Current Genomics, 2022, *23*, 137-146

RESEARCH ARTICLE

MetaConClust - Unsupervised Binning of Metagenomics Data using Consensus Clustering

Dipro Sinha[1,#], Anu Sharma[2,#,*], Dwijesh Chandra Mishra[2,#], Anil Rai[2], Shashi Bhushan Lal[2], Sanjeev Kumar[2], Moh. Samir Farooqi[2] and Krishna Kumar Chaturvedi[2]

[1]*Research Scholar, PG School, ICAR-IARI, New Delhi-110012, India;* [2]*Division of Agriculture Bioinformatics, ICAR-IASRI, New Delhi- 110012, India*

ARTICLE HISTORY

Received: December 22, 2021
Revised: February 16, 2022
Accepted: February 21, 2022

DOI:
10.2174/1389202923666220413114659

Abstract: ***Background*:** Binning of metagenomic reads is an active area of research, and many unsupervised machine learning-based techniques have been used for taxonomic independent binning of metagenomic reads.

***Objective*:** It is important to find the optimum number of the cluster as well as develop an efficient pipeline for deciphering the complexity of the microbial genome.

***Methods*:** Applying unsupervised clustering techniques for binning requires finding the optimal number of clusters beforehand and is observed to be a difficult task. This paper describes a novel method, MetaConClust, using coverage information for grouping of contigs and automatically finding the optimal number of clusters for binning of metagenomics data using a consensus-based clustering approach. The coverage of contigs in a metagenomics sample has been observed to be directly proportional to the abundance of species in the sample and is used for grouping of data in the first phase by MetaConClust. The Partitioning Around Medoid (PAM) method is used for clustering in the second phase for generating bins with the initial number of clusters determined automatically through a consensus-based method.

***Results*:** Finally, the quality of the obtained bins is tested using silhouette index, rand Index, recall, precision, and accuracy. Performance of MetaConClust is compared with recent methods and tools using benchmarked low complexity simulated and real metagenomic datasets and is found better for unsupervised and comparable for hybrid methods.

***Conclusion*:** This is suggestive of the proposition that the consensus-based clustering approach is a promising method for automatically finding the number of bins for metagenomics data.

Keywords: Binning, consensus clustering, coverage, PAM, unsupervised clustering, metagenomics.

1. INTRODUCTION

Metagenomics is an emerging alternative way of analysing the microbial community in complex environmental samples [1]. Knowledge about the genomic constitution is essential to understanding the microbes in the best possible way. Although culturing an individual microorganism is a challenging task, advancements in next-generation sequencing technologies have enabled *in silico* identification of unidentified microbes through metagenomics studies. Metagenome is a huge mixture of genomic reads from different microorganisms. So, it is very challenging to separate individual genomes from the metagenome.

Binning is the process of classifying metagenomic sequences into groups that might be the true representative of an individual genome or genomes from taxonomically related microorganisms [2]. Binning can be majorly performed using taxonomically dependent and independent methods. In a taxonomy dependent method, reference genome is required for the classification of metagenome data [3]. Some of the important taxonomy dependent binning tools are MEGAN [4], MetaPhlan [5], Kraken [6], CLARK [7] and SKraken [8]. Major issues with taxonomy dependent method are unassigned reads, as reference genome information for the majority of the microorganisms is unavailable in the public domain, the comparison part of the algorithm is computationally expensive and gives accurate results only with long reads [9]. Whereas taxonomy independent methods utilize sequence composition information, machine learning, and statistical techniques for binning without any reference genomes. This method is further classified as a composition based [10], abundance-based [11], and hybrid [12].

Various machine learning approaches have been utilized by taxonomically independent approaches in the past

*Address correspondence to this author at the Division of Agriculture Bioinformatics, ICAR-IASRI, New Delhi- 110012, India;
E-mail: anu.sharma@icar.gov.in

 # These authors contributed equally to this work.

1875-5488/22 $65.00+.00

decade. LikelyBin [13] was based on the Markov Chain Monte Carlo process, SCIMM and PHYSCIMM [14] on k-means clustering technique, and CompostBin [12] on k-NN based approach for binning. Dimensionality reduction is an important task for efficient model development in the presence of a large number of composition-based features. In this regard, many techniques like Principal Component Analysis (PCA), weighted PCA [12] and correspondence analysis were found to be very useful. Further, non-linear dimensionality reduction unsupervised machine learning technique like t-distributed stochastic neighbor embedding (t-SNE) [15, 16] has been used for visualization of high-dimensional metagenomics data in a low-dimensional space of two dimensions [17].

Currently, the major focus is on the development of tools based on a hybrid method where the benefits of both approaches are exploited. CONCOCT [18] was developed based on contig coverage and composition. PCA was used for preprocessing, followed by contig clustering using Gaussian Mixture Model. MetaCluster 5.0 [19] was useful for identifying both low and high abundance species in the presence of a large amount of noise due to many extremely low-abundance species. GroopM [20] was designed to cluster multiple related metagenomic samples and was advantageous for its visualization and interactive pipeline. MetaBat [21] used k-medoids clustering, and MaxBin used the distribution of distances within and between genomes for binning. Another binning tool, COCACOLA [22], investigated two types of additional knowledge: the co-alignment to reference genomes and the linkage of contigs provided by paired-end reads. In recent years, MetaCon [9] was introduced, which uses a probabilistic k-mers as the features and k-medoid for binning. CoMet was another promising method for metagenomic binning, which used the DBSCAN clustering technique for initial binning, followed by Dirichlet Process Gaussian Mixture Models in the final stage using tetra-nucleotide frequencies. This method was found to perform not so well in the case of highly sparse coverage data [23].

