Chapter 5
Data Analysis for Gut Microbiota and Health

Xingpeng Jiang and Xiaohua Hu

Abstract In recent years, data mining and analysis of high-throughput sequencing of microbiomes and metagenomic data enable researchers to discover biological knowledge by characterizing the composition and variation of species across environmental samples and to accumulate a huge amount of data, making it feasible to infer the complex principle of species interactions. The interactions of microbes in a microbial community play an important role in microbial ecological system. Data mining provides diverse approachs to identify the correlations between disease and microbes and how microbial species coexist and interact in a host-associated or natural environment. This is not only important to advance basic microbiology science and other related fields but also important to understand the impacts of microbial communities on human health and diseases.

Keywords Microbiome • Data mining • Data analysis • Microbiota • Microbes • Diseases

5.1 Introduction

There are more and more evidences to confirm that human "microbiome" – microbes living in intimate association with us – forms a vital part of our biology and plays an important role in both health and sickness [1]. Metagenomics methods which sequence DNA without directly identifying [2] which organisms they come from and 16s rRNA sequencing [3] which sequence tag DNA for identifying the composition of organisms are two basic way of microbiome analysis.

Recently, huge amount of data are generated from plenty of microbiome projects such as Human Microbiome Project (HMP) [4, 5] and Metagenomics of the Human Intestinal Tract (MetaHIT) [6]. These datasets provide great opportunities to study

X. Jiang (✉)
School of Computer, Central China Normal University, Wuhan, Hubei 430079, China
e-mail: xpjiang@mail.ccnu.edu.cn

X. Hu
School of Computer, Central China Normal University, Wuhan, Hubei 430079, China

College of Computing & Informatics, Drexel University, Philadelphia, PA 19104, USA

B. Shen (ed.), *Healthcare and Big Data Management*, Advances in Experimental Medicine and Biology 1028, DOI 10.1007/978-981-10-6041-0_5

the unknown world of microbes. Analyzing and mining these data will help us to better understand the function and structure of microbial community of the human body, thus the relationships to our health [7, 8].

However, the huge data volume, the complexity of microbial community, and the intricate data properties have introduced challenges for microbiome data analysis and mining [9, 10]. Bioinformaticians including computer scientist, mathematician, and microbiologist work together to develop computational approaches to tackle these challenging issues, roughly focusing on the following computational tasks: (1) dimension reduction and visualization approaches to explore and visualize microbiome data, (2) statistical methods to infer true correlations and relationships between microbes and diseases, (3) computational methods to identify and extract microbial interactions from microbiome datasets, and (4) dynamic modeling and time series analysis to model the ecological system in a holistic way.

Metagenomic data analysis is a timely topic; there is great need for better algorithms to analyze complex microbiome datasets. These efforts undoubtedly will lead to biological insights on how microbes impact human health. We will breifly introduce the current advances in four aspects mentioned above in microbiome data analysis and mining.

5.2 Dimension Reduction and Pattern Identification

After the preprocessing of the metagenomic data, DNA of metagenomic or 16s rRNA sequencing technologies could be summarized by metagenomic profiles [11] which summarize the abundance of functional or taxonomic categorizations in metagenomic sequences. A metagenomic profile matrix typically has hundreds of metabolic pathways, thousands of species or tens of thousands of protein families [12]. Machine learning and multivariate statistics have been employed on the profile matrix to explore and extract the complex patterns and correlations [13]. After dimension reduction, metagenomic profiles are usually represented by several "components" which may facilitate biological interpretation and discovery [14].

For example, PCA has been used frequently in metagenomic profiles to characterize the relationship of metagenomic samples [15]. Another method – MDS – which is based on the dissimilarities of data instead of similarity in PCA has been adopted as a standard technology for visualizing the taxonomic relationships in microbial communities [15]. Recently, a nonnegative matrix factorization (NMF) framework has been used in analyzing metagenomic profiles to gain a different and complementary perspective on relationships between functions, environment, and biogeography of global ocean and soil environment [16–18].

Microbiome datasets can be represented by metabolic paths, taxonomic assignment, or gene families [19]. To integrate information from multiple views, data integration approaches can be used to combine multi-view information simultaneously to obtain a comprehensive view which reveals the underlying data structure shared by multiple views [20]. A novel variant of symmetric nonnegative matrix factorization (SNMF) [21], called Laplacian regularized joint symmetric

nonnegative matrix factorization (LJ-SNMF) has been proposed for this purpose. We conduct extensive experiments on several realistic datasets including Human Microbiome Project (HMP) data [4, 5]. The experimental results show that the proposed method outperforms other variants of NMF, which suggests the potential application of LJ-SNMF in clustering multi-view datasets.

