MetaAB - A Novel Abundance-Based Binning Approach for Metagenomic Sequences

Van-Vinh Le[1,3(✉)], Tran Van Lang[2], and Tran Van Hoai[1]

[1] Faculty of Computer Science and Engineering, HCMC University of Technology, Ho Chi Minh City, Vietnam
vinhlv@fit.hcmute.edu.vn
[2] Institute of Applied Mechanics and Informatics, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[3] Faculty of Information Technology, HCMC University of Technical Education, Ho Chi Minh City, Vietnam

Abstract. Metagenomics is a research discipline of microbial communities that studies directly on genetic materials obtained from environmental samples without isolating and culturing single organisms in laboratory. One of the crucial tasks in metagenomic projects is the identification and taxonomic characterization of DNA sequences in the samples. In this paper, we present an unsupervised binning of metagenomic reads, called MetaAB, which can be able to identify and classify reads into groups of genomes using the information of genome abundances. The method is based on a proposed reduced-dimension model that is theoretically proved to have less computational time. Besides, MetaAB detects the number of genome abundances in data automatically by using the Bayesian Information Criterion. Experimental results show that the proposed method achieves higher accuracy and run faster than a recent abundance-based binning approach. The software implementing the algorithm can be downloaded at http://it.hcmute.edu.vn/bioinfo/metaab/index.htm

Keywords: Metagenomics · Binning · Next-generation sequencing · Bayesian information criterion · Genome abundance

1 Introduction

Since microbes are the most diverse forms on Earth, the understanding of them can bring many benefits to human being [1]. Microbial communities have been studied for many years. However, due to experimental limitations, traditional methods only focus on single species in laboratory culture. A drawback of these methods is that 99% percent of microbes cannot be cultured in the laboratory [2]. Moreover, a clone culture cannot represent the true state of affairs in nature since a sample obtained from a microbial community may contain many species which interact with both each other and their habitats [3]. An alternative research trend

P.C. Vinh et al. (Eds.): ICTCC 2014, LNICST 144, pp. 132–141, 2015.
DOI: 10.1007/978-3-319-15392-6_13

which can overcome the limits of traditional methods is metagenomics. This discipline allows the direct study on genomes from an environmental sample without isolation and cultivation of single organisms. However, it takes many costs to obtain genomic information directly from microbial communities by traditional sequencing technologies (e.g., Sanger sequencing technology). Fortunately, new sequencing technologies (so-called next-generation technologies [4,5]), which can produce millions of reads with small costs, have make metagenomics feasible in practice.

One of the crucial step in a metagenomic project is to classify reads into groups of individual genomes or closely related organisms, which is referred to as *binning problem*. Binning methods can be roughly classified into three main categories: *homology-based*, *composition-based*, and *abundance-based* methods.

Homology-based approaches classify reads by using alignment tools (e.g., Blast, HMMER) to align DNA sequences directly to reference genomes. Among the approaches, MEGAN [6] maps reads by Blast with the nr database of NCBI (National Center for Biotechnology Information), then it assigns labels for the reads using a technique of lowest common ancestor. CARMA [7] is another homology-based method in which data is aligned with a protein database Pfam by either BLAST or HMMER3 homology searches.

Many binning approaches are known as composition-based methods, which use compositional features (e.g., oligonucleotide frequencies, GC-content) for classification. They can be further divided into two kinds of methods: *supervised* and *unsupervised* methods. Supervised methods [8,9] require reference databases which consist of known taxonomic origin sequences. The supervised methods are shown to perform well in case of full-availability of reference databases. However, the majority of microorganisms on Earth remains undiscovered [10]. This makes the methods may be not efficient in practice. To deal with the lack of reference databases, some unsupervised methods were proposed to perform the classification basing on features extracted from analyzed sequences. MetaCluster 2.0 [11], MetaCluster 3.0 [12] and MCluster [13] are recent algorithms which are based on the signature of frequency distribution of tetra-nucleotides. These approaches are shown to be efficient for long sequences ($\geq$ 800kbp), but get low accuracy for short reads (50-400bp). Furthermore, many approaches do not perform well if the abundance levels of genome in data are very different [11].

