Machine Learning Approaches for Metagenomics

Huzefa Rangwala, Anveshi Charuvaka and Zeehasham Rasheed

Department of Computer Science,
George Mason University,
Fairfax, Virginia, USA
rangwala@cs.gmu.edu, {acharuva,zrasheed}@gmu.edu

Abstract. Microbes exists everywhere. Current generation of genomic technologies have allowed researchers to determine the collective DNA sequence of all microorganisms co-existing together. In this paper, we present some of the challenges related to the analysis of data obtained from the community genomics experiment (commonly referred by metagenomics), advocate the need of machine learning techniques and highlight our contributions related to development of supervised and unsupervised techniques for solving this complex, real world problem.

1 Background

Advances in genome-sequencing have transformed the manner of characterizing large populations of microbial communities, that are ubiquitous across several environments. The process of "metagenomics" involves sequencing of the genetic material of all organisms co-existing within ecosystems ranging from ocean, soil and the human body. (can be referred to as community genomics). Orthogonally, proteomics and mass spectrometry allow the study of bio-transformations due to these microbial communities in the form of metaproteomes and metabolomes, respectively. Several researchers and clinicians have embarked on studying the pathogenic role played by the microbiome (i.e., the collection of microbial organisms within the human body) with respect to human health and disease conditions. In a similar effort, other groups of researchers are using the metagenomics technology to characterize different ecological environments across the planet (also referred by "Earth Microbiome").

Annotating microbial sequences (reads or quasi-assembled contigs) within a sample is a challenging task due to the unknown, diverse and complex nature of microbial communities within the different environments. There is a critical need to develop mining methods that can characterize metagenome data in terms of taxonomy, function and metabolic potential, and correlate the multi-modal, microbial data to clinical or environmental metadata.

We present our ongoing efforts that have lead to the development of novel supervised learning approaches and scalable clustering methods to solve these real world challenges.

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 512–515, 2014.

2 Large Scale Metagenome Clustering

The sequencing technologies of today do not provide the complete genome for the micro-organisms, but produce short, contiguous subsequences (referred to as reads) that are fragmented from random positions of the entire genome. The problem of metagenome sequence assembly involves stitching together different reads (e.g., overlapping the prefixes and suffixes of smaller subsequences) to produce organism-specific contiguous genomes. Other challenges are introduced due to the varying abundance, diversity, complexity, genome lengths of previously uncultured (or never sequenced before) microbes within different communities. Genomic technologies also produce large number of sequence reads, and reads that may have varying error idiosyncrasies [5]. As such, the metagenome assembly and analysis problem is complex and challenging [2]. Targeted metagenomics or 16S rRNA gene sequencing provides a first step for the quick and accurate characterization of microbial communities. 16S sequences are marker genes, which exists in most microbial genomes and have a conserved portion for detection (primer development) and a variable portion that allows for categorization within different taxonomic groups [7]. Targeted metagenomics are also effective in detecting species with low abundances. However, they may not be good in discovering unique species (orphans) that have never been sequenced before.

Several algorithms have been developed to analyze targeted metagenomes (16S rRNA marker gene) and whole metagenome samples [5]. Clustering/binning approaches involve the unsupervised grouping of sequences that belong to the same species. Successful grouping of sequence reads has several advantages: (i) it improves the metagenome assembly, (ii) it allows computation of species diversity metrics and (iii) it serves as a pre-processing step by reducing computational complexity within several work-flows that analyze only cluster representatives, instead of individual sequences within a sample.

Contributions: We have developed a locality sensitive hashing (LSH) for binning 16S sequences [11,10] called MC-LSH. We further extended the approach using minwise hashing [1] (called MC-MinH[9]) to operate on unequal length sequences and evaluate the approach for both, 16S and whole metagenome sequences. We also extended the minwise hashing algorithm to develop a scalable Map-Reduce based algorithm for metagenome clustering. We refer to this approach as MrMC-MinH [8]. The key contributions of this work included the development of a distributed map-reduce based implementation of clustering algorithm and the ability to perform hierarchical agglomerative clustering instead of greedy clustering as in MC-LSH and MC-MinH.These developed methods provide key biodiversity estimation metrics that are used by biologists.

3 Multiple Hierarchical Classification

The relationship between the microbial communities and human health (or environment) is characterized by first identifying the content, abundance and variance of the microbes across different samples. For an understanding of the microbial-host interaction, it is also crucial to determine the “functionality” and “metabolic potential” induced by these microbes. As such, there is a need to develop methods that annotate the metagenome in terms of “taxa” content and further characterize the ORFs (predicted from metagenomes [12,4]) and transcript sequences in terms of functional and metabolic activity.

