RESEARCH ARTICLE

Biomarker discovery in inflammatory bowel diseases using network-based feature selection

Mostafa Abbas[1☉], John Matta[2☉], Thanh Le[3], Halima Bensmail[1], Tayo Obafemi-Ajayi[3]*, Vasant Honavar[4], Yasser EL-Manzalawy[4,5]*

1 Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar, **2** Department of Computer Science, Southern Illinois University Edwardsville, Edwardsville, IL, United States of America, **3** Engineering Program, Missouri State University, Springfield, MO, United States of America, **4** College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, United States of America, **5** Geisinger Health System, Danville, PA, United States of America

☉ These authors contributed equally to this work.
* TayoObafemiAjayi@missouristate.edu (TO-A); yelmanzalawi@geisinger.edu (YE-M)

Abstract

Reliable identification of Inflammatory biomarkers from metagenomics data is a promising direction for developing non-invasive, cost-effective, and rapid clinical tests for early diagnosis of IBD. We present an integrative approach to Network-Based Biomarker Discovery (NBBD) which integrates network analyses methods for prioritizing potential biomarkers and machine learning techniques for assessing the discriminative power of the prioritized biomarkers. Using a large dataset of new-onset pediatric IBD metagenomics biopsy samples, we compare the performance of Random Forest (RF) classifiers trained on features selected using a representative set of traditional feature selection methods against NBBD framework, configured using five different tools for inferring networks from metagenomics data, and nine different methods for prioritizing biomarkers as well as a hybrid approach combining best traditional and NBBD based feature selection. We also examine how the performance of the predictive models for IBD diagnosis varies as a function of the size of the data used for biomarker identification. Our results show that (i) NBBD is competitive with some of the state-of-the-art feature selection methods including Random Forest Feature Importance (RFFI) scores; and (ii) NBBD is especially effective in reliably identifying IBD biomarkers when the number of data samples available for biomarker discovery is small.

Introduction

Inflammatory bowel disease (IBD) refers to disorders that involve chronic inflammation in the gastrointestinal tract. The two main types of IBD are ulcerative colitis (UC), which is characterized by continuous ascending inflammation from the rectum into the colon and periods of relapse and remittance [1], and Crohn disease (CD), which is characterized by discontinuous skip lesions affecting any part of the gastrointestinal tract [2]. Recent metagenome-wide

association studies have implicated some changes in the microbial communities in the gut microbiota with the onset and progression of IBD. [3–6]. However, the precise nature of the changes in the gut microbiota in IBD remains to be fully understood [3].

IBD, particularly in children, fails to be correctly diagnosed, or diagnosed in a timely fashion, because of the frequency of nonspecific symptoms at the onset of the disease [7, 8]. Although several non-invasive tests exist for IBD, none has been shown to be capable of diagnosing the two main IBD subtypes with sufficient accuracy [9]. Therefore, a biomarker signature for diagnosing IBD and differentiating between the two major IBD subtypes is highly desirable [8, 10]. Identification of microbial biomarkers is a promising direction, not only for predicting IBD onset but also for predicting IBD risk factors [11].

Identification of disease microbiomarkers from metagenomics data requires effective computational and statistical methods for determining, from a very large number of candidate biomarkers, a minimal subset of biomarkers that can accurately discriminate between two or more phenotypes (e.g., IBD versus healthy). This task presents several challenges in practice [12]: curse of dimensionality; high degree of sparsity of the metagenomics data; complexity of the underlying biology; limitations of sequencing technology and of methods for determining microbial composition and functional profiles from metagenomic data. To date, several statistical methods have been proposed in the literature to compare an abundance of features (e.g., genes or operational taxonomic units (OTUs)) between two groups [13]. Some of these methods have been designed specifically for RNA-Seq data (e.g., DESeq [14] and edgeR [15]) while recent tools such as metagenomeSeq [16] and analysis of composition of microbiomes (ANCOM) [17] have been developed specifically for metagenomics data, which often exhibits greater sparsity than RNA-Seq data. Machine learning methods for feature selection [18] offer a promising approach to identifying, from either RNA-Seq or metagenomics data, an optimal subset of the features (potential biomarkers) that can be used to build predictive models that can effectively diagnose a disease or discriminate between disease subtypes.

Recent analysis of microbial ecology networks (MEN) (where the nodes denote microbial taxa and links denote some measure of correlations between the composition of the corresponding taxa) derived from healthy and type 2 diabetes (T2D) groups has shown topological differences between the two networks at the global, module (i.e., sub-networks or communities), and node levels and found that the differences in cluster membership of the nodes in the two networks can serve as biomarkers for T2D [19]. Motivated by these findings, Abbas et al. [20] hypothesized that MEN corresponding to different phenotypes should exhibit different topologies, and the resulting differences in topology at the node and sub-network levels could be exploited in biomarker discovery. They tested this hypothesis using a framework for network-based biomarker discovery (NBBD). NBBD has two key modules: (i) A network construction module for assembling MEN from the abundance data for microbial taxa (e.g., OTUs); (ii) A node importance scoring module for comparing MEN for the chosen phenotypes and assigning a score to each node based on the degree to which the topological properties of that node differ across two networks. They reported results of experiments with a large data set of new-onset pediatric IBD metagenomics biopsy samples showing that NBBD could effectively discover IBD biomarkers [20].

In this study, we build on and extend the results of Abbas et al. [20] in two aspects: (i) We introduce a novel node importance scoring method based on three different node resilience measures [21] for identifying potential biomarkers. The strength of this approach is that the optimal number of features used to specify a biomarker need not be fixed a priori; (ii) We describe a hybrid approach for integrating network-based and random forest feature importance (RFFI) scores for improving the identification of a minimal subset of features to discriminate between the phenotypes of interest (based on the relative abundance of the microbial

taxa represented by the features). We also report results of extensive experiments with several instantiations of the NBBD framework using five different network inference tools, nine node importance scoring functions, and varying number of data samples used to perform feature selection. Our results demonstrate the viability of the NBBD framework for biomarker identification, not only from extremely sparse and high-dimensional data but also from datasets with small number of samples.

Materials and methods

Data sets

BIOM files (see http://biom-format.org) and meta-data (including age, gender, race, disease severity, behavior, and location) for a large cohort IBD data set [3] were downloaded from the QIITA database (https://qiita.ucsd.edu/) (Study ID: 1939). The data set consists of 1359 metagenomics samples including rectal tissue biopsy and fecal samples and each sample has 786 OTUs at the genus level that were extracted using the summarize_taxa.py QIIME script. We filtered the data by discarding fecal samples and samples corresponding to patients with age greater than 18 years. The resulting data set consists of metagenomic biopsy samples for 657 IBD and 316 healthy control cases, respectively. Thus, each sample (which correspond to a row in the table), is encoded by a tuple of values that represent the relative abundances of the various microbial taxa (indexed by the columns) in the sample. To evaluate our models, we randomly split the data into training and test sets, named DS400 and DS573, such that the training data has 200 healthy and 200 IBD samples, and the test data has 457 healthy and 116 IBD samples. It should be noted that predictive models are often tested on a data distribution that reflects the natural distribution of the different classes. However, in this case, the available IBD and Healthy samples do not reflect the natural distribution of IBD and Healthy cases in the pediatric population. The prevalence of IBD worldwide has been reported to be close to 0.3% of the population [22]. Hence, given the high degree of class imbalance expected in the natural distribution of data, we anticipate that the reported performance of *all* of the methods in our comparison to substantially overestimate the true performance of the predictive models were they to be deployed in a real-world setting. However, this should not impact the validity of the overall conclusions from our study.

The training data is also used for feature selection (i.e., selecting a subset of features that are most relevant for the classification task). In our experiments, we examined the effect of using a small fraction of the training data for performing feature selection. Specifically, we experimented with the following choices of data for feature selection, which we call the feature selection data sets (FSDS): $DS50 \subset DS100 \subset FSD200 \subset DS300 \subset DS400$, each with equal numbers of IBD and healthy samples.

Network-Based Biomarker Discovery (NBBD) framework

We summarize the Network-based Biomarker Discovery (NBBD) framework below: (See Fig 1, adapted from [20]). Given a feature selection data set (FSDS) of metagenomics samples in the form of a labeled OTU table: (i) The network construction module, produces two MEN, one from the healthy samples, and one from the IBD samples, using the chosen network construction tool (e.g., CoNet [23]); (ii) The node importance scoring module compares the two networks and scores each node in terms of its contribution to the differences between the two networks (as measured using one or more network similarity measures); (iii) The k highest scoring nodes provide the k features used to train and evaluate binary classifiers for predicting whether or not a given metagenomic sample belongs to a healthy of IBD individual.

