SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Biomolecular Network Analysis and Application

July 2016, Vol. 59 070104:1–070104:7
doi: 10.1007/s11432-016-5587-8

Identification of the clustering structure in microbiome data by density clustering on the Manhattan distance

Xingpeng JIANG[1], Xiaohua HU[1,2] & Tingting HE[1*]

[1] *School of Computer Science, Central China Normal University, Wuhan* 430079, *China;*
[2] *College of Computing & Informatics, Drexel University, Philadelphia, PA* 19104, *USA*

Received April 15, 2016; accepted May 19, 2016; published online June 16, 2016

Abstract Clustering technology is a method for grouping data points into clusters containing a group of similar data points. In a real dataset such as microbiome data, the data points are presented as profiles or a probability distribution. These data points form the periphery of a cluster, making it difficult to identify the real clustering structure. In this study, we used density clustering on several distance measures to overcome this difficulty. Experiments using a real dataset indicated that the Manhattan distance is an appropriate distance measure for clustering analysis of microbiome data.

Keywords microbiome, information distance, data visualization, density clustering, microbial community

Citation Jiang X P, Hu X H, He T T. Identification of the clustering structure in microbiome data by density clustering on the Manhattan distance. Sci China Inf Sci, 2016, 59(7): 070104, doi: 10.1007/s11432-016-5587-8

1 Introduction

Analysis of the microbiome is a holistic approach that emphasizes the integration and interaction of different elements in a microbial community [1, 2]. Microbiome datasets are often comprised of different representations of data such as metabolic pathways, taxonomic assignments, and gene families [3, 4]. DNA sequences obtained from metagenomic or 16SrRNA sequencing technologies could be summarized using metagenomic profiles, which represent the abundance of functional or taxonomic categories in metagenomic sequences. A metagenomic profile matrix typically contains hundreds of metabolic pathways, thousands of species, or tens of thousands of protein families (see Figure 1 for an illustration) [5]. Machine-learning and multivariate statistics have been employed on the profile matrix to explore and extract the complex structures and correlations [6]. For example, metagenomic samples could be represented by several "components", which may facilitate biological interpretation and discovery [7].

Clustering is an efficient data-mining technique, with many applications for real data. Clustering can be used to identify groups of objects that may reflect significant phenomena such as biological or social grouping. In recent years, the clustering of microbiome data samples has been frequently used for analyzing microbiome data. Arumugam et al. [8] integrated principal components analysis and clustering

[*] Corresponding author (email: tthe@mail.ccnu.edu.cn)

● Taxonomic profiles

$$T_{ts} = \begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_s \\ \hline T_1 & 19 & 4 & \cdots & 9 \\ T_2 & 9 & 8 & \cdots & 0 \\ \vdots & & \cdots & & \\ T_t & 12 & 0 & \cdots & 7 \end{array}$$

● Functional profiles

$$F_{fs} = \begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_s \\ \hline F_1 & 9 & 5 & \cdots & 1 \\ F_2 & 8 & 2 & \cdots & 0 \\ \vdots & & \cdots & & \\ F_f & 2 & 0 & \cdots & 7 \end{array}$$

Figure 1 Illustration of the metagenomic profile. The columns are usually the metagenomic samples; the taxonomic profile contains thousands of rows of taxa, and the functional profile contains many rows of functional categorizations.

analysis and found that the microbial composition in the human gut is not random, but can be classified into at least three enterotypes. Although the origin of different enterotypes remain unknown, it may be related to the response of the human immune systems to bad or good bacteria, or different methods of excreting wastes from the body [9, 10]. Similar to blood type, enterotype is not related to demographic factors such as country, age, gender, race, or other body indices. There are also different opinions related to the concept of enterotypes, with some authors suggesting that it is not practical to collapse enterotype variation into a few discrete clusters [11]. Instead, they argued that enterotype distribution is continuous, and can vary widely within an individual. Thus, the utility of discrete clustering in microbiome analyses remains a topic of active debate.

