Metagenomic Binning Pipelines - the State of the Art

Contents

1 Abstract

New generations of sequencing platforms coupled with numerous bioinformatics tools have led to rapid technological progress in metagenomics 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', a greater number of large data sets are being produced than ever before. Newer algorithms that take advantage of the size of these datasets are continually being developed. Binning algorithms are defined as the grouping of assembled metagenomic contigs by their genome of origin. Selecting the most appropriate binning algorithm can be a daunting task and is influenced by many factors. This review serves as a guide to direct the researcher to the binning algorithm that best suits their needs.

2 Background

The explosion in popularity and success in the field of metagenomics over the last 25 years can be largely attributed to the advances in computing technologies. An example of the outcomes of this can be found in the Human Microbiome Project; a project that has been greatly improved the understanding of the microbila flora involved in human health and disease. These advances have brought with them greater demands for storage, CPU time, and consequently more efficient algorithms. The main function of binning tools is to reconstruct species/biological entities from metagenomic samples. 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, researchers have historically been limited in their capacity to collect and analyse sequencing data. As the cost of sequencing is rapidly falling, this burden has been significantly lessened. Whole Genome Shotgun sequences does not require cultivation. At the time of writing, shotgun metagenomic sequencing costs on average three times as much as 16S sequencing in comparison. The objectives of this review is for the reader to be better informed about the latest algorithms (since 2017) for binning metagenomic samples. The second part of this review is for the reader to be informed about distinguishing factors between the methods. The last part is for the reader to make an informed decision based on those factors for their needs. This review will be broken down into the following sections:

3 Overview of recent methods for metagenomic binning

3.1 Innovations in recent binning strategies

A metagenomic sample is comprised of many organisms and the standard procedure is to retrieve the sequences from the mixture of organisms. The final goal of binning is to reconstruct the sequences from each organism present in the original sample. Currently we can distinguish from 3 different stategies in binning algorithms, read binning, contig binning, and gene binning. Among the binning tools developed in recent years a subset of them are dedicated to cluster reads (read-binning) (MetaBBC-LR, BioBloom Tools, CLAME, LVQ-KKN, Meta VW, HirBin, MEGAN-LR). The main purpose of read-binning tools is to pre-process reads into clusters for a posterior targeted assembly, here we find reference-free and non-reference-free tools, and tools designed for short-read or long-read sequencing technologies. The majority of binning tools we can find are oriented toward clustering contigs (contig-binning) into bins, which may represent the genome from a single biological entity/organism. Contig-binning tools normally rely on coverage information and sequence composition. Binning contigs have played a central role in software development in the field, a review on the benchmarking binning algorithms was done by Yue et al., 2020. Progress in contig-binning algorithms can be seen in the proposals to integrate new sources of information (for example, from scaffold-graphs(Binnacle), paired-end reads(COCACOLA), or 3D contact information(MetaTOR)) and state of the art algorithms in machine learning (CoCoNet, Variational Autoencoders for Metagenomic Binning (VAMB)).

3.1.1 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).

3.1.2 MSPs, binning co-abundant genes

Binning of co-abundant genes represents an alternative proposal to reconstruct species/biological entities from a set of metagenomic samples. Co-abundant gene binning methods assume each gene coming from a shared chromosome will display proportional abundances across samples, if you have enough samples from a common environment you can identify the sets of genes from a common organism of origin (MLGs Chameleon-clust 2012, CAGs and MGSs Canopy 2014, Markovclust-

MGCs Karlsson 2013, MSPs MSPminner 2018). To the extent of our knowledge, in the past few years MSPminer was the only Software developed exploiting this approach. MSPminer introduced a robust proportionality measure detecting co abundant but no necessarily co-occurring. This tools groups co-abundant genes into Metagenomic Species Pan-genomes or Metagenomic Species Pan-genomess (MSPs) and classify genes within an MSP as core, accessory and shared. The factors that impact directly on MSP quality include the sample composition, the sequencing depth, the previous bioinforamtic steps to build the reference gene dataset and to map the reads. A high number of samples with varying phenotypes improve the quality of MSPs. MSPs can be employed for taxonomic profiles of new samples from similar ecosystems, to compare strains between samples building a presence/absence table of accessory genes and for biomarker discovery. By binning contigs carrying genes from the same MSP it is also possible to build a MAG.

3.1.3 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. 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.

3.1.4 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 deep variational autoencoders (Nissen et al., n.d.)... COCONET (Arisdakessian, Nigro, Steward, Poisson, & Belcaid, 2021)...