Some recent unsupervised approaches can perform on short reads by using the information of genome abundances in data. MetaCluster 5.0 [23] separates reads into three groups of different abundance levels (high, low and extremely low level) and applies further classification strategies to each group. AbundanceBin [15] and Olga *et al* [16] are two approaches for binning of reads which only reply on the feature of genome abundances. Those approaches group reads into bins that the reads in the same bin belong to genomes of similar abundance levels. Both approaches is based on an assumption that the occurrences of l-mers (with a sufficient value of l) in data follow Poisson distribution, and then an expectation maximization algorithm is used to estimate genome abundances. Another abundance-based binning approach, MarkovBin [14], models nucleotide

sequences as a fixed-order Markov chain and classifies them into groups of different genome abundances. However, this method still does not support detecting automatically the number of genome abundance levels in data.

This paper proposes a new abundance-based binning algorithm for metagenomic reads without any reference databases, called MetaAB (i.e., Abundance-based Binning of METAgenomic sequences). The proposed method uses a reduced-dimension model to find maximum likelihood estimates of parameter in a statistical model, which can reduce much computational time comparing with other approaches. Furthermore, by the advantage of the proposed model, we applies a new method of estimating the number of bins in data basing on the Bayesian information criterion.

The following sections of this paper are organized as follows. In section 2, a proposed reduced-dimension model is presented, then it is applied within an algorithm which additionally can detect the number of genome abundance levels in data by using the Bayesian information. Section 3 shows experimental results. The last section provides conclusions and future works.

2 Methods

An abundance of a species is the number of individual of the species within a given area or community. An environmental sample may contain many genomes of species with different abundance levels. This work aims to extract the information of genome abundances in a metagenomic dataset in order to classify reads into bins (or clusters) such that reads in each bins belong to genomes of very similar abundances. The proposed method is based on an observation that l-mer frequencies in reads generated from a genome is proportional to the genome abundance [15,16]. Besides, basing on the study of Lander and Waterman [17], an assumption used in this work is that the number of the occurrences of l-mers in a set of reads from a single genome follows a Poisson distribution, and all l-mers appearing in a metagenomic project are considered as mixture of Poisson distributions. Using the assumption, the proposed method firstly tries to find the maximum likelihood estimate of parameters for the model. It then classifies reads into bins basing on the probability of their l-mers belonging to each components.

2.1 Mixture Model of l-mer Frequencies

Given a metagenome dataset which consists of n reads $R = \{r_1, r_2, \ldots, r_n\}$. Let $w_1, \ldots, w_q$ be a set of l-mers in the dataset. We have a data $\mathcal{X}$ with q observations, where $c(w_i), i = \{1, \ldots, q\}$ is the value of the observation ith (i.e., the number of occurrences of w_i in the dataset). From the above assumption, the distribution of l-mers within each genome g_m is governed by a Poisson distribution with parameter λ_m. The probability function of the number of occurrences of an l-mer w_i coming from the genome g_m is

$$p_m(c(w_i)|\lambda_m) = \frac{\lambda_m^{c(w_i)} e^{-\lambda_m}}{c(w_i)!} \tag{1}$$

Assuming that the dataset consists of k species with different abundance levels, and $c(w_i), i = [1, \ldots, q]$ is independent, identically distributed observations. A finite mixture model of the k components is the convex combination, and its probability density function can be written as

$$p(c(w_i)|\Theta) = \sum_{m=1}^{k} \alpha_m p_m(c(w_i)|\theta_m), \tag{2}$$

where $\alpha_1, \ldots, \alpha_k$ are the mixing proportions and must satisfy $\sum_{m=1}^{k} \alpha_m = 1, \alpha_m > 0$. Besides, $\Theta = (\alpha_1, \ldots, \alpha_k, \theta_1, \ldots, \theta_m)$ is the set of parameters of the mixture. Each θ_m is the set of parameters of the mth component. In this context of Poisson model, we have $\theta_m \equiv \lambda_m$. The log-likelihood corresponding to the mixture of k components is:

$$\log \mathcal{L}(\Theta|\mathcal{X}) = \log \prod_{i=1}^{q} p(c(w_i)|\Theta) = \sum_{i=1}^{q} \log \left(\sum_{m=1}^{k} \alpha_m p_m(c(w_i)|\lambda_m) \right). \tag{3}$$

We aim to find the maximum likelihood estimate (MLE) of the parameter Θ, which represents the most likely assignment of the l-mers to the genomes in the dataset.

$$\Theta^* = \arg\max_{\Theta} \log p(\mathcal{X}|\Theta) \tag{4}$$

We note that this model have been also applied in [15,18] for different purposes.

