University of Peradeniya

Grouping Nucleotide Sequences with More Precision Using GraphK-LR

Nethmi Ranasinghe[1], Sathsarani Aththanayaka[1], Jayathri Ranasinghe[1]
Vijini Mallawaarachchi[2] , Damayanthi Herath[1]
[1]Department of Computer Engineering, Faculty of Engineering, University of Peradeniya. [2]Flinders Accelerator for Microbiome Exploration, College of Science and Engineering, Flinders University, Australia.

Abstract - Metagenomics has shown rising interest towards Third-gen sequencing to overcome the limitations of short-reads. This research involves developing GraphK-LR, a novel long-read binning refiner designed to incorporate kingdom level information of different microbial kingdoms, utilizing graph based approaches.

Background - Metagenomics differs from traditional lab culturing by enabling study of genetic material directly from the microbial communities. Metagenomic Binning is the process of grouping nucleotide sequences belonging to similar organisms. Long reads, sequenced with high precision can carry species-specific signals, facilitating direct grouping into microbial kingdoms, enhancing the accuracy of metagenomic binning.

Methodology

In refining initial binning result of a given tool,

- connected nodes with different labels are considered as mis-binned nodes, which undergo refinement.
- single-copy marker genes which occur once in almost every genome are used for annotating nodes at kingdom-level.

