Sequence-based Deep Learning Reveals the Bacterial Community Diversity and Horizontal Gene Transfer

Hao Zheng¹, Jie Yin², Jieruo Gu³* and David Xingfei Deng⁴*

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, CA, USA
²Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN, USA
³Internal Medicine, The 3rd Hospital of Sun Yat-Sen University, Guangzhou, China 510630
⁴Ardent BioMed Guangzhou Inc., Guangzhou, China 510530
*Corresponding authors (daviddeng@ardentbiomed.com)

Abstract - Metagenomics is the application of advanced genomic techniques to the study of microbial communities in their native environments, and does not require lab cultivation nor isolation of individual genomes. Here, we propose a novel approach for metagenomics operational taxonomic unit (OTU) assignment using deep learning. The experimental results demonstrate that a PCA-based convolutional neural network is very powerful and fast for metagenomics OTU assignment and also prioritizing horizontal gene transfer events.

Keywords: OTU, metagenomics, sequencing, deep learning

1 Introduction

Metagenomics studies the whole microbial communities in their native environments by using advanced computational and genomic techniques, and does not need lab cultivation nor isolation of individual genomes. It is a very powerful method to reveal sample relative abundance and the genetic material of a microbe or entire communities of organisms directly from its environment without losing its nativity [1].

NGS-based metagenomics is very hot and has been popularly used in various studies. For example, human gut microbiomes were studied by using fecal samples from individuals of various ages through computational biology approaches, leading to the discovery of age-related DNA commonality and uncertainty [13][17]. Researchers also collected the Seawater samples from the Sargasso Sea and the diversity of microbial communities and gene signatures were analyzed through WGS sequencing [10]. This initial project was further expanded into the Sorcerer II Global Ocean Sampling expedition, in which a total of 44 samples were gathered from the Northwest Atlantic through Eastern Tropical Pacific and sequenced using NGS-based biological approaches to study microbial genomes and the richness in the surface seawater [11][12]; Approaches alike were also explored to study the bacteria lives and their functional potentials in vagina, human saliva, human skin, and soils [6]-[9][14][15].

To determine both the community composition and the potential physiology of the abundant community members in the metagenomics sample, OTU assignment is a
critical step [23]-[25]. Metagenomic OTU assignment is usually done based on the contigs or unassembled reads from the sequencing data to cluster these contigs or reads into closely related populations [27]. Depending on the structure of the microbial community, the quality and depth of reads, OTU assignment can be performed at different levels, e.g. to classify reads of the same superkingdom into high levels like phyla or even down to the finer species levels.

For such OTU assignment, a number of in silico methodologies have been devised recently. These methods can be generally split into two main categories: (a) supervised methods and (b) unsupervised methods [16]. Most supervised methodologies are relying on some sort of taxonomical information. To that end, reference databases are usually needed for the assigning taxonomical levels to contigs or reads. The intrinsic algorithm utilizes either sequence similarity, or genomic signatures such as tandem nucleotide sequencing composition patterns. Such methods include MEGAN, Phylopythia, NBC, PhymmBL, and SPHINX [18]-[22].

In comparison, clustering approaches are usually independent of taxonomy. Such methods generally requires no additional reference databases nor taxonomic information. The basic idea behind taxonomy independent methods is that reads from different species may have some intrinsic patterns. For instance, different α-proteobacteria species may have guanine-cytosine contents spanning <30% to >60% [28]. Popular methods under this category include TETRA, variants of SOMs, CompostBin, AbundanceBin and MetaCluster [29]-[33].

2 Materials and Methods

2.1 Data Sets

Because of the nature of metagenomics, we don’t have the truth for OTU. To benchmark the performance, we used in-house simulated data sets that were generated from available bacterial genomes. The dataset is further partitioned to serve as training and testing based on cross validation schema to be described below.

We also tested the performance of the proposed method on a very comprehensive synthetic metagenomic data set named simHC [26]. SimHC simulated a high-complexity microbial community by the Integrated Microbial Genomes and Microbiomes system of JGI. The lengths of the genomic fragments in the simHC data set span from 130 up to 3,754 bps. And, about 2.6% of nucleotides in simHC were not specified, mimicking the noisiness encountered during next generation sequencing.