Furthermore, linear correlation or regression methods are also employed to investigate the relationships among taxa or functions and their relationships to existing environmental or physiological data (metadata) such as Pearson Correlation and Canonical Correlation Analysis (CCA) [22]. CCA has been proposed for investigating the linear relationships of environmental factors and functional categorizations in global ocean [23].

The vast majority of methods employed in current metagenomics analysis are under the hypothesis that structures and relationships in a microbial community are linear. However, the interactions among microbiota are most likely nonlinear, and the mathematical spaces of microbiota are most likely in a manifold [24] or probabilistic space [25, 26] instead of Euclidean space. We could visualize and explore these structures using only several components which are the intrinsic dimensions discovered by manifold and probabilistic models. This provides a mechanistic understanding of how a microbial community is generated by probabilistic mixing of microbial components as well as a powerful tool for exploring the temporal dynamics of microbiome composition.

Finally, many kinds of nonlinear relationships such as taxa-taxa patterns and function-environment correlations could be investigated using the nonlinear statistical methods. We summarize these steps in a computational framework (see Fig. 5.1). The computational framework is based on our current understanding of metagenomic data, and we will integrate the advanced nonlinear dimension reduction methods and statistical methods to discover novel relationships.

5.3 Relationship and Correlation

Another important problem in microbiome analysis is to identify the biomarkers (i.e., bacterial taxa, microbial genes, or pathways) that are associated with disease, where the microbiome data are summarized as the composition of the bacterial taxa, protein families, or metabolic pathways at different levels [27]. To discover biomarkers for diseases or environmental factors, the most common approaches focus on regression techniques incorporating the complex interaction patterns among species (or gene functions). We have developed a new regression framework called "manifold-constrained regularization" (McRe) [28], which inherits the strength of manifold embedding for regularization of linear regression. This method can incorporate species interaction network as prior information to infer novel relationships.

Several studies consider the regression analysis of microbiome compositional data, where the goal is to identify the biomarkers that are associated with a continuous response such as the body mass index (BMI) [9]. Compositional data are strictly positive and multivariate that are constrained to have a unit sum. Lin