2.2 A Reduced-Dimension Model

Regarding the aspect of computational cost, this study modifies the above mixture model for reducing dimension. Firstly, we present the following lemma:

Lemma 1. *Given two l-mers w_i, w_j, and a component m with parameter of λ_m. If $c(w_i) = c(w_j)$, we have $p_m(c(w_i)|\lambda_m) = p_m(c(w_j)|\lambda_m)$.*

Proof. ccording to expression 1, we have

$$\begin{aligned} p_m(c(w_i)|\lambda_m) - p_m(c(w_j)|\lambda_m) &= \frac{\lambda_m^{c(w_i)} e^{-\lambda_m}}{c(w_i)!} - \frac{\lambda_m^{c(w_j)} e^{-\lambda_m}}{c(w_j)!} \\ &= 0 \text{ (because } c(w_i) = c(w_j)) \end{aligned} \tag{5}$$

That means $p_m(c(w_i)|\lambda_m) = p_m(c(w_j)|\lambda_m)$.

Given a set of all l-mers $w_1, \ldots, w_q$ in the dataset R. Sorting the l-mers into b non-empty groups in which all l-mers $w_i, w_j, i \neq j$ in the same group t have the same number of occurrences and are equal to $c_t, t = \{1, \ldots, b\}$ (i.e., $c(w_i) = c(w_j) = c_t$), and $\forall t, s \in \{1, \ldots, b\}, c_t \neq c_s$. Denoting by $nu_t \geq 1, t = \{1, \ldots, b\}$ the number of l-mers in group t. We have

$$q = \sum_{t=1}^{b} nu_t \tag{6}$$

It is clear that since $nu_t \geq 1$, we always have $b \leq q$.

According to the Lemma 1, two l-mers having the same number of occurrences have the same probability of belonging to components. Thus, the log-likelihood corresponding to the mixture of k components, stated in expression 3, can be reformulated as

$$\log \mathcal{L}(\Theta|\mathcal{X}) = \sum_{t=1}^{b} nu_t \log \left(\sum_{m=1}^{k} \alpha_m p_m(c_t|\lambda_m) \right) \tag{7}$$

In practice, a large proportion of l-mers from the same genomes have the same number of occurrences (i.e., $nu_t \gg 1$). Given the number of l-mers q, the larger value of nu_t it is, the smaller value of b it is (see equation 6). Therefore, by using expression 7, the cost for finding maximum log-likelihood estimate of the parameter Θ can be much reduced.

2.3 Estimating Model Parameters

The Expectation Maximization (EM) algorithm [19] is used to find maximum likelihood estimates of the parameter Θ. The observed data $\mathcal{X}$ is considered to be incomplete data, and the missing data is a set of b labels $\mathcal{Z} = \{z_1, \ldots, z_b\}$ which is associated with the observed data. Each binary vector $z_t = [z_{t1}, \ldots, z_{tk}], t = \{1, \ldots, b\}$, indicates which genome produces the l-mers whose counts are equal to c_t, where $z_{tm} = 1, m = \{1, \ldots, k\}$ if the l-mers whose counts are equal to c_t is from the mth genome, and $z_{tm} = 0$ otherwise. The log-likelihood of the complete data $(\mathcal{X}, \mathcal{Z})$ is

$$\log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Z}) = \sum_{t=1}^{b} nu_t \sum_{m=1}^{k} z_{tm} \log \alpha_m p(c_t|\lambda_m). \tag{8}$$

In the EM algorithm, the unknown set of parameters $\Theta = (\alpha_1, \ldots, \alpha_k, \lambda_1, \ldots, \lambda_m)$ are randomly initialized. The parameters will be updated after each iteration. We denote by $\Theta^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_k^{(s)}, \lambda_1^{(s)}, \ldots, \lambda_m^{(s)})$ the set of parameters obtained after s iterations. Each iteration performs the following two steps (the following represents for iteration $s + 1$):

+ Expectation Step: Calculate the probability of l-mers whose counts are equal to $c_t, t = \{1, \ldots, b\}$ belonging to species mth given parameter $\Theta^{(s)}$, and c_t:

$$p(z_{tm} = 1|c_t, \Theta^{(s)}) = \frac{\alpha_m^{(s)} p_m(c_t|\lambda_m^{(s)})}{\sum_{v=1}^{k} \alpha_v^{(s)} p_v(c_t|\lambda_v^{(s)})} \tag{9}$$

Denoting $p(z_{tm} = 1|c_t, \Theta^{(s)})$ by π_{tm}, and it is called a posterior probability.