2.2 Preprocessing

The raw metagenomics data is usually noisy and may contain DNA sequences of eukaryotic origin. Preprocessing of the raw data is a critical upstream step to ensure meaningful prefiltering includes removal of redundant, low quality sequences and sequences of eukaryotic origin [3][4]. Oligo frequencies patterns up to hexamer are extracted. As the pattern often contains
redundant information, mapping it to a feature vector can get rid of this redundancy and yet preserve most of the intrinsic information content of the pattern. These extracted features have great role in distinguishing input patterns.

2.3 Deep Learning

We used principal component analysis (PCA) for extracting the features to serve as the inputs to deep learning. Based on the eigenvectors produced by the PCA methods, the projection vectors from the training set were obtained and then serves as the inputs to the neural network. Usually, we need a lot of training epochs to learn meaningful weights, or we require related data sets to be used for seeding a fine-tuning of transfer learning network. Here, we turn a PCA into an auto-encoder, by generating an encoder level of the PCA parameters and furthermore adding a decoder level [34].

To extract meaningful and representative features from a high dimensional space is usually challenging. Such a problem is well known as curse of dimensionality. To address this problem, we utilized PCA transformation to serve as the inputs to the neural network [2]. More specifically, we then set the weight matrix of the neutral network as $\Theta$ and as a result the initialization cost of the network depends only on the number of samples serving to obtain the principal components.

Relu activation function was used over traditional sigmoid and tanh function. The Relu function is defined as follows

$$f(x) = \max(0, x) \quad (1)$$

Relu activation function has its unique advantages over traditional sigmoid and tanh function. The training of the network using Relu is much faster, with reduced likelihood of the gradient to vanish. Also, sparse representations resulting from the Relu activation function is normally more beneficial than dense representations.

2.4 Training Schema

57% of the samples are used for training. These are presented to the network during training, and the network is adjusted according to its error. 10% of the samples are used for validation. These are used to measure network generalization, and to halt training when generalization stops improving. The rest 33% of the samples are reserved for testing. These have no effect on training and so provide an independent measure of network performance during and after training. Training will stop by itself at the time generalization stops getting improving. This is indicated by an increase in the cross-entropy error of the validation sample set.

3 Results

The results demonstrate that this sequence-based deep learning method can reveal the bacterial community diversity with high accuracy and indicate potential underlying horizontal gene transfer events.

3.1 Performance

Figures 1-2 illustrate that the performance of the proposed method outperforms two popular existing methods (TETRA and Phylopythia) using both in house simulated
datasets and widely used simHC data set in terms of specificity and sensitivity.

Figure 1. Comparison of Specificity and Sensitivity of the proposed method versus popular existing methods at the superkingdom level using simulated metagenome data.

Figure 2. Comparison of Specificity and Sensitivity of the proposed method versus popular existing methods at superkingdom and Phylum levels using widely used simHC data set.

3.2 Horizontal Gene Transfer

Horizontal gene transfer (HGT) is the transfer of genetic materials between organisms other than by the transmission of it from parent to offspring. HGT is commonly observed among prokaryotes (e.g. from archea to bacteria). Obviously, this phenomenon adds another layer of complexity to the OTU assignment of metagenomic fragments and would confused the classifier in making the decision. As a result, the classification results can in turn reflect the potential underlying HGT events among prokaryotes. We observe that relatively highly misclassification among certain organisms, such as Escherichia coli, Bacillus subtilis, and Methanobacterium thermoautotrophicum, indicating the potential HGT event among these organisms, as supported from literature [35].

4 Conclusions

Metagenomics has gained tremendous attention with the advance of computer engineering and bioinformatics [1][36]. In this paper, we have investigated a new approach for metagenomic OTU assignment using PCA-initialized deep learning using Relu activation. We demonstrated that this proposed method is very efficient to tackle the problem of metagenomics OTU assignment.

5 References


