Authors:

Yun Liu, College of Communication Engineering, Jilin University, Changchun City, Jilin Province, China.

Tao Hou, College of Communication Engineering, Jilin University, Changchun City, Jilin Province, China.

Fu Liu, College of Communication Engineering, Jilin University, Changchun City, Jilin Province, China.

Corresponding author:

Name: Fu Liu

Address: No.5988 Renmin Street, Changchun City, Jilin Province, China

Phone number: +86 13610708679

Email: liufu@jlu.edu.cn

A new unsupervised binning method for metagenomic dataset with automated estimation of number of species

Abstract: One necessary step of metagenome analysis is to assign sequences to classes according to their taxonomic origins. Unsupervised binning method is one of the two binning categories. However existing unsupervised binning methods yield to estimate the species number automatically and accurately. In this paper, a new unsupervised binning method based on an improved fuzzy c-means method (iFCM) is presented for metagenomic dataset. First, a range of the number of bins is obtained by the relationship among sequencing depth, number of reads and average read length. Secondly, iFCM algorithm is implemented several times with the initial number of bins in this range. Finally, the number of bins is determined by a clustering validity, modified partition coefficient. Experimental results show that this method is an effective unsupervised binning method for metagenomic dataset and could estimate the species number more accurately than MetaCluster3.0 and AbundanceBin.

Key words: metagenomics, unsupervised binning, improved fuzzy c-means, partition coefficient

1. Intruduction

Metagenome research uses next-generation sequencing technologies to sequence the entire community of microbial species, including culturable and unculturable species (Droge & McHardy 2012; Mande et al. 2012). The exiting metagenomic projects, such as Acid Mine Drainage Biofilm (AMD) project (Galperin 2004), Human Gut Microbiome (HGM) project (Qin et al. 2010) and gut microbiome in obese and lean twins (Turnbaugh et al. 2009), have

made a deep insight to the microbial communities. Metagenomic dataset has the characteristics of large volume, short read length, large number of species and uneven abundance ratio (Wang et al. 2012a), all making binning a difficult task.

Existing binning methods for metagenomic dataset fall into two categories, namely supervised and unsupervised binning methods (Liao et al. 2013; Mande et al. 2012). Supervised binning methods are nearly all based on reference sequences or pre-computed models (Brady & Salzberg 2009; Huson et al. 2007; MacDonald et al. 2012). However these methods will be ineffective when the reference information is lack, and most bacteria (up to 99%) in environmental samples are unknown (Eisen 2007).

To resolve this problem, binning methods with unsupervised techniques have been developed for the metagenomic dataset containing unknown species in recent years, such as AbundanceBin (Wu & Ye 2011), MetaCluster3.0 (Leung et al. 2011), MetaCluster4.0 (Wang et al. 2012a), and MetaCluster5.0 (Wang et al. 2012b). For unsupervised binning methods, automatic and accurate estimation of the number of species in a metagenomic dataset is an unevadable issue. However, all the above methods fail to do this work very well. AbundanceBin merges two bins if they have identical abundance values or genome sizes (Wu & Ye 2011). So it doesn't work well when metagenomic dataset contains DNA reads sampled from different species but with identical abundance ratio. The series of MetaClusters merge two bins if their distance is less than a threshold after c-means clustering with a relative large number of bins. However, the clustering performance of MetaClusters is extremely sensitive to the selection of threshold (See in section 3).

Fuzzy c-means method (FCM) is one of the most widely used clustering method in area of image segmentation (Cai et al. 2007; Noordam et al. 2002), data analysis (Ball & Hall 1967) and so on. However, FCM tends to cluster data into groups with similar size (J.C. Noordam et al. 2002; Lin et al. 2014), making it not suitable for unbalanced dataset, especially for metagenomic dataset with uneven abundance ratio. An improved version of FCM (iFCM) with cluster size constrains presented in (J.C. Noordam et al. 2002) have achieved successful application in unbalanced dataset. So iFCM is applied as the unsupervised binning method for metagenomic dataset.