+ Maximization Step: In this step, the parameters are updated according to

$$\Theta^{(s+1)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(s)}), \tag{10}$$

where the Q-function is the expectation of the complete data log-likelihood:

$$\begin{aligned} Q(\Theta, \Theta^{(s)}) &= E[log(p(\mathcal{X}, \mathcal{Z}|\Theta))|\mathcal{X}, \Theta^{(s)}] \\ &= \sum_{t=1}^{b} nu_t \sum_{m=1}^{k} \pi_{tm} log(\alpha_m) + \sum_{t=1}^{b} nu_t \sum_{m=1}^{k} \pi_{tm} log(p_m(c_t|\theta_m)) \end{aligned} \tag{11}$$

The parameters can be calculated as follows.

$$\alpha_m^{(s+1)} = \frac{\sum_{t=1}^{b} nu_t \pi_{tm}}{\sum_{t=1}^{b} nu_t}, \lambda_m^{(s+1)} = \frac{\sum_{t=1}^{b} nu_t \pi_{tm} c_t}{\sum_{t=1}^{b} nu_t \pi_{tm}} \tag{12}$$

Once the parameters of the mixture model are estimated. Each read r_j is assigned into a component (or bin) basing on the probability of their l-mers belonging to the components. Denote by f_{im} the probability of an l-mer w_i belonging to bin mth ($i = \{1, \ldots, q, \}, m = \{1, \ldots, k\}$). Choose $t \in \{1, \ldots, b\}$ such that $c(w_i) = c_t$, we set $f_{im} = \pi_{tm}$. Let y_j to indicate in which bin a read r_j is assigned. It is calculated as

$$y_j = \underset{1 \le m \le k}{\arg\max} \frac{\prod_{w_i \in r_j} f_{im}}{\sum_{u=1}^{k} \left(\prod_{w_i \in r_j} f_{iu} \right)}. \tag{13}$$

2.4 Binning Algorithm

The pseudocode for the proposed algorithm is provided in Algorithm 1. The occurrences of l-mers in all reads $r_i \in R, i = \{1, \ldots, n\}$ are firstly calculated. In order to find the number of bins in data, we use the Bayesian information criterion. The method is a penalized likelihood approach which was shown to perform well in many fields [21]. A drawback of the BIC is that it takes much computational time to compute. However, the reduced-dimension model proposed in this study makes it applicable. The BIC is defined as BIC $= log\mathcal{L}(\Theta_M^*|\mathcal{X}) - \frac{d}{2} log(q)$ in which, M is the number of components, $\mathcal{L}(\Theta_M^*|\mathcal{X})$ is the maximum likelihood with M components, and d is the numbers of parameters in the mixture model. With this Poisson mixture model, we have $d = 2M - 1$ for a M-finite Poisson mixture model. To compute the maximum likelihood $\mathcal{L}(\Theta_M^*|\mathcal{X})$, the EM algorithm presented above is used. To choose the best model for the l-mers distribution, the EM algorithm is performed iteratively with the different number of components (or bins) m. The model which have the largest BIC value is chosen. The final step of the algorithm is to assign reads into the bins basing on the probability of their l-mers belonging to the bins. Some empty bins in which there are not any reads assigned will be removed.

Note that, after l-mer counts are computed, some untrusted l-mers whose counts do not correctly reflect the genome abundances exist in data are discarded. The unstrusted l-mers may be produced by: (1) l-mers are repeated within each genome; (2) l-mers are shared by different genomes; (3) and sequencing errors which can produce unreal l-mers.

Algorithm 1. Binning algorithm

Input: List of reads R, the number of reads n, the length of l-mers l, the minimum number of bins k_{min}, the maximum number of bins k_{max}
Output: List of bins C, the number of bins k
1: Compute counts of l-mers in R
2: Discard untrusted l-mers
3: $m = k_{min}$
4: **repeat**
5: Call *EM algorithm* in which the number of components is fixed to m
6: Compute BIC value BIC_m
7: $m = m + 1$
8: **until** $m > k_{max}$
9: $BIC_{max} = max(BIC_m), k_{min} \leq m \leq k_{max}$
10: $k = m$, where $BIC_m = BIC_{max}$
11: Assign $r_i \in R, i \in \{1, ..., n\}$ into bins C using Equation (13)
12: Remove empty bins
13: $k = k-$ the number of empty bins

3 Experiments Results

In those experiments, the proposed method is compared with AbundanceBin [15] (version 1.01, February 2013) on datasets of both with and without sequencing errors. According to the study in [22], the percentage of common l-mers between microbial genomes is less than 1% when $l \leq 20$. Moreover, AbundanceBin was shown to achieve the best performance with l-mer length of 20. Therefore, we also choose $l = 20$ for those experiments. To evaluate the approaches, two commonly used performance metrics, namely, *precision* and *recall* which are defined in [23] are used. The computer used for the experiments is an Intel Xeon with 20GB RAM running at 2.3 GHz.

