing methods, MetaBinG2 is fast, highly accurate, especially for those samples with a significant percentage of unknown organisms.

INTRODUCTION

The amount of metagenome sequencing data can be huge. Many methods are developed to do the taxonomy classification of metagenome sequencing data. Existing methods can be divided into two categories. One is alignment-based, such as PAUDA (1), MAGAN (2), PhymmBL (3). These methods are often very accurate but slow. Another category of methods are composition-based, such as NBC (4) and metaCV (5). This type of methods are faster than the former but with lower accuracy. Existing methods based on k-string search like Kraken (6) and CLARK (7), have excellent performance both in speed and precision but they are heavily dependent on the known species. Their performance faded for samples from an environment with a large number of unknown organisms.

Here we present MetaBinG2, a tool to classify metagenome sequencing data for samples from an environment with a large number of unknown organisms. The previous version of MetaBinG (8) has shown excellent speed performance on metagenomics sequence classification, which is almost 1500-fold faster than Phymm (9). In MetaBinG2, we added a new assumption that a sequence is more likely from an organism if its abundance is higher than the others when the distances between this sequence and several organisms are similar. The accuracy of MetaBinG2 is close to 80% at phylum level for sequencing data with length about 100bps.

In order to evaluate its classification potential for unknown organisms, we used clade exclusion method, which is an effective way to measure the capability to identify source genome when the number of unknown organisms in the samples is large (10).

MATERIALS AND METHODS

A set of simulated dataset SimDataset was created to test the performances of MetaBinG2 and existing methods, such as CLARK, metaCV, and MetaBinG. In addition, a mock dataset, a real world dataset and MetaSUB Inter-City metagenomic dataset were collected to test MetaBinG2.

SimDataset

This dataset was derived from MG-RAST repository - MetaSimHC_100 (10). Metagenome sequencing data were created with NeSSM (11) based on the community structure, and sequencing read length was set to 100bp and 250bp.

Mock dataset

This dataset was selected from HMMC. Its NCBI accession id is SRR072232. 22 species were mixed with different percentages.

Real world data

This dataset was sequenced from cow rumen (12) with read length of 125bps. The accession id of the data in NCBI is PRJNA60251.

MetaSUB Inter-City metagenomic data

We downloaded the metagenomic data from CAMDA contest Challenge 1 – MetaSUB urban microbiome diversity challenges. The sizes of raw data in fastq format from Boston, Sacramento and New York were 220Gbps, 247Gbps, and 341Gbps respectively.

Evaluation

TP represents sequences whose predicted taxonomies are same with their true taxonomies. FP means sequences whose predicted taxonomies are different from their true taxonomies. For MetaBinG and MetaBinG2, the accuracy was calculated as the number of TP/total number of sequences. For a fair comparison, we calculated sensitivity=TP/(TP+FN), precision=TP/(TP+FP), and adjusted accuracy=(sensitivity+precision)/2 for, CLARK and metaCV.

MetaBinG2

(1) Building database

MetaBinG2 converts a complete genome sequence into a state-transitions vector under the kth-order markov model. A state in this Markov model is defined as a sequence with the length of k, and each state can transfer to four kinds of stats, so that there are 4^{k+1} kinds of transition probabilities. The transition probabilities from the state m to the state n of the genome i is calculated as following:

$$kMM_{i,mn} = P_i(O_m|O_n) = \frac{F_i(O_m|O_n)}{F_i(O_m)} \quad (1)$$

where O_m and O_n are oligonucleotides of length k, $F_i(O_m|O_n)$ means the count of events that state O_m transfers to O_n, and $P_i(O_m|O_n)$ represents the transition probability from the O_m to the O_n of the genome i.

(2) Calculate the distance between short sequences and genomes.

We designed MetaBinG2 with a reasonable assumption that a sequence is more likely from an organism which take a larger percentage when the distance between this sequence and several organisms are similar. Those most reliably classified results' distribution is used as a priori knowledge in subsequent analyzes. The similarity between a short sequence with length l and a genome i can be reflected by S_i as following:

$$S_i = \left(-\sum_{j=0}^{l-k-1} \ln\left(p_i(O_j|O_{j+1})\right) \right) * (1 + \alpha\omega_i) \quad (2)$$

where $O_j$and O_{j+1} are oligonucleotides of length k, $p_i(O_j|O_{j+1})$ represents the transition probability from the O_j to the O_{j+1} of the genome i, ω_i stands for the weight of genome i which is calculated as the number of sequences assigned to genome i divided by the total number of sequences, and α is a training value to control the force of weight ω_i. For each sequence, a genome in the database with the minimum score is selected as the source genome.

The vectors for short sequences and genomes are used to calculate the scores between each sequence and each genome through matrix multiplication, which is achieved by cublas in GPU. The score update with the weight and the search of the best score is also running on GPU, while the annotation of the best score is running on CPU (Figure 1).

