Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

CellPress

Review

Disentangling Interactions in the Microbiome: A Network Perspective

Mehdi Layeghifard,[1] David M. Hwang,[2,3] and David S. Guttman[1,4,*]

Microbiota are now widely recognized as being central players in the health of all organisms and ecosystems, and subsequently have been the subject of intense study. However, analyzing and converting microbiome data into meaningful biological insights remain very challenging. In this review, we highlight recent advances in network theory and their applicability to microbiome research. We discuss emerging graph theoretical concepts and approaches used in other research disciplines and demonstrate how they are well suited for enhancing our understanding of the higher-order interactions that occur within microbiomes. Network-based analytical approaches have the potential to help disentangle complex polymicrobial and microbe–host interactions, and thereby further the applicability of microbiome research to personalized medicine, public health, environmental and industrial applications, and agriculture.

Trends

Polymicrobial communities (microbiota) are complex, dynamic, and ubiquitous.

Microbiota play a central role in host health and development. For example, dysbiotic shifts in the composition of the human microbiome have been linked to a wide variety of health issues, such as obesity, diabetes, eczema, heart disease, asthma, colitis, etc.

The complexity of microbiomes motivates a movement from reductionist approaches that focus on individual pathogens in isolation to more holistic approaches that focus on interactions among members of the community and their hosts.

Network theory has emerged as an extremely promising approach for modelling complex biological systems with multifaceted interactions between members, such as microbiota.

Networks enhance the analysis of polymicrobial interactions within microbiota and their role in health, disease, and development.

The Hidden Parallel World of the Microbiome

Our world is dominated by, and wholly dependent on, complex microbial communities (i.e., microbiota) that are not a mere collection of independent individuals but a complex of interconnected ecological communities that communicate, cross-feed, recombine, and coevolve. Microbiome interactions are of course not limited to the immediate microbial community, but also occur between the microbes and their hosts, where they have been shown to play key roles in the development, metabolism, homeostasis, and immunity of their hosts [1,2]. While host-associated microbiomes can show striking variability from one healthy individual to another [3], perturbations or imbalance in the community composition (generally referred to as dysbiosis) are associated with unfavorable host responses, and sometimes serious pathologies. For example, dysbiosis of the human gut microbiome is associated with a wide range of pathologies, including diarrhea [4], diabetes [5], colorectal cancer [6], inflammatory bowel disease [7], irritable bowel syndrome [8], and obesity [9].

Despite their omnipresence, centrality to life, and clear link to health and disease, we are only beginning to understand how microbes interact with each other and their hosts. Ecological interactions within microbiomes, where for example microorganisms compete for resources [3] or exchange genetic material [10], are known to influence microbiome composition and host health [11,12]; nevertheless, the scope and characteristics of these polymicrobial interactions are still largely unexplored [3]. This is in large part due to the complex interplay that occurs among microbial taxa and their hosts, which makes the function of the collective microbiome more than the function of any of its constituent species [13,14]. This diversity and interdependence challenges classical models of single-species infections, where Koch's Postulate can be tested

[1]Department of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada
[2]Department of Pathology, University Health Network Toronto, Ontario, Canada
[3]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
[4]Centre for the Analysis of Genome Evolution & Function, University of Toronto, Toronto, Ontario, Canada

and causality inferred. For example, **microbial synergism** (see Glossary) is reported to cause increased levels of antibiotic resistance, biofilm development, tissue damage, and adaptation to the environment [15,16].

*Correspondence:
david.guttman@utoronto.ca
(D.S. Guttman).

Network-Thinking for Microbiome Research

Given the spectrum of host–microbe interactions associated with health, dysbiosis, as well as polymicrobial and single-agent infections, it is becoming increasingly clear that understanding microbiome interactions is essential for understanding microbiome function. These complex communities need to be viewed with an appreciation of the dynamic ecological and evolutionary processes that drive them. And we postulate that ecological insight into these relationships and processes will help in the development of therapeutic approaches to prevent dysbiosis, microbial infections, and other microbiome-associated pathologies [13].

Systems-oriented, graph-theoretical approaches can facilitate microbiome analysis and enhance our understanding of complex ecological and evolutionary processes. Microbiota are complex in both structure and function due to their dynamic nature, compositional variability, and their ability to self-reproduce and self-organize. This complexity can be well represented and modelled as **networks** (see Box 1 for a summary of network types). A key feature of network theory is that the architectural features of networks appear to be universal to most complex systems, including microbiomes, molecular interaction networks, computer networks, microcircuits, and social networks. This universality paves the way for using expertise developed in well-studied non-biological systems to characterize the intricate interwoven relationships that shape microbial interactions. The unparalleled value of network theory becomes apparent in cases where the goal is revealing patterns behind small- and large-scale ecological and evolutionary processes within high dimensional datasets with complex distributions [17]. The application of network theory to microbiome studies can be used to model the co-occurrence of microorganisms, find microbial relationships essential for community assembly or stability, and deduce the influence of various interactions on the host health. We feel that a more widespread application of network theory in microbiome analysis will provide valuable insights into the organization, function, and evolution of these important communities.