We evaluated the NBBD framework using five network construction methods and nine node importance scoring methods summarized below.

Network construction methods

We experimented with several widely used methods for constructing MEN from metagenomic data. We used the default parameters of each tool, unless noted otherwise. Each of these methods is briefly described as follows.

- SparCC: Sparse Correlations for Compositional data (SparCC) [24] infers a network of associations between the microbial species based on the linear Pearson correlations between the log-transformed components (e.g. OTUs). Since log transformation cannot be applied to zeros, which are frequent in microbiome data, zeros are usually substituted with a small value, called pseudo-count. SparCC makes two main underlying assumptions: (i) the number of nodes (e.g. OTUs) is large; and (ii) the underlying network is sparse. We applied the implementation of SparCC included as part of the SPIEC-EASI tool [25].

- MB: The Meinshausen and Bühlmann (MB) method [26] is another technique for estimating sparse networks based on estimation of the conditional independence restrictions of each individual node in the graph. The MB method determines the direct neighbors of each target node by finding the smallest subset of nodes such that the target node is conditionally independent of the rest of the networks given the direct neighbors so identified. MB is also implemented in SPIEC-EASI [25].

- RMT: Random Matrix Theory (RMT) method uses the Pearson correlation coefficient to add an edge between two OTUs if their correlation is higher than a threshold. Instead of using a user-defined threshold, RMT utilizes a procedure based on the Random Matrix Theory to automatically detect a reliable threshold. The method is implemented in the Molecular Ecological Network Analysis Pipeline [27] available at http://ieg4.rccc.ou.edu/mena. We used the default parameters except for the parameter controlling the number of OTUs that build the network. An OTU was used if it is expressed in at least 25% of the samples. The default value of that parameter is 50% of the samples, and with the parameter set to 50% the method failed to construct the network.

- CoNet: This method infers the association network by combining two complementary approaches to evaluate the significance of the associations [23]. The first approach is an ensemble method of similarity or dissimilarity measures while the second is a novel permutation-renormalization bootstrap method, ReBoot [23]. We followed the procedure described in [28] to construct the networks for the IBD and healthy phenotypes.

- Proxi: Proxi [29] is a Python package for proximity graph construction. In proximity graphs, each node is connected by an edge (directed or undirected) to its nearest neighbors according to some distance metric d. In our experiments, we set the number of neighbors to seven and used the absolute value of Pearson's Correlation between two vectors (subtracted from one) as the distance function between two vectors.

Node importance scoring methods

We considered two approaches for scoring nodes (i.e., features) based on: (i) differences in the topological properties of the nodes in the two networks [20]; (ii) common nodes in the critical attack sets [30] determined from the two networks. The first approach assumes that a biomarker has different patterns of interactions with other OTUs in healthy and IBD samples.

The second approach assumes that biomarkers correspond to a special set of nodes, in the two networks, called a critical attack set [30] such that the removal of nodes in the critical attack set from a graph results in clustering the network into a number of subnetworks (i.e., microbial communities in the case of MEN).

Node scoring using topological properties. Let $G_i(V_i, E_i)$ and $G_j(V_j, E_j)$ be two graphs constructed using two groups of metagenomics samples (e.g., healthy and IBD). The Node Topological Property Scoring (NTPS) method scores each node $v \in V_i \cap V_j$ with respect to a node topological property P as follows: $score^P(v) = |f_P(v, G_i) - f_P(v, G_j)|$, where $f_P(v, G)$ is the value of the property P for a node v in a graph G. In this work, we experimented with the following node properties computed with NetworkX software [31]:

- Betweenness Centrality (btw): Betweenness centrality of a node v is defined as $f_{btw}(v, G) = \sum \frac{\sigma(u,w|v)}{\sigma(u,w)}_{u,v,w \in V}$, where $\sigma(u, w)$ is the total number of shortest paths between u and w and $\sigma(u, w|v)$ is the number of shortest paths between u and w passing through v.

- Closeness Centrality (cls): Closeness centrality of a node v is given by $f_{cls}(v, G) = \frac{n-1}{\sum_{u=1}^{n-1} d(u,v)}$, where $d(u, v)$ is the shortest path distance between u and v and n is the number of nodes that can reach v.

- Average Neighbor Degree (and): The average neighborhood degree of a node v is given by $f_{and}(v, G) = \frac{1}{|N(v)|} \sum_{u \in N(v)} k_u$, where $N(v)$ denotes the set of neighbors of node v and k_u is the degree of node $u \in N(v)$.

- Clustering Coefficient (cc): For unweighted graphs, the clustering coefficient of a node v is given by $f_{cc}(v, G) = \frac{2T(v)}{deg(v)(deg(v)-1)}$, where $T(v)$ is the number of triangles that include node v and $deg(v)$ is the degree of v.

- Node Clique Number (ncn): The node clique number of a node v is the size of the largest maximal clique containing v. A clique is a subset of nodes such that there is an edge between every pair of distinct nodes.

- Core Number (cn): The core number of a node v is the largest value k of a k-core containing v, where a k-core is a maximal subgraph that contains nodes of degree k or more.

Critical Attack Set Scoring. Critical Attack Set Scoring (CASS) is based on a node resilience clustering algorithm, NBR-Clust [21, 30]. We briefly describe below, the node resilience measures (specifically the three utilized in this work) before proceeding to describe how they are used to identify biomarkers.

Node-based resilience measures quantify the resilience of a network in terms of the extent of damage (as measured by disruption of connectivity between otherwise connected components or clusters of nodes) caused to the network by the removal of a set of critical nodes (called the attack set) [32]. Because the nodes in the attack set are crucial for maintaining connectivity across the network, removal of the nodes in the attack set can be expected to partition the network into clusters that are isolated from (i.e., disconnected from) each other. Different node resilience measures yield different attack sets with different degrees of sparseness [30]. In this work, we focused on three measures, namely vertex attack tolerance (VAT), integrity, and tenacity.

- The VAT of an undirected, connected graph $G = (V, E)$ is defined as [32, 33]: $VAT(G) = \min_{S \subset V, S \neq \emptyset} \{ \frac{|S|}{|V - S - C_{max}(V-S)| + 1} \}$, where S is an attack set and $C_{max}(V - S)$ is the largest

connected component in $V - S$. The goal is to identify small attack sets that consist of nodes that are most cricial in preserving network connectivity.

- Integrity is defined as [34]: $I(G) = \min_{S \subset V}\{|S| + C_{max}(V - S)\}$. Integrity balances the size of the attack set with the largest connected component in the network resulting from the removal of the attack set. An increase in attack set size can more easily be offset by a decrease in C_{max}, which means that attack set sizes will tend to be larger than with VAT. Generally, the attack set for integrity S_I will include the most crucial nodes (as generated by VAT), plus additional nodes that if removed, make the graph disconnected.

- Tenacity is defined as [35]: $T(G) = \min_{S \subset V}\{\frac{|S| + C_{max}(V - S)}{\omega(V - S)}\}$, where $\omega(V - S)$ is the number of connected components in $V - S$. This measure identifies nodes that, if removed, result in partitioning the graph into a large number of components. Thus, the tenacity attack set S_T will include almost all nodes that if removed, can make the graph disconnected.

In order to calculate these resilience measures, we utilized a heuristic known as Greedy betweenness centrality (Greedy-BC) [36]. For a given resilience measure, the Greedy-BC heuristic estimates candidate attack sets by iteratively selecting the node with highest betweenness centrality and removing it from the network. This process results in a node-removal ordering, which is used to calculate all three resilience measures. Each node is then, in order, added to the attack set, with a new graph configuration being generated with each iteration. The resilience measure is updated iteratively after each graph configuration update. The goal is to iteratively optimize the resilience measure. This greedy heuristic can be used to optimize VAT, integrity and tenacity with acceptable accuracy [32, 37]. Of the three resilience measures [30], VAT tends to yield the smallest attack set while tenacity yields the largest. A consequence of using the Greedy-BC heuristic is that the three attack sets are related as follows: $S_V \subseteq S_I \subseteq S_T$.

To select features for training IBD classifiers, we apply the NBR-Clust algorithm separately to the the IBD and Healthy networks to obtain the critical attack sets for healthy (G_H) and IBD (G_D) samples. We then select the nodes that are shared by the critical attack sets of both graphs.