3.2 Binning for solving new biological challenges/ for viral genome

Most binning algorithms are designed for prokariotic organisms leaving viruses out of the design. Viruses are a seriuos threath to human health CoCoNet uses deep leaning to model co-ocurrence of contigs from the same viral genome. The method uses a neural network which returns the probability for a pair of contigs comming from the same genome, this probabilities are employed to construct bins representing the species present in the sample. The network was optimized for diverse viral metagenomes, the network learns to model coverage variability within samples, a critical feature in viral metagenomes where DNA amplification methods are needed to increase input genetic material. VirBin clusters contigs for genome reconstruction of viral strains, different strains within viral species may show different biological properties such as transmissibility or virulence. Composition based features are ussually are not enough to separate haplotypes, VirBin recieves contigs as inputs and outputs the estimated number of haplotypes via contig aligment and returns the contigs for each haplotype based on reltive abundance distribution, when the contigs are long enough VirBin produce better results. Pipelines for Endosymbiont organisms binning also have been recently developed. Extracting endosymbiont sequences from their host poses a similar problem as a metagenomic samples. Seqdex is a tool written in R which tries to separate endosymbiont from host sequences, they proposed they could use specific features in endosymbiotic systems to better solve this problem. This tool combines partial taxonomic annotations obtained trough homology searches and sequence composition to predict the contig's organism of origin from host and its endosymbionts and helps the user to understand how effective is the classification.

4 Choosing the most appropriate binning algorithm (Classification by output)

A review on the benchmarking binning algorithms was done by Yue et al., 2020. Resource management is an important factor in the choice of binning algorithm. The tradeoff between number of Central Processing Units (CPUs), memory, and time are important considerations. Newer advances in pipeline technologies have ameliorated these costs. 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.