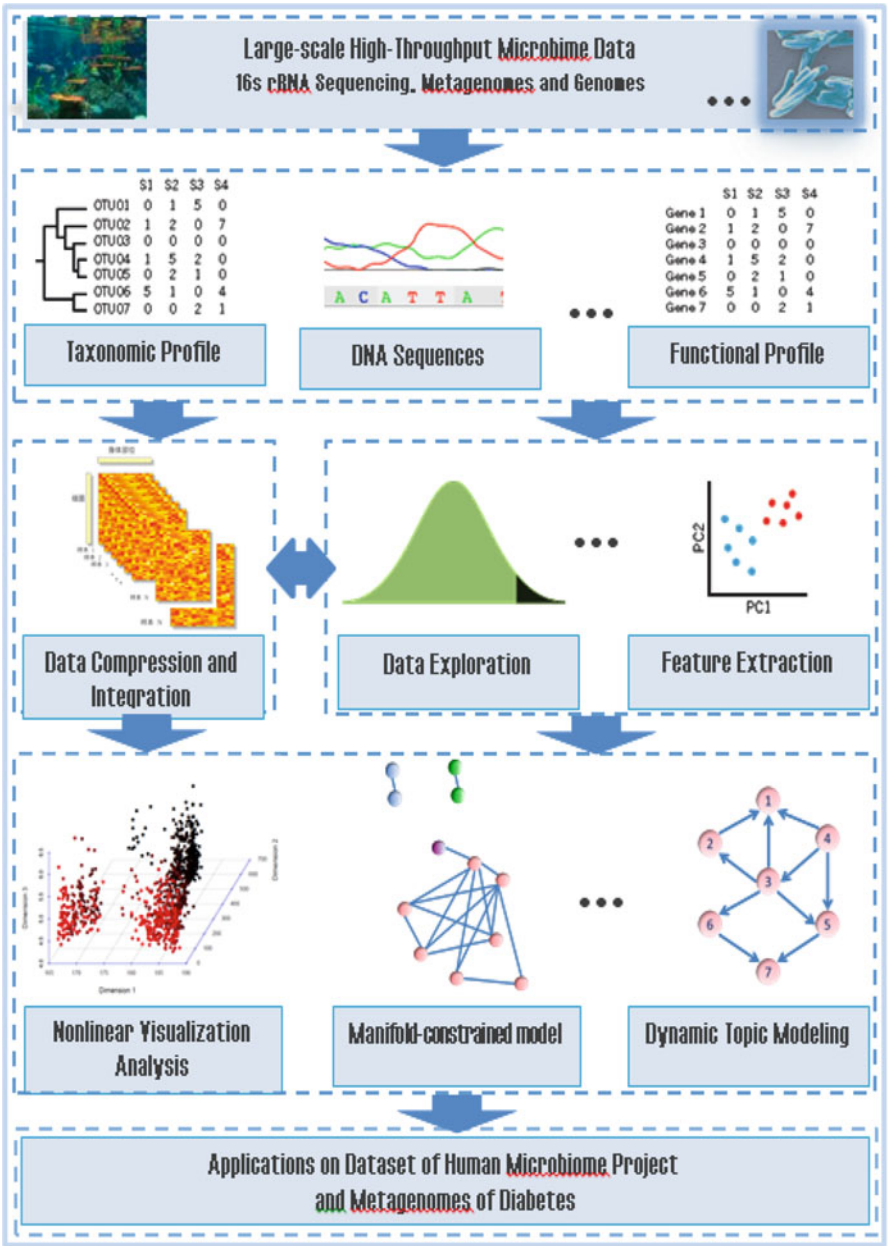

Fig. 5.1 Nonlinear analysis framework for metagenomic profiles

et al. [29] proposed a variable selection procedure for such models in high-dimensional settings and derived the weak oracle property of the resulting estimates [29]. Shi et al. [30] proposed a penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints which

is developed [30]. This provides valid confidence intervals of the regression coefficients and can be used to obtain the p-values which could be used to measure statistical significance [30]. Randolph et al. have formulated a family of regression models that naturally extends the dimension-reduced graphical explorations common to microbiome studies; the method could be viewed as a penalized version of the low-dimensional linear model for compositions [31].

5.4 Networking Microbiome

A network perspective provides unprecedented opportunities for integrating and analyzing big microbiome data for studying the structure and the function of microbial communities [32]. A microbial interaction network (MIN, e.g., species-species interaction network) shapes the structure of a microbial community and forms its ecosystem function and principle, i.e., the regulation of microbe-mediated biogeochemical processes [33]. Deciphering interspecies interaction is challenging in the wet lab due to the difficulties of coculture experiments and the complicated patterns of species interactions [34]. The knowledge of these small-scale microbial interactions such as pairwise competitions is often distributed widely in various media including PubMed literatures, biological databases, Wikipedia documents, etc., making it difficult to integrate and analyze [35]. Researchers have started to infer pairwise interspecies interactions such as competitive and cooperative interactions leveraging to heterogeneous microbial data including metagenomes, microbial genomes, and literature data. These efforts have facilitated the discovery of previously unknown principles of MIN, verified the consistency, and resolved the contradiction of the application of macroscopic ecological theory in microscopic ecology [36].

Species interact in a complex style with many types of interactions unknown. Previous works on species inference based on metabolic methods are based on the following two approaches. Bornstein proposed a computational method for inferring pairwise interactions from reconstructed metabolic network of species with whole-genome sequences available publicly [37]. The method can identify pairwise competitive and cooperative interactions. Another way is using *flux balance analysis* (FBA) models [38] to infer species interaction when the metabolic model of a species (or strain) is available [39].

More than 100 genome-scale metabolic network models were published. Constraint-based modeling (CBM) was already used for the inference of three potential interactions [40]: negative, where two species compete for shared resources; positive, where metabolites produced by one species are consumed by another producing a synergistic co-growth benefit; and neutral, where co-growth has no net effect. By using the FBA simulation community metabolic network, we can find key enzymes and reactions in the metabolic network, thus acting as a potential environmental and physiological fingerprint. In a two species system, the CBM solver aims to explore the type of interactions by comparing the total biomass

Fig. 5.2 A constraint-based modeling to model pairwise interaction

$$v_{BM,AB} = max \sum_{m \in \{A,B\}} v_{BM,m} \quad (1)$$

Subject to:

$$SV = 0$$

$$v_{i,min} \leq v_i \leq v_{i,max}$$

$$v_i \in V$$

production rate (denoted AB) in the pairwise system to the sum of corresponding individual rates recorded in their individual growth (denoted A+B). The CBM model is defined in Fig. 5.2, where ^{V}BM,m is the maximal biomass production rate in a system m, corresponding to species A and B. When AB << A+B, A and B have a competitive relationship.