Identification and evaluation of IBD biomarkers

Given a training data set DS400, a feature selection data set (e.g., DS50), a test data set DS573, a feature selection method (FSM), and the number of selected features $k \in \{10, 20, 30, 40, 50, 60\}$: First, we applied the FSM to the feature selection data to determine top k features. Then, we generated variants of the training and test data with only the selected features and used them to train and estimate the performance of a Random Forest (RF) [38] classifier. In each case, the input to the classifier consists of the relative abundance of the microbial taxa represented by the selected features. In our experiments, we used RF classifiers implemented in Scikit-learn [39] with the number of estimators set to 500 trees.

In addition to our proposed network-based feature selection methods, we considered the following traditional and commonly used feature selection methods: (i) Filter-based feature selection using Information Gain (IG) and F-Statistic (FStat); (ii) Recursive Feature Extraction (RFE) that uses LASSO [40] estimator for estimating the importance of features and removes the lowest ranked 10 features at each iteration; (iii) RF Feature Importance (RFFI) which is an embedded feature selection method where the FS data are used to train a RF classifier with 500 trees, and feature importance scores are then inferred from the learned model as suggested by Breiman [38].

We report the predictive performance of all IBD classifiers considered in this study as measured using Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Matthews Correlation Coefficient (MCC), and Area Under ROC Curve (AUC) [41].

Results

Feature selection improves the predictive performance of RF classifiers

Table 1 reports the performance of top (in terms of highest AUC and smallest number of selected features) RF classifiers using five different feature selection data sets as well as using all input features (FSM = None). For RF classifier without feature selection method, the AUC is 0.74. Using the smallest feature selection data set (DS50), the three traditional feature selection methods yield RF classifiers with better AUC scores. The highest observed AUC corresponds to a RF classifier trained using the top 50 features selected using RFFI method. On the other hand, when using the largest feature selection dataset (DS400), all feature selection methods yield models with AUC better than the baseline model with no feature selection. Interestingly, RFFI seems to benefit substantially by increase in the size of the feature selection data set since it returns only 20 features that are as discriminative as the 50 features determined using DS50.

We observed that some feature selection methods (e.g., IG) are sensitive to changes in the FSDS. For example, the best subset of features returned using the IG filter is with DS300. On the other hand, with DS400 (which includes all instances in DS300), the IG filter fails to

Table 1. Performance of the top (in terms of highest AUC and smallest number of selected features) performing RF classifiers for different choices of feature selection data set and traditional feature selection methods.

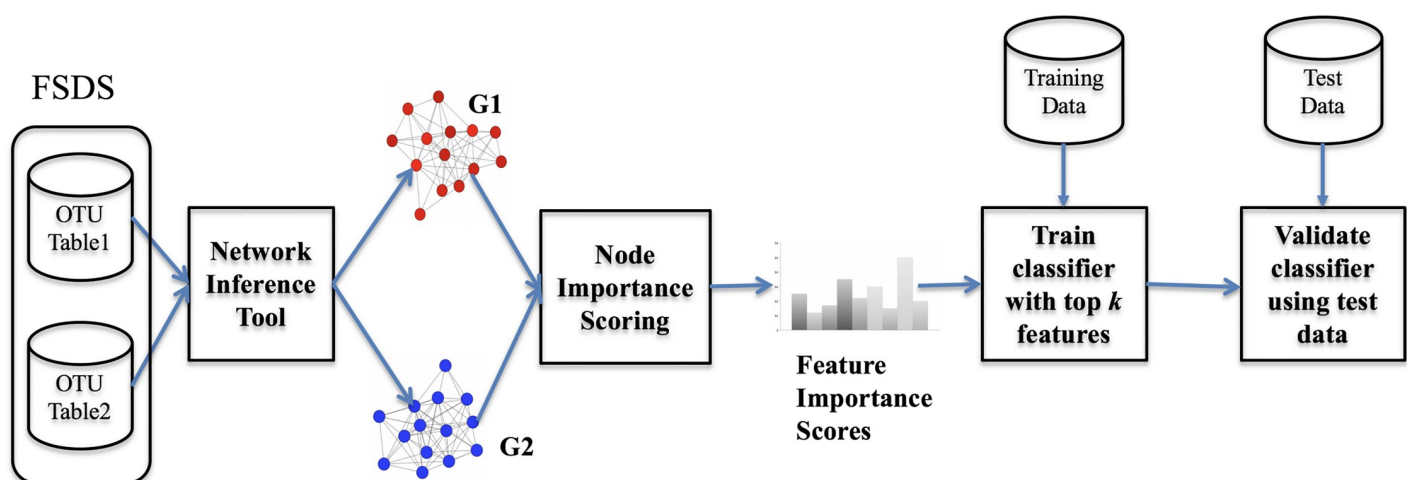| FSDS	FSM	# Features	ACC	Sn	Sp	MCC	AUC
DS50	None	NA	0.66	0.64	0.75	0.31	0.74
	IG	60	0.65	0.62	0.78	0.32	0.76
	FStat	60	0.63	0.64	0.62	0.21	0.69
	RFE	40	0.69	0.66	0.78	0.36	0.79
	RFFI	50	0.68	0.65	0.82	0.38	0.80
DS100	None	NA	0.66	0.64	0.75	0.31	0.74
	IG	60	0.65	0.62	0.74	0.29	0.75
	FStat	20	0.68	0.66	0.72	0.32	0.74
	RFE	50	0.66	0.62	0.81	0.35	0.78
	RFFI	40	0.68	0.65	0.80	0.37	0.79
DS200	None	NA	0.66	0.64	0.75	0.31	0.74
	IG	20	0.69	0.68	0.73	0.34	0.79
	FStat	50	0.68	0.67	0.72	0.31	0.75
	RFE	60	0.65	0.62	0.76	0.30	0.78
	RFFI	20	0.67	0.63	0.81	0.36	0.79
DS300	None	NA	0.66	0.64	0.75	0.31	0.74
	IG	30	0.69	0.66	0.80	0.38	0.80
	FStat	60	0.68	0.67	0.75	0.34	0.76
	RFE	60	0.68	0.65	0.80	0.36	0.79
	RFFI	30	0.68	0.64	0.81	0.37	0.79
DS400	None	NA	0.66	0.64	0.75	0.31	0.74
	IG	60	0.64	0.61	0.73	0.28	0.75
	FStat	40	0.70	0.69	0.72	0.34	0.76
	RFE	60	0.64	0.62	0.73	0.28	0.76
	RFFI	20	0.69	0.68	0.76	0.36	0.80

determine a good subset of selected features. We suspect that the biomarkers identified using such unstable feature selection methods are likely to be unreliable.

Performance of network-based feature selection methods

Results in Table 1 demonstrate the superior performance of RF feature importance for identifying a small subset of discriminative features from metagenomics data which is widely acknowledged in the literature [42, 43]. Here, we report results of experiments (using the framework in Fig 1) designed to address the following questions: (i) which network inference tool learns graphs that could be suitable for our network-based feature selection method?; (ii) how do the results of network-based feature selection using different Node Topological Property Scoring (NTPS) and Critical Attack Set Scoring (CASS) compare to each other as well as to results in Table 1?

First, for each of the five FSDS considered in our experimented and using graphs generated by five NIMs, we evaluated our NBBD framework using six topological properties for NTPS approach and k identified biomarkers for $k \in \{10, 20, 30, 40, 50, 60\}$. A total of 900 experiments were conducted and are reported in Tables A-E in in S1 File. Table 2 summarizes all these tables for results obtained using DS50 by reporting the performance of top performing (in terms of highest AUC and smallest number of selected features) RF classifiers. Table 2 reveals the following interesting observations: (i) Models using networks generated by CoNet, Proxi, and RMT achieve performance comparable to that of best performing models in Table 1 using RFFI and RFE feature selection; (ii) The AUC of the top performing models obtained using RMT graphs are consistently good (i.e., AUC scores in the range 0.77-0.78), while other NIMs yield top performing models with a wider range of AUC scores; (iii) There is no single topological property that can be used to train RF classifiers that outperform their counterparts trained using other properties. However, the topological properties that work best appear to depend on the network construction method used. For example, CoNet and Proxi based models achieve their highest AUC scores using 'and' and 'cn' properties, respectively. Even though RMT based models have almost the same AUC for all six different topological properties, the method seems to work best with 'cc' property since it reaches the highest AUC score of 0.78 using only 20 features whereas it requires at least 50 features using other properties.

