Unsupervised fuzzy binning of metagenomic sequence fragments on three-dimensional Barnes-Hut *t*-Stochastic Neighbor Embeddings

Leandro Ariza-Jiménez[1], O.L. Quintero[2], Nicolás Pinel[3]

***Abstract*— Shotgun metagenomic studies attempt to reconstruct population genome sequences from complex microbial communities. In some traditional genome demarcation approaches, high-dimensional sequence data are embedded into two-dimensional spaces and subsequently binned into candidate genomic populations. One such approach uses a combination of the Barnes-Hut approximation and the *t*-Stochastic Neighbor Embedding (BH-SNE) algorithm for dimensionality reduction of DNA sequence data pentamer profiles; and demarcation of groups based on Gaussian mixture models within human-imposed boundaries. We found that genome demarcation from three-dimensional BH-SNE embeddings consistently results in more accurate binnings than 2-D embeddings. We further addressed the lack of *a priori* population number information by developing an unsupervised binning approach based on the Subtractive and Fuzzy *c*-means (FCM) clustering algorithms combined with internal clustering validity indices. Lastly, we addressed the subject of shared membership of individual data objects in a mixed community by assigning a degree of membership to individual objects using the FCM algorithm, and discriminated between confidently binned and uncertain sequence data objects from the community for subsequent biological interpretation. The binning of metagenome sequence fragments according to thresholds in the degree of membership opens the door for the identification of horizontally transferred elements and other genomic regions of uncertain assignment in which biologically meaningful information resides. The reported approach improves the unsupervised genome demarcation of populations within complex communities, increases the confidence in the coherence of the binned elements, and enables the identification of evolutionary processes ignored in hard-binning approaches in shotgun metagenomic studies.**

I. INTRODUCTION

Given that the vast majority of microorganisms from natural microbial communities are as of yet unculturable [1], microbial ecology studies rely on the reconstruction of community structure and dynamics from community DNA sequencing data. High throughput second generation sequencing platforms such as Illumina can generate billions of base pairs (bp) of DNA sequence, sufficient to promise an ecologically relevant sampling depth for microbial community genomes. Unfortunately, the sequences are currently generated in fragments of at most 600 bp in length each. Even with advanced *de novo* genome sequence assemblers fine tuned to metagenomic data [2], [3], the reconstruction of long stretches of individual genomes remains a formidable challenge. Segregation of individual fragments into discrete, coherent sub-groups (binning) not only aids in the success of genome reconstruction (since the reduction in data complexity improves genome assembly [4]), but is also essential to enable downstream ecological analysis of community taxonomic structure and the functional potential of each individual population. Binning of short sequence fragments remains one of the most challenging tasks in shotgun metagenomic studies [5].

The Barnes-Hut *t*-Stochastic Neighbor Embedding (BH-SNE) is a non-linear dimensionality reduction method introduced as an approach to enable the visualization and subsequent binning of genomic fragments [6]. BH-SNE-based mappings were demonstrated to outperform Principal Component Analysis (PCA) at binning [7]. Although BH-SNE can embed multidimensional data into two and three dimensions [8], in metagenomic studies it is customary to embed data into two-dimensional spaces and then perform a binning process [6], [9]–[12]. The data loss that inevitably results from dimensionality reduction algorithms such as BH-SNE or PCA may compromise the interpretability of the results or even distort the extant natural groupings [13]. Therefore, embedding metagenomic data into three-dimensional spaces could have a positive influence in the subsequent binning process. In this paper we propose the use of three-dimensional BH-SNE-based embeddings of genomic sequence fragments for binning purposes. Additionally, we present an unsupervised group identification method that uses the Subtractive clustering and the Fuzzy *c*-means (FCM) algorithms, in conjunction with internal clustering validity indices. In real communities, group membership can be a distributed and partial attribute. Horizontal gene transfer is a frequent event in microbial communities. At the moment of analyzing the constituents of each of the genome populations (bins), it is important to recognize the uncertainty in group membership. Applying hard binning can result in artifactual attributions of group membership and the consequent errors in functional interpretation of the populations. By performing a soft clustering on the metagenomic sequence fragments, we can assign a degree of membership to individual sequence fragments, and discriminate between confidently binned and uncertain sequence data objects from the community for subsequent biological interpretation.