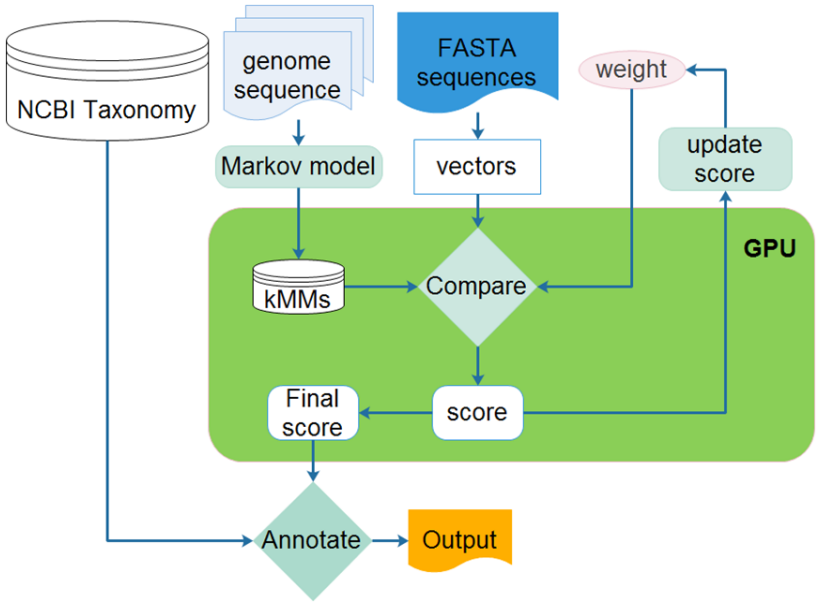

Figure 1. The system diagram of MetaBinG2.
First, MetaBinG2 load the database; Second, the short sequences in ~~fasta~~fast format are also transferred into state-transitions matrix; Then, these two matrix are upload to GPU memory and do matrix multiplication to get the score matrix by cublas function; The score matrix is adjusted with weights, and the source genomes with minimum scores will be selected; Next, the weights are updated according to the percentage of each selected genome back on CPU; when the BC distance between the current genomes percentages and the last genomes percentages is less than the cutoff, the final scores will be annotated with their taxonomy information on CPU and output.

RESULTS

Clade exclusion experiment

MetaBinG2 was compared with CLARK, metaCV and MetaBinG by clade exclusion on SimDataset. Accuracy of MetaBinG2 were more stable with unknown and were better as the length of sequences increase (Figure 2a 2b). These 4 tools were running on the same node with 24 cores and with the parameters make use of CPU and GPU as efficiently as possible. (Figure 2c). Running time of MetaBinG2 is comparable with these methods holding advantage on speed and even better.

Comparison predicted abundance

The species percentage are clear for SimDataset and mock dataset. We calculate the cosine values between predicted abundance of the 4 methods on these two datasets (Figure 2d 2e) and the true abundance to evaluate a method 's capability providing outline of a sample's community structure. The predicted community structure of MetaBinG2 is more similar to the true abundance than the others.

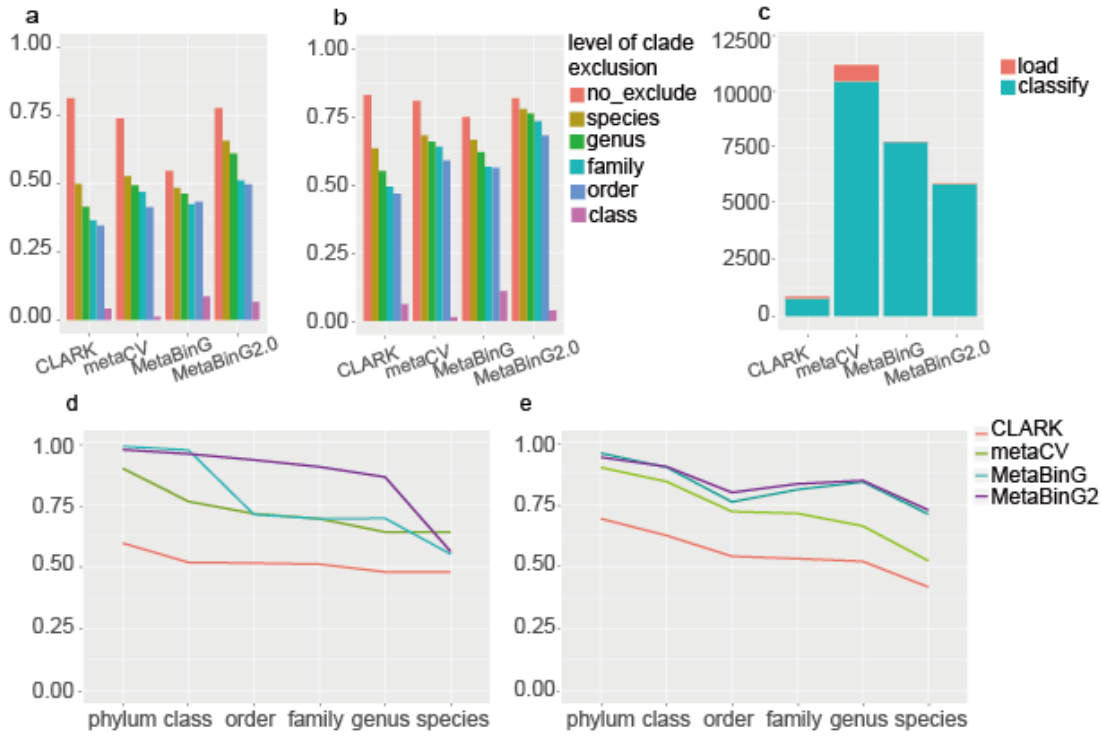

Figure 2. Performance comparison of CLARK, MetaCV, MetaBinG and MetaBinG2.

