Identifying enterotype in human microbiome by decomposing probabilistic topics into components

Xingpeng Jiang*, Jonathan Dushoff†, Xin Chen* and Xiaohua Hu*
*College of Information Science and Technology, Drexel University,
Philadelphia, PA, United States
xpjiang@drexel.edu, xc35@drexel.edu, xh29@drexel.edu
†Department of Biology, McMaster University,
Hamilton, Ontario, CA
dushoff@mcmaster.ca

***Abstract*—Discovering the global structures of microbial community using large-scale metagenomes is a significant challenge in the era of post-genomics. Data-driven methods such as dimension reduction have shown to be useful when they applied on a metagenomics profile matrix which summarize the abundance of functional or taxonomic categorizations in metagenomic samples. Analogously, model-driven method such as probability topic model (PTM) has been used to build a generative model to simulate the generating of a microbial community based on metagenomic profiles. Data-driven methods are direct and simple, they provide intuitive visualization and understanding of metagenomic profiles. Model-driven methods are often complicated but give a generative mechanism of microbial community which is helpful in understanding the generating process of complex microbial ecology. However, results from model-driven methods are usually hard to visualize and there is less an intuitive understanding of them. We developed a new computational framework to incorporate the strength of data-driven methods into model-based methods and applied the framework to discover and interpret enterotype in human microbiome.**

***Keywords*-Dimension reduction; metagenomic profile; probability topic model; non-negative matrix factorization**

I. Introduction

In the last several years, a great amount of metagenomes has been accumulated in public database which bring about huge challenge for both computational and biological scientists to understand and explore them. Computational methods have been employed to reduce the large number of dimensions in a metagenome dataset. There methods are helpful in visualizing and analyzing metagenomic profiles which summarize the abundance of functional or taxonomic categorizations in a metagenomic sample and facilitated the biological interpretation of acquired data. Typically, there are two different class of methods. The first class includes data-driven methods which are usually algebraic or statistical methods applied directly on a dataset. For example, Principle Component Analysis (PCA) has been used frequently in metagenomic studies [1]. Recently a non-negative matrix factorization (NMF) framework has been employed in analyzing metagenomic profiles to gain insights on relationships between functions, environment, and biogeography of global ocean [2]. Canonical Correlation Analysis (CCA) has been proposed for investigating the linear relationships of environmental factors and functional categorizations in global ocean [3]. Another class of methods include model-driven computational technologies. We recently provided a probability topic model framework (in particular, we used Latent Dirichlet Allocation, LDA) to model the generating process of a microbiome based on the functional or taxonomic categorization. The method has been successfully applied in the studies of human microbiome [4], [5].

Data-driven methods such as PCA and NMF are direct and simple, they provide intuitive visualization and understanding of data structure embedded in metagenomic profiles. Model-driven methods are often complicated but provide a generative mechanism of microbial community which is helpful in understanding the generating process of complex microbial ecology. However, the results of model-driven methods usually are hard to be visualized and there is less a intuitive understanding of the them. In recent studies [1], there are some general consensus about the phylogenetic composition in human gut microbiome although their variations across human population is still not clear. Generally, a large fraction of the metagenome can be firstly matched to the reference genome set on the genus and phylum level. Computational methods can be applied to analyze the phylogenetic composition profiles. It has been demonstrated that distinct clusters have been identified in human gut microbiome. In [1], multidimensional cluster analysis and principle component analysis (PCA) are performed on phylogenetic abundance profiles to further cluster 33 samples into 3 distinct clusters (a.k.a. "enterotypes"), which are identified by the levels of one of three genera: Bacteroides, Prevotella and Ruminococcus. These identified enterotypes explain neither the host properties (such as disease status, age,) nor other demographic properties. To provide a gen-

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

erative model of microbial community, we applied LDA firstly on the genus profile of the 33 metagenome samples. Furthermore, we adopted NMF to analyze and visualize the results of LDA to discover the structure of latent topics. Our results suggest that the combined framework has incorporated not only the advantages of LDA model to understand the generative mechanism of complex microbial community, but also the strength of NMF in exploratory analysis and visualizing components in metagenomic datasets.

METHODS

Datasets

We applied the proposed framework on 33 human faecal metagenomes from Sanger sequencing [1]. Phylogenetic annotation of samples was performed by aligning reads against a database of 1,511 reference genomes. Genus abundance was estimated after normalizing for genome size. Unclassified reads are removed in this study. The resulted taxonomic profile has 222 taxa in the level of genus [1].