One of major challenges in applying clustering techniques is to find the optimal number of clusters initially. The major contribution of MetaConClust is to provide a solution to this by automatically finding the optimal number of clusters using a resampling based consensus clustering approach. This approach significantly improves the accuracy of the binning of metagenomic data.

2. MATERIALS AND METHODS

2.1. Materials

The data used for this study was downloaded from the MyCC section of the SOURCEFORGE website. (https://sourceforge.net/projects/sb2nhri/files/MyCC/Data/). Two metagenomics datasets have been taken for study *viz.* 10s [24] and Sharon [25]. 10s is the simulated dataset of ten already known species, whereas Sharon is a real dataset containing 32 unknown species.

2.2. Methods

Let G be a metagenome containing genomes $g_1, g_2, g_3 \ldots g_n$ of n species in a metagenomics sample. Then

$$G = \cup C_{ij}$$

Where C_{ij} is the set of contigs, $i = 1\ to\ n$ and $j = 1\ to\ m_k\ \ k = 1\ to\ n$)

Metagenomics binning deals with finding n distinct clusters of metagenomics contigs/reads such that $\cap\ C_{ij} = \emptyset$ in ideal condition. But it is a difficult proposition to achieve in practical scenarios. It has been observed and demonstrated in previous studies [9, 26] that coverage is directly correlated to the relative abundance of the organisms in a microbial sample and is able to discriminate closely related organisms. This paper describes a novel unsupervised method for binning of metagenomics data into *c* clusters. Two major contributions discussed in this paper are:

i. Coverage based partitioning of metagenomic data into groups
ii. Automatic discovery of the number of clusters (*c*) through the application of a robust and efficient consensus clustering method.

The workflow of the binning algorithm, is depicted in Fig. (**1**).

2.2.1. Phase 1: Partitioning of Metagenomics Data on Coverage Information

In phase one of MetaConClust, the original dataset was partitioned into groups with low, medium, high, and very high coverage information so that the number of contigs in each cluster was more than fifty.

2.2.2. Phase 2: Unsupervised Binning of Metagenomics Contigs

In phase two of this algorithm, four steps were performed, namely: (1) Compositional feature extraction, (2) Finding the optimum number of *c* by using consensus clustering for each group, (3) applying k-medoids clustering algorithm to each group and (4) merging of bins after step three to obtain the final bins.

2.3. Compositional Feature Extraction

GC content and Tetranucleotide Frequency (TNF) have been found useful in delineating genomes and were used as genomic signatures in this algorithm. The abundance level of TNF derived from environmental shotgun sequences also shows similarities within the same genome [12, 27]. GC content is calculated using a Perl program [10].

$$GC = \frac{\text{Guanine (G)} + \text{Cytosine (C)}}{\text{Total base pairs in a sequence}}$$

GC count values were normalized using log transformation, and weighted TNF were calculated by normalising each tetramer frequency by the total tetramer frequency [26].

2.4. Finding the Number of Clusters Automatically

One of the major hurdles in metagenomics data clustering is to find the optimum number of bins. In this algorithm, a more efficient and robust method was used to obtain the optimum bins, namely, the consensus method. The number of clusters (c) for each group formed in phase one, was calculated using the consensus method. As per our knowledge, it was not previously used to identify the optimum number of clusters.

