Low-Density Locality-Sensitive Hashing Boosts Metagenomic Binning

Yunan Luo[1,4], Jianyang Zeng[1], Bonnie Berger[2,3], and Jian Peng[4]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
bab@mit.edu
[3] Department of Mathematics, MIT, Cambridge, MA, USA
[4] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
jianpeng@illinois.edu

1 Introduction

Metagenomic sequencing techniques produce large data sets of DNA fragments (e.g. reads or contigs) from environmental samples. To understand the microbial communities and functional structures within the samples, metagenomic sequence fragments need to be first assigned to their taxonomic origins from which they were derived (also called "binning") to facilitate downstream analyses.

Arguably the most popular metagenomic binning approaches are alignment-based methods. A sequence fragment is searched against a reference database consisting of full genomes of organisms, and the highest scoring organism is assigned as the taxonomic origin. Although efficient sequence alignment algorithms, including BWA-MEM [1], Bowtie2 [2] and (mega)BLAST [3], can readily be used for this purpose, the computational cost of alignment-based methods becomes prohibitive as the size of the sequence dataset grows dramatically, which is often the case in recent studies.

Another completely different binning approach is based on genomic sequence composition, which exploits the sequence characteristics of metagenomic fragments and applies machine learning classification algorithms to assign putative taxnomic origins to all fragments. Since classifiers, such as support vector machines, are trained on whole reference genome sequences beforehand, compositional methods normally are substantially faster than alignment-based methods on large datasets. The rationale behind compositional-based binning methods is based on the fact that different genomes have different conserved sequence composition patterns, such as GC content, codon usage or a particular abundance distribution of consecutive nucleotide k-mers. To design a good compositional-based algorithm, we need to extract informative and discriminative features from the reference genomes. Most existing methods, including PhyloPythia(S) [4, 5], use k-mer frequencies to represent sequence fragments, where k is typically small (e.g. 6 to 10). While longer k-mers, which capture compositional dependency

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 255–257, 2016.
DOI: 10.1007/978-3-319-31957-5

within larger contexts, could potentially lead to higher binning accuracy, they are more prone to noise and errors if used in the supervised setting. Moreover, incorporating long k-mers as features increases computational cost exponentially and requires significantly larger training datasets.

2 Method

We introduce a novel compositional metagenomic binning algorithm, Opal, which robustly represents long k-mers in a compact way to better capture the long-range compositional dependencies in a fragment. The key idea behind our algorithm is built on locality-sensitive hashing (LSH), a dimensionality-reduction technique that hashes input high-dimensional data into low-dimensional buckets, with the goal of maximiz the probability of collisions for similar input data. To the best of our knowledge, it is the first time that LSH functions have been applied for compositional-based metagenomic binning. We propose to use them first to represent metagenomic fragments compactly and subsequently for machine learning classification algorithms to train metagenomic binning models. Since metagenomic fragments can be very long, sometimes from hundreds of bps to tens of thousands of bps, we hope to construct compositional profiles to encode long-range dependencies within long k-mers. To handle large ks, we develop string LSH functions to compactly encode global dependencies with k-mers in a low-dimensional feature vector, as oppose to directly using a 4^k-length k-mer profile vector. Although LSH functions are usually constructed in a uniformly random way, we propose a new and efficient design of LSH functions based on the idea of the low-density parity-check (LDPC) code invented by Robert G Gallager for noisy message transmission [6, 7]. A key observation is that Gallager's LDPC design not only leads to a family of LSH functions but also makes them efficient such that even a small number of random LSH functions can effectively encode long fragments. Different from uniformly random LSH functions, the Gallager LSH functions are constructed structurally and hierarchically to ensure the compactness of the feature representation and robustness when sequencing noise appears in the data. Methodologically, starting from a Gallager design matrix with row weight t, we construct m hash functions to encode high-order sequence compositions within a k-mer. In contrast to the $O(4^k)$ complexity it would take to represent contiguous k-mers, our proposed Gallager LSH adaptation requires only $O(m4^t)$ time. For very long k-mers, we construct the Gallager LSH functions in a hierarchical fashion to further capture compositional dependencies from both local and global contexts. It is also possible to use Opal as a "coarse search" procedure in the compressive genomics manner to reduce the search space of alignment-based methods [8]. We first apply the compositional-based binning classifier to identify a very small subset or group of putative taxonomic origins which are ranked very highly by the classifier. Then we perform sequence alignment between the fragment and the reference genomes of the top-ranked organisms. This natural combination of compositional-based and alignment-based methods provides metagenomic binning with high scalability, high accuracy and high-resolution alignments.

3 Results

To evaluate the performance of Opal, we trained an SVM model with features generated by the Gallager LSH method. When tested on a large dataset of 50 microbial species, Opal achieved better binning accuracy than the traditional method that uses contiguous k-mer profiles as features [4]. Moreover, our method is more robust to mutations and sequencing errors, compared to the method with the contiguous k-mer representation. Opal outperformed (in terms of robustness and accuracy) BWA-MEM [1], the state-of-the-art alignment-based method. Remarkably, we achieved up to two orders of magnitude improvement in binning speed on large datasets with mutations rates ranging from 5 % to 15 % over 20–50 microbial species; moreover, we found Opal to be substantially more accurate than BWA-MEM when the rate of sequencing error is high (e.g., 10–15 %). It is counterintuitive that a compositional binning approach is as robust as or even more robust than alignment-based approaches, particularly in the presence of high sequencing errors or mutations in metagenomic sequence data. Finally, we combined both compositional and alignment-based methods, by applying the compositional SVM with the Gallager LSH coding as a "coarse-search" procedure to reduce the taxonomic space for a subsequent alignment-based BWA-MEM "fine search." This integrated approach is almost 20 times faster than original BWA-MEM and also has substantially improved binning performance on noisy data. The above results indicate that Opal enables us to perform accurate metagenomic analysis for very large metagenomic studies with greatly reduced computational cost.

Acknowledgments. This work was partially supported by the US National Institute of Health Grant GM108348, the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205.

References

1. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprints (2013). arXiv:1303.3997
2. Langmead, B., Salzberg, S.: Fast gapped-read alignment with bowtie 2. Nat. Methods **9**, 357–359 (2012)
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)
4. Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., McHardy, A.C.: Taxonomic metagenome sequence assignment with structured output models. Nat. Methods **8**(3), 191–192 (2011)
5. McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods **4**(1), 63–72 (2007)
6. Gallager, R.: Low-density parity-check codes. IEEE Trans. Inf. Theory **8**(1), 21–28 (1962)
7. MacKay, D., Neal, R.: Near shannon limit performance of low density parity check codes. Electron. Lett. **32**, 1645–1646 (1996)
8. Yu, Y.W., Daniels, N.M., Danko, D.C., Berger, B.: Entropy-scaling search of massive biological data. Cell Syst. **2**, 130–140 (2015)