LDA

The Latent Dirichlet Allocation model [6] is an effective probabilistic topic model firstly introduced in text mining domain to infer latent semantic topics from text documents. The LDA model allows us to study underlying concurrence patterns of the data and extract useful knowledge such as latent semantic topics. The learning process of LDA model is entirely unsupervised and therefore it is suitable for research areas which lack of labeled data. The generative process behind is that gut microbiomes (metagenome samples) are represented as random mixtures over latent topics, where each topic is characterized by a distribution over genus. We have not include the methodological description of LDA in this paper because of page limit, please see relate references for detail [4]–[6]. In our analysis, the R package "topicmodels" is used [7]. We used the "VEM" method in this package for parameter inference.

NMF

We denote that the discovered taxonomic distribution of all topics as a matrix ϕ whose columns are topics and rows are taxa. Entries in ϕ is the probability of a taxon belong to a latent topic. NMF decomposition finds matrices W and H, (with dimension $p \times k$ and $k \times s$, respectively, where k is the *rank* of our factorization) such that $WH \approx \phi$. We search for non-negative approximations that minimize the Kullback–Leibler (KL) divergence between ϕ and WH [8]. We have introduced a concordance index based on the H matrix for choosing an appropriate rank (k) for NMF analysis in the presence of overlap [2], [9]. The concordance index reflects the stability of this matrix across different realizations of the factorization, and is used to select a good decomposition rank k. Finally, we treat the symmetric,

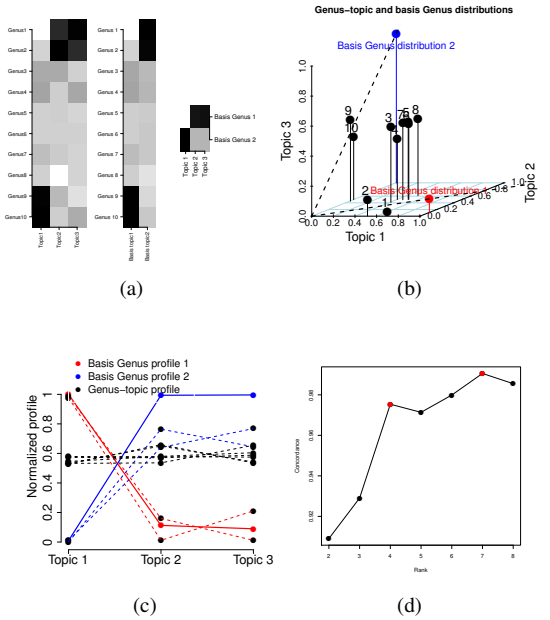


Figure 1. The illustration of our method. (a).The probability distribution of topics (in the example the number of topics is 3) over genera can be viewed as a normalized profile matrix with non-negative entries. We call this matrix genus-topic profile (The left matrix). NMF is applied on this profile matrix and the profile is approximated by the product of two low-dimension matrix (the middle and right matrices, with a rank of 2). White indicate high probability value and black indicate low probability value. (b). The projection of taxonomic profile over topics into a low-dimension space spanned by two basis taxa profiles. Each basis taxa profile belong to a topic component (the blue and red points). (c). We select representative taxas for each topic component by how "close" is a taxa to a basis taxa. In the paper, the "closeness" is calculated based on Pearson correlation coefficient. (d). The selection of appropriate rank in the 100 topics' distribution of 33 gut metagenomes based on the concordance plot of topic distribution. Rank 4 and rank 7 are selected by the model respectively because they are local peaks in the figure.

positive, similarity matrix $S = \hat{H}^T \hat{H}$ as a weighted graph-adjacency matrix, and apply spectral reordering after an "affinity" transformation. Choosing the scale r of the affinity transformation is a complex problem, we chose the value of r that minimized the Laplacian distance criterion for the untransformed matrix. We have proposed a method based on correlation coefficient [2] to select representative taxa. In the "correlation" method, we used the Pearson correlation coefficient for measuring the closeness between component and taxa [2]. Then, a threshold of θ is used to select taxas whose profiles across samples are most correlate to a component.