However, iFCM doesn't have the ability to determine the number of clusters. A common method is to implement iFCM several times with a series of number of clusters, and then select the best clustering structure according to a cluster validity (Tibshirani & Walther 2005). Partition coefficient (PC) was firstly presented in (Bezdek 1973a) as a cluster validity and have been developed greatly (Wu & Yang 2005), which is monotonous with the cluster number however. Here, a new modified partition coefficient (mPC) is presented to avoid this monotonicity . By combining iFCM and mPC, method in this paper could handle the metagenomic dataset with uneven abundance ratio effectively and output the number of species automatically as well.

In this paper, a new unsupervised binning method, MetaBin2.0, is presented, which consists of four steps: (1) calculate the feature matrix of metagenomic dataset by k-mer frequencies; (2) determine the value range of the initial number of bins for iFCM; (3) implement iFCM algorithm to cluster metagenomic sequences several times with the initial

number of bins determined in step (2); (4) choose the most appropriate clustering result according to mPC.

2. Method

MetaBin2.0 contains four steps, which will be described in detail in this part.

2.1 Construction of the feature matrix by k-mer frequencies

Unsupervised binning methods usually utilize k-mer frequencies to construct feature matrix for metagenomic dataset (Liao et al. 2013). K-mer is a substring of DNA read with length k, so there are 4^k kinds of k-mers in a DNA read. For a DNA read with length l, there are total $l-k+1$ k-mers. Therefore, a DNA read can be represented by a feature vector $f=[f_1, f_2, ..., f_{4^k}]$ which should meet the following condition:

$$\sum_{i=1}^{4^k} f_i = l-k+1 \tag{1}$$

where f_i is the occurrence number of i th mer in a DNA read. Previous research in (Chor et al. 2009; Zhou et al. 2008) showed that $k=4$ is the best choice for metagenomic binning. So the feature vector will be 256-dimensional.

For a metagenomic dataset contains N DNA reads, the feature matrix $\boldsymbol{F}_{256\times N}$ is constructed, where f_{ij} represents the occurrence number of i th mer in j th read. Then $\boldsymbol{F}$ is normalized as below:

$$x_{ij} = \frac{f_{ij} - \min\limits_{j=1,...,N} f_{ij}}{\max\limits_{j=1,...,N} f_{ij} - \min\limits_{j=1,...,N} f_{ij}} \tag{2}$$

2.2 Determination of the value range of initial number of bins

The number of clusters is often a necessary initial condition for unsupervised binning

methods (Cai et al. 2007; de Carvalho & Tenório 2010). In this part, a value range of initial number of clusters will be determined according to the characteristics of metagenomic dataset.

Sequencing depth (d) is an evaluation index of high-throughput sequencing capacity, which is defined as the mean number of times that a nucleotide is sequenced (C.Wooley et al. 2010). It has the following relationship with number of reads, average read length and total genome length of a metagenomic dataset:

$$d = \frac{N \times l}{G} \tag{3}$$

where N is the number of reads, l is the average length of read and G is the total genome length.

In a metagenomic dataset contains c species, the total genome length G is:

$$G = \sum_{i=1}^{c} G_i \tag{4}$$

where G_i is the genome length of i the species. Suppose that g is the average genome length of c species. Then