Fig 1. Overview of the NBBD framework. Feature Selection Dataset (FSDS) which is a subset of, or the same as, training data set in the form of two OTU tables corresponding to two groups of metagenomics samples are first used to construct two networks. The node importance scoring modules compares topological properties of shared nodes in the two networks and outputs scores to prioritize the input features. Top selected features are then used to train and evaluate a classifier.

Table 2. Performance of top (in terms of highest AUC and smallest number of selected features) performing RF classifiers for combinations of different choices of Network Inference Method (NIM) and network-based feature selection using different properties for Node Topological Property Scoring. All results were obtained using DS50 as the feature selection dataset.

NIM	FSM	# Features	ACC	Sn	Sp	MCC	AUC
CoNet	and	50	0.69	0.66	0.81	0.38	0.79
	btw	30	0.67	0.64	0.78	0.34	0.78
	cc	60	0.67	0.65	0.76	0.33	0.77
	cls	60	0.66	0.63	0.77	0.32	0.74
	cn	60	0.67	0.64	0.79	0.35	0.78
	ncn	60	0.66	0.64	0.73	0.31	0.75
MB	and	50	0.64	0.63	0.69	0.26	0.71
	btw	50	0.65	0.62	0.77	0.31	0.74
	cc	50	0.61	0.59	0.72	0.24	0.71
	cls	50	0.63	0.61	0.69	0.24	0.72
	cn	60	0.62	0.61	0.67	0.23	0.69
	ncn	60	0.62	0.57	0.80	0.30	0.75
Proxi	and	60	0.64	0.61	0.76	0.30	0.75
	btw	50	0.69	0.67	0.78	0.36	0.77
	cc	60	0.62	0.61	0.68	0.24	0.70
	cls	50	0.57	0.53	0.72	0.20	0.67
	cn	40	0.65	0.62	0.77	0.31	0.78
	ncn	60	0.67	0.65	0.75	0.33	0.77
RMT	and	50	0.66	0.64	0.75	0.32	0.78
	btw	50	0.67	0.64	0.79	0.35	0.78
	cc	20	0.68	0.66	0.78	0.35	0.78
	cls	60	0.67	0.65	0.78	0.35	0.77
	cn	60	0.68	0.66	0.75	0.33	0.77
	ncn	60	0.68	0.66	0.76	0.34	0.78
SparCC	and	60	0.61	0.57	0.73	0.25	0.69
	btw	60	0.68	0.66	0.75	0.34	0.75
	cc	40	0.60	0.57	0.71	0.23	0.70
	cls	60	0.66	0.65	0.72	0.29	0.73
	cn	50	0.66	0.64	0.72	0.29	0.72
	ncn	60	0.63	0.60	0.74	0.28	0.71

Second, we repeated the experiments described in the preceding paragraph but using CASS based on three node resilience measures as the Node Importance Scoring module in our NBBD framework. The performance of the resulting RF classifiers are reported in Table F in S1 File and summarized in Table 3 for DS50. Table F in S1 File shows that the highest AUC score of 0.79 can be reached using DS100 and graphs learned using Proxi (and 28 features) or SparCC (and 51 features) as well as using DS400 and graphs obtained using SparCC (and 54 features). Results in Table 3 suggest that the three CASS methods seem to need larger feature selection data sets in order to reach a predictive performance comparable to those obtained using traditional feature selection methods or NTPS methods. Unlike all other feature selection methods considered in this work, CASS methods do not require the user to provide the number of features to be selected from the input data as a parameter.

In summary, our results suggest that the five NIMs, except MB [26], can be successfully used in our NBBD framework for identifying discriminative features (i.e., potential IBD biomarkers) from metagenomics data. Our results also show that network-based feature selection

Table 3. Performance (highest AUC attained, and the smallest number of features chosen) by the top performing RF classifiers for combinations of different choices of Network Inference Method (NIM) and network-based feature selection using three resilience measures for Critical Attack Set Scoring (CASS). Results obtained using DS50 as the feature selection data set.

NIM	FSM	# Features	ACC	Sn	Sp	MCC	AUC
AUC CoNet	CASS_I	21	0.66	0.64	0.75	0.31	0.77
	CASS_T	35	0.67	0.64	0.76	0.33	0.76
	CASS_V	21	0.66	0.64	0.75	0.31	0.77
MB	CASS_I	6	0.51	0.47	0.68	0.12	0.61
	CASS_T	33	0.57	0.53	0.73	0.21	0.65
	CASS_V	NA	NA	NA	NA	NA	NA
Proxi	CASS_I	11	0.65	0.63	0.72	0.28	0.72
	CASS_T	39	0.65	0.64	0.70	0.27	0.73
	CASS_V	1	0.25	0.07	0.96	0.04	0.51
RMT	CASS_I	8	0.49	0.46	0.58	0.03	0.52
	CASS_T	12	0.56	0.52	0.72	0.19	0.64
	CASS_V	3	0.64	0.64	0.61	0.21	0.62
SparCC	CASS_I	117	0.66	0.64	0.74	0.31	0.76
	CASS_T	125	0.66	0.64	0.72	0.30	0.76
	CASS_V	NA	NA	NA	NA	NA	NA

https://doi.org/10.1371/journal.pone.0225382.t003

methods are comparable to some commonly used traditional feature selection methods including the widely used RFFI. Moreover, with small size feature selection data sets, network-based feature selection methods applied to RMT graphs outperform traditional feature selection methods.

Performance of hybrid feature selection methods

Preliminary results reported in an early version of this work (Fig 4 in Abbas et al. [20]) show that only 12 OTUs were shared among the three subsets of 30 biomarkers determined using RFFI and two instances of the NBBD framework. Therefore, we hypothesize that the feature importance scores estimated using RFFI and the best instances of our NBBD framework are complementary with each other. To test this hypothesis, we developed a hybrid feature selection method that returns the product of RFFI and NBBD based on NTPSs as combined feature importance. Results for the hybrid method are reported for each of the five FSDS using graphs generated by five NIMs and instances of the NBBD framework using six topological properties for the NTPS approach and the top $k \in \{10, 20, 30, 40, 50, 60\}$ biomarkers in Tables G-K in S1 File and the top performing RF classifiers using DS50 are reported in Table 4.

Table H in S1 File reports the results for RF classifier using hybrid feature selection based on instances of the NBBD framework applied to MB graphs and shows that the two best performing RF classifiers with AUC scores of 0.82 and 0.81 are obtained using the 'and' property and the top 10 and 20 features (respectively). Interestingly, these two classifiers were trained using features determined using MB graphs inferred from DS50. This is a substantial improvement in performance compared with the RF model trained using RFFI and features determined using DS50 (see Table 1) which has an AUC score of 0.80 using 50 features. In addition, several RF models with AUC scores higher than 0.80 were obtained using Proxi, RMT, and SparCC graphs (see Tables G-K in S1 File).

Table 4 summarizes the results in Tables G-K in S1 File by reporting the top performing RF classifiers obtained using DS50 (i.e., the smallest feature selection dataset). In this table, two RF classifiers using MB and RMT graphs have equal AUC scores of 0.82. Several RF classifiers

Table 4. Performance of the top performing RF classifiers (with the highest AUC and using the smallest number of features) for combinations of different choices of Network Inference Method (NIM) and hybrid feature selection based on RFFI and different properties for Node Topological Property Scoring. All results were obtained using DS50 as the feature selection dataset.

GIM	FS Method	# Features	ACC	Sn	Sp	MCC	AUC
CoNet	and	30	0.68	0.65	0.79	0.36	0.79
	btw	20	0.65	0.61	0.79	0.33	0.78
	cc	60	0.66	0.64	0.74	0.31	0.74
	cls	40	0.62	0.59	0.77	0.29	0.74
	cn	40	0.66	0.64	0.75	0.32	0.76
	ncn	40	0.67	0.65	0.75	0.32	0.76
MB	and	20	0.73	0.72	0.76	0.40	0.82
	btw	40	0.66	0.64	0.78	0.33	0.78
	cc	20	0.66	0.62	0.79	0.34	0.77
	cls	40	0.65	0.64	0.72	0.29	0.77
	cn	10	0.69	0.68	0.74	0.34	0.76
	ncn	20	0.65	0.61	0.80	0.34	0.79
Proxi	and	50	0.68	0.66	0.77	0.35	0.78
	btw	30	0.69	0.66	0.82	0.39	0.79
	cc	50	0.65	0.62	0.77	0.31	0.78
	cls	50	0.67	0.63	0.83	0.37	0.79
	cn	40	0.62	0.60	0.70	0.24	0.73
	ncn	40	0.68	0.65	0.80	0.36	0.79
RMT	and	60	0.68	0.64	0.80	0.36	0.79
	btw	40	0.64	0.60	0.80	0.32	0.78
	cc	40	0.69	0.65	0.81	0.38	0.82
	cls	50	0.69	0.66	0.80	0.37	0.80
	cn	40	0.64	0.60	0.80	0.32	0.76
	ncn	50	0.68	0.65	0.81	0.37	0.80
SparCC	and	30	0.67	0.64	0.78	0.34	0.80
	btw	40	0.70	0.68	0.78	0.37	0.79
	cc	30	0.66	0.63	0.79	0.34	0.78
	cls	30	0.67	0.64	0.80	0.36	0.78
	cn	50	0.67	0.63	0.82	0.36	0.80
	ncn	40	0.66	0.62	0.81	0.34	0.79

https://doi.org/10.1371/journal.pone.0225382.t004

reached an AUC score of 0.80, but only the model based on SparCC graphs is using a small number of features. Comparing results in Tables 2 and 4 suggests that the RF classifiers using hybrid feature selection outperform counterpart RF classifiers using NTPS only in terms of predictive performance and/or number of features used to train the models.