With the increasing size of datasets in microbiome studies, it is now possible to use probabilistic models to investigate the enterotype. Based on the functional elements derived from the non-redundant coding DNA sequence catalogue of the human gut microbiome, we have demonstrated that the configuration of functional groups in metagenome samples can be inferred using probabilistic topic modeling [12]. Each microbial sample (assuming that the relative abundances of functional elements are already known from a homology-based approach) can be considered as a "document" which is a mixture of functional groups, while each functional group (considered the "latent topic") is a weighted mixture of functional elements (including taxonomic levels, and indicators of gene orthologous groups and KEGG pathway mappings). In the analogy, the functional elements can be considered as the "words". Estimating the probabilistic topic model can uncover the configuration of functional groups (the latent topic) in each sample. The results derived from this approach were found to be consistent with recent discoveries in a fecal microbiota study of patients with inflammatory bowel disease [12]. The latent topics estimated from human gut microbial samples were verified by the recent discoveries in the fecal microbiota study, which demonstrated the effectiveness of the probabilistic topic model.

In this paper, we propose using the Manhattan distance(L_1 distance) instead of the common L_2 distance for microbiome clustering. We also compare two information distance measures that are frequently used in microbiome studies. We employed an efficient density clustering method for clustering data points. Finally, experiments on a real microbiome dataset indicated that the proposed approach has potential merits for application in microbiome studies.

2 Methods

Dataset: We used the dataset obtained from the Human Microbiome Project (HMP [13, 14]). After filtering out body sites with less than 15 samples, the dataset contained 637 samples drawn from seven body sites, including one vagina(posterior fornix), one gut(stool), one nasal(anterior nares), and three oral (supragingival plaque, tongue dorsum, and buccal mucosa) sites. Table 1 shows a summary of the dataset. The phylogenetic profile containing the relative abundances of microorganisms was estimated using the software MetaPhlAn at the species level (710 × 637). The datasets were all downloaded from the HMP data site: http://hmpdacc.org/ [5].

Several density clustering approaches have been proposed in recent years, including K-means, K-medoids, mean-shift, density-based spatial clustering of applications with noise, and others [15]. These methods are useful for different types of datasets, but they also have drawbacks in dealing with particular

Table 1 Summary of the HMP samples

ID	Body sites	Number of samples
1	Stool	134
2	Posterior_fornix	49
3	Anterior_nares	86
4	Buccal_mucosa	106
5	Plaque	122
6	R_Retroauricular_crease	17
7	Tongue_dorsum	123

data types. A novel density-based approach has been proposed using a fast search and find of density peaks [16]. This powerful method has a simple assumption that the cluster centers are surrounded by neighbors with a lower local density and that they are at a relatively large distance from any points with a higher local density. We employed this method in the microbiome data analysis for this study owing to its efficiency over other methods.

Distance matrix: The input of the method is a distance matrix; for example, a Euclidean distance matrix is often used in many applications. However, we found that the L_1 distance (also called the Manhattan distance or city block distance) may be more suitable for analyzing microbiome data. Considering two data points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, the L_2 distance is defined as

$$d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2.$$

Furthermore, the L_1 distance is defined as

$$d(x,y) = \sum_{i=1}^{n} \left| x_i - y_i \right|.$$

The differences in the L_1 and L_2 distance derive from the fact that in the L_1 distance, dimensions with large differences are less important than for the L_2 distance. We observed that microbiome profiles often have some dimensions with large values, whereas most of the dimensions have small values. We expect that large-value dimensions have equal importance to the small-value dimensions. Thus, the L_1 distance appears to be a better candidate than the L_2 distance in this case.

Kolmogorov distance: The Kolmogorov distance is the maximal distance between the cumulated spectra [17]. The function returns this distance and the corresponding frequency. This is an adaptation of the statistic computed by the non-parametric Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test does not require the assumption that the population is normally distributed. The empirical distribution function F_n for n observations of X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i),$$

where $I_{[-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i < x$ and equal to 0 otherwise. The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x \left| F_n(x) - F(x) \right|,$$

where $\sup_x$ is the supremum of the set of distances. By the Glivenko-Cantelli theorem, if the sample comes from distribution $F(x)$, then D_n almost surely converges to 0 in the limit when n goes to infinity.