$$G = cg \tag{5}$$

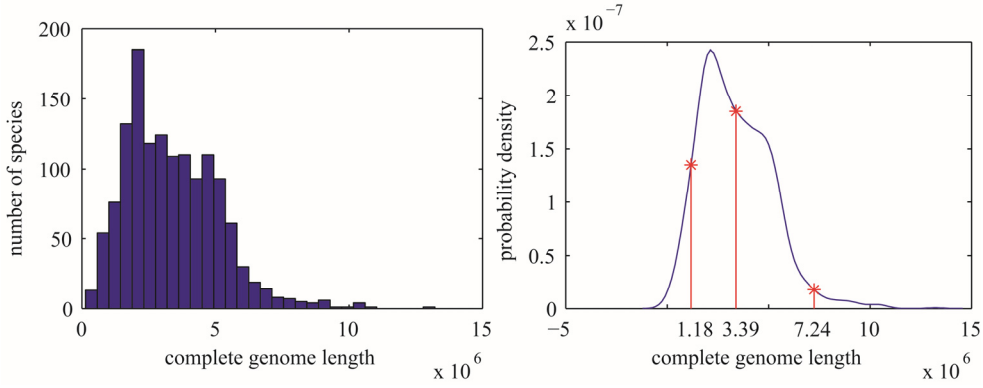
By combining formula (3) and (5), the number of species can be estimated by:

$$c = \frac{N \times l}{d \times g} \tag{6}$$

For a metagenomic dataset, N and l are easy to get. Sequencing depth can be calculated by:

$$d = \frac{sequenced\ data\ size}{original\ data\ size} \tag{7}$$

Different species has different genome size. To estimate g, we counted complete genome length of 1379 bacteria species. The statistical histogram is showed in Figure 1 (a). Figure 1 (b) is the probability density curve of complete genome size estimated by kernel density method. Here the function *ksdensity* in MATLAB is utilized to do this work.



(a) Statistical histogram (b) Probability density curve

Figure 1 The statistical histogram and probability density curve of complete genome length of 1379 bacteria species

Let $f(g)$ be the probability density function of complete genome size plotted in Figure 1 (b), $\bar{g}$ be the average complete genome length. Then two important values, $g_{\min}$ and $g_{\max}$, will be calculated through formula (8) and formula (9).

$$\int_{\bar{g}}^{g_{\max}} f(g)dg \geq 0.45 \tag{8}$$

$$\int_{g_{\min}}^{\bar{g}} f(g)dg \geq 0.45 \tag{9}$$

The results are $g_{\min} = 1.18 \times 10^6$, $g_{\max} = 7.24 \times 10^6$. So the range $[g_{\min}, g_{\max}]$ could include 90% of those 1379 bacteria.

Finally, for a metagenomic dataset, the value range of initial number of bins $[c_{\min}, c_{\max}]$ could be determined by

$$c_{\min} = \frac{N \times l}{d \times g_{\max}} \tag{10}$$

and

$$c_{\max} = \frac{N \times l}{d \times g_{\min}} \tag{11}$$

2.3 Cluster progress using iFCM

2.3.1 Brief introduction to FCM

For a metagenomic dataset with N DNA reads, suppose that $\boldsymbol{X} = \{x_{ij}\}_{256\times N}$ is the normalized feature matrix computed by formula (2). FCM partitions these N DNA reads into c clusters through an iterative minimization process of an objective function $J(\boldsymbol{U},\boldsymbol{V})$ (Bezdek 1973b), which is defined as:

$$J(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{q} d(\boldsymbol{x}_i,\boldsymbol{\theta}_j) \tag{12}$$

where $\boldsymbol{x}_i$ is the ith data point of $\boldsymbol{X}$, $\boldsymbol{\theta}_j$ is the jth cluster center, $u_{ij} \in [0,1]$ is the membership value of $\boldsymbol{x}_i$ to $\boldsymbol{\theta}_j$ with a constrain $\sum_{j=1}^{c} u_{ij} = 1$, $q \in [1,+\infty)$ is the fuzziness degree, and $d(\cdot)$ is the similarity measure.

