Metagenomic Binning via Graph Representation Learning and Clustering

Presenter :Wei Zhou

Supervisor : Dr Yu Lin

Australian National University

Presentation Outline

01	Background & Goal	03
02	Challenge	09
03	Methodology	11
04	Experiment	19
05	Visualization	23
06	Further Analysis	24
07	Q & A	25

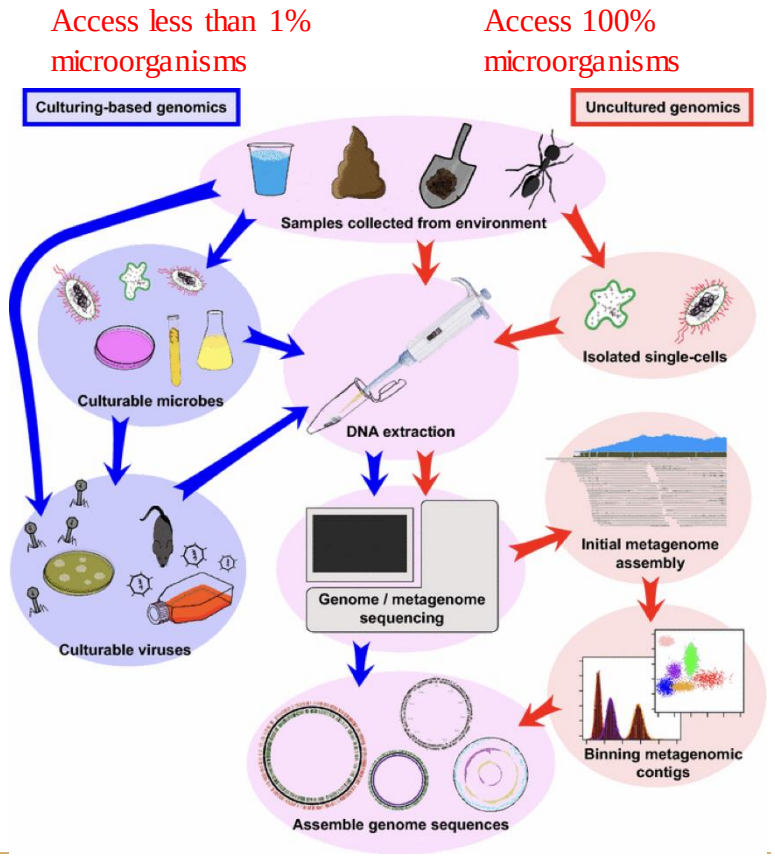
01 Background

What is Metagenomic?

- Defined as the study of genetic materials that are collected directly from various natural environments
- No need for isolation and lab cultivation of individual species
- Culture-independent method
- Allow analyse of 100% genetic materials

Application

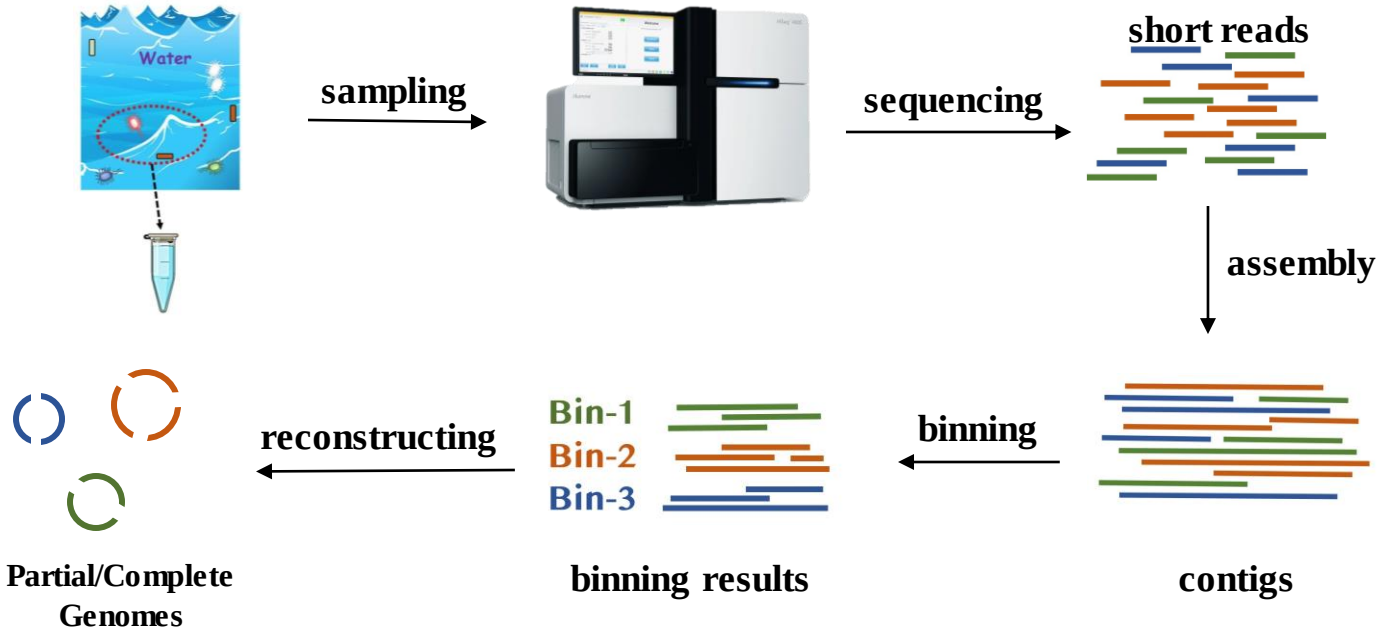
- Medicine
- Engineering
- Agriculture
- Ecology



From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Garza, D.R. & Dutilh, B.E. Cellular and Molecular Life Sciences (2015) 72: 4287

