

## RESEARCH ARTICLE

# TACO: Taxonomic prediction of unknown OTUs through OTU co-abundance networks

Zohreh Baharvand Irannia<sup>1</sup> and Ting Chen<sup>1,2,\*</sup>

<sup>1</sup> Program in Computational Biology and Bioinformatics, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

<sup>2</sup> Bioinformatics Division, TNLIST, Tsinghua University, Beijing 100084, China

\* Correspondence: tingchen@tsinghua.edu.cn

Received December 28, 2015; Revised March 22, 2016; Accepted March 23, 2016

**Background:** A main goal of metagenomics is taxonomic characterization of microbial communities. Although sequence comparison has been the main method for the taxonomic classification, there is not a clear agreement on similarity calculation and similarity thresholds, especially at higher taxonomic levels such as phylum and class. Thus taxonomic classification of novel metagenomic sequences without close homologs in the biological databases poses a challenge.

**Methods:** In this study, we propose to use the co-abundant associations between taxa/operational taxonomic units (OTU) across complex and diverse communities to assist taxonomic classification. We developed a Markov Random Field model to predict taxa of unknown microorganisms using co-abundant associations.

**Results:** Although such associations are intrinsically functional associations, we demonstrate that they are strongly correlated with taxonomic associations and can be combined with sequence comparison methods to predict taxonomic origins of unknown microorganisms at phylum and class levels.

**Conclusions:** With the ever-increasing accumulation of sequence data from microbial communities, we now take the first step to explore these associations for taxonomic identification beyond sequence similarity.

**Availability and Implementation:** Source codes of TACO are freely available at the following URL: <https://github.com/baharvand/OTU-Taxonomy-Identification> implemented in C++, supported on Linux and MS Windows.

**Keywords:** metagenomics; 16s rRNA gene; taxonomic profiling; taxonomic prediction; Markov Random Field; OTU co-abundance network

## INTRODUCTION

Microorganisms are present in almost every habitat on Earth, such as soil, sea water, fresh water, human body, and air [1]. They play fundamental roles in every aspect of life, functioning as an essential component in nutrient cycles, breaking down toxic wastes to safe materials, changing the climate, and affecting human health and disease [2]. In most natural environments, microorganisms form communities with complex interdependencies that have posed significant challenges for scientists investigating these organisms with traditional methods [3]. With the advent of new high-throughput DNA sequencing technologies, scientists were able to study the genetic material of these microorganisms [4].

Metagenomics, known as the culture-independent sequencing approach, bypasses the difficulties in studying genetic diversity, population structure, and ecological roles of uncultivable microorganisms by directly sequencing genetic materials from the environment [5,6]. This approach has been applied to the sequencing of selected marker genes, such as 16s rRNA genes, hypervariable regions within marker genes, and even the whole metagenome. In these sequencing projects, one of the main goals is to obtain the composition of the microbial community within a given environment [7–9]. 16s rRNA gene sequencing is the primary technique to taxonomic profiling. Computational methods have been developed to analyze sequencing data in order to obtain the taxonomic composition of microbial communities. There are two

groups of approaches: comparison-based methods and composition-based methods [10]. In this study, we focus on the comparison-based methods, which are more widely used.

Comparison-based methods utilize homology information obtained by searching sequencing reads for microbial sequences having known taxonomic origins. About one million 16s rRNA gene sequences have been collected and organized as a tree into such databases as RDP [11], Silva [12], and Greengenes [13]. Using database search tools like BLAST [14], taxonomic origins of 16s rRNA sequencing reads can be inferred based on statistically significant BLAST hits. This approach has been widely used in programs such as MG-RAST [15], and MEGAN [16]. In detail, sequence comparison produces a similarity score, called sequence identity, between a read and one sequence in the database, by which we assign a certain taxonomic level to the read. This process requires (i) calculation of sequence identity [17] and (ii) a similarity threshold for each taxonomic level [18].

Unfortunately, the calculation of sequence identity can vary significantly because of the choice of the following factors: the usage of global or local alignment, the scoring function used in the dynamic programming, counting the number of gaps by each gap individually, merging consecutive gaps into one gap, or ignoring gaps, the denominator used to calculate sequence identity, and regions of hypervariable regions of 16s rRNA genes. There is no agreement on the best choice. Thus there is variation among existing programs in computing sequence identity. As a result, different databases or programs may produce contradicting taxonomic assignments.

The similarity threshold for each taxonomic level varies significantly too. Yarza *et al.* (2014) observed that the median and minimum sequence identities are 96.4% and 94.8%, respectively, at the genus level; 92.25% and 87.65%, respectively, at the family level; 89.2% and 83.55%, respectively, at the order level, 86.3% and 80.38%, respectively at the class level, and 83.68% and 77.43% respectively at the phylum level. Therefore, given two sequences and their sequence identity ( $< 85\%$ ), it may be difficult to determine at which taxonomic level they share a common origin, especially at the high taxonomic levels such as phylum and class.