Utilizing Lagrange Multiplier, the objective function $J(\boldsymbol{U},\boldsymbol{V})$ is minimized and the membership matrix $\boldsymbol{U}$ would be:

$$u_{rs} = \frac{1}{\left(d\left(\boldsymbol{x}_r,\boldsymbol{\theta}_s\right) / \sum_{j=1}^{c} d\left(\boldsymbol{x}_r,\boldsymbol{\theta}_j\right) \right)^{2/(q-1)}}, r = 1,2,\ldots,N, s = 1,2,\ldots,c \tag{13}$$

And the cluster center is:

$$\boldsymbol{\theta}_j = \frac{\sum_{i=1}^{N} u_{ij}^{q} \boldsymbol{x}_i}{\sum_{i=1}^{N} u_{ij}^{q}}, 1 \le j \le c \tag{14}$$

So FCM algorithm can be summarized as below:

(1) Construct membership matrix $\boldsymbol{U}$ with random decimal fraction.

(2) Compute cluster centers using formula (14).

(3) Update $\boldsymbol{U}$ using formula (13).

(4) Compute objective function J value using formula (12).

(5) Repeat step (2) to (4) until $\left|J^{(t)}-J^{(t-1)}\right|<\varepsilon$, where ε is a very small number.

2.3.2 Improved FCM (iFCM)

As an improved FCM algorithm, the clustering strategy of iFCM is to weaken the contribution of DNA read belonging to larger cluster, while maintain the contribution of DNA read belonging to small cluster (Lin et al. 2014). In each iterative process, a condition value f_i for every DNA read x_i is computed after defuzzification:

$$f_i=\frac{1-S_j}{1-\min S_j}, i=1,2,...,N, j=1,2,...,c \tag{15}$$

where $S_j=N_j/N$, N_j is the number of DNA reads in cluster j. With this definition, reads assigned to same cluster would have the identical condition value and this value is only depending on the size of their assigned cluster. By combining condition value and membership matrix, the formula (13) can be rewrote as:

$$u_{rs}=\frac{f_r}{\left(d\left(\boldsymbol{x}_r,\boldsymbol{\theta}_s\right)/\sum_{j=1}^{c}d\left(\boldsymbol{x}_r,\boldsymbol{\theta}_j\right)\right)^{2/(q-1)}}, r=1,2,\ldots,N, s=1,2,\ldots,c \tag{16}$$

So the pipeline of iFCM can be described as below:

(1) Construct membership matrix $\boldsymbol{U}$ with random decimal fraction.

(2) Compute cluster centers using formula (14).

(3) Defuzzification and compute condition value using formula (15).

(4) Update $\boldsymbol{U}$ using formula (16).

(5) Compute objective function J value using formula (12).

(6) Repeat step (2) to (5) until $\left|J^{(t)} - J^{(t-1)}\right| < \varepsilon$, where ε is a very small number.

In this paper, iFCM algorithm is implemented by $c_{\max} - c_{\min} + 1$ times with different initial number of clusters belonging to $[c_{\min}, c_{\max}]$, which is determined in Section 2.2.

2.4 Choose the best clustering result using mPC

In Section 2.3, we have obtained $c_{\max} - c_{\min} + 1$ clustering results with different number of clusters. In this section, mPC is utilized to select the best one from $c_{\max} - c_{\min} + 1$ clustering results, which is the final output of MetaBin2.0.

PC is defined as (Bezdek 1973a; Wu & Yang 2005):

$$PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^2 \tag{17}$$

A more widely used and simplified version of PC is defined as:

$$PC = \frac{1}{N}\sum_{i=1}^{N} \max_{j=1,\ldots,c} u_{ij} \tag{18}$$

However, when it is used to validate the clustering result for metagenomic dataset, we find that the value of PC is monotonic with c (see in section 3). As a result, the output number of bins will always be $c_{\min}$.

To solve this problem, a modified PC, mPC, is presented to validate the clustering result, which is defined as:

$$mPC = \frac{1}{N}\sum_{i=1}^{N}\left(\max_{j=1,\ldots,c} u_{ij} - \frac{1}{\sum_j u_{ij}}\right) \tag{19}$$

Experimental results illustrate that the performance of mPC is better than PC (see in Section

3).

3. Results

In this section, we will evaluate MetaBin2.0 on several datasets, and compare its performance with MetaCluster3.0, AbundanceBin and MetaCluster5.0.