RESULTS AND DISCUSSION

We applied the proposed framework to discover and interpret enterotype in human microbiome. Firstly, LDA is applied on 33 gut metagenomic profiles from Sanger sequencing [1] to discover the latent "topics" in the microbial community. Then a NMF is applied on the resulted topic distribution over taxa which characterized the probability

of taxa ("terms") in each latent topic. Our previous experimental results have demonstrated the effectiveness of LDA in analyzing metagenomes of human gut [4], [5]. However, the estimation of the number of topics is a computationally extensive problem. In our new framework we don't need to estimate the number of topics in a PTM model, we simply assume that there are 100 topics and then apply NMF on the 100 topics' distribution over genus. The results of NMF give the global view of the organization structure of these 100 topics. The NMF decomposition can be thought of as an empirical attempt to describe observed taxonomic patterns in terms of a small number of taxonomic "components" (see Figure 1.a). Each component is associated with a "basis topic distribution" describing the average relative abundance of each genus in the component, and a "basis genus distribution", describing how strongly the component is represented at each topic (Figure 1.b). Thus, the observed genus-topic distribution at a topic is approximated as a weighted sum of the basis topic distribution of our components (Columns of W), with each component's profile weighted by its basis genus distribution at that topic (rows of H). In explicit terms, we approximate the observed $p \times s$ genus-topic matrix (ϕ) as the product of: a $p \times k$ matrix whose k columns are basis topic distribution for our components (W); and a $k \times s$ matrix whose k rows are the corresponding basis genus distributions (H). The demonstrative example in Figure 1, uses a factorization of rank 2 (k) to reduce a 10×3 matrix of genus abundances (X) into 10×2 matrix of basis topic profiles (W) and a 2×3 matrix of basis genus distributions (H).

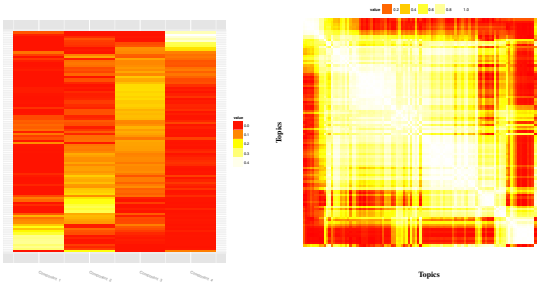
The model selection of NMF suggest that rank 4 and 7 are local peaks in the concordance plot (Figure 1.d). Because of the limit of pages, we only took 4 for this analysis although it will be interesting to see how the topics are hierarchically organized in rank 7. The "basis genus distributions" (H) for each of the four components are shown in Figure 2.a. The order of these topics is determined by spectral reordering (see methods and [9]). Each component has one or more sets of "characteristic", which have relatively high probability within that component and low abundance in other components. However, unlike some traditional clustering approaches, NMF does not restrict topics to be assigned to a single component, and in fact some topics are found in high concentration in multiple components. For example, Figure 2.a shows areas of overlap occurring on the same row between the component 2 and 3 (Figure 2.a). We then constructed a topic similarity matrix, using NMF as a *filter* (Figure 2.b). This filtered similarity matrix shows clearer patterns of clustering than we find using "direct" similarity or PCA-based filtering (data not shown). These clusters naturally overlap in many cases, illustrating the advantages of not relying on a strict clustering algorithm. The top and the last clusters (corresponding to component 1 and 4) along the dialog of the similarity matrix correspond to topic blocks dominated by a single component. However, the two and third clusters have a lot of overlapping. This can be seen by comparison with Figure 2.a.

Dually, we can visualize the structure of "basis topic distribution" (matrix W, see Figure 3). Interestingly, the clustering patterns of genus based on the distribution of basis topics are more clear than the clustering structure of topics (comparing Figure 2.b to Figure 3.b). This suggest that topics have more overlapping structure than genus indicating that LDA-discovered topics are naturally overlapped thus the NMF-filtering is necessary and helpful to understand these topics.

To better understand the functional relevance of the NMF components, we identified genera that were strongly associated with each component. We selected representative genera based on their *correlations* among their spatial distribution and the "basis genus distribution" of each component (see Figure 1.c). We found that component 1 and 3 had a suite of strongly associated taxa whose distribution across topics was strongly correlated with the basis genus distribution of the component. Component 2 has only one representative genus–Bacteroides with a very high correlation coefficient (0.91). In component 4, Prevotella also has a very high correlation (0.97). This is not surprising when we noticed that Bacteroides and Prevotella are two robust enterotypes in the previous report [1]. Another reported enterotype is related to Ruminococcus, which ranked 11 in component 3. Component 3 has not reported in the previous analysis of the same dataset [1]. Further investigation of these genus suggest that this new enterotype seems related to disease status. For example, the top correlated genus Klebsiella which is believed to be a cause of a wide range of disease states, notably pneumonia, urinary tract infections, septicemia, and soft tissue infections [10]. The second one–Shigella, causes disease in primates [11]. These results suggested that the proposed computational framework can identify not only robust enterotypes but also novel enterotypes with biologically significance.