Analysis of top performing models and the identified IBD biomarkers

Table 5 compares the performance of the top RF classifiers obtained using traditional feature selection and hybrid feature selection methods evaluated in our experiments. Using a hybrid

Table 5. Performance comparison of top three RF classifiers obtained using traditional feature selection and hybrid feature selection methods.

NIM	FSDS	FS Method	# Features	ACC	Sn	Sp	MCC	AUC
NA	DS400	RFFI	20	0.69	0.68	0.76	0.36	0.80
MB	DS50	RFFI × and	20	0.73	0.72	0.76	0.40	0.82
RMT	DS50	RFFI ×cc	40	0.69	0.65	0.81	0.38	0.82

https://doi.org/10.1371/journal.pone.0225382.t005

Table 6. AUC scores for top three RF classifiers obtained using RFFI feature selection and two hybrid feature selection methods, MB_and and RMT_cc, using different feature selection datasets.

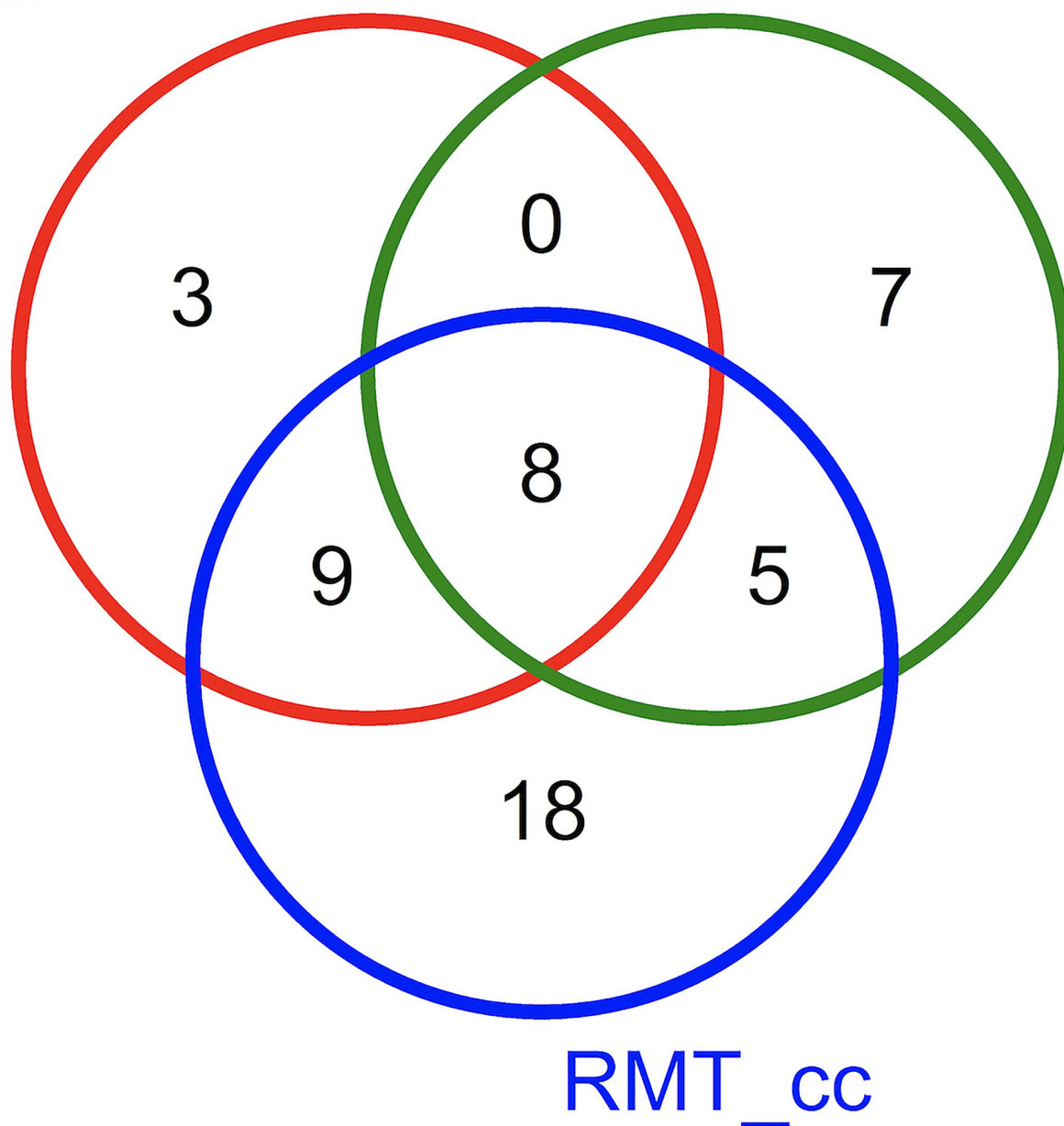| FSDS	RFFI	MB_and	RMT_cc
DS50	0.76	0.82	0.82
DS100	0.78	0.76	0.79
DS200	0.79	0.75	0.77
DS300	0.77	0.79	0.77
DS400	0.80	0.78	0.75

https://doi.org/10.1371/journal.pone.0225382.t006

scoring method combining RFFI (estimated from DS50) and 'and' scores (determined from MB graphs), a RF classifier trained using the top 20 features outperforms the best RF developed using RFFI (estimated from DS400) in four out of five performance metrics. Table 6 shows the AUC scores for these three models using different FSDSs. Since feature selection data sets are nested (i.e., $DS50 \subset DS100 \subset FSD200 \subset DS300 \subset DS400$), we expect feature selection methods to return the same or better subset of features as we increase the size of the FSDS used. Our expectation is almost realized using the RFFI method, except that there is a drop in AUC score when DS300 is used. On the other hand, our expectation is violated using the hybrid feature selection methods. The highest AUC score is observed using DS50, and increasing the size of the FSDS leads to a drop in classifier performance. This suggests that NIMs such as MB and RMT might be highly unstable to changes in the input data. In other words, networks constructed from DS50 and DS400 (as an example) are substantially different. For instance, Fig A in S2 File compares the four MB graphs generated using the MB method from IBD and healthy samples in DS50 and DS400. We found that MB constructs two networks (over the same set of nodes) but with a minimal overlap in edges from DS50 and DS400 data. In the absence of the ground truth, we can not determine which network is closer to reality. However, our results show that graphs inferred from DS50 allow our NBBD framework to identify a better set of features.