In this study, we focus on taxonomic prediction of unknown sequences or OTUs, which are defined as those with borderline sequence identities with known sequences measured at either phylum or class levels. By “known” we mean sequences with known taxonomic origins, and by “unknown” we mean that database search fails to assign this sequence to a taxon. Our prediction is based on an important fact that microorganisms live in a complex community with highly interdependent relationships [19–

22]. Some of these relationships are reflected in associations between two microorganisms that are co-abundant across multiple environmental perturbations. Exploring these associations may help us to infer the taxonomic origins of unknown microorganisms. To the best of our knowledge, this is the first attempt to predict the taxonomic origins of unknown microorganisms that are dissimilar to any of those found on databases such as RDP, Silva, and Greengenes. Although co-abundant associations are intrinsically functional correlations, it should be noted that co-abundant clusters are more likely to be composed of microorganisms from the same phylum or class, suggesting that functional associations and taxonomic associations strongly overlap.

To further understand the relationship between these two kinds of associations, we construct a global co-abundant OTU network using multiple samples [19–24], and we then apply a statistical graphical model, termed Markov Random Field (MRF) [25], to characterize the taxonomic associations in this network. MRF has been applied to solve different problems, such as image noise restoration and protein function prediction [26]. We used this model in a Bayesian framework to calculate the most probable configuration of taxonomic profiling of a given network which gives us the most probable taxonomic profile labels for an unknown OTU.

## RESULTS

### Datasets

We applied our method, as detailed in the sections below, to three different datasets, including human intestine, human skin, and soil. The human intestine dataset includes 70 samples of intestinal microbiota in patients with inflammatory bowel disease (IBD) sequenced from the V4 region of 16s rRNA genes with average length of 90 bp for a total of 560,000 reads [27]. The human skin dataset consists of 86 samples, each with approximately 3,500 reads for a total of 300,350 reads sequenced from the V2 region of the 16S rRNA with an average length of 250 bp [28]. The soil dataset includes 49 samples with a total of 277,363 reads sequenced from the V4 regions of the 16s rRNA gene with average length of 150 bp [29].

### Data preprocessing and network construction

#### Data preprocessing

For each dataset, we trim reads, pool all the samples together, and cluster all reads into OTUs using CROP (Clustering 16S rRNA for OTU prediction) [30] with pairwise distance  $\leq 3\%$ . CROP is an unsupervised

Bayesian clustering method based on the Gaussian Mixture Model. CROP reports each OTU with a center sequence, along with all other member sequences. We remove unreliable small OTUs with fewer than 5 reads to obtain a sample-OTU matrix in which each element represents the number of reads from a specific sample belonging to a particular OTU. We normalize the matrix by dividing the value of each element by the total number of reads in the corresponding sample.

#### Network construction

Using the sample-OTU matrix, we compute the Spearman's correlation coefficient between every pair of OTUs across all samples, and we retain those with correlation coefficient value  $> 0.5$  and  $p$ -value  $< 0.05$ , using the permutation test. As a result, we obtain a co-abundance network with nodes as OTUs and edges connecting pairs of co-abundant OTUs.

#### Taxonomic labeling

We use the RDP classifier [31] to annotate each OTU using the center sequence reported by CROP. The output of the RDP classifier is the taxonomic annotation at different taxonomic levels within the taxonomic hierarchy, each with a confidence level. In this study, we investigate the taxonomic relationships between OTUs at the level of class. The distribution of the number of OTUs belonging to each class for each dataset is shown in Supplementary Figures S1, S2 and S3. For taxonomic prediction, we focus on classes with at least 1% of the total number of OTUs in the network. We applied our Bayesian method to predict the probability that an unknown OTU belongs to each taxonomic class. Combining the results from all classes, we chose the most probable class for each unknown OTU.

### Characteristics of OTU co-abundance networks

#### Soil network

The number of OTUs in the soil dataset after clustering was 13,490, and after preprocessing, we obtained a network with 572 nodes and 4,996 edges. Using the RDP classifier, we labeled nodes with a RDP confidence level below 50% as unknown. As a result, we obtained a network with 403 known and 169 unknown nodes at the class level. This is consistent with our understanding that soil is a very heterogeneous environment with a large number of rare species. As shown in Table 1, a set of measures, such as average number of neighbors, average clustering coefficient, and modularity [32], have been

calculated to describe the topology of the networks. In this calculation, we generated 1,000 random networks by permuting edges, while preserving the distribution of node degree [33], and for each random network, we calculated average clustering coefficient, and modularity. The clustering coefficient and modularity of the soil network are 0.211 and 0.827, respectively, both of which are much higher than the average clustering coefficient (0.029) and modularity (0.29) of random networks. The  $p$ -values for the clustering coefficient and modularity of the soil network are both  $< 0.001$ . This demonstrates the fact that the soil network is more organized than what would be expected by a random network.

