Metagenomic Binning Pipelines - the State of the Art

1 Abstract

- *Decision tree graphical abstract for the choice of binning algorithm*
- *Features that distinguish binning algorithms*
- *Some guidelines for choosing the correct binning techniques appropriate for a given study*

New generations of sequencing platforms coupled to numerous bioinformatics tools have led to rapid technological progress in metagenomics and metatranscriptomics to investigate complex microorganism communities. Nevertheless, a combination of different bioinformatic tools remains necessary to draw conclusions out of microbiota studies. As sequencing costs have dropped at a rate above 'Moore's law', bigger data sets are available, and proportional costs of analysis have risen as a consequence. Binning is the grouping of assembled metagenomic contigs by their genome of origin. Algorithms for binning are a rapidly evolving field. The number of these algorithms are growing over time. Selecting the most appropriate binning algorithm can be a daunting task and is influenced on computational resources available and experimental variables relating to the sequencing. This review serves as a roadmap to direct the researcher to the binning algorithm that best suits their needs.

2 Background

- *General introduction to history of binning*
- *increase in popularity of the field of metagenomics*
- *Talk about reduced cost in sequencing and scheduling efficiency in pipelines*
- *Talk about HMP*

Compared to amplicon, shotgun metagenome can provide functional gene profiles directly and reach a much higher resolution of taxonomic annotation. However, due to the high demands on computational resources, cost, and expertise necessary to perform this analysis, this method has been used The substantial decrease in cost of sequencing however has lessened this burden. At the time of writing, shotgun metagenomic sequeng costs on average three times as much as 16S sequencing comparitavely. A review on the benchmarking binning algorithms was done by Yue et al., 2020.

3 Methods for metagenomic binning

- *Not sure if to focus on this or the appropriateness*
- *The increased impact of machine learning in analysis*
- *Short section - just for past-present-future completeness*

3.1 Metagenome Assembled Genomes

A Metagenome-Assembled Genome (MAG) is a single-taxon assembly based on one or more binned metagenomes that has been asserted to be a close representation to an actual individual genome (that could match an already existing isolate or represent a novel isolate).

3.2 Metagenomic Species Pan-genomes

Microbial pan-genomes are gene repertoires composed of core genes present in all strains and accessory genes present in only some of them (Medini et al., 2005). In a shotgun metagenomic sequencing context, we define as shared the genes detected in some samples where the species is not present.

A strain found in a sample is an instance of the species pan-genome: it is made of all the species (shared) core genes and of a subset of (shared) accessory genes. Core genes are suitable for taxonomic profiling at species-level while accessory genes can be used to compare strains across samples. Genes tagged as shared should be used carefully as they contain false positives counts or are subject to horizontal transfer.

We assumed that core genes of a microbial species should be consistently detected in samples where the species is present if sequencing depth allows (co-occurrence) and should yield directly proportional mapped reads counts across samples (co-abundance). Remarkably, a core and an accessory gene should have proportional counts only in the subset of samples carrying a strain with that accessory gene

3.3 Megagenome Assembled Genome

3.4 Binning microbial genomes with deep learning

The integration of deep learning techniques into the field of metagenomics has revolutionised the field of metagenomics. The VAMB pipeline was developed to take advantage of variational autoencoders; a generative machine learning model that uses a combination Improved metagenome binning and assembly using deep variational autoencoders Nature biotechnology - 4th Jan 2021 the VAMB pipeline (Nissen et al., n.d.)

4 Chosing the most appropriate binning algorithm

Resource management is an important factor in the choice of binning algorithm. The tradeoff between number of CPU's, memory, and time are important considerations. Pipeline vs standalone? Alignment based or alignment free An analysis pipeline is defined as a program that combines several programs in a defined order to complete a complex analysis. Improperly developed, validated, and/or monitored pipelines may generate inaccurate results.

4.1 Binning for viral genomes

New insights from uncultivated genomes of the global human gut microbiome Nature - 13th March 2019 (Nayfach, Shi, Seshadri, Pollard, & Kyrpides, 2019)

5 Conclusion

- *New and open areas of research in which the application of metagenomic pipelines are relevant*
- *HMP and other*
- *The increased impact of machine learning in analysis*