Fig 2 shows the Venn diagram of unique and shared OTUs among the three subsets of features used for training the top three models in Table 5. We found that the number of unique OTUs in each subset is 7, 3, and 18 for RFFI, MB_and, and RMT_cc sets, respectively. Interestingly, 17 out of the 20 features in MB_and are also in RMT_cc and 8 out of these 17 common OTUs are also shared with RFFI. Table L in S1 File lists the OTUs in these three sets of selected features. We further conducted downstream statistical analysis of the common 8 OTUs which are highlighted in bold in Table L in S1 File. More precisely, we assessed the significance of the difference between the medians of sample normalized relative abundance of these OTUs in IBD and healthy populations using the Kruskal-Wallis nonparametric test (Figs B-F in S2 File). Analysis of DS400 (Fig F in S2 File) shows significantly higher abundance of (Aggregatibacter, Fusobacterium, and Sutterella) in IBD samples relative to healthy samples. The increase of Aggregatibacter genus in IBD samples has been reported in several recent studies [44, 45]. Also, the high abundance of Fusobacterium in IBD samples has been suggested as a biomarker in several studies [3, 46]. Sutterella spp. have been frequently associated with several human diseases including autism and IBD [47, 48]. However, other studies [49, 50] have suggested that Sutterella spp. are unlikely to play a role in the pathogenesis of IBD. Fig F in S2 File also shows significant decreases in Roseburia, Dialister, and Clostridiales. These three biomarkers have been repeatedly reported in previous studies [51–53]. Finally, results of our statistical analysis reported in Fig F (S2 File) suggest that two of our top identified genera biomarkers, Bacteroides and Oscillospira, have no significant

Fig 2. Venn diagram of unique and shared features selected using RF Feature Importance (RFFI), network-based feature selection applied to MB (RMT) networks and using 'and' ('cc') for node importance scoring.

differences in IBD and control samples. Bacteroides is a dominant and biologically important bacteria genus in the microbiota of the human gastrointestinal tract [54] and Oscillospira is an under-studied bacterial genus that is hard to cultivate but is consistently being identified in several human gut microbiota association studies [55]. This highlights the need for developing more sophisticated differential abundance tests that take into account the sparsity and compositional nature of metagenomics data.

Sensitivity analysis of Kruskal-Wallis and Mann-Whitney nonparametric tests against the number of samples analyzed has been conducted using all variants of FSDS. The complete results of this analysis is reported in Figs B-K in S2 File. Surprisingly, both tests failed to show any significant differences between IBD and healthy groups using DS50. Overall, the results from the two nonparametric tests are in agreement with each other, and our results suggest that at least 100 samples are needed for each group in order to demonstrate significant differences in the abundances of six out of the top eight identified biomarkers.

Comparative analysis of IBD and healthy networks

We show that interesting biological findings can be derived from comparative analysis of the IBD and healthy ecological networks inferred from the DS400 training set using the RMT program [27]. Our choice of RMT network inference tool is motivated by the superior performance of the classifiers based on RMT networks (see Tables 2 and 5). In this analysis, we compared the two networks at the global, module, and node levels and interpreted some of the findings in the context of existing literature. Table M in S1 File reports several simple topological properties of the IBD and health networks. Surprisingly, the values of these simple global properties look similar in the two networks. Next, we consider differences between the two networks at the module and node levels.

For module level comparisons, we used the MCODE clustering algorithm [56], implemented as a Cytoscape [57] plugin, to extract top highly-connected modules in IBD and healthy networks, respectively. Fig L in S2 File shows the highest scored module from each network. We found that the top IBD module has 29 nodes while the top healthy module has only 20 nodes. Interestingly, nodes in the healthy module have more interactions among each other (i.e., average node degree is 16.7) while nodes in the IBD modules have less interactions (i.e., average node degree is 14.1). In the healthy module, the top interacting genera are Coprobacillus, Staphylococcus, Turicibacter, Clostridium. Each of them have 19 interacting neighbor. In the IBD module, the top interacting node (genus) is Anaerostipes with 20 interacting neighbors. Anaerostipes genus is one of the major bacteria that utilize lactate to produce butyrate, which is beneficial to colonic health [58]. Finally, we used DyNet [59] to compare the top IBD and healthy modules. The results are shown in Fig M in S2 File, where: green edges represent edges present only in healthy module; red edges represent edges present only in IBD module; and gray edges represent edges present in both networks. The results show that the two modules share five known genera. Fig N in S2 File shows the sub-network including these five genera. All possible pairwise interactions among these five genera are observed in Fig N in S2 File except for two pairs, (Parvimonas, Porphyromonas) and (Epulopiscium, Desulfovibrio), where the interactions are exclusive in IBD and healthy modules, respectively. Interestingly Porphyromonas and Parvimonas can be found in the oral cavity and the colon and were found to be more abundant in patients with colorectal cancer (CRC) [60, 61].

We also analyzed the differences between IBD and healthy networks at the node level. Following Abbas and El-Manazalawy [19], we limited this analysis to nodes corresponding to our identified potential biomarkers. Specifically, we generated biomarkers sub-networks including the 8 common biomarkers identified in this study (see Table L in S1 File). Fig O in S2 File shows that these eight OTUs seem to have more interactions among each other in the IBD population (11 edges) than in the healthy control samples (6 edges). A major difference between IBD and healthy biomarkers sub-networks is the Dialister node. In the healthy biomarkers sub-network, Dialister is an isolated node, while in the IBD biomarker sub-network, Dialister is interacting with the Bacteroides node. The preceding analyses set the stage for

experimental validation and detailed assessment of the biological significance of the identified interactions to be pursued as topics of future research.

Discussion

The past decade has witnessed a revolution in microbiology and microbiome research. Advances in sequencing technologies and computational techniques coupled with large scale collaborative efforts such as Human Microbiome Project (HMP) [62] and American Gut Project [63] have generated unprecedented amounts of metagenomics data. Analysis and interpretation of such data presents many statistical and computational challenges [64, 65]. One such challenge has to do with the reliable identification of biomarkers (in the form of species, genes, or pathways) that differentiate between two or more phenotypes [12].

To address this challenge, we have developed NBBD, a novel metagenomics system biology framework for microbial biomarker discovery. The NBBD framework integrates network analysis and machine learning approaches for reliable identification of biomarkers from metagenomics data. Given two OTU tables corresponding to two phenotypes, NBBD uses any existing tool for constructing phenotype specific networks from the data. Depending on the tool used, these networks model the interactions, the correlations, or the proximity relationships between microbes. Next, the nodes are scored using different scoring methods that quantify the extent to which the nodes contribute to differences in the topological properties of the nodes in the two networks. The k top-scoring nodes are used as the set of selected features to train and test classifier using machine learning. We conducted extensive experiments to evaluate the NBBD framework, configured using five different network inference tools and nine different node importance scoring methods, using a large data set from a cohort of 657 IBD and 316 healthy healthy pediatric metagenomics biopsy samples, respectively.

Although several tools for constructing microbial ecology networks from metagenomics data have been developed, they leave considerable room for improvement [12, 66]. For example, Weiss et al. [66] benchmarked the performance of eight correlation detection strategies on simulated and real metagenomics data and showed significant inconsistency (in terms of number of edges) among graphs inferred using different tools. Using simulated data, they showed that all of the tools exhibited extremely low precision (below 0.20). That is, for every identified true edge, there are at least four false positive edges in the constructed network. While the five network construction tools considered in our study are among the top performing tools in Weiss et al. [66], they are far from perfect. It is indeed remarkable that the noisy networks produced by such tools can be used to reliably identify discriminative features and to identify potential IBD biomarkers.

In this study, we performed experiments to examine the sensitivity of classifiers to the number of samples in the feature selection data set. To facilitate fair comparison between classifiers, we used the entire training data for training the classifiers using the features determined based on different subsets of the training data. Our results suggest that traditional feature selection methods fail to determine a minimal subset of discriminative features from small feature selection datasets. Interestingly, we found that several network-based feature selection methods returned a minimal subset of discriminative features using the smallest feature selection data set, DS50.

In this work, we showed how to develop network-based feature selection methods using arbitrary topological properties of vertices in the network (e.g., betweenness centrality, membership in attack sets, etc.) It would be interesting to explore other network based feature selection methods, e.g., using vertex similarity algorithm (e.g., SimRank [67] and ASCOS [68]), network similarity algorithms (e.g., DeltaCon [69]), and network-based anomaly detection

methods [70]. Work in progress is aimed at exploring the utility of these algorithms for developing more sophisticated Node Importance Scoring (NIS) modules for the NBBD framework.

Our sensitivity analysis also revealed that the microbial ecology networks constructed using state-of-the-art network construction methods are highly sensitive to the data samples used to construct the network. Needless to say, this lack of stability of network construction algorithms has serious implications for subsequent biological interpretation of microbial ecology networks, and in the contest of our work, the reliability of the biomarkers discovered from analysis of microbial ecology networks. In order for the predictive models trained using the features selected using network-based feature selection methods) to be reliable, we need to ensure the feature selection methods have a high degree of stability with respect to changes in the underlying network. Note that the stability of feature selection algorithms is a function of both the properties of the algorithm itself as well as the data supplied to the algorithm. Hence, improvements are needed on both fronts.