01 Background

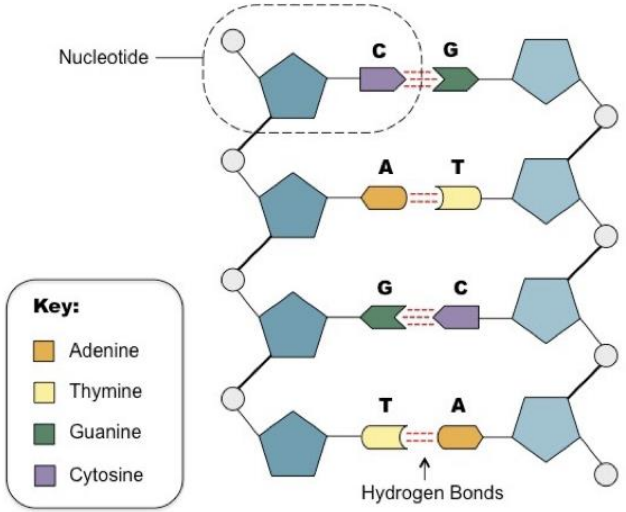
Pipeline of Metagenomic Analysis



01 Background

Nucleotide sequences

- Carrying genetic information of organisms
- Composing of 4 nucleotide bases
  - adenine (A)
  - thymine (T)
  - cytosine (C)
  - guanine (G)



01 Background

High-throughput Sequencing(HTS)

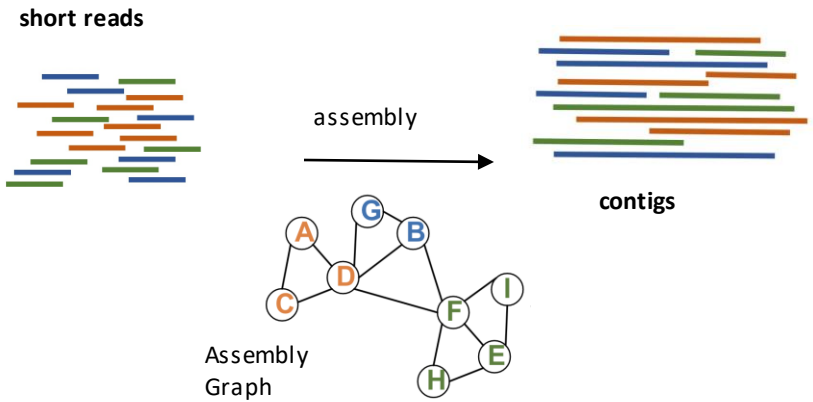
- Input sampled DNA fragments
- Produce short reads with 100-300 base pairs length
- About 0.1% error rate
- Low costs
- High throughput

https://www.omnia-health.com/product/next-generation-sequencing-platforms

01 Background

Assembly

- Reads are too short to produce reliable binning results
- Assemble short reads into longer contigs
- Obtain assembly graph