Difference between two cumulative frequency spectra (DiffCumSpec): Two distributions (e.g., two frequency spectra) are compared by computing the difference between two cumulative frequency

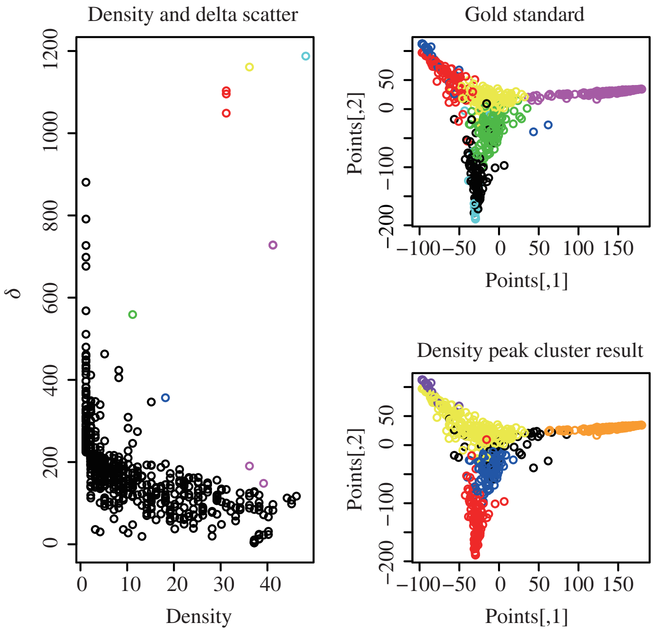

Figure 2 Density clustering on the Manhattan distance (L_1 distance).

spectra [18]. Both spectra are transformed into cumulative distribution functions (CDFs). The spectral difference is then computed according to

$$D = \frac{1}{n}|X - Y|$$

with X and Y indicating the spectrum CDFs, and $0 < D < 1$.

Density clustering: In density-based clustering [16], the clusters are defined as areas of higher density than the remainder of the dataset. Objects in these sparse areas, which are required to effectively separate clusters, are usually considered to be noise and border points. Density-based clustering computes two quantities for each data point: the local density ρ_i and its distance δ_i from points of higher density. The local density ρ_i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c),$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise.

Then, δ_i is defined bymeasuring the minimum distance between the point i and any other point with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}).$$

3 Results

We compared the results of density clustering based on the L_2 and L_1 distance using a real human microbiome dataset. We calculated the distance matrix for each dataset, and applied density clustering. Nonmetric multidimensional scaling was used to investigate the clustering results; specifically, the function is oMDS in the R package MASS was used for this purpose [19].

We found that the density clustering based on the L_1 distance showed better performance than that based on the L_2 distance in the metabolic profile. According to the decision graphs (left panels of Figures 2 and 3), six cluster centers were selected in both cases. The multidimensional scaling (right panels in Figures 2 and 3) showed that the results for the L_1 distance identified more of the true cluster structure than those for L_2 (there were 7 clusters in the real data). A large skin microbiome cluster in the L_2 distance was divided into several skin clusters, which is close to the real data.