Fundamentally, constructing microbial ecology networks from metagenomic data requires determining the correlation or similarity between (abundances of) microbial taxa from a relatively small number of metagenomic samples. This problem is not fundamentally different from the problem of determining gene co-expression networks from gene expression data [71], or that of determining functional brain networks from fMRI data [72]. All of these applications present some shared challenges: In most cases, the number of features (genes, brain regions, microbial taxa) far exceed the number of data samples; It is generally impossible, without making additional assumptions or incorporating domain knowledge, to distinguish between direct and indirect correlations; The choice of the correlation or similarity measure is often application-dependent. Methods for microbial ecology network estimation from metagenomic data could benefit greatly from recent advances in high dimensional correlation matrix estimation [73–76]. Work in progress is aimed at evaluating the applicability of such methods in constructing stable microbial ecology networks from metagenomic data.

Conclusions

We have proposed a novel Network-Based Biomarker Discovery (NBBD) framework for detecting disease biomarkers from metagenomics data. NBBD consists of two major customizable modules: A network inference module, for constructing microbial ecology networks from OTU tables extracted from the metagenomic data for the phenotypes of interest; and a node importance scoring module, which compares the resulting phenotype-specific networks and scores the nodes based on different measures of the node's contribution to the differences between the networks.

We have evaluated the proposed NBBD framework, using five different network construction methods, in combination with nine different node importance scoring methods, on a large data set from a cohort of 657 IBD and 316 healhy pediatric metagenomics biopsy samples. Our results show that NBBD, when used to train predictive models for IBD diagnosis from metagenomic data, is very competitive with some of the state-of-the-art feature selection methods including the widely used method based on random forest feature importance scores. Our results further show that a hybrid approach that combines NBBD scores and the random forest feature importance scores yields further improvements in performance. Furthermore, the proposed method is able to achieve its best observed performance using only 50 samples for feature selection. Work in progress is aimed at further improving the two key components of NBBD, e.g., by incorporating recent advances in high dimensional correlation matrix estimation [73–76] to improve the reliability and the stability of the resulting networks, exploring improved node scoring methods. Other promising directions for future research include

systematic evaluation of the NBBD framework for biomarker discovery from different types of omics data, integrative analyses of multi-omics data [77, 78], e.g., using information-preserving low-dimensional network embeddings [79].

Supporting information

S1 File. Detailed results using different network inference methods, different feature selection methods, and different datasets.
(XLSX)

S2 File. Supplementary results for: Kruskal-Wallis and Mann-Whitney statistical tests applied to different feature selection datasets; and comparative network analysis.
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: Tayo Obafemi-Ajayi, Yasser EL-Manzalawy.

Formal analysis: Mostafa Abbas, John Matta, Thanh Le, Halima Bensmail, Tayo Obafemi-Ajayi, Vasant Honavar, Yasser EL-Manzalawy.

Methodology: Yasser EL-Manzalawy.

Writing – original draft: Mostafa Abbas, John Matta, Tayo Obafemi-Ajayi, Yasser EL-Manzalawy.

Writing – review & editing: Mostafa Abbas, John Matta, Thanh Le, Halima Bensmail, Tayo Obafemi-Ajayi, Vasant Honavar, Yasser EL-Manzalawy.

References

1. Schmidt C, Stallmach A. Etiology and pathogenesis of inflammatory bowel disease. Minerva gastroen-terologica e dietologica. 2005; 51(2):127–145. PMID: 15990703

2. Van Assche G, Dignass A, Panes J, Beaugerie L, Karagiannis J, Allez M, et al. The second European evidence-based consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis. Journal of Crohn's and Colitis. 2010; 4(1):7–27. https://doi.org/10.1016/j.crohns.2009.12.003

3. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell host and microbe. 2014; 15(3):382–392. https://doi.org/10.1016/j.chom.2014.02.005

4. Kamada N, Seo SU, Chen GY, Nunez G. Role of the gut microbiota in immunity and inflammatory disease. Nature Reviews Immunology. 2013; 13(5):321. https://doi.org/10.1038/nri3430 PMID: 23618829

5. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology. 2014; 146(6):1489–1499. https://doi.org/10.1053/j.gastro.2014.02.009 PMID: 24560869

6. Manichanh C, Reeder J, Gibert P, Varela E, Llopis M, Antolin M, et al. Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. Genome research. 2010. https://doi.org/10.1101/gr.107987.110 PMID: 20736229

7. Ruemmele FM, Targan SR, Levy G, Dubinsky M, Braun J, Seidman EG. Diagnostic accuracy of serological assays in pediatric inflammatory bowel disease. Gastroenterology. 1998; 115(4):822–829. https://doi.org/10.1016/s0016-5085(98)70252-5 PMID: 9753483

8. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, et al. A microbial signature for Crohn's disease. Gut. 2017; p. gutjnl–2016. https://doi.org/10.1136/gutjnl-2016-313235

9. Holtman GA, Lisman-van Leeuwen Y, Reitsma JB, Berger MY. Noninvasive tests for inflammatory bowel disease: a meta-analysis. Pediatrics. 2016; 137(1):e20152126. https://doi.org/10.1542/peds.2015-2126

10. Viennois E, Zhao Y, Merlin D. Biomarkers of inflammatory bowel disease: from classical laboratory tools to personalized medicine. Inflammatory bowel diseases. 2015; 21(10):2467–2474. https://doi.org/10.1097/MIB.0000000000000444 PMID: 25985250

11. Shanahan F, Quigley EM. Manipulation of the microbiota for treatment of IBS and IBD: challenges and controversies. Gastroenterology. 2014; 146(6):1554–1563. https://doi.org/10.1053/j.gastro.2014.01.050

12. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome biology. 2011; 12(6):R60. https://doi.org/10.1186/gb-2011-12-6-r60 PMID: 21702898

13. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017; 5(1):27. https://doi.org/10.1186/s40168-017-0237-y PMID: 28253908

14. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010; 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

16. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature methods. 2013; 10(12):1200. https://doi.org/10.1038/nmeth.2658 PMID: 24076764

17. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease. 2015; 26(1):27663.PMID: 26028277

18. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003; 3(Mar):1157–1182.

19. Abbas M, EL-Manzalawy Y. Predictive and Comparative Network Analysis of the Gut Microbiota in Type 2 Diabetes. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2017. p. 313–320.

20. Abbas M, Le T, Bensmail H, Honavar V, El-Manzalawy Y. Microbiomarkers discovery in inflammatory bowel diseases using network-based feature selection. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2018. p. 172–177.

21. Matta J, Obafemi-Ajayi T, Borwey J, Wunsch D, Ercal G. Robust graph-theoretic clustering approaches using node-based resilience measures. In: Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE; 2016. p. 320–329.

22. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet. 2017; 390(10114):2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0

23. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology. 2012; 8(7):e1002606. https://doi.org/10.1371/journal.pcbi.1002606 PMID: 22807668

24. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS computational biology. 2012; 8(9):e1002687. https://doi.org/10.1371/journal.pcbi.1002687 PMID: 23028285

25. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology. 2015; 11(5):e1004226. https://doi.org/10.1371/journal.pcbi.1004226 PMID: 25950956

26. Meinshausen N, Bühlmann P, et al. High-dimensional graphs and variable selection with the lasso. The annals of statistics. 2006; 34(3):1436–1462. https://doi.org/10.1214/009053606000000281

27. Deng Y, Jiang YH, Yang Y, He Z, Luo F, Zhou J. Molecular ecological network analyses. BMC bioinformatics. 2012; 13(1):113. https://doi.org/10.1186/1471-2105-13-113 PMID: 22646978

28. Faust K, Lima-Mendez G, Lerat JS, Sathirapongsasuti JF, Knight R, Huttenhower C, et al. Cross-biome comparison of microbial association networks. Frontiers in microbiology. 2015; 6:1200. https://doi.org/10.3389/fmicb.2015.01200 PMID: 26579106

29. El-Manzalawy Y. Proxi: a Python package for proximity network inference from metagenomic data. bioRxiv. 2018; p. 357764.

30. Matta J, Obafemi-Ajayi T, Borwey J, Sinha K, Wunsch D, Ercal G. Node-Based Resilience Measure Clustering with Applications to Noisy and Overlapping Communities in Complex Networks. Applied Sciences. 2018; 8(8):1307. https://doi.org/10.3390/app8081307

31. Hagberg A, Swart P, Chult DS. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.

32. Matta J, Ercal G, Borwey J. The vertex attack tolerance of complex networks. RAIRO-Operations Research. 2017; 51(4):1055–1076. https://doi.org/10.1051/ro/2017008

33. Ercal G. On Vertex Attack Tolerance in Regular Graphs. arXiv preprint arXiv:14092172. 2014.

34. Barefoot CA, Entringer R, Swart H. Vulnerability in graphs—a comparative survey. J Combin Math Combin Comput. 1987; 1(38):13–22.