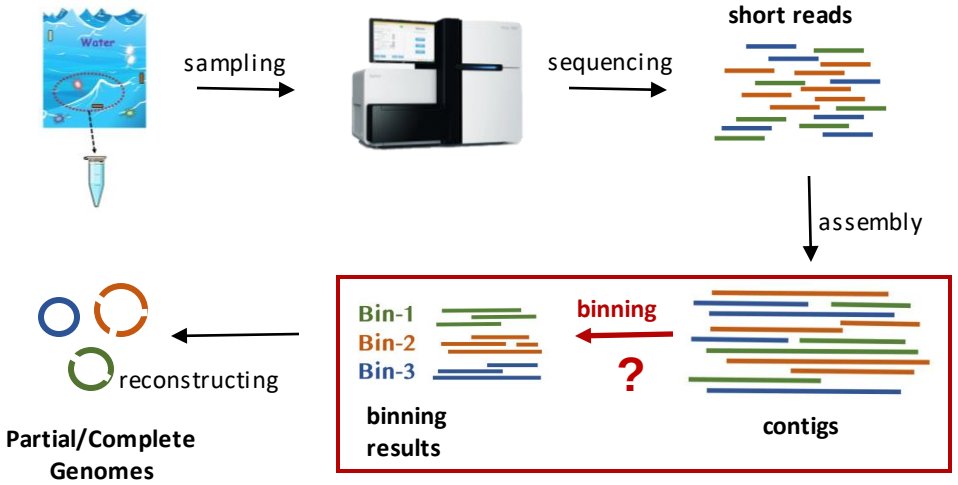


01 Background & Goal

Metagenomic Binning

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

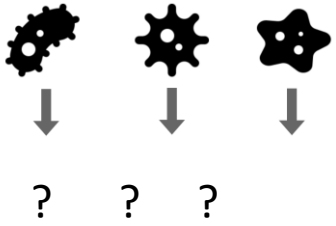
Metagenomic Workflow

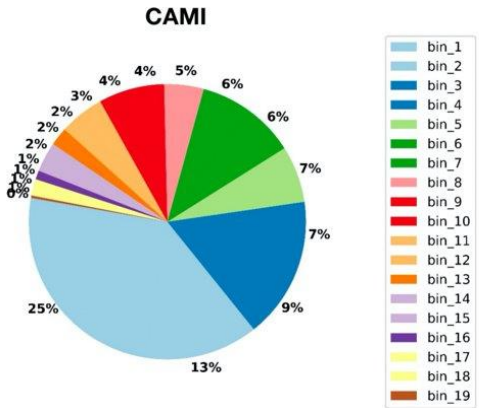


02 Challenge in Metagenomic Binning

1. How to learn the homophilous features of contigs?

2. How to mine the heterophilous relations among marker genes?

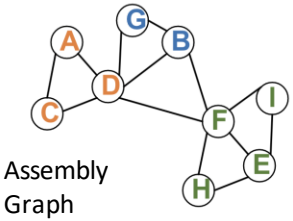




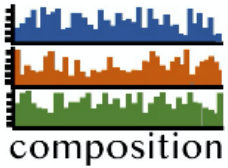
02 Challenge in Metagenomic Binning

1. How to learn the homophilous features of contigs?

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

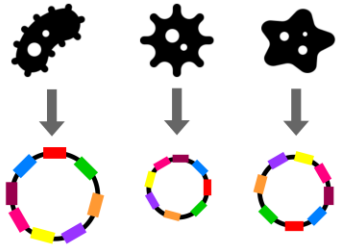


Assembly Graph


composition

2. How to mine the heterophilous relations among marker genes?

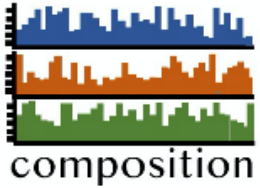
- solve by graph matching and clustering with single-copy marker genes contained in assembled contigs