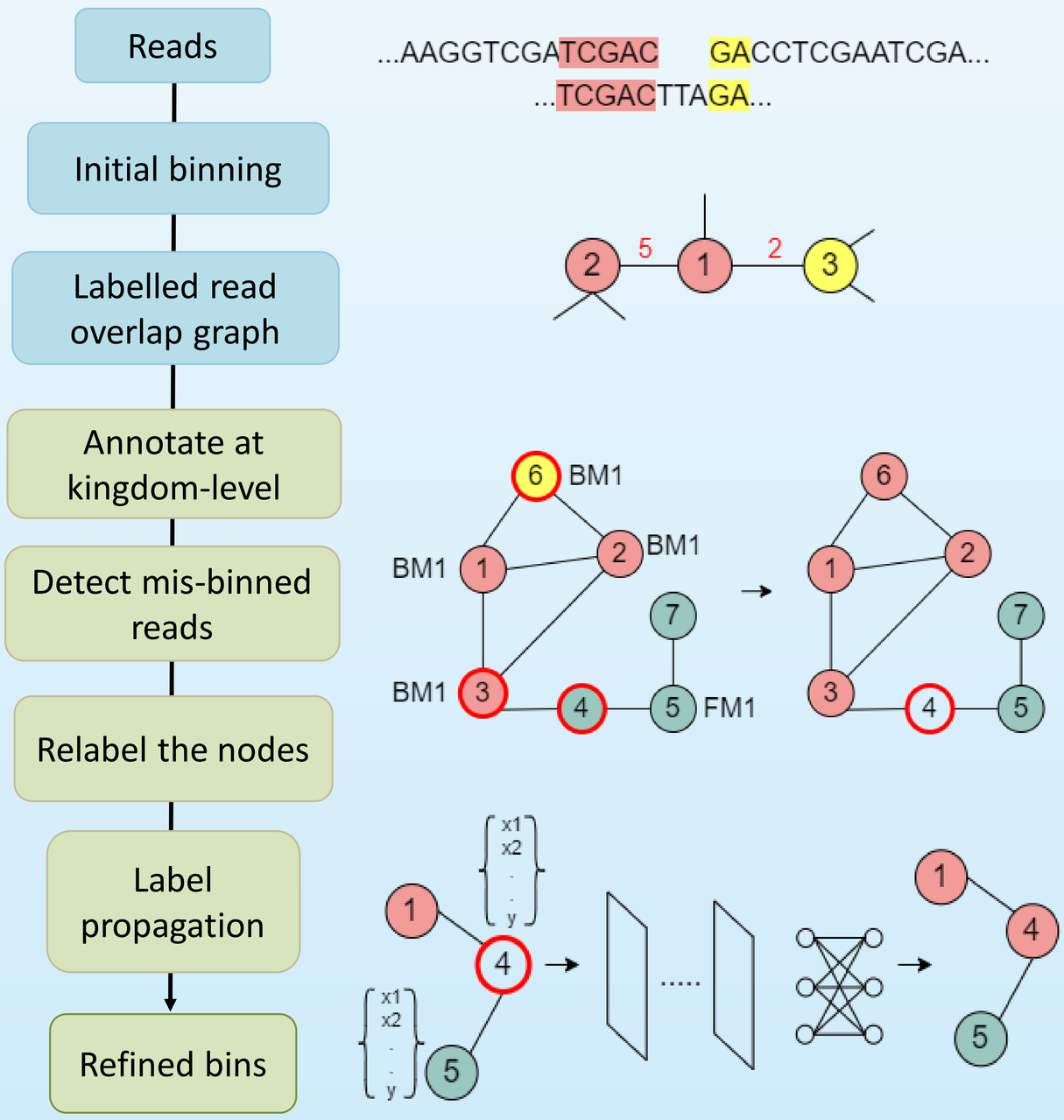

Figure 1: Work Flow of GraphK-LR

GraphSAGE, is a graph neural network that learns node representations by sampling and aggregating information from neighbors. It enables efficient label propagation in large scale graphs.

Existing Long-read Binning Tools

- Composition - Normalized frequency of short substrings of a particular read
- Coverage - Number of reads that overlaps with a specific region in a reference genome

Tool	Clustering Algorithm
MetaBCC-LR[1]	DBSCAN (density based algorithm)
LRBinner[2]	Distance based statistical algorithm
OBLR[3]	HDBSCAN (density based hierarchical algorithm)

Table 1 : Existing tools comparison

Results

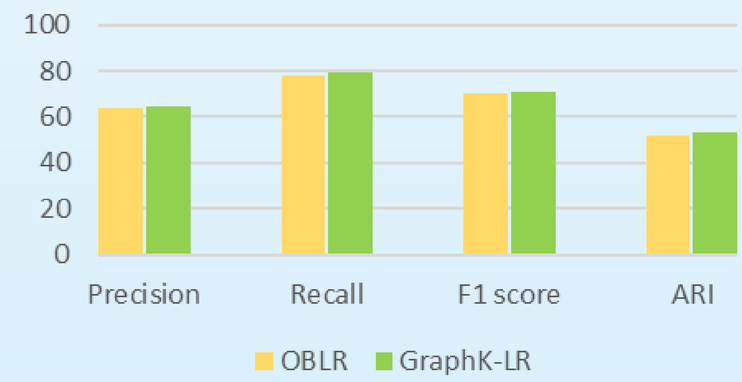

Figure 2: Evaluation Results

- ERR97765782; a mock community dataset with 71 species. Recovered 35 bins at initial binning.
- Improved initial binning results by a modest 1%

Practical Use

Accurately dissecting metagenomic samples allows quality reconstruction of genomes. This helps in,

- Uncovering new cancer therapeutic strategies
- Exploring important microbes for sustainable agriculture

etc.

References

[1] Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). MetaBCC-LR: metagenomics binning by coverage and composition for long reads. Bioinformatics (Oxford, England), 36(Suppl_1), i3–i11.
[2] Wickramarachchi, A., & Lin, Y. (2022). Binning long reads in metagenomics datasets using composition and coverage information. Algorithms for molecular biology : AMB, 17(1), 14.
[3] Wickramarachchi, A., & Lin, Y. (2022, May). Metagenomics binning of long reads using read-overlap graphs. In RECOMB International Workshop on Comparative Genomics (pp. 260-278). Cham: Springer International Publishing.

Contact details
Name : Dr. Damayanthi Herath
Tel. No.: +94779667468
Email : damayanthiherath@eng.pdn.ac.lk

Multidisciplinary AI Research Centre (MARC)
University Research Council
University of Peradeniya
Peradeniya, 20400, Sri Lanka