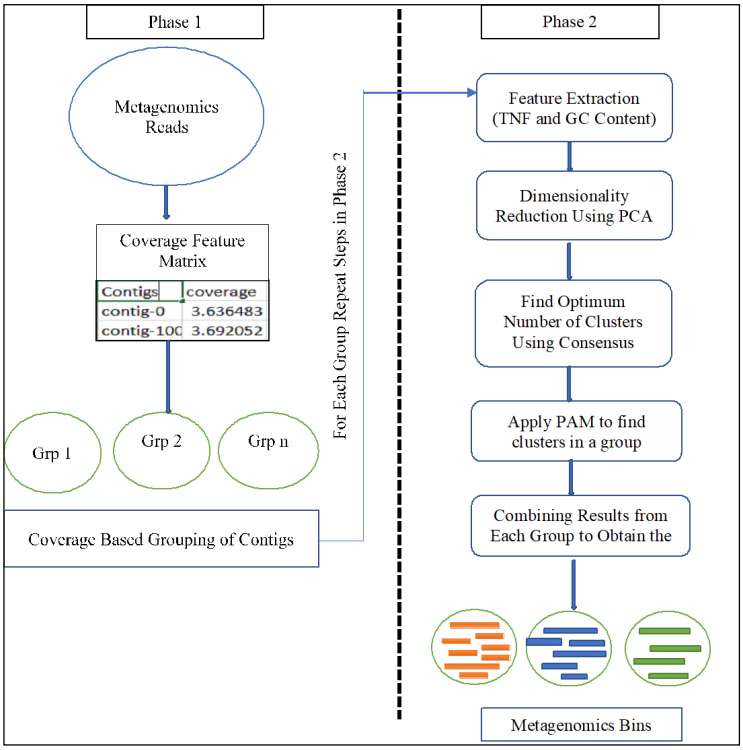

Fig. (1). Workflow of MetaConClust. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The Consensus clustering method involves subsampling from a set of items, such as metagenomics reads [28]. Consensus clustering determines the number and membership of likely clusters quantitatively in a dataset efficiently. The Consensus clustering method involves subsampling from a set of items and then performing clustering of specified cluster counts (k). Then, pairwise consensus values, the proportion that two items occupied the same cluster out of the number of times they occurred in the same subsample, are calculated and stored in a symmetrical consensus matrix for each k.

Algorithm: Consensus clustering algorithm for finding the optimal number of c

Input: a set of items D = $\{e_1, e_2, e_3 \ldots e_N\}$, a clustering algorithm like k-means, PAM, DBSCAN *etc*., a resampling scheme Resample, number of resampling iterations H, set of cluster numbers to try, K = $\{k_1, k_2, k_3 \ldots k_{max}\}$

Output:

1. for k ∈ K, do
2. M←∅ {set of connectivity matrices, initially empty}
3. for h= 1, 2…, H do
4. D(h) ←Resample(D) {generate perturbed version of D}
5. M(h) ←Cluster(D(h),K) {cluster D(h) into K clusters}
6. M←M∪M(h)
7. end {for h}
8. M(K)← compute consensus matrix from M ={M(1),..., M(H)}
9. end {for K}
10. k ← best k∈ K based on consensus distribution of M(K)'s
11. Return {M(K) :k∈K}

The R package, ConsensusClusterPlus [26], was used for performing consensus clustering. Delta area plot generated from this package was used to determine the relative increase in consensus and determine *k* at which there was no appreciable increase in area under plot. This method is very useful in estimating the prior value for *k*.

2.5. Binning of Metagenome Data using PAM

PAM or k-medoids algorithm was used for each group in the first phase using TNF and log(GC) to obtain the final bins. PAM is found to be useful in handling outliers during clustering. PAM clustering is performed using the following steps:

i. The first cluster center is picked at random between all data points
ii. Other cluster centers are picked as far as possible from previous clusters centers
iii. Associate each data point to the closest cluster center by computing median
iv. Re-compute cluster centers based on new clusters
v. Iterate until the clusters remain unchanged.

The results obtained after applying this algorithm on synthetic and real data are explained in the results and discussion section.

3. EVALUATION MEASURES

The developed approach was evaluated using recall, specificity, accuracy and silhouette index as evaluation measures. The recall is the proportion of positively labelled instances that were predicted as positive.

$$Recall = \frac{TP}{TP + FN}$$

Specificity is the proportion of negatively labelled instances that were predicted as negative.

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy is the percentage of predictions that are correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. The silhouette coefficient or silhouette Index is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

$$\text{Silhouette Index} = \frac{b - a}{\max(a, b)}$$

where

a= average intracluster distance (average distance between each point within a cluster).

b= average intercluster distance (the average distance between all clusters). Silhouette Index was used to assess the cohesiveness in clusters.

Rand Index is a measure of similarity between two data clusters. Its value ranges from 0 to 1

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

Given a set of n elements S= $\{o_1, o_2, \ldots o_n\}$, and two partitions of S to compare X=$\{X_1, X_2, \ldots, X_n\}$, a partition of S into R subsets, and Y=$\{Y_1, Y_2, \ldots, Y_n\}$, a partition of S into s subsets, define the following:

a = the number of pairs of an element of S that are in the same subset of X and in the same subset of Y.

b = the number of pairs of an element of S that are in the different subset of X and in the different subset of Y.

c = the number of pairs of an element of S that are in the same subset of X and in the different subset of Y.

d = the number of pairs of an element of S that are in the different subset of X and in the same subset of Y.

4. RESULTS

4.1. Pre-processing of Metagenome Data

The contigs having length more than or equal to 1000bp in size were taken into consideration as an exhaustive literature study has shown that long contigs contains more information about the genome in comparison to short contigs [9]. Metagenomic data were partitioned into groups based on coverage values as low coverage (1 to 60), high coverage (60-180) and very high coverage (greater than 180). The distribution of contigs in formed groups for 10s and Sharon dataset is given in Table **1** and Table **2**, respectively.

Groups having more than fifty contigs were taken into consideration. Considering this criterion, the 10s dataset is partitioned into one group (coverage 1-60) and the Sharon dataset into two groups (coverage<1 and coverage (1-60)).