single-copy marker genes

03 Methodology

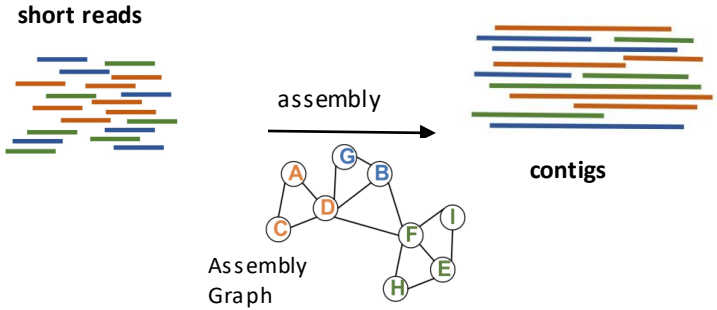
Composition Information

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



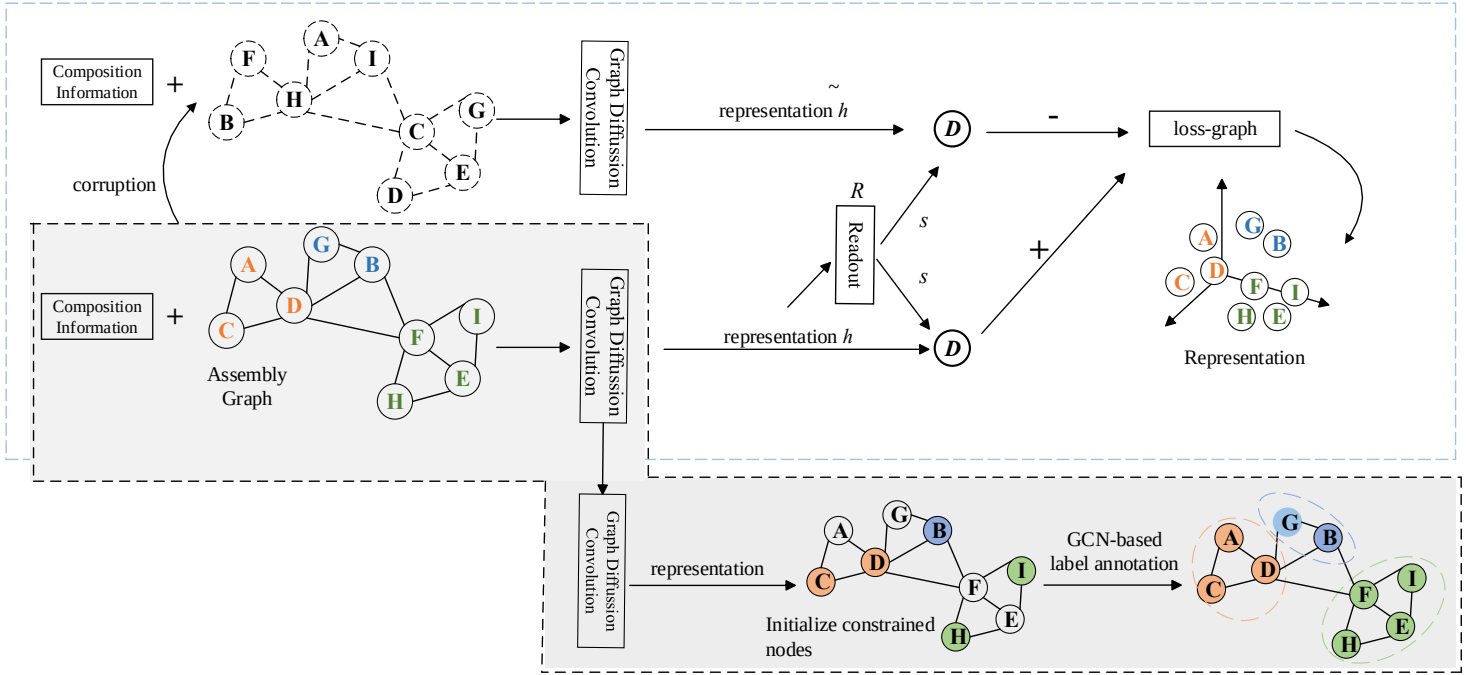
Assembly Graph

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