5.5 Dynamics and Time Series Analysis

Microbial abundance dynamics along the time axis can be used to explore complex interactions among microorganisms [41]. It is important to use time series data for understanding the structure and function of a microbial community and its dynamic characteristics with the perturbations of the external environment and physiology [42]. Current studies usually use time sequence similarity [43], or clustering time series data for discover dynamic microbial interactions; these methods often do not take the full advantage of the time sequences. Thus the interactions among microorganisms cannot be accurately predicted. We have explored a vector autoregression (VAR) model [44] to lift the limitations of traditional methods. VAR models and interaction inference: Due to the high-dimensional nature of microbiomics data, the number of samples is far greater than the number of microorganisms; direct interaction inference by VAR is not feasible. In our previous studies, we have designed several graph regularization-based VAR (GVAR) methods for analyzing the human microbiome. We found that our approach improves the modeling performance significantly on several microbiome dataset. The experimental results indicate that graph regularization achieves better performance than other sparse VAR model based on elastic net regularization. However, the interpretation of the inference results is hard and far from complete. Furthermore, graph regularization, despite a classic manifold regularization method, suffers some problems because of its weak extrapolating ability. A novel regularization – Hessian regularization [45] – which fits the data perfectly and extrapolates nicely to unseen data will be utilized to overcome the issue.

In the future, state-space model [46] or probabilistic Boolean network model [47] could be used for modeling large-scale microbiome data for application. We will extend these methods by integrating specific information of the microbiomics data. The state-space model is a powerful method for simulation of dynamical

systems, and it is widely used in engineering control systems which is a dynamic time-domain model to imply time as the independent variable. It is possible to extend the state-space model by considering the species delay in the regulatory network of relationships, not just describe the level of species richness' impact on the internal state, and assume that the internal state can independently evolve. Species with time delay regulatory network of relationships are better suitable for microbial interactions, because the regulation between microorganisms is often a slow process with delay, rather than an instantaneous process.

5.6 Conclusion

The data from Human Microbiome Project (HMP) [4, 5], which includes more than 5000 samples with profiles of hundreds of taxonomic or functional categorizations, are constructed from 15 or 18 distinct body sites of 242 individuals. Methodological development is still in its infancy for effectively analyzing and mining the data. Many microbiome dataset are also from various studies focusing on disease, diets, and other investigations. These data have created a great opportunity for understanding and also a tremendous computational and theoretical challenge. There is a great need to develop novel mathematical and computational methods for finding nonlinear signal and patterns in human-associated microbial metagenomes.

The identification of complex structures and patterns of microbial communities is at the essential part of studies in microbial ecology. The expected method helps shed light on discovering the complex relationships among microbes. In the future, nonlinear methods should be considered as an important tool in analyzing metagenomics, not only because microbial function can be viewed at multi-scales, from individual genomes to communities to global cycles, but also the complex interaction across scales.

References

1. Shreiner AB, Kao JY, Young VB (2015) The gut microbiome in health and in disease. Curr Opin Gastroenterol 31(1):69–75
2. Comparative metagenomics of microbial communities. Science. [Online]. Available: http://science.sciencemag.org/content/308/5721/554. Accessed 04 Feb 2017
3. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and Eucarya. PNAS 87(12):4576–4579
4. T. H. M. P. Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486(7402):207–214
5. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI (2007) The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature 449 (7164):804–810
6. Qin J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65