4.2. Dimensionality Reduction Using PCA

Another major issue in the binning of metagenomics data is the high dimensional space, which leads to enhanced resource requirements in terms of computing power. In our case, the feature matrix has 256 features corresponding to tetranucleotide frequencies and one for GC contents. PCA was applied to the feature matrix for dimensionality reduction [12]. A Scree plot was used for selecting the number of dimensions representing maximum variation for both datasets, as shown in Figs. (**2** and **3**).

For both datasets, the first six dimensions were found to be explaining the maximum variation in data and were selected as the final examination.

Table 1. Distribution of contigs in groups' for10s dataset.

Domain	Species Name	Number of Contigs		
		Coverage <1	Coverage from 1 to 60	Coverage >60
Bacteria	*Neisseria meningitidis*	0	383	9
Bacteria	*Rhodopseudomonas palustris*	0	444	3
Bacteria	*Bacillus clausii*	0	106	0
Bacteria	*Thiobacillus enitrificans*	0	110	0
Bacteria	*Escherichia coli*	0	320	4
Bacteria	*Lawsonia intracellularis*	0	48	1
Bacteria	*Listeria welshimeri*	0	84	1
Archeae	*Methanococcus maripaludis*	0	69	3
Bacteria	*Staphylococcus aureus*	0	140	12
Bacteria	*Crocosphaera subtropica*	0	429	6

Table 2. Distribution of contigs in groups' for Sharon dataset.

Species Label	Number of Contigs		
	Coverage<1	Coverage from 1 to 60	Coverage >60
Species 1	102	6	0
Species 2	0	14	0
Species 3	2	73	0
Species 4	0	1	0
Species 5	7	52	0
Species 6	0	14	0
Species 7	0	27	0
Species 8	0	23	0
Species 9	0	1	0
Species 10	0	23	1
Species 11	0	1	0
Species 12	2	6	0
Species 13	18	41	0
Species 14	0	21	1
Species 15	6	13	0
Species 16	0	0	9
Species 17	0	0	2
Species 18	3	0	0
Species 19	11	2	0
Species 20	0	1	0
Species 21	0	1	0
Species 22	0	1	0
Species 23	5	5	0
Species 24	0	7	1
Species 25	0	2	1
Species 26	251	92	0
Species 27	391	0	0
Species 28	357	1	0
Species 29	211	0	0
Species 30	258	0	0
Species 31	253	0	0
Species 32	8	0	0

4.3. Finding Optimum Number of Bins (c)

The consensusClusterPlus function of R package was used for finding the optimum number of clusters for both datasets. This function was called with parameter values as k-means method for clustering with k as 10, the number of subsamples as 100, distance measure as Euclidean and default values for other parameters. Fig. (**4**) shows the delta plot for the 10s dataset whereas Figs. (**5** and **6**) show the delta plot obtained after applying consensus clustering on two groups for the Sharon dataset.

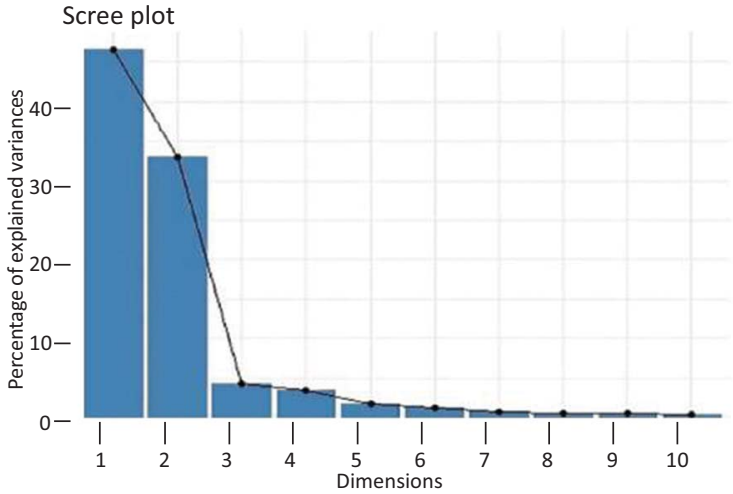

Fig. (2). Scree plot after PCA for 10s dataset. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

