COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge

Yang Young Lu, Ting Chen, Jed A. Fuhrman, and Fengzhu Sun
ylu465@usc.edu, fsun@usc.edu

Abstract

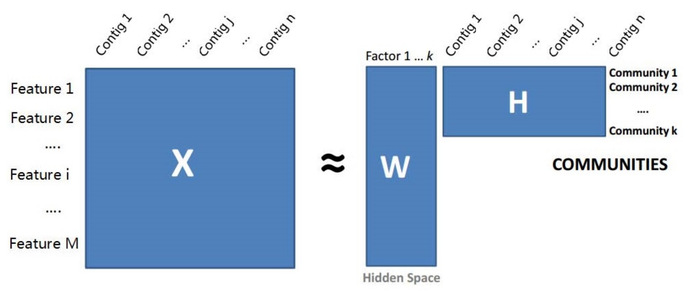
The advent of next-generation sequencing (NGS) technologies enables researchers to sequence complex microbial communities directly from environment. Since assembly typically produces only genome fragments, also known as contigs, instead of entire genome, it is crucial to group them into operational taxonomic units (OTUs) for further taxonomic profiling and down-streaming functional analysis. OTU clustering is also referred to as binning.

We present COCACOLA, a general framework automatically bin contigs into OTUs based upon sequence composition and coverage across multiple samples. The effectiveness of COCACOLA is demonstrated in both simulated and real datasets in comparison to state-of-art binning approaches such as CONCOCT [1], GroopM [3] and MetaBAT [4]. The superior performance of COCACOLA relies on two aspects. One is employing L_1 distance instead of Euclidean distance for better taxonomic identification during initialization. More importantly, COCACOLA takes advantage of both hard clustering and soft clustering by sparsity regularization.

In addition, the COCACOLA framework seamlessly embraces customized prior knowledge to facilitate binning accuracy. In our study, we have investigated two types of additional knowledge, in particular, the co-alignment to reference genomes and linkage of contigs provided by paired-end reads. We find that both co-alignment and linkage information further improve binning in the majority of cases. COCACOLA is scalable and in parallel, the running time on binning is faster than MetaBAT and much faster than CONCOCT and GroopM.

Software: https://github.com/younglululu/COCACOLA

Feature Matrix Representation of Contigs



There are two possible types of feature:

- `Abundance Profile`
- `Composition Profile` (tetra-mer)

Problem Formulation

According to above illustration, given a feature matrix X, we want to find two matrices W and $\mathbb{H}$ satisfying:

$$X \approx W\mathbb{H}$$
$$s.t.\ \ W \geq 0,\ \ \mathbb{H} \in \{0,1\}^{K\times N}, \|\mathbb{H}_{\cdot n}\|_0 = 1\ for\ n = 1, 2, \cdots, N$$

The matrices W and $\mathbb{H}$ are obtained by minimizing a certain objective function. We use Frobenius norm, commonly known as the sum of squared error:

$$\arg\min_{W,\mathbb{H}} \|X - W\mathbb{H}\|_F^2 \tag{1}$$
$$s.t.\ \ W \geq 0,\ \ \mathbb{H} \in \{0,1\}^{K\times N}, \|\mathbb{H}_{\cdot n}\|_0 = 1\ for\ n = 1, 2, \cdots, N$$

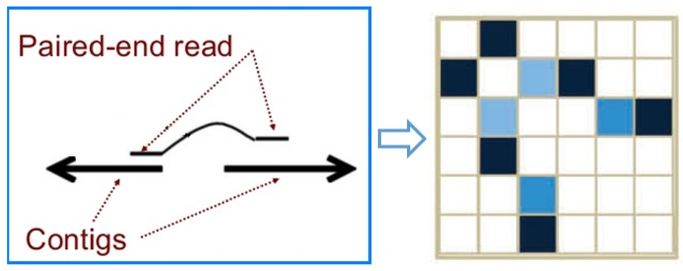
Eq. (1) is NP-hard to solve. We relax the binary constraint of $\mathbb{H}$ with numerical values. Hence Eq. (1) is reformulated as the following minimization problem:

$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2 \tag{2}$$

Relaxation of binary constraint makes the interpretation from hard clustering to soft clustering, where hard clustering means that a contig can only be assigned to one OTU, while soft clustering allows a contig to be assigned to multiple OTUs. It has been observed that by imposing sparsity on each column of H, the hard clustering behavior can be facilitated [7]. Therefore, Eq. (2) is further modified through the Sparse Nonnegative Matrix Factorization (SNMF) form [7]:

$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2 + \alpha\sum_{n=1}^{N}\|H_{\cdot n}\|_1^2 \tag{3}$$

Additional Information (optional)



We consider two possible additional information, which is encoded in laplacian matrix $\mathcal{L}$:

- `paired-end reads linkage`
- `co-alignment to reference genomes`

By incorporating the regularization item of additional information, the objective function changes to the following form:

$$\arg\min_{W,H\geq 0} \|X - WH\|_F^2 + \alpha\sum_{n=1}^{N}\|H_{\cdot n}\|_1^2 + \beta Tr(H\mathcal{L}H^T) \tag{4}$$

Optimization

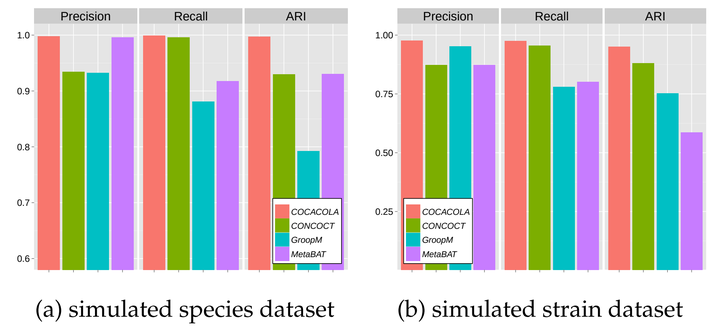
We use Alternating Nonnegative Least Squares (ANLS) [5, 7, 8] to solve Eq. (4), iteratively handle two nonnegative least square (NNLS) subproblems in Eq. (5) until convergence.

$$H \leftarrow \arg\min_{H\geq 0} \|X - WH\|_F^2 + \alpha\sum_{n=1}^{N}\|H_{\cdot n}\|_1^2 + \beta Tr(H\mathcal{L}H^T) \tag{5a}$$
$$W \leftarrow \arg\min_{W\geq 0} \left\|X^T - H^T W^T\right\|_F^2 \tag{5b}$$

By using block coordinate descent (BCD) [6], Eq. (5a) can be further reorganized into:

$$\arg\min_{H\geq 0} \left\| \begin{pmatrix} X \\ 0_{1\times N} \\ \sqrt{\beta}H^{old}\mathcal{A} \end{pmatrix} - \begin{pmatrix} W \\ \sqrt{\alpha}e_{1\times K} \\ \sqrt{\beta}I_K \end{pmatrix} H \right\|_F^2 \tag{6}$$

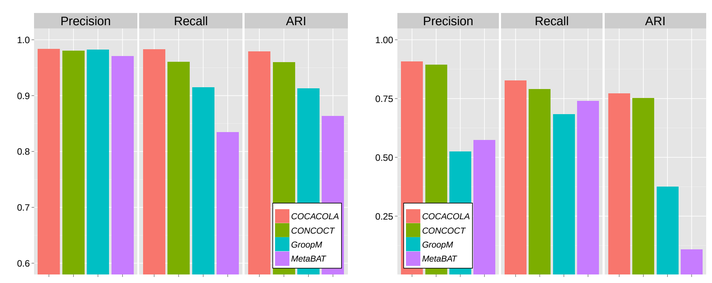
Performance on Simulated Datasets



(a) simulated species dataset (b) simulated strain dataset

The simulated species dataset consisted of 101 different species across 96 samples. A total of 37,628 contigs remain for binning after co-assembly and filtering.
The simulated strain dataset consisted of 20 different species or strains from the same species across 64 samples, including five different *E. coli* strains, five different *Bacteroides* species, five different species from different *Clostridium* genera, and five different *gut* bacteria. A total of 9,417 contigs remain for binning after co-assembly and filtering.