3.1 Datasets

Due to the lack of standard metagenomic datasets, simulated datasets are widely used to evaluate the performance of binning algorithms. A tool used for generating metagenomic reads is MetaSim [24] which allows us to select a sequencing model and control considered parameters (e.g., read length, genome coverage, error rate). We simulate metagenomic datasets based on the bacterial genomes which are downloaded from the NCBI (National Center for Biotechnology Information) database. We generate samples which can be classified into two groups. The first group which is denoted by from S1 to S7 contains reads without sequencing errors. The second group denoted by from T1 to T7 contains reads of sequencing errors. The error-free sequencing sequences (with length of 150bp) are created by the exact simulator setting of MetaSim, while error sequencing sequences (with length of 80bp) follow the Illumina error profile with an error rate of 1%. The samples in the two groups (from S1 to S7, and from T1 to T7)

have the same the number of species, the number of abundance levels, abundance levels and the list of used species or strains, respectively.

3.2 Results on Error-Free Sequencing Reads

MetaAB firstly is compared with AbundanceBin on the samples from S1 to S7. The parameters of AbundanceBin were set default. Table 1 presents the *precision* and *recall* of the two approaches. It can be seen from the table, by using the BIC, the proposed approach is able to estimate correctly the number of bins for most of the samples (6 out of 7 cases), while AbundanceBin fails to estimate correctly the number of bins for 3 out of 7 cases. Note that each bin consists of reads from one or many species which have similar abundances. In addition, MetaAB can achieve better both *precision* and *recall* for most the tested cases. On computational performance, the proposed approach needs smaller computing time than that of AbundanceBin in many cases, especially the samples of the large number of reads.

Table 1. The *precision* and *recall* of AbundanceBin and MetaAB on samples from S1 to S7

ID	# actual bins	AbundanceBin				MetaAB			
		# bins	Precision	Recall	Running time (s)	# bins	Precision	Recall	Running time (s)
S1	2	2	96.57%	96.57%	**94**	2	96.57%	96.57%	116
S2	3	3	94.9%	95.58%	**305**	3	**95.83%**	**95.58%**	328
S3	3	4	90.72%	86.84%	556	**3**	**95.4%**	**95.06%**	**483**
S4	4	4	96.96%	96.96%	**745**	4	**97.61%**	**97.08%**	812
S5	4	3	65.43%	**94.69%**	507	**4**	**85.72%**	85.24%	**489**
S6	5	4	85.54%	**88.41%**	795	**5**	**86.18%**	77.63%	**782**
S7	6	**6**	**94.46%**	94.46%	2808	2	73.12%	**99.16%**	**2519**

3.3 Results on Error Sequencing Reads

Binning approaches should have ability to deal with sequencing errors since there are no any current sequencing technologies which could generate reads without errors. The proposed approach is tested on the datasets with sequencing errors from T1 to T7, and is compared with AbundanceBin. In order to reduce the bad effects of the errors, and for a fair comparison, both approaches are set to discard the l-mers which appear only once from the binning process. Table 2 compares the accuracy and computational time of the two approaches. Obviously, MetaAB can work well with the error sequencing reads and outperforms AbundanceBin for most of the tested samples. The proposed approach can estimate correctly the number of bins in each sample for 5 out of 7 cases, whereas AbundanceBin detects correctly the number of bins for only one sample (sample T1). Because of sequencing errors, AbundanceBin seems to return the estimated number of bins which are much less than the actual ones. This helps it to achieve high *recall*

values, but its get very low *precision* for the samples. It is very interesting that MetaAB get higher *precision* than that of AbundanceBin for all the tested samples. Furthermore, the proposed approach is much faster than AbundanceBin.