3.1 Evaluation criteria

To evaluation the performance of MetaBin2.0, three commonly used criteria, $Precision$, $Sensitivity$ and $F\_measure$, are utilized.

Suppose that a metagenomic dataset containing N DNA reads from c species is clustered into c' groups, then a clustering result matrix $\boldsymbol{R}_{c'\times c}$ will be constructed, where $r_{ij}(1 \le i \le c', 1 \le j \le c)$ represents the number of DNA reads from species j that are partitioned into group i.

$Precision$ is defined as:

$$Precision = \frac{\sum_{i=1}^{c'} \max_{j=1,\ldots,c} r_{ij}}{N} \tag{20}$$

$Sensitivity$ is defined as:

$$Sensitivity = \frac{\sum_{j=1}^{c} \max_{i=1,\ldots,c'} r_{ij}}{N} \tag{21}$$

$F\_measure$ is defined as:

$$F\_measure = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{22}$$

$Precision$ measures the purity of each group, $Sensitivity$ represents the concentration of DNA reads from same species, and $F\_measure$ evaluates the overall performance of

clustering result.

3.2 Datasets

3.2.1 Simulated datasets

The simulated datasets in this paper are based on complete genome from National Center for Biotechnology Information (NCBI) database (www.ncbi.nlm.nih.gov) and a metagenomic sequencing simulator software MetaSim (Richter D C et al. 2008). Error rate at read start is 0.01, while at read end is 0.02. The insertion error rate and deletion error rate are both 0.2.

A. 5 species with 200bp read length

Dataset A contains 50 thousand DNA reads with 200bp read length from 5 species. The abundance ratio of Dataset A is 1:2:3:4:5. The detail information is listed in Table 1.

Table 1 Information of Dataset A

	Species name	Number of reads
1	Bacillus subtilis subsp. subtilis str. 168 chromosome	7343
2	Chlamydophila pneumoniae AR39	4248
3	Methanothermobacter thermautotrophicus str. Delta H chromosome	9080
4	Streptococcus pyogenes M1 GAS chromosome	13005
5	Thermotoga maritima MSB8 chromosome	16324

Table 2 Information of Dataset B

	Species number	Number of reads
1	Aeropyrum pernix K1	942
2	Bacillus halodurans C-125 chromosome	4958
3	Campylobacter jejuni subsp. jejuni NCTC 11168 chromosome	2917
4	Deinococcus radiodurans R1 chromosome chromosome 1	6248
5	Escherichia coli O157:H7 str. EDL933 chromosome	16239
6	Haemophilus influenzae Rd KW20 chromosome	6626
7	Lactococcus lactis subsp. lactis Il1403	10085
8	Mesorhizobium loti MAFF303099 chromosome	33449
9	Pyrococcus horikoshii OT3 chromosome	9158
10	Thermoplasma acidophilum DSM 1728 chromosome	9368

B. 10 species with 200bp read length

Dataset B contains 100 thousand DNA reads with 200bp read length from 10 species. The abundance ratio is 1:2:3:4:5:6:7:8:9:10. The detail information of Dataset B is listed in Table 2.

Table 3 Information of Dataset C

	Species name	Number of reads
1	Bacillus subtilis subsp. Subtilis str. 168 chromosome	8372
2	Chlamydophila pneumoniae AR39	4932
3	Halobacterium sp. NRC-1 chromosome	12138
4	Mycobacterium leprae TN chromosome	26046
5	Mycoplasma genitalium G37	5824
6	Staphylococcus aureus subsp. Aureus Mu50	34824
7	Thermoplasma volcanium GSS1 chromosome	22324
8	Ureaplasma parvum serovar 3 str. ATCC 700970	11928
9	Vibrio cholerae O1 biovar eltor str. N16961 chromosome II	19468
10	Xylella fastidiosa 9a5c8372	54144