#### Human skin network

The number of OTUs in the human skin dataset after clustering was 645, and after preprocessing, we obtained a network with 414 nodes and 3,982 edges. Using the RDP classifier, we labeled 360 OTUs as known and 54 as unknown. Table 1 shows that the clustering coefficient and modularity of the human skin network are 0.326 and 0.647, respectively, both of which are much higher than the average clustering coefficient (0.164) and modularity (0.192) of random networks. The  $p$ -values for the clustering coefficient and modularity of the human skin network are both  $< 0.001$ . This demonstrates the fact that the human skin network is more organized than what would be expected by a random network.

#### Human intestine network

The number of OTUs in the human intestine dataset after clustering was 4,422, and after preprocessing, we obtained a network with 4,000 edges and 541 nodes. Using the RDP classifier, we obtained 313 known OTUs and 228 unknown OTUs. Table 1 shows that the clustering coefficient and modularity of the human intestine network are 0.265 and 0.767, respectively, both of which are much higher than the average clustering coefficient (0.042) and modularity (0.256) of the random networks. The  $p$ -values for the clustering coefficient and modularity of the human intestine network are both  $< 0.001$ . This demonstrates the fact that the human intestine network is more organized than what would be expected by a random network.

#### Clustering coefficients for each taxonomic label

How well OTUs belonging to the same taxa are connected with each other in the network affects the accuracy of predicted taxonomic labels of unknown OTUs. Therefore, to precisely measure the degree to which OTUs from the

**Table 1. Network statistics for microbial co-abundance networks from soil, human skin, human intestine, and for the random networks (-R).**

	Soil	Skin	Intestine	Soil-R	Skin-R	Intestine-R
Number of nodes	572	414	541	572	414	541
Number of edges	4996	3982	4000	4996	3982	4000
Avg. Degree	11	16	14	11	16	14
Clustering coefficient	0.211	0.326	0.265	0.029±0.003	0.164±0.008	0.042±0.004
Modularity	0.827	0.647	0.767	0.298±0.01	0.192±0.005	0.256±0.005

Avg. stands for average, and Soil/Skin/Intestine-R represents the average value of the statistics of 1,000 randomly generated networks using the preserved degree distribution random network generator algorithm for each microbial dataset.

same taxon cluster together, we defined a new measure called taxonomic clustering coefficient (TCC). Given a taxonomic label  $j$ , the taxonomic clustering coefficient of an OTU is defined as the proportion of the edges between the OTUs with the taxonomic label  $j$  within its immediate neighborhood divided by the total number of edges that could possibly exist between all those OTUs. The taxonomic clustering coefficient of the taxonomic label  $j$  is then defined as the average of the TCC values of all OTUs with label  $j$ .

The taxonomic clustering coefficients of all major taxa in our datasets are shown in Tables 2–4. We used the same measure to compare each taxon in each network with the average clustering coefficient of the same taxon in 1,000 randomly generated networks. The results demonstrate strong intra-taxon associations within the classes in the tables. For example, Table 2 shows that Bacilli, Clostridia, Gammaproteobacteria, Alphaproteobacteria, and Deltaproteobacteria in soil tend to cluster together. Table 3 shows that Clostridia, Deinococci, Betaproteobacteria, Actinobacteria, Bacilli and Sphingobacteria in human skin cluster together, and Table 4 shows that Clostridia, Negativicutes, and Bacteroidia in human intestine cluster together.

**Table 2. Clustering coefficient of each class label in the soil network and the average random network.**

Class label	Soil-CC	Soil-RN-CC
Bacilli	0.688677	0.001027
Clostridia	0.450079	0.002208
Gammaproteobacteria	0.303088	0.00425
Alphaproteobacteria	0.125	0.001857
Deltaproteobacteria	0.105263	0.001178

The column “Soil-CC” shows the taxa clustering coefficient of different classes in soil, and the column “Soil-RN-CC” shows the average taxa clustering coefficient of 1,000 randomly generated networks using the preserved degree distribution random network generator algorithm for soil data.

### Evaluating the accuracy of taxonomic predictions at the class level

Supplementary Figures S1–3 show the distribution of

**Table 3. Clustering coefficient of each class label in the human skin network and average random network.**

Class label	Skin-CC	Skin-RN-CC
Clostridia	0.647505	0.030902
Deinococci	0.375	0
Betaproteobacteria	0.352941	0.009074
Bacteroidia	0.301282	0.017062
Actinobacteria	0.293367	0.02438
Bacilli	0.269231	0.013244
Sphingobacteria	0.172414	0.019276

The column “Skin-CC” shows the taxa clustering coefficient of major classes in skin data, and the column “Skin-RN-CC” shows the average taxa clustering coefficient for 1,000 randomly generated networks for skin data using the preserved degree distribution random network generator algorithm.