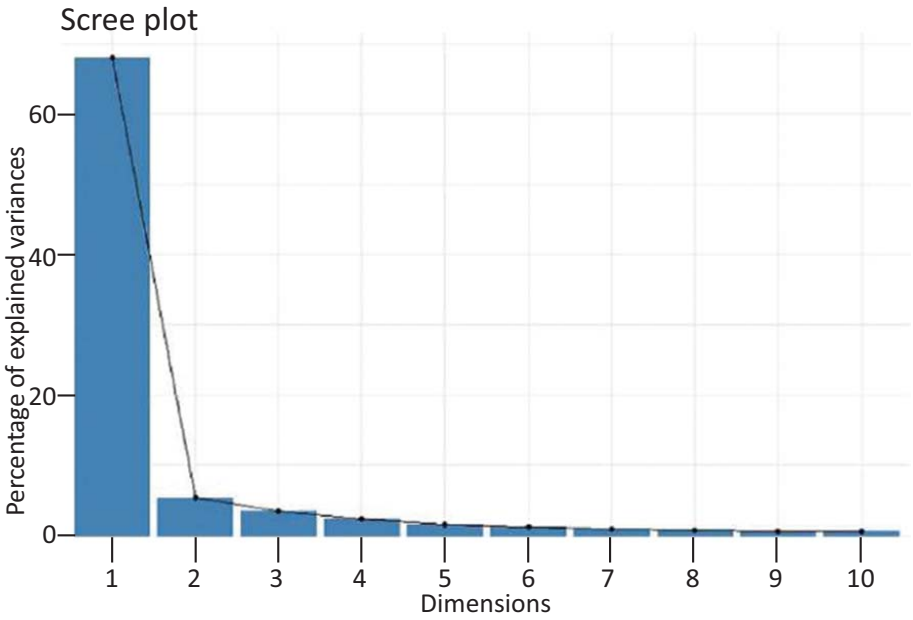

Fig. (3). Scree plot after PCA for Sharon dataset. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

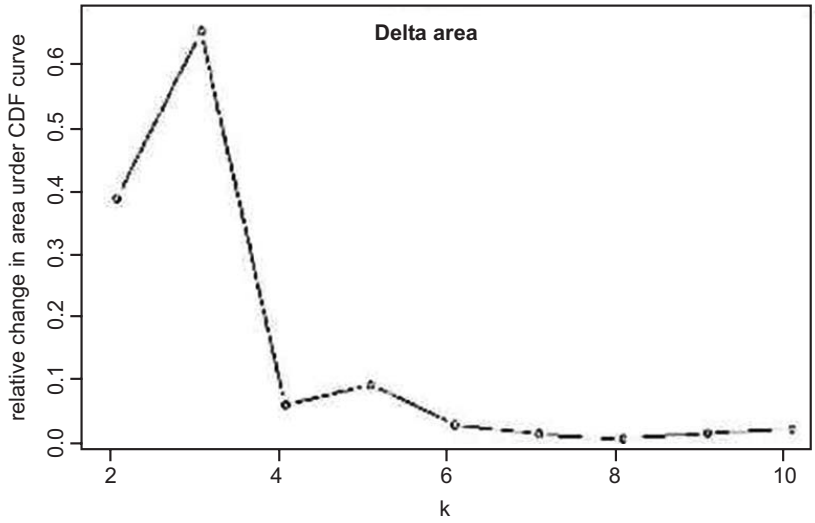

Fig. (4). Delta plot used to find the optimum value of k for the 10s dataset for group 1.

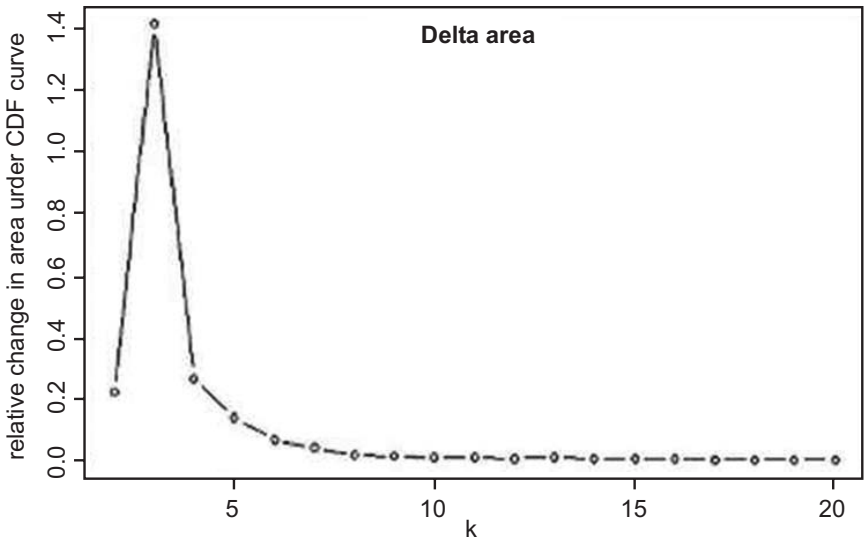

Fig. (5). Delta plot used to find the optimum value of k for the Sharon dataset for group 1.

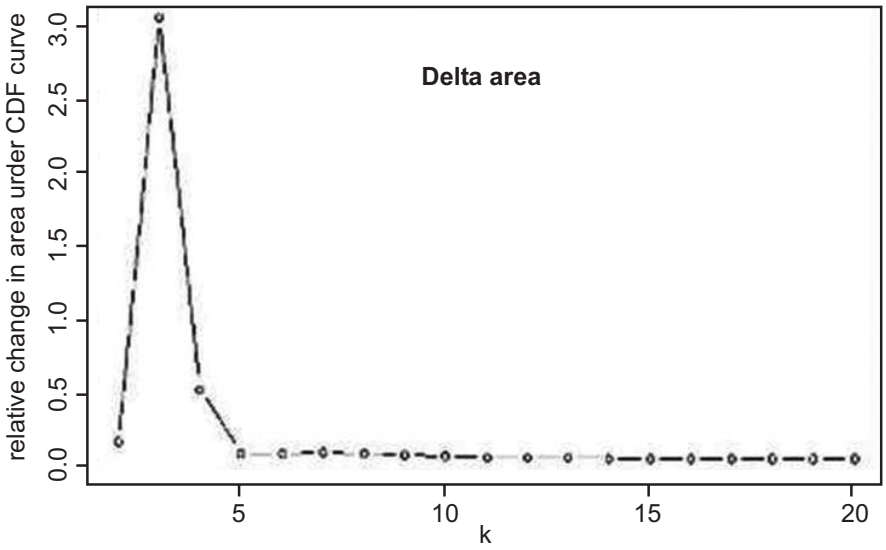

Fig. (6). Delta plot used to find the optimum value of k for the Sharon dataset for group 2.

