Metagenomic Binning via Graph Representation Learning and Clustering

Wei Zhou

Supervisor : Dr Yu Lin

Australian National University

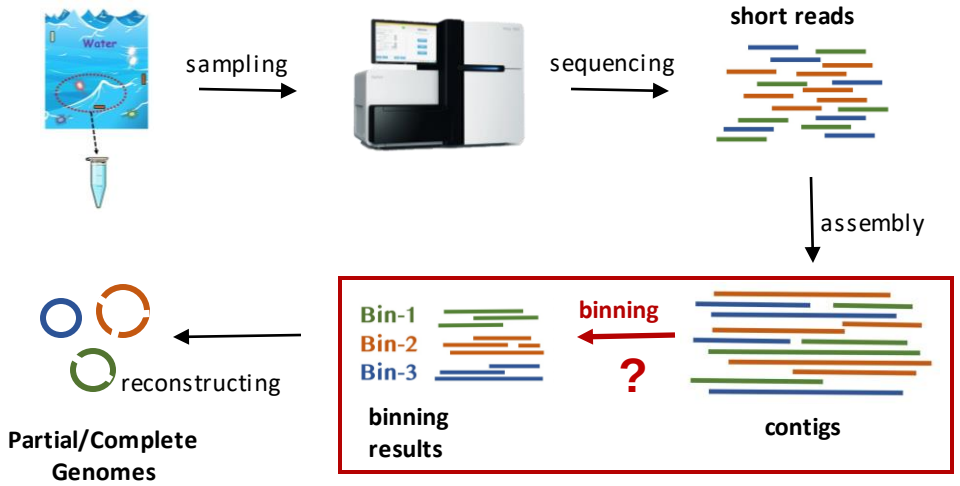
Presentation Outline

01 Background & Goal

Metagenomic Binning

1. Microorganism samples are mixed
2. Goal : Bin assembled contigs correctly
3. Gain valuable insights about the complex microbial communities
4. Identify association between diseases and human microbiome

Metagenomic Workflow

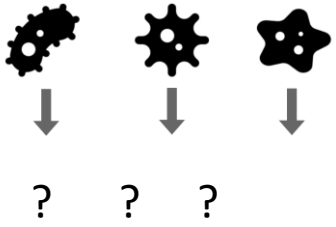


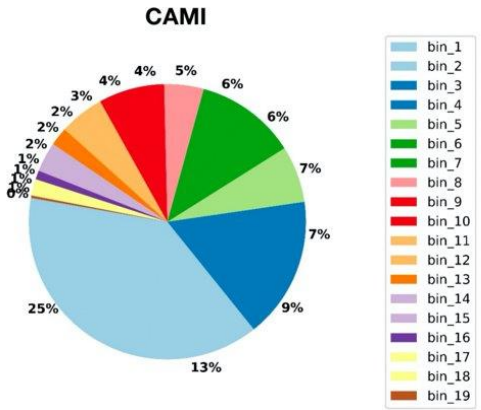
Hansheng Xue. (2022)"RepBin: Constraint-based Graph Representation Learning for Metagenomic Binning" [PowerPoint presentation].

02 Challenge in Metagenomic Binning

1. **Unknown features of bins**

2. **Unknown number of bins**

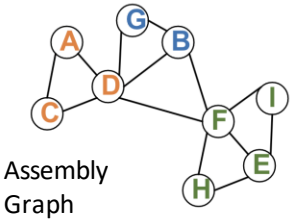




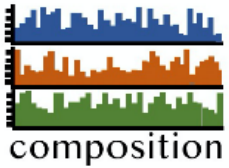
02 Challenge in Metagenomic Binning

1. Unknown features of bins

- solve by constraint-based learning with both assembly graph and composition information of assembled contigs