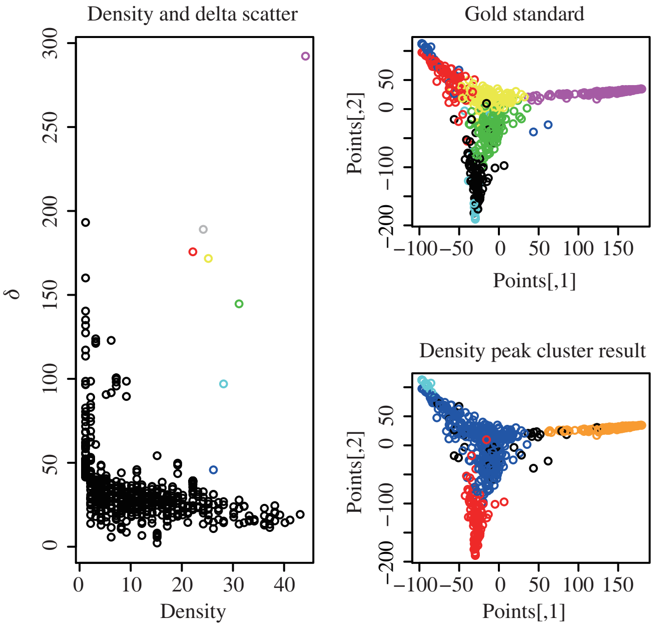

Figure 3 Density clustering on the L_2 distance.

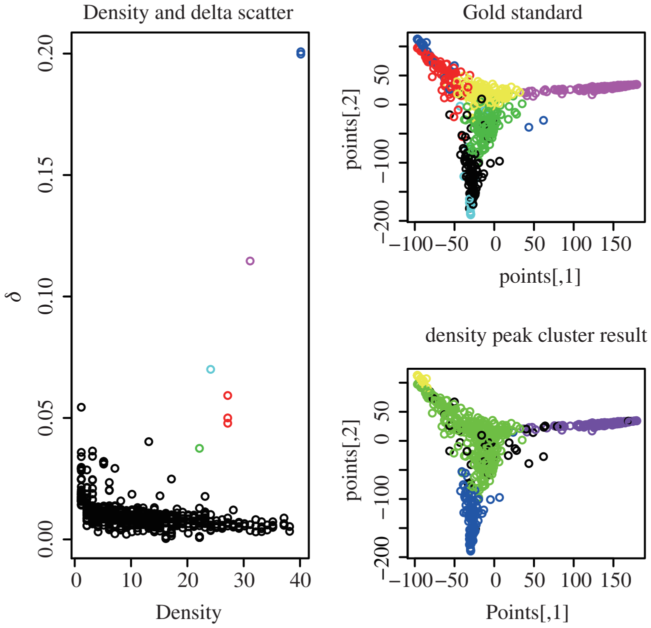

Figure 4 Density clustering on DiffCumSpec distance.

Because an information distance measure is frequently used in microbiome clustering, we tested two information distances—DiffCumSpec and Kolmogorov distance—on the same dataset. Figures 4 and 5 show the results of DiffCumSpec and Kolmogorov distance, respectively. Density clustering on the L_1 distance selected the correct cluster numbers and showed better performance for distinguishing clusters.

4 Conclusion

Although current efforts in bioinformatics have led to great progress in obtaining high-throughput microbiome sequencing data such as sequence matching, assembly of short sequences, sequence storing, indexing, and management, there has been relatively less progress in the methods for analyzing the re-

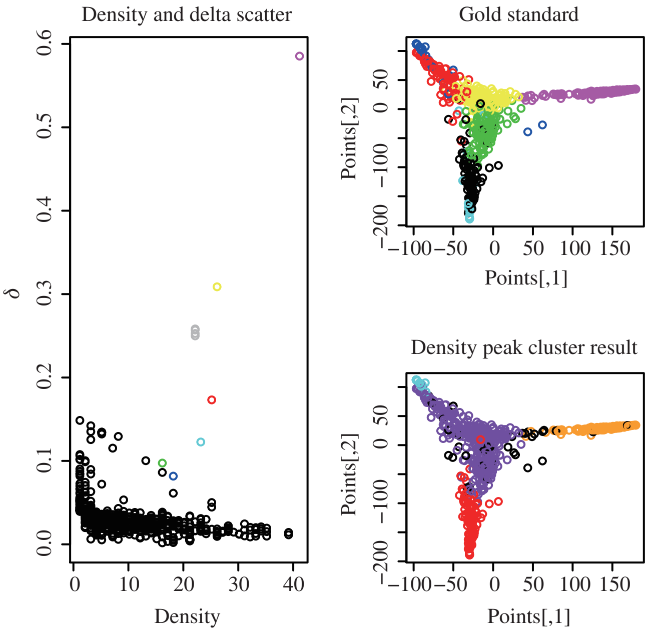

Figure 5 Density clustering on Kolmogorov distance.