Table 4 Information of real dataset

Species	Number of contigs
Ruminococcus bromii L2-63	793
Subdoligranulum variabile DSM 15176	518
Butyrivibrio crossotus DSM 2876	432
Faecalibacterium prausnitzii SL3/3	436
Clostridium sp. CAG:127	205
Prevotella copri DSM 18205	669
Alistipes shahii WAL 8301	897
Alistipes finegoldii DSM 17242	615
Odoribacter splanchnicus DSM 20712	681
Bacteroides vulgatus ATCC 8482	324
Bacteroidales bacterium ph8	1473
Subdoligranulum sp. CAG:314	510
Total	7553

C. 10 species with 100bp read length

D. To test the performance of MetaBin2.0 on database with short read length pair-end reads, Dataset C is constructed, which contains 200 thousand DNA reads with 100bp read length from 10 species. Metacluster5.0 is also implemented on this dataset as it

could only handle reads with less than 128bp read length. The species abundance ratio in Dataset C is also 1:2:3:4:5:6:7:8:9:10. The detail information of is listed in 10 species with 200bp read length

Dataset B contains 100 thousand DNA reads with 200bp read length from 10 species. The abundance ratio is 1:2:3:4:5:6:7:8:9:10. The detail information of Dataset B is listed in Table 2.

Table 3.

3.2.2 Real dataset

Qin et al. (Qin et al. 2010) had collected metagenomic samples from feces of 124 European adults. We selected 2 samples, MH0001 and MH0002, as the real metagenomic dataset. Then, BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was utilized to get the origin info of each contig. Finally, conitgs which are sampled from the most abundant 12 species were selected and 7553 contigs were obtained. The information of this dataset is listed in Table 4.

3.3 Experimental results

3.3.1 Monotonicity of PC

The PC and mPC values with different cluster number of Dataset A, B and C are picted in Figure 2. Experiments on those 3 datasets show that PC is monotonic with cluster number. As a result, PC could't be used as clustering criterion in this paper. With an ingenious modification, mPC solves the defect successfully and could achieve a maximal value in the interval of $[k_{\min}, k_{\max}]$.

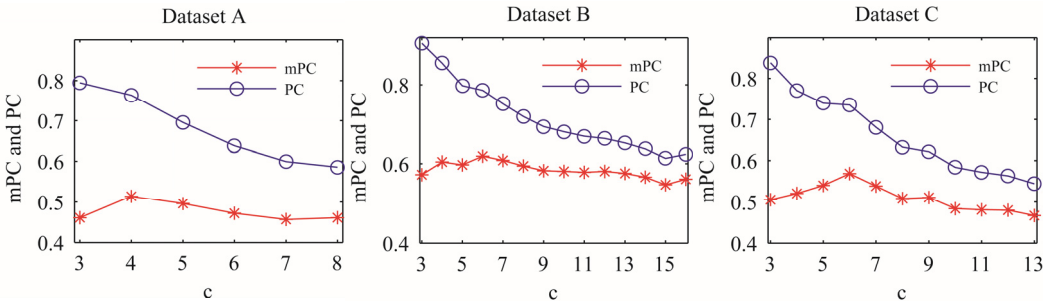


Figure 2 The curves of PC and mPC of Dataset A, B and C in this paper

3.3.2 Experimental results of simulated datasets

The experimental results of 3 simulated datasets are listed in Table 5, Table 6 and Table 7. We find that the performance of MetaCluster3.0 is sensitive to threshold α, a parameter used to determine whether merge two clusters or not (Leung et al. 2011), and setting $\alpha = 0.95$ will achieve the best result. In addition, AbundanceBin bins all of the reads (Dataset A and B) into one group, so the output number of bins of MetaBin2.0 is given to it as an input parameter.

When binning metagenomic dataset with 5 species, MetaBin2.0 has 2.0% higher precision, sensitivity and F_measure than MetaCluster3.0 and 36% higher sensitivity and 26% higher F_measure than AbundanceBin (Table 5).