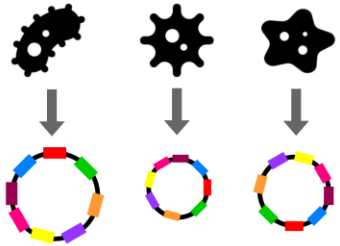


Assembly Graph

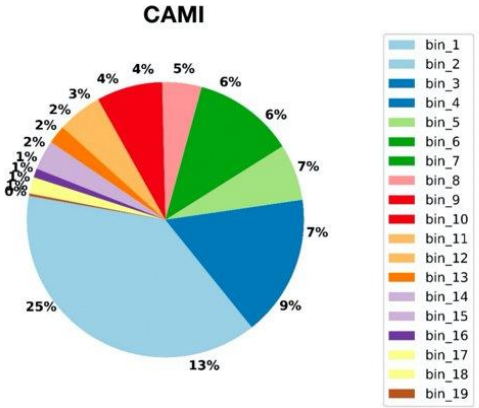


2. Unknown number of bins

- solve by graph matching and clustering with single-copy marker genes contained in assembled contigs



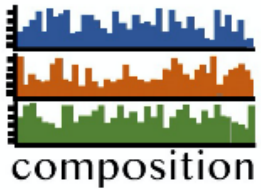
single-copy marker genes



03 Methodology

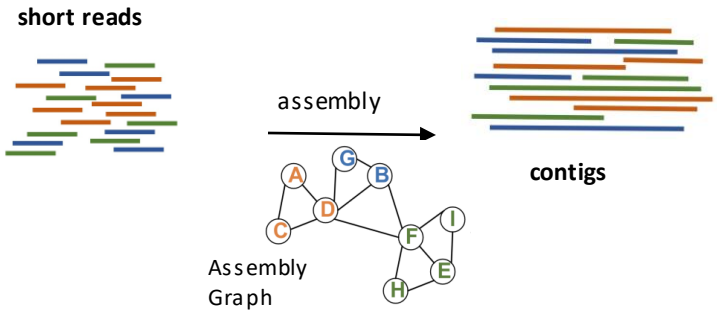
Composition Information

- Biology information
- Contigs of same species have high similarity in composition