Table 1: Comparison of binning algorithms

Software/Algorithm	Year	Description/purpose	Comment/Highlight	Doi	PubmedID
CoCoNet	2021	Deep learning tool for Viral Metagenome Binning	Reconstucts viral genomes	10.1093/bioinformatics/btab213	33822891
Binnacle	2021	Using scaffolds to improve Metagenomic bin quality	Incorporates scaffold information	10.3389/fmicb.2021.638561	33717033
VAMB	2021	Metagenome binning using deep variational autoe...	Autoencoder algorithm, fast processing	10.1038/s41587-020-00777-4	33398153
phyloFlash	2020	ssrRNA profiling and MAG assembly	incorporates ssrRNA profiling info into MAG ass...	10.1128/mSystems.00920-20	33109753
MetaBCC-LR	2020	Metagenomic binning for Long-Reads	Suitable for Long Reads sequencing technology	10.1093/bioinformatics/btaa441	32657364
BioBloom Tools	2020	Reads binning for targeted assembly, alignment ...	Data preparation for targeted assembly, using s...	10.1073/pnas.1903436117	32641514
GraphBin	2020	Refined binning of metagenomic contigs using as...	Incorporates assembly graphs information	10.1093/bioinformatics/btaa180	32167528
MetaSIPSim	2020	Simulating metagenomic stable isotope probing d...	Augment binning resolution with extra experimen...	10.1186/s12859-020-3372-6	32000676
MetaCon	2019	Unsupervised binning k-mers and coverage, focus...	Focus different lengths contigs	10.1186/s12859-019-2904-4	31757198
VirBin	2019	Binning viral haplotypes from assembled contigs	Viral haplotypes MAGs	10.1186/s12859-019-3138-1	31684876
MAGO (*only tool pipeline)	2019	Framework for Production and analysis of MAGs	pipeline	10.1093/molbev/msz237	31633780
SeqDex	2019	Genome separation of Endosymbionts from mixed s...	Identification of endosymbiont	10.3389/fgene.2019.00853	31608107
MetaTOR	2019	High quality MAGs from mammalian guts using met...	Incorporates 3D contact information	10.3389/fgene.2019.00753	31481973
MetaBAT 2	2019	Adatptative binning algorithm for genome recons...	Eliminates manual parameter tuning from previou...	10.7717/peerj.7359	31388474
MetaBMF	2019	Scalable binning algorithm for large scale meta...	Employs sample X contigs cf mapped read counts	10.1093/bioinformatics/btz577	31347687
PolyCRACKER	2019	Method for partitioning polyploid sub genomes b...	Haplotypes for polyploid genomes	10.1186/s12864-019-5828-5	31299888
SolidBin	2019	Improving metagenome binning with semi-supervis...	NaN	10.1093/bioinformatics/btz253	30977806
Autometa	2019	extraction of microbial genomes from individual...	Handles eukaryotic contamination	10.1093/nar/gkz148	30838416
MLBP MrGBP (Algorithm)	2019	Signal processing method for alignment free met...	Alternative description of sequences designed f...	10.1038/s41598-018-38197-9	30770850
CLAME	2018	Aligment based algorithm allowed description of...	Aligment based for reads	10.1186/s12864-018-5191-y	30537931
3D BH SNE (Algorithm)	2018	Fuzzy binning of metagenomic sequence fragments	Horizontal gene transfer and regions of uncerta...	10.1109/EMBC.2018.8512529	30440633
LVQ-KNN	2018	Composition based RNA or DNA binning of short s...	Classify into DNA or RNA sequence	10.1016/j.virusres.2018.10.002	30291874
MSPminer	2018	Abundance based reconstitution of microbial pan...	Pan genome reconstitution	10.1093/bioinformatics/bty830	30252023
MetaWRAP*	2018	Flexible pipeline for genome resolved metagenom...	Hybrid bin extraction algorithm	10.1186/s40168-018-0541-1	30219103
MetaVW	2018	Large scale Machine Learning Sequence classific...	Machine learning for reads based on Khmer profile	10.1007/978-1-4939-8561-6_2	30030800
Opal (algorithm*)	2018	Metagenomic binning through low density binning	Improvement at higher taxonomic levels, discove...	10.1093/bioinformatics/bty611	30010790
BMC3C	2018	Binning contigs using codon usage sequence comp...	Add codon usage information	10.1093/bioinformatics/bty519	29947757
AMBER tool	2018	Assessment of Metagenome Binners	NaN	10.1093/gigascience/giy069	29893851
DAS Tool	2018	Derreplication aggregation and scoring strategy	Combines several binning algorithm results	10.1038/s41564-018-0171-1	29807988
MEGAN-LR	2018	Long Read/ contigs taxonomic binning	Aligment of long reads against reference sequences	10.1186/s13062-018-0208-7	29678199
CoMet	2018	Binning workflow using contain coverage and com...	Single sample, include gc content and 4mer fre...	10.1186/s12859-017-1967-3	29297295
?	2017	Metagenomic binning and association of plasmids...	Plasmid banning at strain level using methylati...	10.1038/nbt.4037	29227468
MetaGen	2017	reference-free learning with multiple metagenom...	Requires multiple samples	10.1186/s13059-017-1323-y	28974263
d2sBin add onn	2017	Improved formula for calculate oligonucleotide ...	Math formula to calculate oligo sequence dissim...	10.1186/s12859-017-1835-1	28931373
BusyBee Web	2017	Bootstrapped supervises binning and annotation	2d interactive scatterplots supervised binning	10.1093/nar/gkx348	28472498
ICoVer	2017	Interactive visualisation tool for verification...	Interactive visualisation tool	10.1186/s12859-017-1653-5"	28464793
HirBin*	2017	High resolution identification of differentiall...	Supervised annotation, unsupervised clustering ...	10.1186/s12864-017-3686-6	28431529
BinSanity	2017	Unsupervised clustering using coverage and affi...	Reduce bias for high/low abundance	10.7717/peerj.3035	28289564
Binning_refinner	2017	Improve genome bins through the combination of ...	Combination of different binning algorithms	10.1093/bioinformatics/btx086	28186226
IFCM add on	2016	Improved binning using Fuzzy C-Means Method	Add estimated distribution of real genome lengths	10.1109/TCBB.2016.2576452	27295684
COCACOLA	2016	binning contigs using composition, read coverag...	Adds paired end read and coaligment information	10.1093/bioinformatics/btw290	27256312
GroopM (2)	2014	Tool for automatic recovery of population genom...	Adds differential coverage to complement compos...	10.7717/peerj.603	25289188

4.1 Identify start point variables

4.2 Identify endpoint

4.3 Tools that are complementary

5 Conclusion

Until now binning methods perform poorly in samples that contain similar strains. Also do not perfor great assignning 16S sequences to bins maybe due to high copy number within a genome. Binning has been focused mainly in prokariotic organisms. Binning of organisms outside prokatiotes need more development, lately some advances have been observed in viral genomes (cite viral catalogue and viral binning organims) but the huge diveristy in viral genomes still poses a challenge for currrent methodologies. eukariotic microscopic organisms does not appear in the current picture. The continouisly increasing number of sequences available require more efficient/faster algorithms and new strategies to recontruct single organisms from enviromental samples. Delvelopment of Machine learning algorithms have started in the field and we expect to see more development soon New and open areas of research in which the application of metagenomic pipelines are relevant The increased impact of machine learning in analysis Short section - just for past-present-future completeness Future developments for metagenomic analysis

References

Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G., & Belcaid, M. (2021). Coconet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics*.

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.