From Delta plot, the optimum number of k for 10s dataset was found as 8 and in the case of Sharon dataset the value of k for group 1 was 6 and 5 for group 2.

4.4. Clustering and Merging

PAM clustering was applied to each group using the R-package *cluster* available at https://cran.r-project.org/web/packages/cluster/index.html. The clusters obtained from each group are combined for the Sharon dataset to get the final clusters.

4.5. Comparison with Existing Binning Tools

The developed method, MetaConClust, was compared with unsupervised Coverage and composition based binning of Metagenomes (CoMet) tool for a single metagenomics sample and semi-supervised tools MetaBat2 and Maxbin2.0. For the evaluation of clusters, silhouette index, accuracy, specificity, precision and recall were used. The results are provided in Table **3** (for 10s dataset), Table **4** (for Sharon dataset), and Table **5**.

Table 3. Evaluation of MetaConClust for 10s dataset.

-	Rand Index	Recall	Accuracy	Specificity
MetaConClust	**0.935**	**0.898**	**0.828**	**0.975**
CoMet	0.779	0.302	0.25	0.916
MetaBat2	0.986	0.946	0.966	0.996
MaxBin2.0	0.975	0.951	0.920	0.993

MetaConClust predicted 8 clusters out of the expected 10 for the 10s dataset and 11 out of 32 in the Sharon dataset, whereas CoMet produced 16 and 15 clusters. Although, CoMet can predict a greater number of bins in the Sharon dataset MetaConClust performed well with respect to evaluation measures such as silhouette index, accuracy, specificity, precision, and recall. MetaConClust performed over CoMet in most aspects (Tables **3**-**5**). Further, MetaBat2 and MaxBin2.0 predicted 10 and 9 bins, respectively, for 10s

dataset as compared to 8 by MetaConClust, but MetaBat2 uses pre-trained probabilistic models during binning, and MaxBin2.0 makes use of single-copy marker genes. For Sharon's data set, the number of bins predicted by MetaBat2.0, MaxBin2 and MetaConClust are 7, 4 and 11.

Table 4. Evaluation of MetaConClust for Sharon dataset.

-	Rand Index	Recall	Accuracy	Specificity
MetaConClust	**0.885**	**0.761**	**0.638**	**0.963**
CoMet	0.785	0.13	0.153	0.9383
MetaBat2	0.861	0.631	0.812	0.987
MaxBin2.0	0.559	0.936	0.3801	0.979

Table 5. Evaluation using silhouette index for unsupervised methods.

-	Silhouette Index	
	10s Dataset	**Sharon**
MetaConClust	**0.49**	**0.265**
CoMet	-0.24	-0.07

4.6. Phylogeny Based Evaluation on 10s Dataset

To further check the accuracy of obtained bins, another approach based on phylogeny was used to find the similarity among species clustered together in the same bin. To investigate this, phylogenetic analysis was performed using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) based on the assumption of the molecular clock. This analysis was performed using DAMBE software [23]. Phylogram obtained for the 10s dataset is shown in Fig. **7**.

It was found that the majority of the contigs clustered in a single bin belong to the same clade in the phylogenetic tree. The clade-wise distribution of the contigs in the obtained clusters (10s dataset) is given in Table **6**.

5. DISCUSSION

The mystery of genomes characteristics of the microbial communities is yet to unveil. Microbes are grown together *in vivo,* and it is very difficult to isolate themselves individually as in a microbial community, several strains of a species exists. An exhaustive literature review reveals that it was a challenging task to do so. In this research, an attempt has been made to develop a suitable method for binning metagenomics data with diversified contig coverage.

In this proposed algorithm, the dataset is divided into groups based on their coverage. With the analysis of data, it can be seen that the contigs from the same species have more or less equal coverage. Another challenging task is to find the optimum number of bins (to predict the actual number of organisms present in the data). For this Consensus clustering method is used where the optimum number of clusters is decided based on multiple resampled units of the given data that tend to give a more accurate number of clusters. For binning, PAM or k-medoid clustering technique is used. PAM is well known for its capability to handle outlier data. As metagenome data is highly diversified, PAM outperformed other clustering techniques. The proposed algorithm is also outperforming the existing unsupervised binning technique, CoMet. In the first phase of CoMet, DBSCAN clustering is used, which is more suitable for dense dataset. However, in the case of parse data, DBSCAN gives clusters with mixed instances. The datasets used in this study represent different microbial communities and