**Table 4. Clustering coefficient for each class label in the human intestine network and the average random network.**

Class label	Intestine-CC	Intestine-RN-CC
Clostridia	0.444464	0.0201
Negativicutes	0.377273	0.00030303
Bacteroidia	0.282927	0.006898
Actinobacteria	0.000303	0

The column “Intestine-CC” shows the taxa clustering coefficient of different classes in human intestine, and the column “Intestine-RN-CC” shows the average taxa clustering coefficient for 1,000 randomly generated networks for human intestine data using the preserved degree distribution random network generator algorithm.

classes identified in the three datasets: 31 in the human skin dataset, 49 in the soil dataset and 19 in the human intestine dataset. In this study, we were specifically interested in dominant classes occupying at least 1% of the nodes in the co-abundance networks. By this definition, we found 15 dominant classes in the human skin network, 14 in the soil network, and 10 in the human intestine network. We then applied the MRF model to each class in these networks. (i) We first labeled each node (OTU) in the network as “1” if it belongs to this class, “0” if it belongs to other classes, and “unknown” if it is

unknown. (ii) We trained MRF parameters for this class using the node labels and the edges in the network. (iii) We predicted the “0/1”-class membership with a probability for each unknown OTU using the Gibbs sampling algorithm. At the end, we combine predictions from all classes and label each unknown node to the class with the highest probability. To assess the accuracy of our taxonomic predictions, we used the Leave-One-Out strategy, using known OTUs in the network.

#### Estimated parameters in the MRF model

MRF parameters are estimated using the quasi-likelihood approach, as explained in Methods. Table 5 shows the estimated parameters of the dominant classes in the skin dataset, and the parameters for the soil and human intestine datasets are shown in Supplementary Tables S1–2. Parameter  $\alpha = \log(\frac{\pi}{1-\pi})$  should be negative, because parameter  $\pi$  which is the representative of the fraction of OTUs having the taxonomic label of interest is generally a small number. In addition, parameter  $\beta-1$  which represents the contribution of an associated OTU not belonging to the given taxonomic label of interest should be negative and parameter  $\gamma-\beta$  should be positive, because it represent the contribution of

**Table 5. Estimated parameters for the human skin dataset.**

Class labels	$\alpha$	$\beta-1$	$\gamma-\beta$
Bacteroidia	-2.33537	-0.04245	0.247362
Actinobacteria	-1.83828	-0.0435	0.198456
Clostridia	-1.65726	-0.02351	0.214375
Bacilli	-2.37669	-0.04131	0.261569
Betaproteobacteria	-2.77259	-0.04109	0.491779
Acidobacteria_Gp4	-4.09767	0.025845	-0.09265
Sphingobacteria	-2.25672	-0.02481	0.134167
Acidobacteria_Gp3	-4.32413	-0.27638	-30.2856
Flavobacteria	-3.38777	0.027906	-0.95404
Epsilonproteobacteria	-4.32413	-0.04917	-29.7953
Deinococci	-3.61765	-0.04201	0.651011
Cyanobacteria	-3.61765	-0.12877	1.07966
Chloroplast	-4.32413	-0.05814	-30.0519
Deltaproteobacteria	-3.49651	0.074062	-0.26545
Alphaproteobacteria	-2.96527	0.051567	-0.27868

The parameter  $\alpha = \log(\frac{\pi}{1-\pi})$ , where  $\pi$  represents the relative abundance of a class in the samples, should be negative because  $\pi$  is generally much smaller than 0.5. For a specific taxonomic class,  $\beta-1$ , which represents the contribution of an associated OTU not belonging to the given class, should be negative, while  $\gamma-\beta$ , which represents the contribution of an associated OTU belonging to the given class, should be positive.

an associated OTU belonging to the given taxonomic label of interest. Most taxonomic labels in the three datasets follow this pattern, except for Flavobacteria and Opitutae in the soil network, Deltaproteobacteria and Alphaproteobacteria in the skin network, and Bacilli in the human intestine network, resulting from the low abundance of these classes in their networks. Moreover, it is not expected that all taxa will have positive intra-dependency among themselves. For example, the class Flavobacteria in the soil dataset is abundant, but its OTUs do not tend to co-occur; therefore, its parameters do not follow the aforementioned pattern.

#### Prediction accuracy

To determine the accuracy of prediction by MRF for each class, we calculated the area under the curve (AUC) value, as well as the true positive rate (TPR) and false positive rate (FPR) defined as

$$\text{TPR} = \frac{\text{\#predictions matched with the label}}{\text{\#predictions}}$$

$$\text{FPR} = \frac{\text{\#predictions NOT matched with the label}}{\text{\#predictions}}$$

Table 6 shows the AUC values for a number of classes in the soil, human skin and human intestine datasets. As expected, the AUC values show better results for classes which tend to cluster together (with high taxonomic clustering coefficients). For example, the class of Bacilli in the soil network has the highest AUC value of 0.76, and also the highest taxonomic clustering coefficient of 0.689. The class of Clostridia in the human skin network has the highest AUC value of 0.72, and also the highest taxonomic clustering coefficient of 0.648. The class of Clostridia in the human intestine network has the second highest AUC value of 0.62, and the highest taxonomic clustering coefficient of 0.444. Another interesting example is Negativicutes in the human intestine dataset. Compared to Clostridia and Bacteroidia, Negativicutes, a class of Firmicutes bacteria, has lower abundance, but shows stronger intra-class associations.