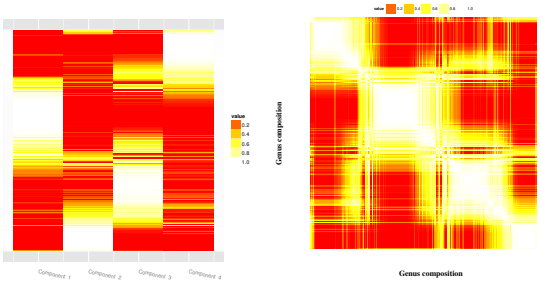
CONCLUSIONS

Interpretation of large-scale metagenomic datasets is technically challenging but it can provide important biological and biogeographic insights. The identification of complex structures and patterns in microbial communities is still at the essential part of studies in microbial ecology. Approaches has been used extensively for discovering structures in metagenomic profiles and investigating the relationship between microbial function and metadata. These methods include both data-driven methods such as PCA, NMF and CCA [1]–[3] and model-driven methods such as LDA. LDA provide a generating process for microbial communities thus provide the potential to understand the generative mechanism of microbial ecology. However,there



(a) (b)

Figure 2. (a). The "basis genus distribution" H and (b). The similarity matrix among topics based on the NMF-filtered genus distribution. The order of topics is determined by spectral reordering the similarity matrix.



(a) (b)

Figure 3. (a). The "basis topic distribution" W and (b). The similarity matrix among genus based on the filtered topic distribution. The order of genera is determined by spectral reordering the genus similarity matrix.

are several problems in this model. Firstly model selection of LDA is usually very hard and computationally extensive because of its probabilistic nature. The proposed framework in this paper is based on the combination of LDA and NMF, thus overcame these issues by taking advantage of the ability of NMF in model selection. Secondly, the results of LDA cannot be intuitively visualized and explored. The proposed method is designed to explore the complex structure in metagenomic profiles by exploring the organization of latent topics. We find that the discovered topics have distinguished clustering structure. Further analysis of the results indicate that we discovered not only robust enterotypes which are supported by previous reports but also new enterotypes which may related to disease status and worth further investigation. The extension of topic models could be used to understand temporal relationships. In future, it will be interesting to extend the proposed method to explore enterotype dynamics over time and investigate the correlations between topics and particular disease status.

ACKNOWLEDGMENT

This work was supported by the Defense Advanced Projects Research Agency under grants HR0011-05-1-0057, HR0011-09-1-0055; NSF CCF 0905291, NSF CCF 1049864, NSF IIP 1160960, NSFC 90920005 and NSFC 61170189.

REFERENCES

[1] M. Arumugam *et al.*, "Enterotypes of the human gut microbiome." *Nature*, vol. 473, pp. 174–80, 2011 May 12.

[2] X. Jiang *et al.*, "Functional biogeography of ocean microbes revealed through non-negative matrix factorization," *PLoS One, Accepted*, 2012.

[3] J. Raes *et al.*, "Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data," *Molecular Systems Biology*, vol. 7, p. 473, Mar. 2011.

[4] X. Chen, X. Hu, X. Shen, and G. Rosen, "Probabilistic topic modeling for genomic data interpretation," *IEEE BIBM*, pp. 149–152, 2010.

[5] X. Chen *et al.*, "Inferring functional groups from microbial gene catalogue with probabilistic topic models," *IEEE BIBM*, pp. 3–9, 2011.

[6] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[7] B. Grün and K. Hornik, "topicmodels: An r package for fitting topic models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 5 2011.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–91, 1999 Oct 21.

[9] X. Jiang, J. S. Weitz, and J. Dushoff, "A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data," *Journal of Mathematical Biology*, vol. 64, no. 4, pp. 697–711, 2012.

[10] R. PODSCHUN and U. ULLMANN, "Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors," 1998.

[11] K. Ryan, C. Ray, and J. Sherris, *Sherris Medical Microbiology: An Introduction to Infectious Diseases*, ser. Lange Basic Science. McGraw-Hill, 2004.