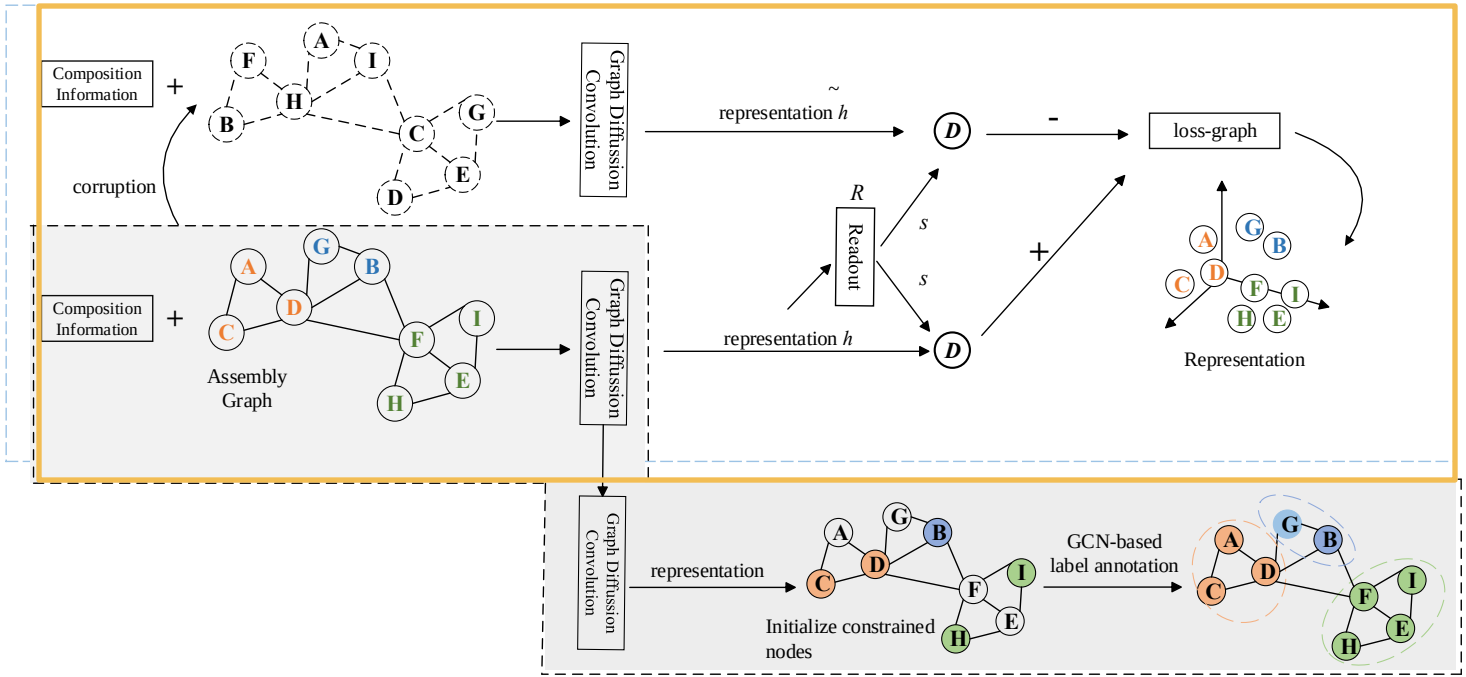
03 Methodology

MixBin framework



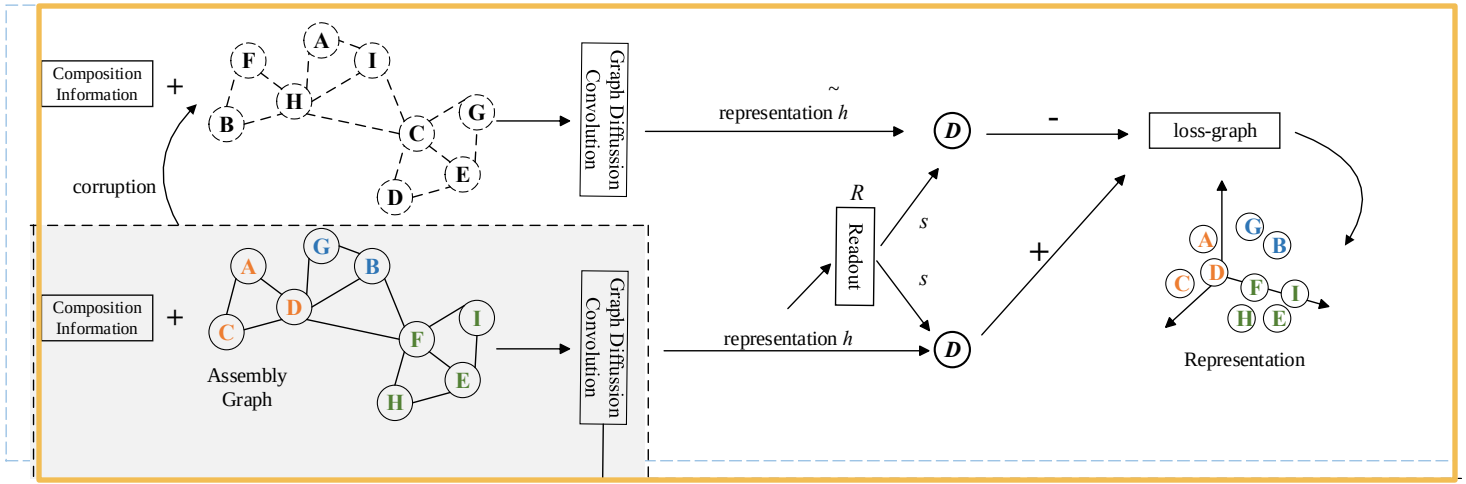
03 Methodology

MixBin Part 1 : Contrastive graph representation learning



03 Methodology

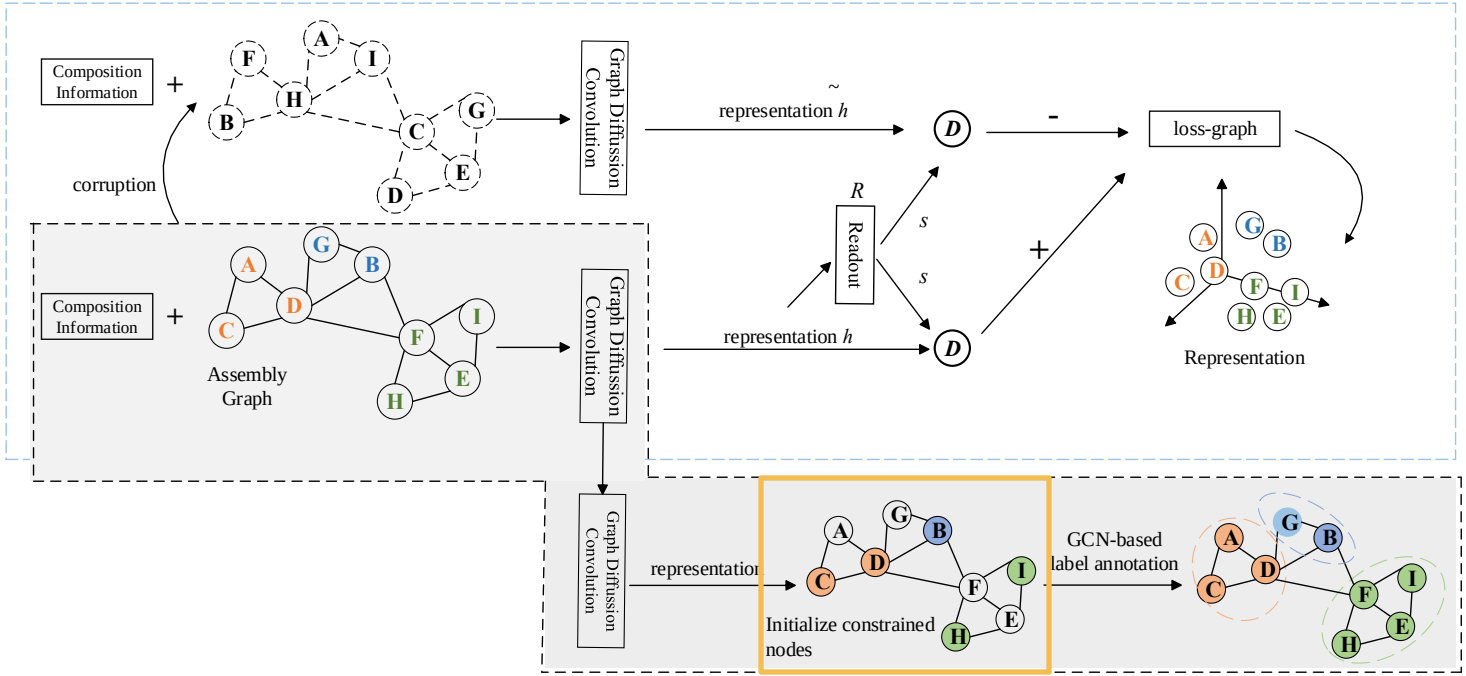
MixBin Part 1 : Contrastive graph representation learning