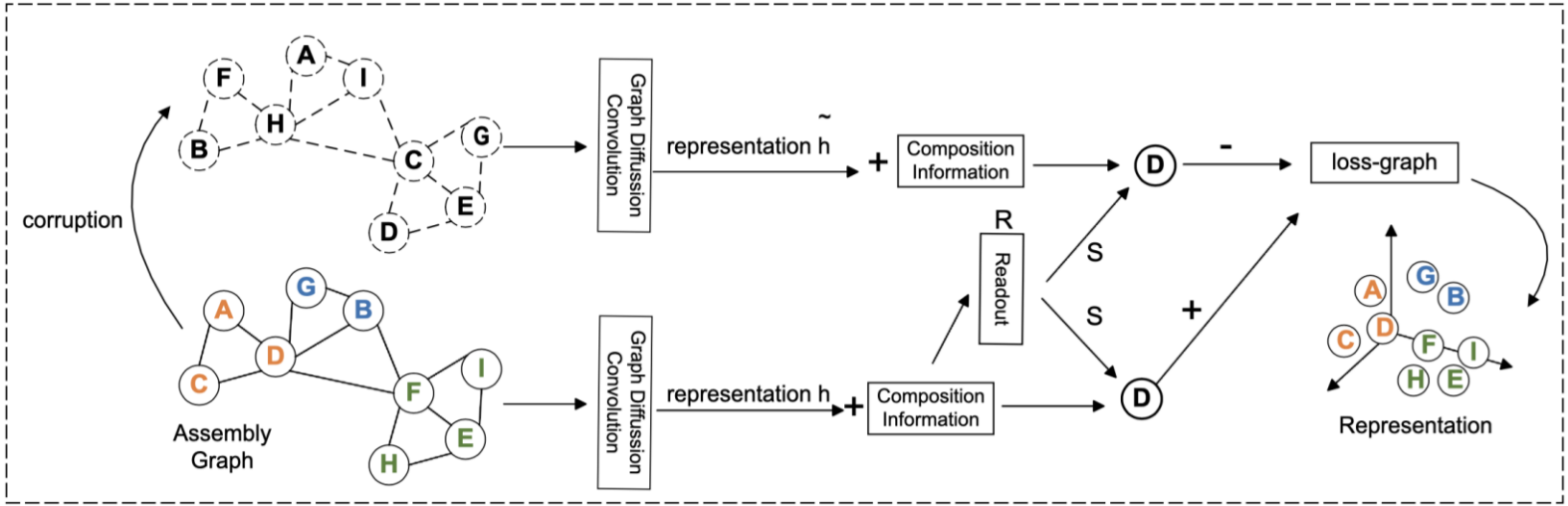
Assembly Graph

- Contigs as nodes
- Majority linked contigs belong to same species



03 Methodology

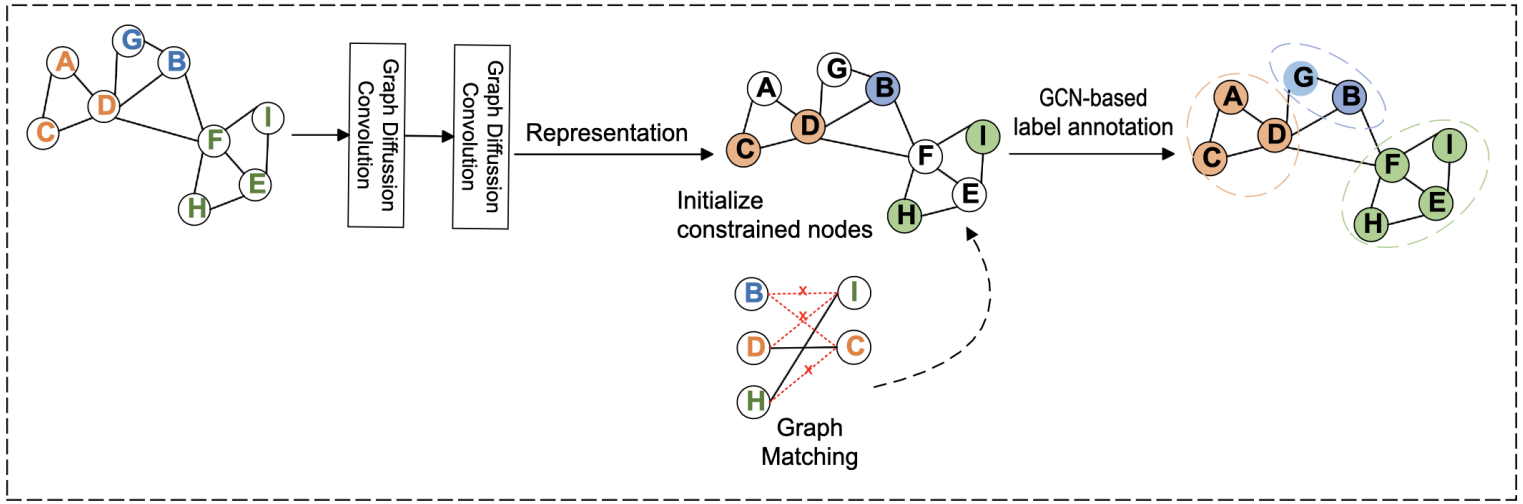
Part 1 : Contrastive Graph Learning



1. Generate negative graph with corruption function
2. Learn h and h _tilde using Graph Diffusion Convolution
3. Concatenate with composition information
4. Obtain global representation S by readout function R
5. Maximize the mutual information with discriminator D
6. Obtain representations

03 Methodology

Part 2 : Constraint-based Clustering



04 Experiments & Result

Datasets		CONCOCT	MaxBin2	MetaBAT2	RepBin	My method
Sim-5G	Precision	91.60	91.13	100	99.69	99.61
	Recall	40.50	46.69	6.61	99.69	99.61
	F1	56.16	56.16	12.4	99.69	99.61
Sim-10G	Precision	86.99	86.99	100	99.22	99.43
	Recall	39.72	39.72	6.39	99.55	99.66
	F1	54.54	54.54	12.1	99.37	99.55
Sim-20G	Precision	84.02	84.03	96.77	97.31	98.74
	Recall	42.27	42.27	7.73	96.98	96.80
	F1	56.24	56.24	14.32	97.15	97.76

- Better performance
- More simplified model than RepBin

Reference

Hansheng Xue. (2022)“RepBin: Constraint-based Graph Representation Learning for Metagenomic Binning” [PowerPoint presentation].

Wang, J. and Jia, H., 2016. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology*, *14*(8), pp.508-522.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E. and Bremges, A., 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, *14*(11), pp.1063-1071.

Klicpera, J., Weißenberger, S. and Günnemann, S., 2019. Diffusion improves graph learning. *arXiv preprint arXiv:1911.05485*.

Australian National University

Thank you

Any question?

Australian
National
University