The overall AUC varies by different environmental datasets. This happens because each microbial community has its distinct characteristics of interactions within and between classes. Consequently, better prediction accuracy is obtained for the soil and human skin datasets than that for the human intestine dataset. Supplementary Figure S3 shows the distribution of classes in human intestine. In this network, the Clostridia, a highly polyphyletic class of Firmicutes, dominate with lower intra-taxon relationship, which, in turn, affects the overall prediction accuracy.

**Table 6. AUC values for major classes in soil, human skin, and human intestine datasets.**

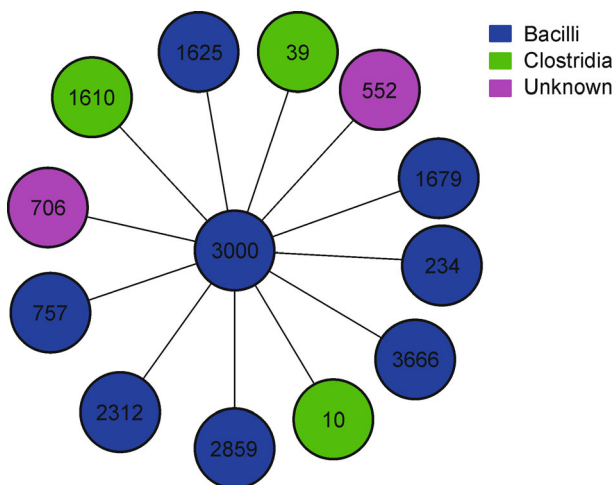
Soil	AUC	Skin	AUC	Intestine	AUC
Bacilli	0.76	Clostridia	0.72	Negativicutes	0.70
Betaproteobacteria	0.65	Bacteroidia	0.67	Clostridia	0.62
Deltaproteobacteria	0.59	Betaproteobacteria	0.66	Bacteroidia	0.60
Alphaproteobacteria	0.59	Deinococci	0.59	Erysipelotrichia	0.59
		Gammaproteobacteria	0.59	Actinobacteria	0.56
		Actinobacteria	0.54		
		Alphaproteobacteria	0.58		
		Bacilli	0.57		

### Combining taxonomic predictions with sequence comparisons

Our approach can be used under the following situation. Through database search, an OTU  $\alpha$  is found to be distantly similar ( $< 85\%$ ) to a known taxon  $\beta$ . Based on the sequence similarity alone, it would be difficult to know whether  $\alpha$  and  $\beta$  belong to the same class or phylum or not. Then our network based approach, can be applied to compute the probability that  $\alpha$  belongs to the same phylum/class as  $\beta$ . We show how this works in the following example.

In the soil samples, there is an OTU #3000 which is unclassified by RDP, as RDP did not find it similar to any sequence in the database with at least 50% confidence level at the rank class. Through the co-abundance network, TACO predicts that it belongs to Bacilli with probability 98%. We extracted a subnetwork around this node and show it in Figure 1.

The clustering coefficient table in our paper shows that Bacilli is one of those classes of which OTUs tend to co-occur together in different environments and samples.



**Figure 1.** An unknown OTU (#3000) was predicted to be in the class of Bacilli according to the taxa of its neighbors in the network.

Based on the subnetwork around this node, it is highly probable that it belongs to class Bacilli.

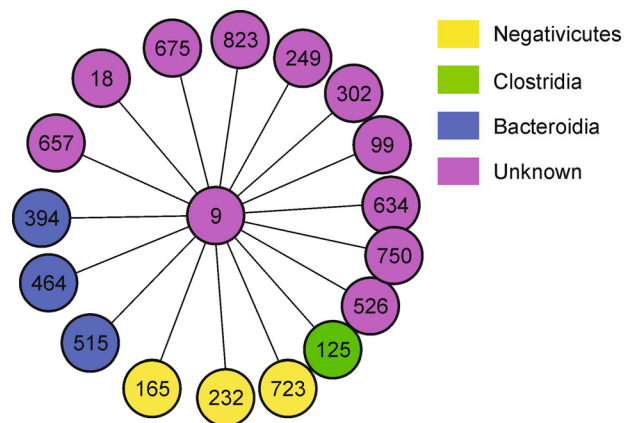
To validate our prediction, we searched this sequence in Silva. The returning information about this sequence using the Silva database and Silva aligner shows that the LCA (lowest common ancestor) prediction for this sequence is Firmicutes (phylum) and Bacilli (class) with 67% similarity.