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

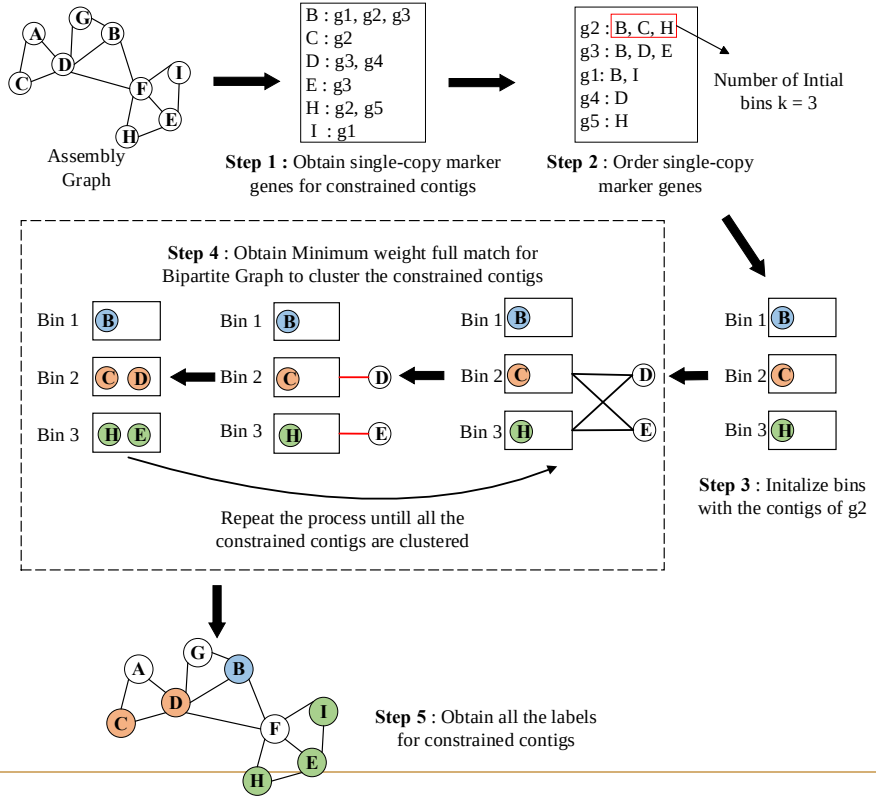
03 Methodology

MixBin Part 2 : Constrained bipartite graph matching



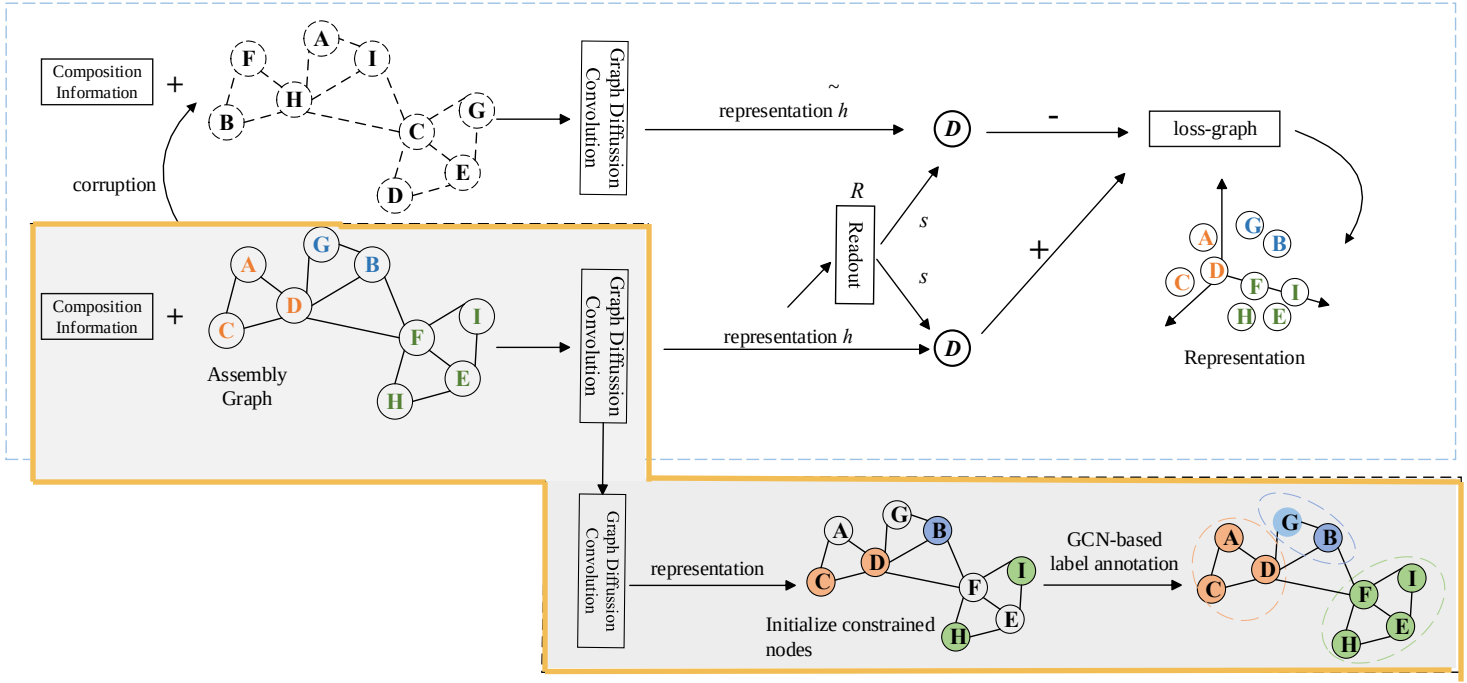
03 Methodology

MixBin Part 2 : Constrained bipartite graph matching



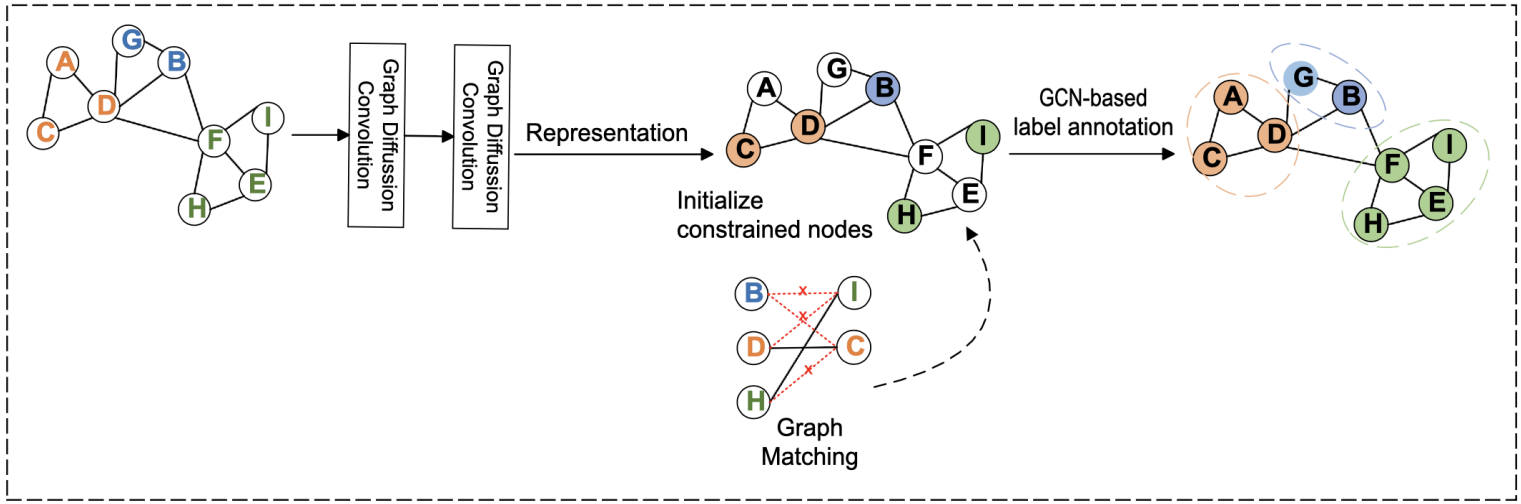
03 Methodology

MixBin Part 3 : GCN-based label propagation



03 Methodology

MixBin Part 3 : GCN-based label propagation



04 Experiment Setup

Datasets: 5 simulated data sets

Datasets	Read length (bp)	Number of paired-end reads	Number of assembled contigs	Mean contigs length (bp)	Number of links	Number of constraints	Number of species in ground truth
Sim-5G	300	2,000,000	519	51,723	2,488	91	5
Sim-10G	300	6,999,998	920	47,279	4,210	67	10
Sim-20G	300	15,000,001	1,452	48,021	6,531	75	20
Sim-50G	300	20,477,955	5,088	28,680	18,808	85	50
Sim-100G	300	51,167,221	15,729	19,978	62,518	85	100

Baselines: **4** unsupervised GNNs, **4** graph clustering models, and **8** binning tools.