II. MATERIALS AND METHODS

A. Metagenomic data

Experiments were conducted using sequence data from the Human Microbiome Project (HMP; [14]). Twenty

[1]L. Ariza-Jiménez, Mathematical Modelling Research Group, Universidad EAFIT, Colombia (larizaj@eafit.edu.co)

[2]O. L. Quintero, Mathematical Modelling Research Group, Universidad EAFIT, Colombia (oquinte1@eafit.edu.co)

[3]N. Pinel, Biodiv., Evol., Cons. Res. Group, Universidad EAFIT, Colombia (npinelp@eafit.edu.co)

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

seven genome sequence assemblies of intestinal microorganisms were obtained from the HMP data portal (http://hmpdacc.org/HMRGD/). The pairwise average amino acid identity (AAI) among all of the sequences was calculated using the on-line AAI calculator from the Kostas Laboratory (http://enve-omics.ce.gatech.edu/aai/). The sequences were hierarchically clustered based on their pairwise AAI values. "Synthetic" communities (SC1 to SC4) with 3, 5 and 10 population genomes were constructed to include a wide range of phylogenetic relatedness (inferred from the AAI values) among the member populations. Sequence data sets were constructed by generating the desired number of randomly distributed contigs (sequence fragments), 1000 bp in length each, from the genome sequences. Each simulated community included uniformly abundant populations. Because of differing genome sizes, uniformity in population abundance distribution did not represent uniformity in sequence data abundance. The communities illustrated in this paper ranged from 1000 to 20000 contigs. The pentamer (five-letter strings) frequency profiles in each of the contigs of the assembled communities were calculated, considering direct sequences and their reverse complements as the same (representing therefore a simplified *5-mer* dictionary). The counts of all possible 5-mers were initiated with 0.1 pseudocounts to avoid zero values in the frequency table. A centered log-ratio transformation was applied to the frequency data prior to processing with the BH-SNE algorithm. Experiments were also conducted on two pairs of simulated metagenomic datasets, EqualSet1-Equalset2 and Dataset1-Dataset2, already used in [6] and [15], respectively. These additional datasets have similar characteristics with the first group of synthetic communities (number of contigs, 1000 bp in length, and uniformly abundant populations) and were preprocesed as described above.

B. Dimensionality reduction via the BH-SNE algorithm

The BH-SNE [8] is a variant of the *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) [16], a non-linear dimensionality reduction method, in which approximations based on the Barnes-Hut algorithm [17] are introduced to decrease the computational complexity, thereby allowing the applicability of the BH-SNE to large data sets with millions of objects. Generally speaking, both BH-SNE and *t*-SNE algorithms follow a probabilistic approach to capture the neighborhood structure of a set of data objects $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in the original space $\mathbb{R}^r$ by learning an *s*-dimensional embedding in which each object is represented by a point of a set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i \in \mathbb{R}^s$ and $s \ll r$. Here, the BH-SNE algorithm was used to embed the high-dimensional metagenomic data into low-dimensional spaces wherein a binning algorithm can be applied. Two- and three-dimensional mappings were generated to investigate how increasing the dimension of the embedding space affects the subsequent binning of genomic fragments. As in [6], default parameter values for the BH-SNE algorithm were used. The effective number of local neighbors based on which neighborhood structure is captured, known as perplexity, was set to 30. The speed-accuracy trade-off parameter θ, which ranges between 0 and 1, was set to 0.5. Preliminary tests with different values for θ (0.25, 0.50, 0.75) did not reveal significant differences in binning performance, and thus the default value of 0.5 was retained. Finally, the preliminary PCA-based dimensionality reduction step that the algorithm performs by default was omitted to avoid any possible influence of this procedure in the separation of the groups in the embedded map as reported in [18].

C. Unsupervised method for metagenomic binning

The Subtractive clustering algorithm [19] was used to find the centers of clusters in the embedded data sets. Subtractive assumes each data object has the potential of being a cluster center based on the density of its surrounding data objects. The algorithm uses a particular pair of positive radii r_a and r_b, with values between 0 and 1, to condition the effect of data objects around potential cluster centers in the density measure. The second radius was set as $r_b = 1.25 r_a$, since r_b should be greater than r_a in order to ensure the identification of cluster centers that are sufficiently separated. The Subtractive algorithm was run for values of r_a from 0.1 to 0.8 with increments of 0.05, to obtain estimates of the number of clusters and their corresponding locations in the embedding space as a function of the radius r_a. Each of the embeddings was clustered using the FCM algorithm (with a fuzzification parameter $m = 2$) according to each of the estimates of centers provided by Subtractive, thereby obtaining a fuzzy partition matrix whose entries described the degree of membership of each data object to the different discovered groups. To obtain a hard clustering, each data object was assigned to the cluster with the largest measure of membership. The goodness of each hard clustering was estimated using an internal validity index. The clustering that optimized the validity index was selected as the best one for the embedded metagenomic data. The following cluster validity indices were evaluated: Silhouette (Silh), Calinski-Harabasz (CH), and Dunn. This selection was based on the best performing indices in [20].