### Validating the taxonomic predictions of unknown OTUs

To validate our taxonomic predictions of unknown OTUs, which are based on the RDP classification results, we searched them against the SILVA rRNA database [12] because SILVA uses different methods to annotate OTUs.

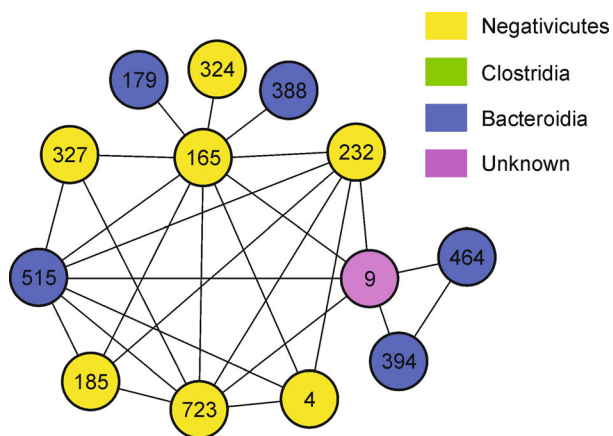
After this search, many unknown OTUs remained, including 20 OTUs in the human intestine dataset, 15 OTUs in the human skin dataset, and 20 OTUs in the soil dataset. These OTUs had no prediction at the class level from SILVA.

We investigated these OTUs in the human intestine dataset, and we show the result for OTU#9 in the human intestine network as an example. We can see the unique



**Figure 2.** The neighborhood of node 9 in human intestine network has equal numbers of Bacteroidia (3) and Negativicutes (3).

advantage of our Bayesian method which takes into consideration the global connections of nodes in a network. Because in Figure 2, node 9 is surrounded with many unknown nodes and also there are three Bacteroidia and three Negativicutes around it which makes it difficult to identify the taxonomic class which node 9 belong to it. We further examine the second level of association around node 9 in the Human Intestine Network and find that because of strong association between the yellow Negativicutes, it is most likely that node 9 belongs to a Negativicutes. Not all of the second level association has been shown in Figure 3 to be easier to see the strong associations between Negativicutes nodes.



**Figure 3. The first and second level neighbors of node 9 in the human intestine network.** There is a high connectivity among nodes with class label Negativicutes (yellow). The probability of node 9 belonging to class Negativicutes is the highest.

### DISCUSSION

We developed a statistical approach to predict taxonomic identities of unknown OTUs. We focus on a key property of microbial communities, i.e., their co-abundance associations, which are used to annotate the unknown taxa using those known taxa. Our method works well for annotations at the class and phylum levels, and depends upon and complements to the sequence comparison based methods which predicts known OTUs as input to our model. However, we have explained that our method would likely be the only choice when sequence comparison methods fail give a confident prediction.

The result of taxonomic predictions proves that there is significant overlap between the co-abundance network and the taxonomic relationships, although the network intrinsically consists of functional associations. We believe that the degree of such overlap varies in different communities and thus has a significant impact on the

prediction accuracy. Using the soil dataset as an example, each soil sample is a mixture of many small communities in different locations within a certain sampling distance, and thus its complexity is very high. As a result, the co-abundance network may not reflect true functional associations, and thus has less overlap with the taxonomic relationships. This explains the lower taxonomic prediction accuracy of the soil dataset.

We also suspect that our model may not work well at the low level taxa such as the genus and species levels, at which the number of known OTUs may be too few for parameter estimation in our MRF model.

Overall, we provide a very useful MRF model for taxonomic prediction. The posterior probability computed by the model shows the confidence of the prediction. The underlying Bayesian approach is a global approach as it considers all the associations between nodes, among the same taxa, and between different taxa across the whole network.

### METHODS

Assume that we are given multiple 16s rRNA sequencing datasets. We pool these datasets together and cluster the data into  $N$  OTUs. Then we construct an OTU co-abundant network  $G=(V,E)$ , where nodes represent OTUs and edges represent co-abundant associations between OTUs. Let  $(V_1,V_2,\dots,V_k)$  be the unknown taxonomic origins and let  $(V_{k+1},V_{k+2},\dots,V_{k+m})$  be the OTUs of known taxonomic origins and  $N=k+m$ . Consider a specific taxonomic level, such as class, which consists of  $M$  taxonomic labels  $(TX_1,TX_2,\dots,TX_M)$ . Each known OTU in the network is assigned exactly one label among  $(TX_1,TX_2,\dots,TX_M)$ .