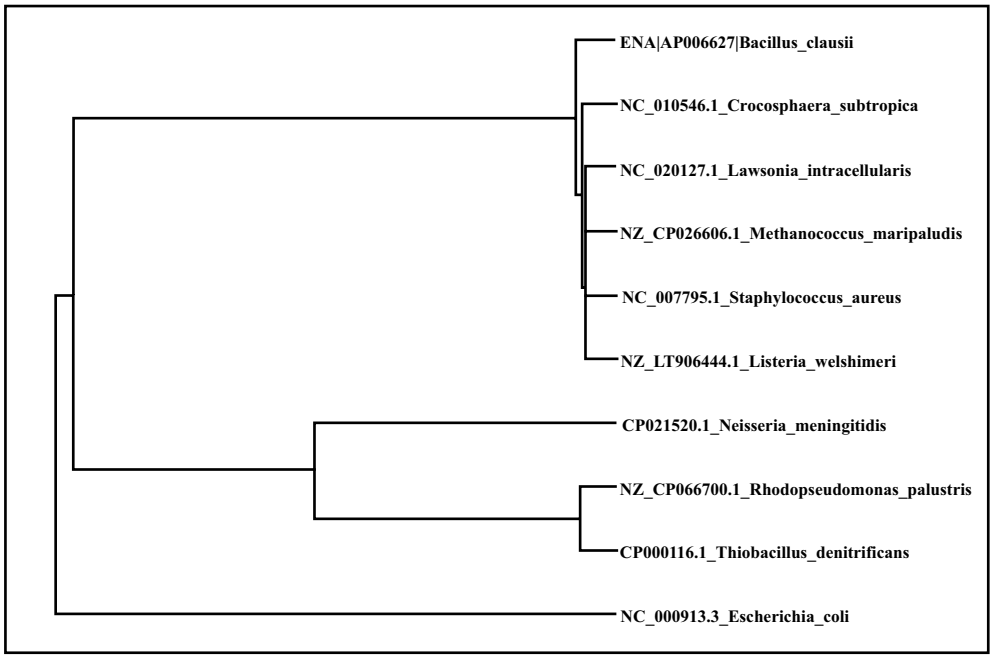

Fig. (7). Phylogram obtained using DAMBE on 10s dataset.

experimental setups. The performance of MetaConClust on both datasets was better than CoMet. However, still, there was a noticeable number of contigs present in a cluster from different species. To investigate this issue, a phylogenetic analysis was performed within a cluster. It has been found that the contigs belonging to a cluster are mostly belonging to the same phylogenetic group. This suggested to us that although contigs in one bin were from different species but are related as they belong to the same clade. This further strengthens our claim on the proposed method.

Table 6. Clade wise contig distribution in clusters.

Cluster	Clade 1	Clade 2	Clade 3
1	3	537	4
2	129	2	19
3	3	26	268
4	100	55	24
5	5	277	4
6	71	37	0
7	135	0	0
8	430	3	1

The performance of MetaConClust was also compared with MetaBat 2.0 and MaxBin 2. Both are semi-supervised and perform better than the proposed algorithm, which detected eight bins for 10s data. But for the Sharon dataset, MetaConClust performs better in the case of predicting bins than MetaBat 2.0 and MaxBin 2.

CONCLUSION

The MetaConClust is a unsupervised binning method based on the compositional genomic features for clustering highly diversified coverage metagenomic data. However, some improvements can still be made in areas where the abundance of a species is very low. It is unable to cluster organisms having low contig contribution in the dataset.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available at the URL: sb2nhri - Browse /MyCC/Data at SourceForge.net [24, 25]. Further, all the intermediate results are available with the authors and will be made available on request.

FUNDING

This work is done at ICAR-IASRI, New Delhi, under a research project.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The authors are highly thankful to the reviewers for their valuable comments and suggestions that significantly helped improve the paper and study. We are grateful for the reviewers' time and hard work that has gone into these comments. Further, we are very thankful to ICAR-IASRI for allowing us to work in this area. Thanks are due to all those who helped directly or indirectly in the execution of this work.

REFERENCES

[1] Handelsman, J. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.,* **2004**, *68*(4), 669-685. http://dx.doi.org/10.1128/MMBR.68.4.669-685.2004 PMID: 15590779

[2] Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; Wilkening, J.; Edwards, R.A. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics,* **2008**, *9*(1), 386-393. http://dx.doi.org/10.1186/1471-2105-9-386 PMID: 18803844

[3] Sedlar, K.; Kupkova, K.; Provaznik, I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.,* **2016**, *15*, 48-55. http://dx.doi.org/10.1016/j.csbj.2016.11.005 PMID: 27980708

[4] Huson, D.H.; Auch, A.F.; Qi, J.; Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res.,* **2007**, *17*(3), 377-386. http://dx.doi.org/10.1101/gr.5969107 PMID: 17255551

[5] Segata, N.; Waldron, L.; Ballarini, A.; Narasimhan, V.; Jousson, O.; Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods,* **2012**, *9*(8), 811-814. http://dx.doi.org/10.1038/nmeth.2066 PMID: 22688413

[6] Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.,* **2014**, *15*(3), R46. http://dx.doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807

[7] Ounit, R.; Wanamaker, S.; Close, T.J.; Lonardi, S. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics,* **2015**, *16*(1), 236. http://dx.doi.org/10.1186/s12864-015-1419-2 PMID: 25879410

[8] Qian, J.; Marchiori, D.; Comin, M. Fast and sensitive classification of short metagenomic reads with skraken. In: *Biomedical Engineering Systems and Technologies*; Springer: **2017**, pp. 212-226.