D. Performance evaluation

The F-measure, a commonly used performance metric in metagenomic binning studies [21]–[23] when the true number of microbial populations is known, was used to evaluate the outcomes of the proposed binning methodology. Because of the non-deterministic nature of the BH-SNE algorithm, 11 replicate embeddings were generated for each synthetic community, and the proposed unsupervised clustering method was applied to all the replicates. Median F-measure values are reported.

III. RESULTS AND DISCUSSION

Fig. 1 illustrates the binning process applied to a BH-SNE embedding of a synthetic community (SC) with 5 genomic populations. The upper panel shows the number of cluster centers estimated by Subtractive as a function of r_a. The bottom panel depicts the behavior of the different cluster validity

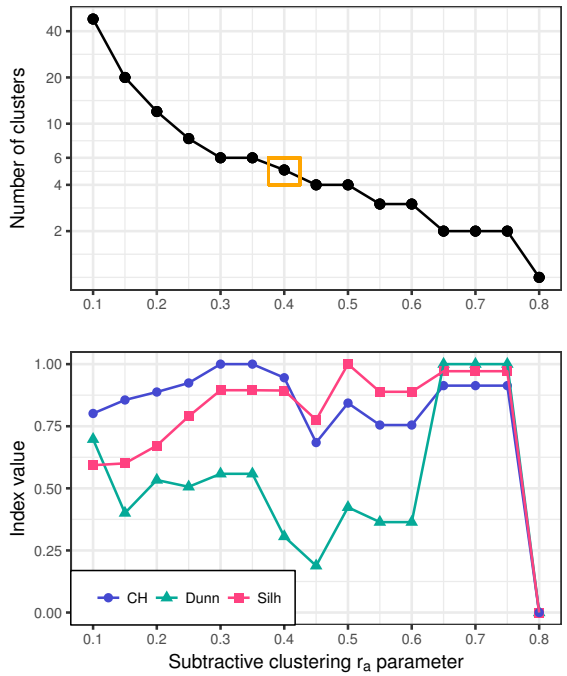

Fig. 1. Unsupervised binning upon BH-SNE embeddings of pentamer profiles from metagenomic sequence data. The upper panel's vertical axis is plotted on a logarithmic scale. Cluster validity indices in the bottom panel are normalized.

TABLE I

BINNING PERFORMANCE ON SYNTHETIC METAGENOMIC COMMUNITIES

		Dunn		Silh		CH		MC3	
SC	Space	*C*	F	*C*	F	*C*	F	*C*	F
SC1	BHSNE2	**3**	**0.97**	2	0.83	**3**	**0.97**	**3**	0.82
(3 pop.)	BHSNE3	11	0.48	2	0.83	**3**	**0.97**		
SC2	BHSNE2	**5**	0.81	4	0.82	6	0.84	**5**	0.80
(5 pop.)	BHSNE3	6	0.71	**5**	0.84	**5**	**0.90**		
SC3	BHSNE2	9	0.51	9	**0.61**	13	0.57	7	0.61
(10 pop.)	BHSNE3	13	0.57	9	**0.61**	11	0.60		
SC4	BHSNE2	5	0.66	4	0.69	**10**	0.77	8	0.78
(10 pop.)	BHSNE3	11	0.72	7	**0.82**	7	**0.82**		
Dataset1	BHSNE2	22	0.20	**2**	**0.99**	**2**	**0.99**	**2**	0.97
(2 pop.)	BHSNE3	**2**	**0.99**	**2**	**0.99**	**2**	**0.99**		
Dataset2	BHSNE2	**3**	**0.96**	**3**	**0.96**	**3**	**0.96**	**3**	0.91
(3 pop.)	BHSNE3	4	0.88	**3**	**0.96**	**3**	**0.96**		
EqualSet2	BHSNE2	13	0.40	4	0.71	4	0.71	2	**0.75**
(3 pop.)	BHSNE3	**3**	0.60	4	0.72	4	0.72		
EqualSet1	BHSNE2	5	0.80	3	0.70	6	0.82	8	**0.83**
(9 pop.)	BHSNE3	14	0.71	5	**0.83**	5	**0.83**		