7. Rieder R, Wisniewski PJ, Alderman BL, Campbell SC (2017) Microbes and mental health: a review. Brain Behav Immun, In Press
8. Dzutsev A, Badger JH, Perez-Chanona E et al (2017) Microbes and Cancer. Annu Rev Immunol 35:199–228
9. Tsilimigras MCB, Fodor AA (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 26(5):330–335
10. 2015 Microbiome, metagenomics, and high-dimensional compositional data analysis. Ann Rev Stat Appl 2(1):73–94
11. Xiao K-Q et al (2016) Metagenomic profiles of antibiotic resistance genes in paddy soils from South China. FEMS Microbiol Ecol 92(3), fiw023
12. Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19(7):1141–1152
13. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087830. Accessed 04 Feb 2017
14. Jiang X, Hu X, Xu W, He T, Park EK (2013) Comparison of dimensional reduction methods for detecting and visualizing novel patterns in human and marine microbiome. IEEE Trans Nanobioscience 12(3):199–205
15. Tyler AD, Smith MI, Silverberg MS (2014) Analyzing the human microbiome: a 'How To' guide for physicians. Am J Gastroenterol 109(7):983–993
16. Bartram AK et al (2014) Exploring links between pH and bacterial community composition in soils from the Craibstone experimental farm. FEMS Microbiol Ecol 87(2):403–415
17. Jiang X et al (2012) Functional biogeography of ocean microbes revealed through non-negative matrix factorization. PLOS ONE 7(9):e43866
18. Jiang X, Weitz JS, Dushoff J (Mar. 2012) A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. J Math Biol 64(4):697–711
19. Arumugam M et al (2011) Enterotypes of the human gut microbiome. Nature 473 (7346):174–180
20. Personalized microbial network inference via multi-view clustering of oral metagenomics data – TiFN. [Online]. Available: http://www.tifn.nl/publication/personalized-microbial-network-inference-via-multi-view-clustering-of-oral-metagenomics-data/. Accessed 04 Feb 2017
21. Kuang D, Ding C, Park H (2012) Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of the 2012 SIAM international conference on data mining (0 vols). Society for Industrial and Applied Mathematics. pp 106–117
22. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P (2011) Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol Syst Biol 7 (1):n/a–n/a
23. Patel PV, Gianoulis TA, Bjornson RD, Yip KY, Engelman DM, Gerstein MB (2010) Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. Genome Res 20(7):960–971
24. He X, Cai D, Yan S, Zhang H-J (2005) Neighborhood preserving embedding. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 2:1208–1213. Vol. 2
25. Chen X, Hu X, Shen X, Rosen G (2010) Probabilistic topic modeling for genomic data interpretation. In: 2010 I.E. International Conference on Bioinformatics and Biomedicine, BIBM 2010, Hong Kong, China, December 18–21, 2010, Proceedings, pp 149–152
26. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing | bioRxiv. [Online]. Available: http://biorxiv.org/content/early/2016/09/22/076836. Accessed 04 Feb 2017
27. Dietert RR, Silbergeld EK (2015) Biomarkers for the 21st century: listening to the microbiome. Toxicol Sci 144(2):208–216
28. Jiang X, Hu X, Xu W, Wang Y (2013) Manifold-constrained regularization for variable selection in environmental microbiomic data. In: 2013 I.E. International Conference on Bioinformatics and Biomedicine, Shanghai, China, December 18–21, 2013, pp 86–89

29. Lin W, Shi P, Feng R, Li H (2014) Variable selection in regression with compositional covariates. Biometrika 101(4):785–797
30. Shi P, Zhang A, Li H (2016) Regression analysis for microbiome compositional data. arXiv:1603.00974 [stat]
31. Randolph TW, Zhao S, Copeland W, Hullar M, Shojaie A (2015) Kernel-penalized regression for analysis of microbiome data. arXiv:1511.00297 [stat]
32. Faust K, Raes J (2012) Microbial interactions: from networks to models. Nat Rev Micro 10 (8):538–550
33. Fuhrman JA (2009) Microbial community structure and its functional implications. Nature 459 (7244):193–199
34. Fritz JV, Desai MS, Shah P, Schneider JG, Wilmes P (2013) From meta-omics to causality: experimental models for human microbiome research. Microbiome 1:14
35. @MInter: automated text-mining of microbial interactions | Bioinformatics | Oxford Academic. [Online]. Available: https://academic.oup.com/bioinformatics/article-abstract/32/19/2981/2196520/MInter-automated-text-mining-of-microbial?redirectedFrom=fulltext. Accessed 04 Feb 2017
36. Cordero OX, Datta MS (2016) Microbial interactions and community assembly at microscales. Curr Opin Microbiol 31:227–234
37. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation | BMC Bioinformatics | Full Text. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0588-y. Accessed 04 Feb 2017
38. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28 (3):245–248
39. Constructing and analyzing metabolic flux models of microbial communities | KBase
40. Shoaie S, Nielsen J (2014) Elucidating the interactions between the human gut microbiota and its host through metabolic modeling. Front Genet 5
41. Gerber GK (2014) The dynamic microbiome. FEBS Lett 588(22):4131–4139
42. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. [Online]. Available: http://www.pnas.org/content/108/Supplement_1/4554.short. Accessed 04 Feb 2017
43. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates | BMC Systems Biology | Full Text." [Online]. Available: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-5-S2-S15. Accessed 2017
44. Jiang X, Hu X, Xu W, Park EK (2015) Predicting microbial interactions using vector autoregressive model with graph regularization. IEEE/ACM Trans Comput Biology Bioinform 12(2):254–261
45. Ma Y, Hu X, He T et al (2016) Hessian regularization based symmetric nonnegative matrix factorization for clustering gene expression and microbiome data[J]. Methods 111:80–84
46. Rangel C et al (2004) Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics 20(9):1361–1372
47. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks | Bioinformatics | Oxford Academic. [Online]. Available: https://academic.oup.com/bioinformatics/article/18/2/261/225574/Probabilistic-Boolean-networks-a-rule-based. Accessed 04 Feb 2017