GNNs : GraphSAGE GAT , DGI, VGAE .

Graph Clustering : O2MAC, AGC, CSC , DCC.

Binning Tools : MetaWatt, CONCOCT, MaxBin2, BusyBeeWeb, MetaBAT2, SolidBin, VAMB, RepBin,

Metrics: **F1**, **ARI**, and **NMI** for ML-based baselines;
Precision, **Recall**, and **F1** for Binning tools

04 Experiments & Result

Benchmarking against GNNs

Datasets		Graph Neural Networks				MixBin-*Learning*
		GSAGE	GAT	DGI	VGAE	
Sim-5G	F1	88.0±0.6	94.5±1.6	79.9±3.4	85.7±1.4	95.04±0.09
	ARI	72.9±0.9	86.9±1.7	54.6±6.8	70.1±2.6	91.00±0.17
	NMI	81.6±0.8	87.7±0.7	68.2±3.3	80.1±3.2	90.92±0.42
Sim-10G	F1	76.6±0.2	73.7±0.5	68.1±1.9	71.4±1.9	92.23±2.19
	ARI	59.3±0.7	54.0±2.6	39.3±2.4	46.0±2.6	84.65±5.56
	NMI	75.9±0.6	74.3±0.4	61.8±2.4	68.9±1.1	91.17±2.14
Sim-20G	F1	77.4±0.6	79.8±1.0	63.9±2.5	72.2±1.7	85.19±2.40
	ARI	61.5±1.8	65.4±1.3	40.1±2.3	54.8±2.1	73.20±3.74
	NMI	81.5±0.3	83.6±0.7	65.8±2.5	75.9±1.2	86.37±1.77
Sim-50G	F1	58.8±0.1	64.2±0.8	51.4±0.3	62.1±1.2	69.38±3.27
	ARI	40.9±1.7	44.7±1.8	37.1±1.1	41.1±0.5	53.68±3.44
	NMI	72.6±0.2	75.9±0.2	64.2±0.7	68.7±0.8	79.48±1.98
Sim-100G	F1	46.7±0.5	50.1±0.1	20.8±0.5	37.6±1.1	54.10±2.23
	ARI	31.2±0.8	26.9±0.3	21.7±0.8	25.8±0.6	37.90±2.88
	NMI	68.2±0.2	66.8±0.8	51.6±0.4	62.3±0.9	71.70±1.30

04 Experiments & Result

Benchmarking against Graph Clustering methods

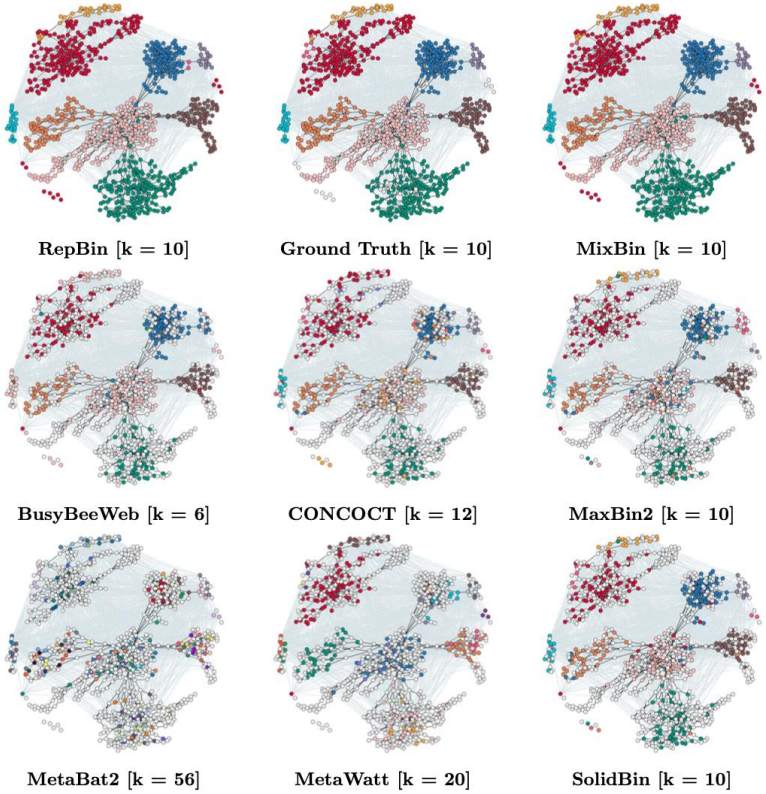
Datasets		Graph Clustering				MixBin
		O2MAC	AGC	CSC	DCC	
Sim-5G	F1	74.9±3.5	80.9±0.4	96.7±0.0	90.9±0.0	99.69±0.18
	ARI	63.6±2.1	92.7±0.8	87.9±0.0	94.0±1.0	99.11±0.42
	NMI	72.5±3.1	90.4±0.5	91.5±0.0	88.6±1.1	98.75±0.68
Sim-10G	F1	65.8±1.4	78.3±0.3	90.9±1.3	92.1±2.9	99.55±0.00
	ARI	53.5±2.1	87.9±0.3	77.9±4.2	83.3±3.0	99.39±0.08
	NMI	69.1±1.7	89.6±0.7	85.1±1.7	77.3±2.3	99.20±0.05
Sim-20G	F1	61.0±3.4	67.0±0.2	83.0±1.4	82.1±1.9	97.78±0.05
	ARI	52.0±2.7	75.9±1.1	63.2±4.3	65.3±2.8	96.20±0.11
	NMI	71.1±1.8	82.0±0.5	83.1±1.1	75.1±3.3	97.06±0.02
Sim-50G	F1	37.1±0.5	54.9±0.2	86.8±0.8	64.4±5.7	87.62±1.91
	ARI	16.5±0.3	44.4±0.8	77.0±2.6	48.7±4.1	85.49±2.83
	NMI	59.8±0.8	79.3±0.1	90.3±0.5	51.3±4.0	91.54±0.99
Sim-100G	F1	29.0±2.0	50.5±0.4	72.5±0.7	57.1±3.3	71.97±0.73
	ARI	8.5±0.1	21.1±0.5	37.5±7.9	40.9±7.9	58.58±0.80
	NMI	58.9±0.1	72.0±0.2	81.0±1.7	55.4±0.9	82.51±0.37