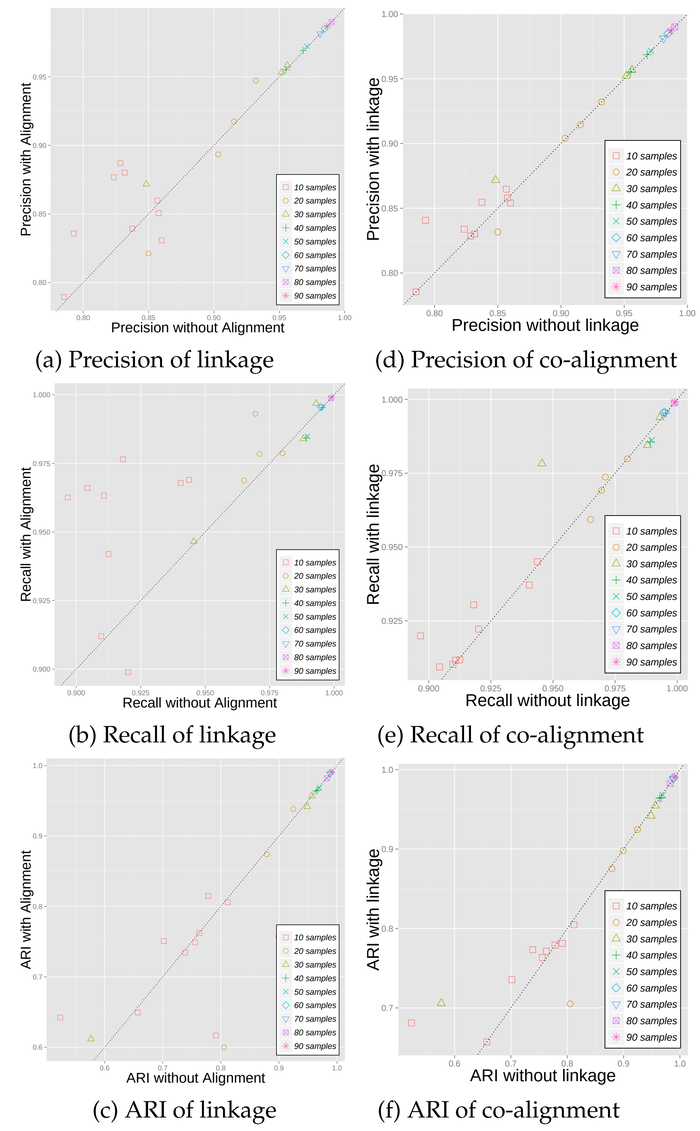
Performance on Real Datasets



(a) real Sharon dataset (b) real MetaHIT dataset

We use a time-series study of 11 fecal microbiome samples from a premature infant [10], denoted as the Sharon dataset. Since the true species that contigs belong to are not known, we assign the class labels by annotating contigs using the TAXAassign script [2]. As a result, $2,614$ out of $5,579$ contigs are unambiguously labeled on the species level for evaluation. Another real dataset embody 264 samples from the MetaHIT consortium [9] (SRA:ERP000108), the same dataset used in MetaBAT [4], denoted as the MetaHIT dataset. $17,136$ out of $192,673$ co-assembled contigs are unambiguously labeled on the species level for evaluation.

The Effect of Additional Information



(a) Precision of linkage (d) Precision of co-alignment

(b) Recall of linkage (e) Recall of co-alignment

(c) ARI of linkage (f) ARI of co-alignment

The simulated species dataset comprises 96 samples overall. Thus we choose sub-samples of size ranging from 10 to 90, with 10 as increment. To avoid duplicate contribution from a particular sample, we choose sub-samples without overlapping. Therefore, the numbers of sub-samples are $9, 4, 3, 2, 1, 1, 1, 1, 1$, respectively.
We compare the binning result by COCACOLA incorporating additional information against the result without additional information. When the sample size exceeds $K = 30$, the regularization effect nearly diminishes, therefore we focus on the 16 cases from $K = 10$ to $K = 30$.

References

[1] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, 2014.

[2] U. Ijaz and C. Quince. TAXAassign v0.4, June 2009.

[3] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, 2014.

[4] D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

[5] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[6] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[7] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. *Technical Report GT-CSE-08-01*, 2008.

[8] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM)*, pages 353–362, 2008.

[9] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[10] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120, 2013.

Last modified: Feb 17, 2016