indices also as a function of r_a, with local maxima indicating the best clustering attained. All three indices present local maxima around 6 populations, close to the true value of 5. However, only for CH are these absolute maxima, a result consistent in all evaluations. Table I contains binning results from the proposed method, and their comparison to results from MetaCluster 3.0 (MC3) [15], an unsupervised binning method commonly used in metagenomic analyses. For each clustering result, median values of the estimated number of fuzzy clusters (*C*) and the accuracy of the clustering in terms of the F-measure (F) are reported. The best results in a given community are in bold. Except for the last two datasets, the results indicate that our unsupervised method for metagenomic binning outperforms MetaCluster 3.0, which previously has demonstrated to achieve better perfomance than other existing unsupervised algorithms on datasets with 1000 bp read length [15]. Binning on 3-D BH-SNE embeddings results in higher F-measure values than binning on 2-D BH-SNE embeddings (Table I). Embedding into a 3-D space represents an increase in CPU time of 2.0-2.5 times over 2-D embeddings. In [7], a *t*-SNE-based approach for cluster identification in metagenomic data yields the best results according to accuracy measures and the Dunn index. Our findings suggests the Dunn index as unreliable for unsupervised cluster number selection, due the discrepancies between absolute and local maxima (Fig. 1). In contrast, the CH index consistently presents local maxima in the region close to the true number of populations in the community. Compared to other binning approaches that also use center-based clustering methods [5], [10], our method estimates the number of genome populations by evaluating different values of r_a over a fixed range, rather than performing an evaluation based on a variable such as the number of fuzzy clusters *C*. The latter approach relies on guesses or prior knowledge about the structure of the community. It is important to note that microbial genomes may contain regions that display individual characteristics in pentamer profiles, and thus may bin as separate groups from the rest of the population genome. This phenomenon may be observed in organisms whose genomes consist of multiple chromosomes of different evolutionary origins. Or, it may reflect horizontally acquired genomic regions. Previous work has found that genomic elements under peculiar evolutionary dynamics, such as the 16S ribosomal gene, can map together and apart from all other groups even while belonging to different organisms [6].

Fig. 2 illustrates a population genome binning strategy that incorporates the concept of the degree of group membership. The upper left panel depicts a 2-D mapping of a synthetic community with 5 populations, with colors and glyphs representing the known structure. The upper right panel depicts the same embedding with the hard clustering and group centers suggested by our unsupervised binning approach. A sixth group is present (gray), resulting from the mixed populations in the bottom left part of the map. Multiple elements in the regions between populations are assigned to the wrong population genome. Applying a degree of membership threshold as a criterion for group assignment discriminates between confidently binned and uncertain sequence data objects (black) from this microbial community for subsequent biological interpretation. This opens the door to the

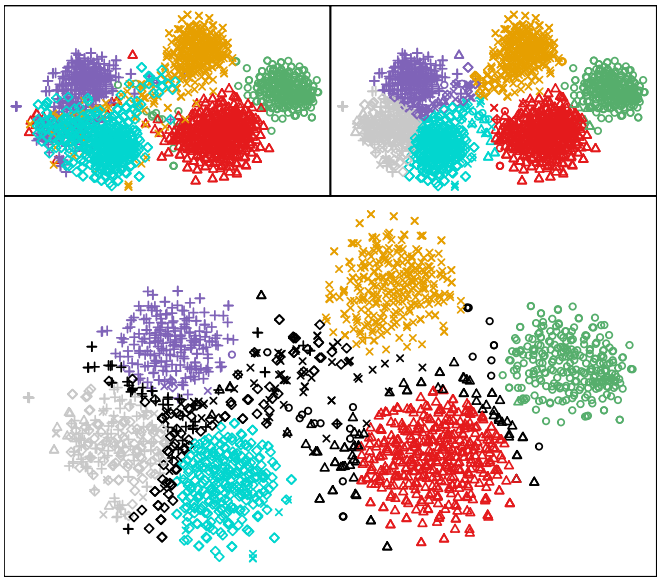

Fig. 2. Fuzzy-based strategy to deal with group membership uncertainty. Top-left panel represents with colors and glyphs true group membership in a 2-D BH-SNE embedding. Top-right panel indicates with colors group assignment from a hard-clustering approach. Bottom panel displays in black elements with degree of membership below the chosen threshold of 0.55.

identification of potential horizontally transferred elements, and of genomic regions experiencing special dynamics of molecular change. A soft binning approach thus presents opportunities over hard binning for refining the biological and ecological interpretation of metagenomic data, a topic that will be the focus of future work.