04 Experiments & Result

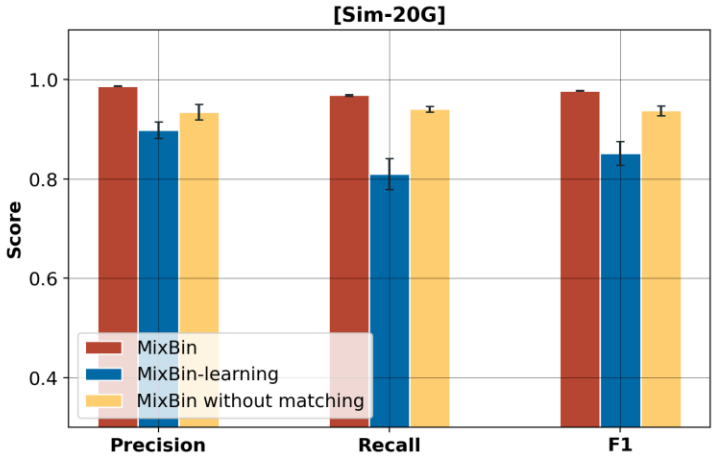
Benchmarking against metagenomic binning tools

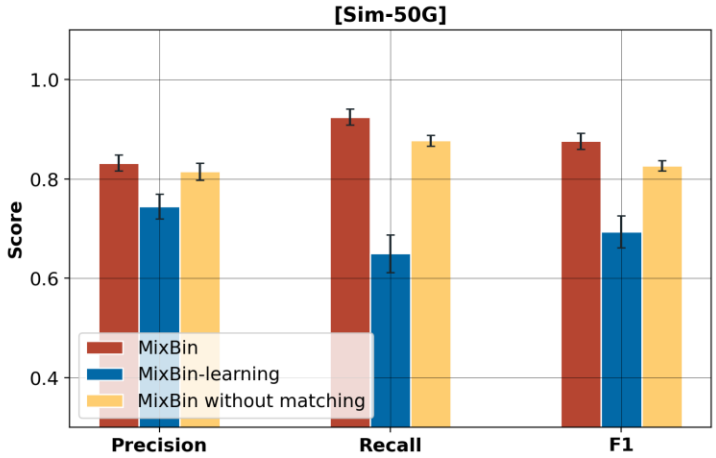
Datasets		MetaWatt	CONCOCT	MaxBin2	BusyBee Web	MetaBAT2	SolidBin	VAMB	RepBin	MixBin
Sim-5G	Precision	100.00	91.60	91.13	86.57	100.00	90.00	100.00±0.00	99.69±0.10	99.69±0.18
	Recall	24.59	40.50	46.69	49.79	6.61	46.49	33.92±0.90	99.69±0.10	99.69±0.18
	F1	39.47	56.16	61.75	63.22	12.40	61.31	50.66±1.02	99.69±0.10	99.69±0.18
	Pred. bins	12	7	5	4	34	5	6	5	5
Sim-10G	Precision	99.29	86.99	89.43	84.47	100.00	91.58	99.93±0.15	99.20±0.00	99.52±0.05
	Recall	26.13	39.72	40.30	45.53	6.39	41.70	33.80±0.20	99.55±0.08	99.57±0.05
	F1	41.38	54.54	55.56	59.17	12.01	57.30	50.51±0.23	99.37±0.04	99.55±0.00
	Pred. bins	20	12	10	6	56	10	11	10	10
Sim-20G	Precision	96.85	84.02	88.25	77.39	96.77	96.51	99.35±0.10	97.31±0.31	98.72±0.03
	Recall	32.01	42.27	41.69	44.51	7.73	85.04	36.88±0.60	96.98±0.69	96.86±0.13
	F1	48.12	56.24	56.63	56.52	14.32	90.41	53.79±0.64	97.15±0.61	97.78±0.05
	Pred. bins	33	22	21	12	88	20	22	20	20
Sim-50G	Precision	79.26	63.22	66.78	8.58	78.41	77.52	84.22±0.73	80.31±0.48	83.24±1.63
	Recall	17.65	38.76	40.89	4.21	5.67	38.67	39.32±0.45	90.59±2.01	92.49±2.42
	F1	41.42	47.65	51.23	5.65	11.32	51.60	55.37±1.56	84.55±1.80	87.62±1.91
	Pred. bins	75	56	53	12	98	45	48	50	47
Sim-100G	Precision	63.22	52.31	54.78	50.95	67.32	77.93	65.31±1.21	66.42±1.72	64.39± 0.72
	Recall	15.73	22.61	27.62	41.69	4.81	12.22	32.29±0.39	83.79±2.02	81.63± 2.43
	F1	32.34	34.59	36.73	45.86	9.63	21.12	45.21±0.79	74.08±0.74	71.97±0.73
	Pred. bins	157	132	127	28	256	84	87	100	92

04 Visualization



04 Further Analysis





Thank you

Any question?

Australian
National
University