The past decade has also seen an explosion in the number of diverse databases that are curated and maintained by different researchers with varying expertise and interests. These databases have a unique characteristic i.e., the data is structured as a hierarchy. We seek to develop approaches that can benefit from jointly learning the prediction models for taxonomy, function and metabolic potential. Different hierarchical databases have implicit similarities between them. For example, the Gene Ontology database has 27 different mappings from other annotation databases, defined using manual and semi-automated procedures. The basis for these mapping include use of sequence, literature search, evolutionary information and structure information.

Contributions: Towards this end, we have developed regularized multi-task learning models [3] that leverage the existing hierarchical structure present in the annotation databases. The models also leverage the implicit relationships between the different databases available for the same annotation problem (e.g., KEGG and MetaCyc for metabolic potential).

Given, multiple hierarchical source databases; within our formulation the objective is to classify an instance accurately across all the multiple hierarchies.

For each of the different classes across multiple hierarchies, we define a binary classification task. These classification tasks predict whether an example belongs to the particular class or not. However, instead of training each of these tasks independently (single task learning), the training for all these tasks are combined using the MTL approach [3] The rationale for the proposed approach is that each of the binary tasks are related “within” the hierarchy due to the explicit structure in the databases. Across the hierarchical sources, it is expected that if the underlying relationships are modeled well, it will benefit the generalization performance for individual annotation problems. Further, the MTL approach is also suited for tasks (classes) that have scarce training examples. This MTL approach leveraged the underlying relationships between the multiple hierarchies and significantly outperformed traditional prediction models for classifying sequence data within multiple hierarchical annotation databases. We also extended this approach to classify text documents across large archives like WikiPedia and DMOZ Web directory [6].

4 Conclusion and Future Work

In summary, we presented a set of clustering and classification approaches to analyze metagenome associated data. Using the annotated sequences, we plan to extract an aggregated taxonomic, functional and metabolic activity profile for the microbial samples. These profiles will allow us to compare metagenome samples and correlate the information to clinical or environmental metadata; and train supervised phenotypic classifiers. All the developed tools are freely available and integrated within bioinformatics work-flows that are easy to use for the biology community.

Acknowledgment. Support by NSF Career to HR, IIS 1252318. All software are linked via project webpage at `http://www.cs.gmu.edu/~hrangwal/microbiomeCareer`

References

1. Broder, A.Z., Charikar, M., Frieze, A.M., Mitzenmacher, M.: Min-wise independent permutations. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 327–336. ACM (1998)
2. Charuvaka, A., Rangwala, H.: Evaluation of short read metagenomic assembly. BMC genomics12 (suppl. 2), S8 (2011)
3. Charuvaka, A., Rangwala, H.: Multi-task learning for classifying proteins with dual hierarchies. In: IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, pp. 834–839. IEEE (December 2012)
4. Delcher, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L.: Identifying bacterial genes and endosymbiont dna with glimmer. Bioinformatics 23(6), 673–679 (2007)
5. Hugenholtz, P., Tyson, G.W.: Microbiology: metagenomics. Nature 455(7212), 481–483 (2008)
6. Naik, A., Charuvaka, A., Rangwala, H.: Classifying documents within multiple hierarchical datasets using multi-task learning. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 390–397. IEEE (2013)
7. Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A., Versalovic, J.: Metagenomic pyrosequencing and microbial identification. Clinical Chemistry 55(5), 856–866 (2009)
8. Rasheed, Z., Rangwala, H.: A map-reduce framework for clustering metagenomes. In: Proceedings of the 12th IEEE International Workshop on High Performance Computational Biology (HiCOMB), Boston, MA. IEEE (May 2013)
9. Rasheed, Z., Rangwala, H.: Mc-minh: Metagenome clustering using minwise based hashing. In: SIAM International Conference in Data Mining (SDM), Austin, TX. SIAM (May 2013)
10. Rasheed, Z., Rangwala, H., Barbara, D.: Efficient clustering of metagenomic sequences using locality sensitive hashing. In: SIAM International Conference in Data Mining, Anaheim, CA, pp. 1023–1034. SIAM (April 2012)
11. Rasheed, Z., Rangwala, H., Barbara, D.: LSH-Div:species diversity estimation using locality sensitive hashing. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Philadelphia, USA. IEEE (October 2012)
12. Zhu, W., Lomsadze, A., Borodovsky, M.: Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38(12), e132 (2010)