(a) (b) Comparison of accuracy at phylum level on SimDataset with different levels ofclade exclusion with read length 100bp (a) and 250bp (b).The accuracy of each software is tending to decrease as the level of clade exclusion moves from no_exclude to class. (c) Comparison of time cost for the 4 methods, including the database loading time and the time for classification.(d)(e) Comparison the cosine value between predicted abundance and the true abundance on SimDataset (d) and mock dataset (e).

Performance of MetaBinG2

Researchers (12) assembled 15 genome bins from the cow rumen metagenomic sequences and assigned them into 4 phylogenetic orders - Bacteroidales, Clostridiales, Myxococcales, and Spiochaetales. In the result of MetaBinG2, the first four phyla and four classes with highest percentage are the same with the assembled result. On the order level, Bacteroidales, Clostridiales, and Spiochaetales also have relatively high percentage. For these real-life dataset with plenty of unknown species, MetaBinG2's performance is satisfying. We also applied MetaBinG2 on MetaSUB metagenomic data and identified city specific organism abundances for different cities (Figure 3).

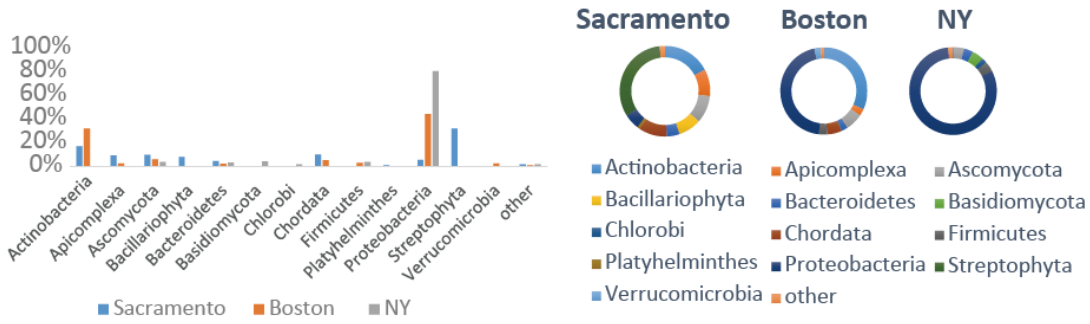

Figure 3. Community structure on phylum of three cities.

DISCUSSION

MetaBinG2 require much smaller storage space - 72M compared with 24G for CLARK and 50G for metaCV when database is built from 2606 genomes. It has more potential to use more additional genomes as training set than the others.

The development of third generation sequencing technology represented by PacBio is good for MetaBinG2 due to the increase of reads length is benefit for its classification accuracy.

MetaBinG2’s scope of application is large. For example, the contamination analysis. Like specific organism abundances of different cities can be identified by MetaBinG2 conveniently, the community structure of any corner in laboratory can be identified with MetaBinG2 too. The contamination analysis is very useful for experiment repeatability. Furthermore, MetaBinG2 may be helpful to shorten the cycle of pathogenic bacteria identification which is fatal in hospital to avoid the worse tendency.

AVAILABILITY

MetaBinG2kit is an open source collaborative initiative available in the GitHub repository (https://github.com/mengmayang/MetaBinG2kit)

REFERENCES

1. Huson, D. H., & Xie, C. (2013). A poor man's BLASTX-high-throughput metagenomic protein database search using PAUDA. Bioinformatics, btt254.

2. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome research, 17(3), 377-386.

3. Brady, A., & Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nature methods, 8(5), 367-367.

4. Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics, 27(1), 127-129.

5. Liu, J., Wang, H., Yang, H., Zhang, Y., Wang, J., Zhao, F., & Qi, J. (2012). Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic acids research, gks828.

6. Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), R46.

7. Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC genomics, 16(1), 236.

8. Jia, P., Xuan, L., Liu, L., & Wei, C. (2011). MetaBinG: Using GPUs to accelerate metagenomic sequence classification. PloS one, 6(11), e25353.

9. Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature methods, 6(9), 673-676.

10. Peabody, M. A., Van Rossum, T., Lo, R., & Brinkman, F. S. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC bioinformatics, 16(1), 362.

11. Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., & Wei, C. (2013). NeSSM: a next-generation sequencing simulator for metagenomics. PLoS One, 8(10), e75448.

12. Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., ... & Mackie, R. I. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science, 331(6016), 463-467.