Table 5 Performance on Dataset A

Methods	Species discovered	Given number of bins	Precision	Sensitivity	F_measure
MetaCluster3.0	4	-	0.7885	0.7157	0.7503
Abundancebin	1	-	-	-	-
Abuncancebin	-	4	0.7901	0.3750	0.5086
Metabin2.0	4	-	**0.8069**	**0.7361**	**0.7699**

Table 6 Performance on Dataset B

Methods	Output number of bins	Given number of bins	Precision	Sensitivity	F_measure
MetaCluster3.0	4	-	0.8940	0.4275	0.5784
Abundancebin	1	-	-	-	-
Abuncancebin	-	6	0.6013	0.3861	0.4702
Metabin2.0	6	-	**0.6100**	**0.5673**	**0.5874**

As the increase of number of species, the performances of all those three methods decrease significantly. For instance, the F_measure of MetaBin2.0 decreases from 0.7699 to 0.5874 (Table 5 and Table 6). In spite of this, MetaBin2.0 has 1% higher precision, 18% higher sensitivity and 11% higher F_measure than AbundanceBin. MetaCluster3.0 has higher precision, while lower sensitivity and F_measure than MetaBin2.0. Furthermore, MetaBin2.0 discovers 6 species, while MetaCluster3.0 only discovers 4 species (Table 6).

In dataset C, MetaBin2.0 discovers 6 species, while MetaCluster5.0 only discovers 5species (Table 7). Notably, MetaCluster5.0 has high precision while low sensitivity, illustrating that it bins most of the reads into one group according to the definition of precision (formula (20)) and sensitivity (formula (21)). Finally, MetaBin2.0 has 1% higher F_measure than MetaCluster5.0.

Table 7 Performance on Dataset C

Methods	Output number of bins	Precision	Sensitivity	F_measure
MetaCluster5.0	5	0.8656	0.2894	0.4337
MetaBin2.0	6	**0.4509**	**0.4328**	**0.4416**

Table 8 Performance on real dataset

Methods	Output number of bins	Precision	Sensitivity	F_measure
MetaCluster3.0	8	0.6399	0.5250	0.5767
MetaBin2.0	11	**0.7074**	**0.5579**	**0.6238**

Table 9 Performance of MetaBin2.0 on dataset D

Groups	Major species	Precision
Group 1	Clostridium sp. CAG:12	80%
Group 2	Subdoligranulum variabile DSM 15176	78.08%
Group 3	Ruminococcus bromii L2-63	76.72%
Group 4	Subdoligranulum sp. CAG:314	59.51%
Group 5	Prevotella copri DSM 18205	58.32%
Group 6	Faecalibacterium prausnitzii SL3/3	56.98%
Group 7	Bacteroidales bacterium ph8	53.48%

Group 8	Odoribacter splanchnicus DSM 20712	51.89%
Group 9	Butyrivibrio crossotus DSM 2876	42.54%
Group 10	Alistipes shahii WAL 8301	40.78%
Group 11	Bacteroides vulgatus ATCC 8482	39.73%

3.3.3 Experimental results of real datasets

The performance of MetaCluster3.0 and MetaBin2.0 on real metagenomic dataset is showed in Table 8. In this dataset, MetaBin2.0 discovers 11 of the 12 species, while MetaCluster3.0 only finds 8 of them. Furthermore, MetaBin2.0 has 7% higher precision, 3% hagher sensitivity and 5% higher F_measure than MetaCluster3.0.

The species information found by MetaBin2.0 and their precision values are listed in Table 9.

4. Conclusion

In this paper, a new unsupervised binning tool for metagenomic dataset, MetaBin2.0, with the ability to determine the number of species in a dataset automatically and accurately, is presented. Experimental results illustrate that MetaBin2.0 has better performance than Abundancebin and MetaCluster3.0, and could discover more species than Abundancebin, MetaCluster3.0 and MetaCluster5.0.