IV. CONCLUSIONS AND FUTURE WORK

This paper presented a strategy for unsupervised, fuzzy binning of metagenomic sequence fragments from BH-SNE embeddings. The integration of Subtractive clustering with internal clustering validity indices (e.g., Calinski-Harabasz index), enabled the selection of group centers without *a priori* knowledge about community structure. Binning from 3-D embeddings outperformed binning from 2-D embeddings. A soft-clustering approach to deal with the biological reality of distributed or uncertain group membership was developed. This methodology will improve the accuracy of subsequent genome content analysis and the ecological interpretations derived therefrom. As future work, we plan to perform rigorous comparisons against state of the art binning approaches. We also plan to extend the analysis to more complex microbial communities, and to implement our method as a public tool for use in microbial ecology studies. Although our unsupervised method is intended for metagenomic binning, we believe it can be adapted and applied to other problems and applications in data science.

REFERENCES

[1] O. Overmann, B. Abt, and J. Sikorski, "Present and Future of Culturing Bacteria," *Annual Review of Microbiology*, vol. 71, no. July, pp. 711–730, 2017.

[2] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data," *Bioinformatics*, vol. 27, no. 13, pp. i94–i101, 2011.

[3] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "MetaSPAdes: A new versatile metagenomic assembler," *Genome Research*, vol. 27, no. 5, pp. 824–834, 2017.

[4] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork, "Assessment of metagenomic assembly using simulated next generation sequencing data," *PLoS ONE*, vol. 7, no. 2, 2012.

[5] S. Girotto, C. Pizzi, and M. Comin, "MetaProb: Accurate metagenomic reads binning based on probabilistic sequence signatures," *Bioinformatics*, vol. 32, no. 17, pp. i567–i575, 2016.

[6] C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes, "Alignment-free visualization of metagenomic data by nonlinear dimension reduction." *Scientific Reports*, vol. 4, p. 4516, 2014.

[7] A. Gisbrecht, B. Hammer, B. Mokbel, and A. Sczyrba, "Nonlinear dimensionality reduction for cluster identification in metagenomic samples," *Proceedings of the International Conference on Information Visualisation*, pp. 174–179, 2013.

[8] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245, 2014.

[9] C. C. Laczny, T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. Margossian, S. Coronado, L. der Maaten, N. Vlassis, and P. Wilmes, "VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data," *Microbiome*, vol. 3, no. 1, p. 1, 2015.

[10] M. Lux, A. Sczyrba, and B. Hammer, "Automatic discovery of metagenomic structure," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[11] S. Kouchaki, S. Tirunagari, A. Tapinos, and D. L. Robertson, "Local Binary Patterns as a Feature Descriptor in Alignment-Free Visualisation of Metagenomic Data," in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, 2016, pp. 1–6.

[12] H.-H. Lin and Y.-C. Liao, "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes," *Scientific Reports*, vol. 6, no. 1, p. 24175, 2016.

[13] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: a review." *IEEE reviews in biomedical engineering*, vol. 3, pp. 120–54, 2010.

[14] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[15] H. C. M. Leung, S. M. Yiu, B. Yang, Y. Peng, Y. Wang, Z. Liu, J. Chen, J. Qin, R. Li, and F. Y. L. Chin, "A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio," *Bioinformatics*, vol. 27, no. 11, pp. 1489–1495, 2011.

[16] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[17] J. Barnes and P. Hut, "A hierarchical O(n-log-n) force calculation algorithm," *Nature*, vol. 324, pp. 446–449, 1986.

[18] A. Mahfouz, M. van de Giessen, L. van der Maaten, S. Huisman, M. Reinders, M. J. Hawrylycz, and B. P. F. Lelieveldt, "Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings," *Methods*, vol. 73, pp. 79–89, 2015.

[19] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, no. 3, pp. 267–278, 1994.

[20] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[21] Y. Wang, H. C. Leung, S. M. Yiu, and F. Y. Chin, "Metacluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, pp. 356–362, 2012.

[22] L. V. Vinh, T. V. Lang, L. T. Binh, and T. V. Hoai, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology*, vol. 10, no. 1, pp. 1–12, 2015.

[23] S. Girotto, M. Comin, and C. Pizzi, "Metagenomic reads binning with spaced seeds," *Theoretical Computer Science*, vol. 698, pp. 88–99, 2017.