Given a specific taxonomic label  $TX_j$ , we want to assign it to unknown OTUs. To accomplish this, let  $(X_1,\dots,X_N)$  be the taxonomic labels for all the OTUs, where  $(X_1,X_2,\dots,X_k)$  are unknown and  $(X_{k+1},\dots,X_{k+m})$  are known. Let  $X_i=1$  if the  $i^{th}$  OTU belongs to taxonomic label  $TX_j$ , and  $X_i=0$  otherwise. Let  $(x_1,x_2,\dots,x_N)$  be a realization of  $X$ . Our aim is to infer the posterior distribution of configuration of  $(X_1,\dots,X_k)$ ,  $P(X_1,X_2,\dots,X_k|X_{k+1},X_{k+2},\dots,X_{k+m})$ , given the known taxonomic labels  $(X_{k+1},X_{k+2},\dots,X_{k+m})$ , using a Bayesian approach.

Without considering the network, the probability of a random OTU having label  $TX_j$  is

$$\pi = \frac{\text{Number of known OTUs with label } TX_j}{m} \tag{1}$$

Given a configuration of  $X$ , the probability of the OTUs is  $x$ ,

$$\prod_{i=1}^N \pi^{x_i} (1-\pi)^{1-x_i} = \left(\frac{\pi}{1-\pi}\right)^{N_1} (1-\pi)^N \tag{2}$$

where  $N_1 = \sum_{i=1}^N x_i$  equals the number of OTUs having the label  $TX_j$ .

Now, considering the network, we have three different types of associations:

Type I: (1↔1) in which both OTUs belong to  $TX_j$ ;

Type II: (1↔0) in which one OTU belongs to  $TX_j$  and the other does not; and

Type III: (0↔0) in which neither one of the OTUs belongs to  $TX_j$

In this network, we define

$$\begin{aligned}
 N_{11} &= \sum_{1 \leq i < j \leq N} x_i x_j : \text{the number of type I associations} \\
 N_{10} &= \sum_{1 \leq i < j \leq N} (1-x_i)x_j + x_i(1-x_j) : \text{the number of type II associations} \\
 N_{00} &= \sum_{1 \leq i < j \leq N} (1-x_i)(1-x_j) : \text{the number of type III associations}
 \end{aligned}$$

Therefore, the probability of the network conditional on the known OTUs is proportional to

$$\exp(\beta N_{01} + \gamma N_{11} + N_{00}). \tag{3}$$

Combining the probabilities of the configuration of  $X$  in Equation (2) and the network in Equation (3), we obtain the total probability of the taxonomic labeling of the network that is proportional to  $\exp(-U(x))$  where

$$U(X) = -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00}, \tag{4}$$

with  $\alpha = \log \frac{\pi}{1-\pi}$ .

Here,  $U(X)$  is called the potential function. Define  $\theta = (\alpha, \beta, \gamma)$ . In the general theory of MRF, the potential function defines a Gibbs distribution of the entire network, as

$$\Pr(X|\theta) = \frac{1}{Z(\theta)} \exp(-U(x)), \tag{5}$$

where  $Z(\theta) = \sum_{x \in X} \exp(-U(x))$ .

The final goal is to calculate the joint probability distribution  $P(X_1, X_2, \dots, X_k | X_{k+1}, X_{k+2}, \dots, X_{k+m})$  using the Bayesian rule and, consequently, the marginal probability distribution for each single random variable  $X_i$  by summing over all configurations of the network. Because calculating the posterior probability for a large set of unknown OTUs is very difficult, we have to utilize the Gibbs sampler to approximate the posterior probability in this study.

### Gibbs sampling

Let  $X_{[-i]} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$  be the set of OTUs excluding  $X_i$ , and let  $M_0^{(i)}$  and  $M_1^{(i)}$  be the numbers of 0-labeled and 1-labeled OTUs that are adjacent to  $X_i$  in the network, respectively. We approximate  $P(X_1, X_2, \dots, X_k | X_{k+1}, X_{k+2}, \dots, X_{k+m})$  by sampling from conditional probability  $P(X_i | X_{[-i]}, \theta)$  for all unknown OTUs  $\{X_i : 1 \leq i \leq k\}$  using Equation (6).

$$\begin{aligned}
 P(X_i = 1 | X_{[-i]}, \theta) &= \frac{P(X_i = 1, X_{[-i]} | \theta)}{P(X_i = 1, X_{[-i]} | \theta) + P(X_i = 0, X_{[-i]} | \theta)} \\
 &= \frac{e^{\alpha + (\beta-1)M_0^{(i)} + (\gamma-\beta)M_1^{(i)}}}{1 + e^{\alpha + (\beta-1)M_0^{(i)} + (\gamma-\beta)M_1^{(i)}}}.
 \end{aligned} \tag{6}$$

The first step in sampling is based on the prior probability that an unknown OTU belongs to the taxonomic label, and thus initial values of 0 or 1 are assigned to the unknown random variables using the Bernoulli distribution with probability  $\pi$  which has been defined in Equation (1). The Gibbs sampler iterates through all of the random variables  $X_i$  and computes their current value by drawing a sample from probability  $P(X_i = 1 | X_{[-i]}, \theta)$  using Equation (6).