Table 2. The *precision* and *recall* of AbundanceBin and MetaAB on samples from T1 to T7

ID	# actual bins	AbundanceBin				MetaAB			
		# bins	Precision	Recall	Running time (s)	# bins	Precision	Recall	Running time (s)
T1	2	2	94.6%	94.5%	135	2	**98.04%**	**98.04%**	**107**
T2	3	2	92%	98.52%	315	2	**92.8%**	**99.53%**	**282**
T3	3	1	49.3%	**100%**	1524	**3**	**96.58%**	96.56%	**422**
T4	4	3	61.22%	**95.39%**	858	**4**	**94.35%**	93.81%	**643**
T5	4	2	63.55%	**94.25%**	670	**4**	**71.43%**	71.89%	**417**
T6	5	2	62.72%	89.99%	1630	**4**	**89.17%**	**91.65%**	**612**
T7	6	2	71.34%	**97.56%**	6789	**6**	**94.27%**	85.33%	**2224**

4 Conclusion

The development of next-generation sequencing, which allows to produce a mass of data, brings computational challenge in metagenomic projects. This study focuses on the challenge in which a reduce-dimension model is proposed. By taking the advantage of the model, a method of detecting the number of bins in data based on the Bayesian information criterion is applied. Our experiments demonstrates that the proposed approach not only achieves higher accuracy but also consumes less computational time than a recent abundance-based binning approach. In future works, we aim to apply the proposed approach for the improvement of compositional-based binning methods.

Acknowledgments. This research is funded by the Ho Chi Minh city University of Technology (Project code: TNCS-2013-KHMT-10).

References

1. Handelsman, J.: The New Science of Metagenomics: Revealing the Secrets of Out Microbial Planet. The National Academies Press, Washington, DC (2007)
2. Aann, R.I., Ludwig, W., Schleifer, K.H.: Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. (1995)
3. Wooley, J.C.: A primer on metagenomics. PloS Computational Biology (2010)
4. Shendure, J., Ji, H.: Next-generation dna sequencing. Nature Biotechnology (2008)
5. Qin, J., Li, R., Wang, J.: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464** (2010)
6. Huson, D.H.: Megan analysis of metagenomic data. Genome Research (2007)
7. Gerlach, W.: Taxonomic classification of metagenomic shotgun sequences with carma3. Nucleic Acids Research (2011)

8. Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., Nattkemper, T.W.: Tacoa: Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics (2009)
9. Yi, W., et al.: Metacluster-ta: taxonomic annotation for metagenomic databased on assembly-assisted binning. BMC Genomics **15** (2014)
10. Eisen, J.A.: Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. PLoS Biol. **5**(3) (2007)
11. Yang, B., Peng, Y., Qin, J., Chin, F.Y.L.: MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. In: ACM BCB (2010)
12. Leung, H.C., Yiu, F.M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chin, F.Y.: A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics **27**(11), 1489–1495 (2011)
13. Liao, R., Zhang, R., Guan, J., Zhou, S.: A new unsupervised binning approach for metagenomic sequences based on n-grams and automatic feature weighting. IEEE/ACM Transaction on Computational Biology and Bioinformatics (2014)
14. Nguyen, T.C., Zhu, D.: Markovbin: An algorithm to cluster metagenomic reads using a mixture modeling of hierarchical distributions. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics
15. Wu, Y.W., Ye, Y.: A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. Journal of Computational Biology **18**(3), 523–534 (2011)
16. Tanaseichuk, O., Borneman, J., Jiang, T.: A probabilistic approach to accurate abundance-based binning of metagenomic reads. In: Raphael, B., Tang, J. (eds.) WABI 2012. LNCS, vol. 7534, pp. 404–416. Springer, Heidelberg (2012)
17. Lander, E.S., Waterman, M.S.: Genomic mapping by fingerprinting random clones: a mathematic alanalysis. Genomic (1988)
18. Li, X., Waterman, M.S.: Estimating the repeat structure and length of dna sequences using -tuples. Genome research **13**(8), 1916–1922 (2003)
19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society.SeriesB (Methodological) **39**(1), 1–38 (1977)
20. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern analysis and machine intelligence **24**(3), 381–396 (2004)
21. Hirose, K., Kawano, S., Konishi, S., Ichikawa, M.: Bayesian information criterion and selection of the number of factors in factor analysis models. Journal of Data Science **9**(2), 243–259 (2011)
22. Wang, Y., Leung, H.C., Yiu, S.M., Chin, F.Y.: Metacluster 4.0: a novel binning algorithm for ngs reads and huge number of species. Journal of Computational Biology **19**(2), 241–249 (2012)
23. 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 **28**(18), 356–362 (2012)
24. Richter, D.C., Ott, F., Auch, A.F., Schmid, R., Huson, D.H.: Metasim - a sequencing simulator for genomics and metagenomics. PLoS ONE (2008)