35. Cozzens M, Moazzami D, Stueckle S. The tenacity of a graph. In: Proc. Seventh International Conference on the Theory and Applications of Graphs, Wiley, New York; 1995. p. 1111–1122.

36. Matta J, Ercal G, Borwey J. The vertex attack tolerance of complex networks. RAIRO-Operations Research. 2017; 51(4):1055–1076. https://doi.org/10.1051/ro/2017008

37. Matta J. A Comparison of Approaches to Computing Betweenness Centrality for Large Graphs. In: International Workshop on Complex Networks and their Applications. Springer; 2017. p. 3–13.

38. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011; 12(Oct):2825–2830.

40. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; p. 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

41. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16(5):412–424. https://doi.org/10.1093/bioinformatics/16.5.412 PMID: 10871264

42. Ditzler G, Morrison JC, Lan Y, Rosen GL. Fizzy: feature subset selection for metagenomics. BMC bioinformatics. 2015; 16(1):358. https://doi.org/10.1186/s12859-015-0793-8 PMID: 26538306

43. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS computational biology. 2016; 12(7):e1004977. https://doi.org/10.1371/journal.pcbi.1004977 PMID: 27400279

44. Sokol H, Leducq V, Aschard H, Pham HP, Jegou S, Landman C, et al. Fungal microbiota dysbiosis in IBD. Gut. 2017; 66(6):1039–1048. https://doi.org/10.1136/gutjnl-2015-310746 PMID: 26843508

45. Menon R, Ramanan V, Korolev KS. Interactions between species introduce spurious associations in microbiome studies. PLoS computational biology. 2018; 14(1):e1005939. https://doi.org/10.1371/journal.pcbi.1005939 PMID: 29338008

46. Strauss J, Kaplan GG, Beck PL, Rioux K, Panaccione R, DeVinney R, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. Inflammatory bowel diseases. 2011; 17(9):1971–1978. https://doi.org/10.1002/ibd.21606 PMID: 21830275

47. Wang L, Christophersen CT, Sorich MJ, Gerber JP, Angley MT, Conlon MA. Increased abundance of Sutterella spp. and Ruminococcus torques in feces of children with autism spectrum disorder. Molecular autism. 2013; 4(1):42. https://doi.org/10.1186/2040-2392-4-42 PMID: 24188502

48. Lavelle A, Lennon G, O'sullivan O, Docherty N, Balfe A, Maguire A, et al. Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. Gut. 2015; p. gutjnl–2014. https://doi.org/10.1136/gutjnl-2014-307873 PMID: 25596182

49. Mukhopadhya I, Hansen R, Nicholl CE, Alhaidan YA, Thomson JM, Berry SH, et al. A comprehensive evaluation of colonic mucosal isolates of Sutterella wadsworthensis from inflammatory bowel disease. PLoS One. 2011; 6(10):e27076. https://doi.org/10.1371/journal.pone.0027076 PMID: 22073125

50. Hiippala K, Kainulainen V, Kalliomäki M, Arkkila P, Satokari R. Mucosal Prevalence and Interactions with the Epithelium Indicate Commensalism of Sutterella spp. Frontiers in microbiology. 2016; 7:1706. https://doi.org/10.3389/fmicb.2016.01706 PMID: 27833600

51. Machiels K, Joossens M, Sabino J, De Preter V, Arijs I, Eeckhaut V, et al. A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. Gut. 2014; 63(8):1275–1283. https://doi.org/10.1136/gutjnl-2013-304833 PMID: 24021287

52. Joossens M, Huys G, Cnockaert M, De Preter V, Verbeke K, Rutgeerts P, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut. 2011; p. gut–2010. https://doi.org/10.1136/gut.2010.223263

53. Tye H, Yu CH, Simms LA, de Zoete MR, Kim ML, Zakrzewski M, et al. NLRP1 restricts butyrate producing commensals to exacerbate inflammatory bowel disease. Nature communications. 2018; 9(1):3728. https://doi.org/10.1038/s41467-018-06125-0 PMID: 30214011

54. Delday M, Mulder I, Logan ET, Grant G. Bacteroides thetaiotaomicron ameliorates colon inflammation in preclinical models of Crohn's disease. Inflammatory bowel diseases. 2018; 25(1):85–96. https://doi.org/10.1093/ibd/izy281

55. Konikoff T, Gophna U. Oscillospira: a central, enigmatic component of the human gut microbiota. Trends in microbiology. 2016; 24(7):523–524. https://doi.org/10.1016/j.tim.2016.02.015 PMID: 26996766

56. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics. 2003; 4(1):2. https://doi.org/10.1186/1471-2105-4-2 PMID: 12525261

57. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13 (11):2498–2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

58. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome biology. 2012; 13(9):R79. https://doi.org/10.1186/gb-2012-13-9-r79 PMID: 23013615

59. Goenawan IH, Bryan K, Lynn DJ. DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics. 2016; 32(17):2713–2715. https://doi.org/10.1093/bioinformatics/btw187 PMID: 27153624

60. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature communications. 2017; 8(1):1784. https://doi.org/10.1038/s41467-017-01973-8 PMID: 29209090

61. Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, et al. The oral microbiota in colorectal cancer is distinctive and predictive. Gut. 2018; 67(8):1454–1463. https://doi.org/10.1136/gutjnl-2017-314814 PMID: 28988196

62. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007; 449(7164):804. https://doi.org/10.1038/nature06244 PMID: 17943116

63. Debelius JW, Vázquez-Baeza Y, McDonald D, Xu Z, Wolfe E, Knight R. Turning participatory microbiome research into usable data: lessons from the American Gut Project. Journal of microbiology & biology education. 2016; 17(1):46. https://doi.org/10.1128/jmbe.v17i1.1034

64. Waldor MK, Tyson G, Borenstein E, Ochman H, Moeller A, Finlay BB, et al. Where next for microbiome research? PLoS Biology. 2015; 13(1):e1002050. https://doi.org/10.1371/journal.pbio.1002050 PMID: 25602283

65. Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. Microbiome data science: understanding our microbial planet. Trends in microbiology. 2016; 24(6):425–427. https://doi.org/10.1016/j.tim.2016.02.011 PMID: 27197692

66. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal. 2016; 10(7):1669. https://doi.org/10.1038/ismej.2015.235 PMID: 26905627

67. Jeh G, Widom J. SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2002. p. 538–543.

68. Chen HH, Giles CL. ASCOS: an asymmetric network structure context similarity measure. In: Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE; 2013. p. 442–449.

69. Koutra D, Vogelstein JT, Faloutsos C. Deltacon: A principled massive-graph similarity function. In: Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM; 2013. p. 162–170.

70. Goldstein M, Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one. 2016; 11(4):e0152173. https://doi.org/10.1371/journal.pone.0152173 PMID: 27093601

71. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in bioinformatics. 2017; 19(4):575–592.

72. He Y, Evans A. Graph theoretical modeling of brain connectivity. Current opinion in neurology. 2010; 23 (4):341–350.PMID: 20581686

73. Fan J, Fan Y, Lv J. High dimensional covariance matrix estimation using a factor model. Journal of Econometrics. 2008; 147(1):186–197. https://doi.org/10.1016/j.jeconom.2008.09.017

74. Bickel PJ, Levina E, et al. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008; 36(1):199–227. https://doi.org/10.1214/009053607000000758

75. Avella-Medina M, Battey HS, Fan J, Li Q. Robust estimation of high-dimensional covariance and precision matrices. Biometrika. 2018; 105(2):271–284. https://doi.org/10.1093/biomet/asy011 PMID: 30337763

76. Ravikumar P, Wainwright MJ, Raskutti G, Yu B, et al. High-dimensional covariance estimation by minimizing ℓ1-penalized log-determinant divergence. Electronic Journal of Statistics. 2011; 5:935–980. https://doi.org/10.1214/11-EJS631

77. EL-Manzalawy Y, Hsieh TY, Shivakumar M, Kim D, Honavar V. Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. BMC Medical Genomics. 2018; 11(3):71. https://doi.org/10.1186/s12920-018-0388-0 PMID: 30255801

78. EL-Manzalawy Y. CCA based multi-view feature selection for multi-omics data integration. In: 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2018. p. 1–8.

79. Sun Y, Bui N, Hsieh TY, Honavar V. Multi-View Network Embedding Via Graph Factorization Clustering and Co-Regularized Multi-View Agreement. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE; 2018. p. 1006–1013.