processed profiles from microbiomic data [20, 21]. Current microbiomic data analysis methods do not consider these data properties and often make some unrealistic assumptions such as linear, Euclidean space, metric-space, continuous data type, which conflict with the true data properties. Current analyses of microbiome data often adopt the Euclidean distance or information distances as the measure of dissimilarity. In this paper, we propose using density clustering based on the L_1 distance to analyze microbiome data. The novel application of the L_1 distance for microbiome data analysis has not been reported previously. We showed that this method improved the clustering performance in density clustering. In future work, we will investigate the clustering accuracy of the proposed method and conduct a systematic evaluation of whether the proposed method achieves better performance with more real datasets.

Acknowledgements This research was supported by National Natural Science Foundation of China (Grant No. 61532008), International Cooperation Project of Hubei Province (Grant No. 2014BHE0017), and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (Grant No. CCNU16KFY04, CCNU14A02008).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Cani P D. Gut microbiota and obesity: lessons from the microbiome. Brief Funct Genom, 2013, 12: 381–387

2 DeWeerdt S. Microbiome: a complicated relationship status. Nature, 2014, 508: S61–S63

3 Bornigen D, Morgan X C, Franzosa E A, et al. Functional profiling of the gut microbiome in disease-associated inflammation? Genom Med, 2013, 5: 65

4 Caporaso J G, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth, 2010, 7: 335–336

5 Gevers D, Pop M, Schloss P D, et al. Bioinformatics for the human microbiome project. PLoS Comput Biol, 2012, 8: e1002779

6 Goodrich J K, Di Rienzi S C, Poole A C, et al. Conducting a microbiome study. Cell, 2014, 158: 250–262

7 La Rosa P S, Shands B, Deych E, et al. Statistical object data analysis of taxonomic trees from human microbiome data. PLoS ONE, 2012, 7: e48996

8 Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature, 2011, 473: 174–180

9 Wang J, Linnenbrink M, Kunzel S, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. Proc Nat Acad Sci USA, 2014, 111: E2703–E2710

10 Viaene L, Thijs L, Jin Y, et al. Heritability and clinical determinants of serum indoxyl sulfate and p-cresyl sulfate, candidate biomarkers of the human microbiome enterotype. PLoS ONE, 2014, 9: e79682

11 Knights D, Ward T L, McKinlay C E, et al. Rethinking "enterotypes". Cell Host Microbe, 2014, 16: 433–437

12 Chen X, Hu X H, Lim T Y, et al. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 980–991

13 Gevers D, Knight R, Petrosino J F, et al. The Human Microbiome Project: a community resource for the healthy human microbiome, PLoS Biol, 2012, 10: e1001377

14 Peterson J, Garges S, Giovanni M, et al. The NIH human microbiome project. Genome Res, 2009, 19: 2317–2323

15 Aggarwal C C, Reddy C K. Data Clustering: Algorithms and Applications. Boca Raton: CRC Press, 2013

16 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344: 1492–1496

17 Kurzyński P, Kaszlikowski D. Information-theoretic metric as a tool to investigate nonclassical correlations. Phys Rev A, 2014, 89: 012103

18 Lellouch L, Pavoine S, Jiguet F, et al. Monitoring temporal change of bird communities with dissimilarity acoustic indices. Meth Ecol Evol, 2014, 5: 495–505

19 Simpson G. CRAN task view: analysis of ecological and environmental data. 2014. https://cran.r-project.org/web/views/Environmetrics.html

20 Bourguet D, Chaufaux J, Seguin M, et al. Frequency of alleles conferring resistance to Bt maize in French and US corn belt populations of the European corn borer, Ostrinia nubilalis. Theor Appl Genet, 2003, 106: 1225–1233

21 Allen V M, Tinker D B, Hinton M H, et al. Dispersal of micro-organisms in commercial defeathering systems. Brit Poult Sci, 2003, 44: 53–59