Table 1: Comparison of binning algorithms

Software/Algorithm	Year	Description/purpose	Reference free	Comment/Highlight	Software category/ Topic	Input/Output data typ
CoCoNet	2021	Deep learning tool for Viral Metagenome Binning	Yes	Reconstucts viral genomes	MAG binning	Contigs, coverge (read
Binnacle	2021	Using scaffolds to improve Metagenomic bin quality	Yes	Incorporates scaffold information	MAG binning-refiner*	Contigs, coverage, bins
VAMB	2021	Metagenome binning using deep variational autoe...	Yes	Autoencoder algorithm, fast processing	MAG binning	Assembled contigs/ bi
phyloFlash	2020	ssrRNA profiling and MAG assembly	No	incorporates ssrRNA profiling info into MAG ass...	SSU rRNA assembly/ connect SSUrRNA to MAG	Reads and reference da
MetaBCC-LR	2020	Metagenomic binning for Long-Reads	Yes	Suitable for Long Reads sequencing technology	LongReads binning	Long reads/ Bin identi
BioBloom Tools	2020	Reads binning for targeted assembly, alignment ...	No	Data preparation for targeted assembly, using s...	Reads binning	Reads and reference se
?	2020	Binning unassembled short reads based on k-mer ...	Yes	Data preparation for targeted assembly, low abu...	Reads binning	NaN
GraphBin	2020	Refined binning of metagenomic contigs using as...	NaN	Incorporates assembly graphs information	MAG binning-refiner*	NaN
MetaSIPSim	2020	Simulating metagenomic stable isotope probing d...	NaN	Augment binning resolution with extra experimen...	Data Simulation/ DNA-SIP	NaN
MetaCon	2019	Unsupervised binning k-mers and coverage, focus...	NaN	Focus different lengths contigs	MAG binning	NaN
VirBin	2019	Binning viral haplotypes from assembled contigs	NaN	Viral haplotypes MAGs	Haplotype/MAG binning*	NaN
MAGO (*only tool pipeline)	2019	Framework for Production and analysis of MAGs	NaN	pipeline	MAG binning-refiner pipeline	NaN
SeqDex	2019	Genome separation of Endosymbionts from mixed s...	NaN	Identification of endosymbiont	MAG binning*	NaN
MetaTOR	2019	High quality MAGs from mammalian guts using met...	NaN	Incorporates 3D contact information	MAG binning	NaN
MetaBAT 2	2019	Adatptative binning algorithm for genome recons...	NaN	Eliminates manual parameter tuning from previou...	MAG binning	NaN
MetaBMF	2019	Scalable binning algorithm for large scale meta...	NaN	Employs sample X contigs cf mapped read counts	MAG binning	NaN
PolyCRACKER	2019	Method for partitioning polyploid sub genomes b...	NaN	Haplotypes for polyploid genomes	Haplotype binning	NaN
SolidBin	2019	Improving metagenome binning with semi-supervis...	NaN	NaN	MAG binning	NaN
Autometa	2019	extraction of microbial genomes from individual...	NaN	Handles eukaryotic contamination	MAG binning / pipeline	NaN
MLBP MrGBP (Algorithm)	2019	Signal processing method for alignment free met...	NaN	Alternative description of sequences designed f...	MAG binning*	NaN
CLAME	2018	Aligment based algorithm allowed description of...	No	Aligment based for reads	Reads binning*	NaN
3D BH SNE (Algorithm)	2018	Fuzzy binning of metagenomic sequence fragments	NaN	Horizontal gene transfer and regions of uncerta...	MAG binning /Fuzzy binning	NaN
LVQ-KNN	2018	Composition based RNA or DNA binning of short s...	NaN	Classify into DNA or RNA sequence	Reads binning	NaN
MSPminer	2018	Abundance based reconstitution of microbial pan...	NaN	Pan genome reconstitution	MSP binning	NaN
MetaWRAP*	2018	Flexible pipeline for genome resolved metagenom...	NaN	Hybrid bin extraction algorithm	MAG binning-refiner pipeline	NaN
MetaVW	2018	Large scale Machine Learning Sequence classific...	NaN	Machine learning for reads based on Khmer profile	Reads binning	NaN
Opal (algorithm*)	2018	Metagenomic binning through low density binning	NaN	Improvement at higher taxonomic levels, discove...	MAG binning	NaN
BMC3C	2018	Binning contigs using codon usage sequence comp...	NaN	Add codon usage information	MAG binning	NaN
AMBER tool	2018	Assessment of Metagenome Binners	NaN	NaN	Benchmark tool for binning software	NaN
DAS Tool	2018	Derreplication aggregation and scoring strategy	NaN	Combines several binning algorithm results	MAG binning-refiner	NaN
MEGAN-LR	2018	Long Read/ contigs taxonomic binning	No	Aligment of long reads against reference sequences	NaN	NaN
CoMet	2018	Binning workflow using contain coverage and com...	NaN	Single sample, include gc content and 4mer fre...	MAG binning	NaN
?	2017	Metagenomic binning and association of plasmids...	NaN	Plasmid banning at strain level using methylati...	MAG binning	NaN
MetaGen	2017	reference-free learning with multiple metagenom...	NaN	Requires multiple samples	MAG binning	NaN
d2sBin add onn	2017	Improved formula for calculate oligonucleotide ...	NaN	Math formula to calculate oligo sequence dissim...	NA, only formula for dissimilarity	NaN
BusyBee Web	2017	Bootstrapped supervises binning and annotation	NaN	2d interactive scatterplots supervised binning	refiner	NaN
ICoVer	2017	Interactive visualisation tool for verification...	NaN	Interactive visualisation tool	MAG refiner	NaN
HirBin*	2017	High resolution identification of differentiall...	NaN	Supervised annotation, unsupervised clustering ...	Reads binning	NaN
BinSanity	2017	Unsupervised clustering using coverage and affi...	NaN	Reduce bias for high/low abundance	MAG binning	NaN
Binning_refinner	2017	Improve genome bins through the combination of ...	NaN	Combination of different binning algorithms	MAG binning-refiner	NaN
IFCM add on	2016	Improved binning using Fuzzy C-Means Method	NaN	Add estimated distribution of real genome lengths	MAG binning	NaN
COCACOLA	2016	binning contigs using composition, read coverag...	NaN	Adds paired end read and coaligment information	MAG binning	Paired end, contigs, co
GroopM (2)	2014	Tool for automatic recovery of population genom...	NaN	Adds differential coverage to complement compos...	MAG binning	NaN

- *Short section - just for past-present-future completeness*
- *Future developments for metagenomic analysis*

References

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., ... others (n.d.). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, 1–6.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., ... Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and cami datasets. *BMC bioinformatics*, *21*(1), 1–15.