However, when the number of species increases, the performance of MetaBin2.0 witnesses a significant degradation. So how to solve this problem so that MetaBin2.0 could handle dataset with more species should be the next step of research.

Acknowledgement: This work has been supported by No. 51105170 of National Natural Science Foundation of China (NSFC). Thanks Mr. Xuecheng An and Shoukun Jiang for providing the programming support. We are also grateful the authors of AbundanceBin and

MetaClusters for providing the tools so that we can do comparative experiments.

Conflict of interest: We declare that there is no conflict of interest.

Download: The software of MetaBin2.0 can be freely downloaded from http://lf-lab.com/news/html/?419.html.

References

Ball GH, and Hall DJ. 1967. A clustering technique for summarizing multivariate data. *Behavioral science* 12:153-155.

Bezdek JC. 1973a. CLUSTER VALIDITY WITH FUZZY SETS. *Journal of Cybernetics* 3:58-73.

Bezdek JC. 1973b. Cluster validity with fuzzy sets. *Cybernetics and Systems: An International Journal* 3:58-73.

Brady A, and Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6:673-U668.

C.Wooley J, Godzik A, and Friedberg I. 2010. A Primer On Metagenomics. *Computational Biology* 6:1-13.

Cai W, Chen S, and Zhang D. 2007. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition* 40:825-838.

Chor B, Horn D, Goldman N, Levy Y, and Massingham T. 2009. Genomic DNA k-mer spectra: models and modalities. *Genome Biology* 10:R108.

de Carvalho FdAT, and Tenório CP. 2010. Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems* 161:2978-2999.

Droge J, and McHardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform* 13:646-655.

Eisen JA. 2007. Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. *Plos Biology* 5:384-388.

Galperin MY. 2004. Metagenomics: from acid mine to shining sea. *Environmental Microbiology* 6:543-545.

Huson DH, Auch AF, Qi J, and Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Research* 17:377-386.

J.C. Noordam, W.H.A.M. van den Broek, Buydens LMC, and Bao Z. 2002. Multivariate image segmentation with cluster size insensitive Fuzzy C-means. *Chemometrics and Intelligent Laboratory Systems* 64:65-78.

Leung HC, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, Chen J, Qin J, Li R, and Chin FY. 2011. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics* 27:1489-1495.

Liao R, Zhang R, Guan J, and Zhou S. 2013. A New Unsupervised Binning Approach for Metagenomic Sequences Based on N-grams and Automatic Feature Weighting. *TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 11:42-54.

Lin P-L, Huang P-W, Kuo CH, and Lai YH. 2014. A size-insensitive integrity-based fuzzy c-means method for data clustering. *Pattern Recognition* 47:2042-2056.

MacDonald NJ, Parks DH, and Beiko RG. 2012. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res* 40:e111.

Mande SS, Mohammed MH, and Ghosh TS. 2012. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform* 13:669-681.

Noordam JC, van den Broek W, and Buydens LMC. 2002. Multivariate image segmentation with cluster size insensitive fuzzy C-means. *Chemometrics and Intelligent Laboratory Systems* 64:65-78.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Meta HITC, Bork P, Ehrlich SD, and Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-65.

Richter D C, Ott F, and F AA. 2008. MetaSim-A Sequencing Simulator for Genomics and Metagenomics. *PloS one* 3:1-12.

Tibshirani R, and Walther G. 2005. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* 14:511-528.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, and Gordon JI. 2009. A core gut microbiome in obese and lean twins. *Nature* 457:480-484.

Wang Y, Leung HC, Yiu SM, and Chin FY. 2012a. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *J Comput Biol* 19:241-249.

Wang Y, Leung HC, Yiu SM, and Chin FY. 2012b. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* 28:i356-i362.

Wu KL, and Yang MS. 2005. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters* 26:1275-1291.

Wu YW, and Ye Y. 2011. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* 18:523-534.

Zhou F, Olman V, and Xu Y. 2008. Barcodes for genomes and applications. *BMC Bioinformatics* 9:546.