### Parameter estimation

All of these procedures are possible if we know the model parameter  $\theta = (\alpha, \beta, \gamma)$ . We can estimate the parameters of the MRF model by using the known part of the network. We have used the standard linear logistic regression model to estimate the parameters based on the assumption that the random variables are independent [26]. However, while it is clear that the random variables are not independent, the MRF model has shown that treating random variables independently does give a good approximation to the Maximum Likelihood Estimation model. From Equation (6), we have

$$\log \frac{P(X_i = 1 | X_{[-i]}, \theta)}{1 - P(X_i = 1 | X_{[-i]}, \theta)} = \alpha + (\beta-1)M_0^{(i)} + (\gamma-\beta)M_1^{(i)}. \tag{7}$$

### Summary of the MRF method

Following these steps will give the posterior probabilities

for all nodes for each taxonomic label:

1. Compute  $\pi$  based on the known OTUs
2. Estimate parameters  $\theta = (\alpha, \beta, \gamma)$  using Equation(7)
3. Set the initial values of the unknown OTUs using

Bernoulli( $\pi$ )

4. Compute the posterior probability of each unknown OTU  $X_i$  using Equation(6) iteratively

5. Repeat Step 4 until all posterior probabilities are convergent.

We specify the burn-in period as 100 iterations which is the time we wait until the Markovian process is stabilized. We assign the value 10 to the lag-in period to eliminate the dependence of the Markovian process. As a result, we average the probability results in the steps of the lag-in period to approximate the final probability.

## SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-016-0073-2.

## ACKNOWLEDGEMENTS

The author thank Professor Fengzhu Sun for his helpful suggestions. This research was partially supported by NIH Center of Excellence in Genomic Sciences (NIH/HG 2 P50 HG002790-06), NIH/NHGRI 1U01 HG006531-01, NSF/DMS ATD 7031026, and NSFC 91019016.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Zohreh Baharvand Irrannia and Ting Chen declare that they have no conflict of interests.

## REFERENCES

1. Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, 68, 669–685
2. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464, 59–65
3. Amann, R. I., Ludwig, W. and Schleifer, K. H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59, 143–169
4. Eisen, J. A. (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol.*, 5, e82
5. Hugenholtz, P., Goebel, B. M. and Pace, N. R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.*, 180, 4765–4774
6. Riesenfeld, C. S., Schloss, P. D. and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38, 525–552
7. Wooley, J. C. and Ye, Y. (2010) Metagenomics: facts and artifacts, and computational challenges. *J. Comput. Sci. Technol.*, 25, 71–81
8. Thomas, T., Gilbert, J. and Meyer, F. (2012) Metagenomics — a guide from sampling to data analysis. *Microb. Inform. Exp.*, 2, 3
9. Teeling, H. and Glöckner, F. O. (2012) Current opportunities and challenges in microbial metagenome analysis — a bioinformatic perspective. *Brief. Bioinform.*, 13, 728–742
10. Mande, S. S., Mohammed, M. H. and Ghosh, T. S. (2012) Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.*, 13, 669–681
11. Maidak, B. (1996) The Ribosomal Database Project (RDP). *Nucleic Acids Res.*, 24, 82–85
12. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41, D590–D596
13. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72, 5069–5072
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410
15. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., *et al.* (2008) The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386
16. Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, 17, 377–386
17. Schloss, P. D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.*, 6, e1000844
18. Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R. and Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.*, 12, 635–645
19. Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R. and Ruppin, E. (2010) The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.*, 38, 3857–3868
20. Chaffron, S., Rehrauer, H., Pernthaler, J. and von Mering, C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.*, 20, 947–959
21. Barberán, A., Bates, S. T., Casamayor, E. O. and Fierer, N. (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.*, 6, 343–351
22. Faust, K. and Raes, J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbiol.*, 10, 538–550
23. Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., *et al.* (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.*, 5, 1414–1425
24. Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A. C., Knight, R., Joint, I., *et al.* (2012) Defining seasonal marine microbial community dynamics. *ISME J.*, 6, 298–308
25. Kindermann, R. and Snell, J. L. (1980) Markov Random Fields and Their Applications. V. 1. Of Contemporary Mathematics. Rhode Island: American Mathematical Society

26. Deng M., Zhang K., Mehta S., Chen T., Sun F. (2004) Prediction of protein function using protein-protein interaction data. *J. Comp. Biol.* 10, 947–960
27. Human-Intestine-NCBI, <http://www.ncbi.nlm.nih.gov/bioproject/204926>
28. Human-Skin NCBI, <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB3280>
29. Soil-NCBI, <http://www.ncbi.nlm.nih.gov/bioproject/PRJEB4349>
30. Hao, X., Jiang, R. and Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics*, 27, 611–618
31. Lan, Y., Wang, Q., Cole, J. R. and Rosen, G. L. (2012) Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS One*, 7, e32491
32. Newman, M. E. J. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103, 8577–8582
33. Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science*, 296, 910–913