[9] Qian, J.; Comin, M. MetaCon: Unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage. *BMC Bioinformatics,* **2019**, *20*(Suppl. 9), 367. http://dx.doi.org/10.1186/s12859-019-2904-4 PMID: 31757198

[10] Teeling, H.; Waldmann, J.; Lombardot, T.; Bauer, M.; Glöckner, F.O. TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics,* **2004**, *5*(1), 163-169. http://dx.doi.org/10.1186/1471-2105-5-163 PMID: 15507136

[11] Wu, Y.W.; Ye, Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J. Comput. Biol.,* **2011**, *18*(3), 523-534.

http://dx.doi.org/10.1089/cmb.2010.0245 PMID: 21385052

[12] Chatterji, S.; Yamazaki, I.; Bai, Z.; Eisen, J.A. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. *arXiv*, **2008**, *2008*, 0708.3098. http://dx.doi.org/10.1007/978-3-540-78839-3_3

[13] Kislyuk, A.; Bhatnagar, S.; Dushoff, J.; Weitz, J.S. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics,* **2009**, *10*(1), 316-331. http://dx.doi.org/10.1186/1471-2105-10-316 PMID: 19799776

[14] Kelley, D.R.; Salzberg, S.L. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics,* **2010**, *11*(1), 544-555. http://dx.doi.org/10.1186/1471-2105-11-544 PMID: 21044341

[15] Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S.S.; Lian, C.; Kwon, H.; Wong, B.M. A machine learning approach for predicting defluorination of per-and polyfluoroalkyl substances (PFAS) for their efficient treatment and removal. *Environ. Sci. Technol. Lett.,* **2019**, *6*(10), 624-629. http://dx.doi.org/10.1021/acs.estlett.9b00476

[16] Perez, H.; Tah, J.H. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. *Mathematics,* **2020**, *8*(5), 662. http://dx.doi.org/10.3390/math8050662

[17] Lin, H.H.; Liao, Y.C. Accurate binning of metagenomic contigs *via* automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.,* **2016**, *6*(1), 24175. http://dx.doi.org/10.1038/srep24175 PMID: 27067514

[18] Alneberg, J.; Bjarnason, B.S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U.Z.; Quince, C. CONCOCT: Clustering contigs on coverage and composition. *Genomics,* **2013**, *1312*, 1-28.

[19] Wang, Y.; Leung, H.C.; Yiu, S.M.; Chin, F.Y. MetaCluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics,* **2012**, *28*(18), i356-i362. http://dx.doi.org/10.1093/bioinformatics/bts397 PMID: 22962452

[20] Imelfort, M.; Parks, D.; Woodcroft, B.J.; Dennis, P.; Hugenholtz, P.; Tyson, G.W.; Groop, M. GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ,* **2014**, *2*, e603. http://dx.doi.org/10.7717/peerj.603 PMID: 25289188

[21] Kang, D.D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ,* **2015**, *3*, e1165. http://dx.doi.org/10.7717/peerj.1165 PMID: 26336640

[22] Lu, Y.Y.; Chen, T.; Fuhrman, J.A.; Sun, F. COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics,* **2017**, *33*(6), 791-798. PMID: 27256312

[23] Xia, X.; Xie, Z. DAMBE: Software package for data analysis in molecular biology and evolution. *J. Hered.,* **2001**, *92*(4), 371-373. http://dx.doi.org/10.1093/jhered/92.4.371 PMID: 11535656

[24] Mende, D.R.; Waller, A.S.; Sunagawa, S.; Järvelin, A.I.; Chan, M.M.; Arumugam, M.; Raes, J.; Bork, P. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One,* **2012**, *7*(2), e31386. http://dx.doi.org/10.1371/journal.pone.0031386 PMID: 22384016

[25] Sharon, I.; Morowitz, M.J.; Thomas, B.C.; Costello, E.K.; Relman, D.A.; Banfield, J.F. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.,* **2013**, *23*(1), 111-120. http://dx.doi.org/10.1101/gr.142315.112 PMID: 22936250

[26] Herath, D.; Tang, S.L.; Tandon, K.; Ackland, D.; Halgamuge, S.K. CoMet: A workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics,* **2017**, *18*(Suppl. 16), 571. http://dx.doi.org/10.1186/s12859-017-1967-3 PMID: 29297295

[27] Gelfand, M.S.; Koonin, E.V. Avoidance of palindromic words in bacterial and archaeal genomes: A close connection with restriction enzymes. *Nucleic Acids Res.,* **1997**, *25*(12), 2430-2439. http://dx.doi.org/10.1093/nar/25.12.2430 PMID: 9171096

[28] Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.,* **2003**, *52*(1), 91-118. http://dx.doi.org/10.1023/A:1023949509487