Constructing Microbiome Networks

A wide range of methods have been used to construct ecological networks based on microbiome data. These approaches vary in their efficiency, accuracy, speed, and computational requirements, and span from simple pairwise Pearson or Spearman correlation measures, to more complex multiple regression and **Gaussian graphical models**. Some of the methods, such as correlation-based methods, are quite popular due to their speed and ease of use [18], while others, such as **probabilistic graphical models**, have not been used extensively to address biological questions, but have seen good success in other fields and gained a reputation for accuracy and minimal bias. Here we will discuss some of the different methods used to construct microbiome networks.

Dissimilarity-Based Methods

The simplest and fastest way to construct co-occurrence networks from operational taxonomic unit (OTU) microbiome data is to use a pairwise dissimilarity index such as Bray–Curtis or Kullback–Leibler. Usually, the significance of pairwise (dis-)similarity scores is evaluated through a permutation test [19], and all significant pairwise connections are aggregated to construct a microbiome network. Faust and colleagues [19] developed a pipeline based on an ensemble approach to predict interactions in the oceanic plankton community [20]. This pipeline combines a number of measures of dependency, such as correlation (e.g., Spearman), similarity (e.g., mutual information), and dissimilarity (e.g., Kullback–Leibler).

Correlation-Based Methods

A popular alternative to dissimilarity-based network inference is correlation-based techniques. These methods detect significant pairwise interactions between OTUs using a correlation coefficient such as Pearson's product-moment correlation coefficient or Spearman's nonparametric rank correlation coefficient. Correlation-based network inference has been successfully used to study human gut [21], soil [22], and phyllosphere [23] microbiomes. For example, Arumugam and colleagues [21] identified three robust human gut enterotypes that were not nation- or continent-specific by analyzing a combined dataset of newly sequenced and published fecal metagenomes of individuals from four countries. However, the use of correlation coefficients to detect dependencies between members of a microbiome suffers from limitations such as detecting spurious correlations among low-abundance OTUs in zero-inflated data or being sensitive to compositionality [24]. Weiss *et al*. [25] evaluated the performance of eight correlation methods on both synthetic and real data in response to challenges specific to microbiome studies and assessed their ability to distinguish signals from noise and detect a range of ecological and time-series relationships. They reported the performance level and shortcomings of each method and provided specific recommendations for correlation technique usage.

Regression-Based Methods

Network inference methods based on pairwise association metrics such as Bray–Curtis and Pearson coefficient are not able to capture more complex forms of polymicrobial interactions [19]. One obvious alternative is to use multiple regression analysis to infer the abundance of one species from the combined abundances of other taxa. Although the method is simple and frequently used, the meaning and interpretation of regression results can be more difficult. For instance, the successfully predicted links might not always mean that there is an underlying biological basis for the association. Moreover, regression-based methods suffer from overfitting that increases with the number of predictor variables, and is associated with an commensurate increase in the number of false positives. Overfitting can be remedied by using sparse regression and cross-validation. One example of using logistic regression models in microbiome network analysis is the study of independent associations between bacteria, viruses, and risk factors in the upper respiratory tract of young children [26]. Using this approach, van den Bergh and colleagues found that *Streptococcus pneumoniae* colonization was positively correlated with the presence of *Haemophilus influenzae*, *Moraxella catarrhalis*, human rhinoviruses, and enteroviruses, and negatively correlated with the presence of *Staphylococcus aureus*. They also observed a strong positive association between *S. aureus* and influenza viruses and a negative association between human rhinoviruses and coronaviruses [26].

Probabilistic Graphical Models

As a recently developed framework, probabilistic graphical models (PGMs) employ ideas from discrete data structures in computer science to efficiently measure uncertainty in high dimensional data using probability theory. In other words, PGMs deal with uncertainty and complexity through the use of probability theory and graph theory, respectively. **Bayesian networks** (also called belief networks) and Markov networks (MN; also called **Markov random fields, MRFs**) are the most popular graphical models used. PGMs use graphs as the foundation for both measuring joint probability distributions (from which we can obtain marginal and conditional probabilities) and representing sets of conditional dependence and independence relations within data in a compact fashion. PGMs can be categorized as directed versus undirected, static versus dynamic, and probabilistic versus decisional. In microbiome networks, links between OTUs represent symmetric undirected associations unless networks are built from time series data, where the change of one factor may temporally lead or follow another factor. Static PGMs model a set of variables represented at a certain point in time, whereas dynamic PGMs model a set of variables across various time points. Finally, while probabilistic models only include

Glossary

Bayesian network (BN): a probabilistic graphical model to represent conditional dependencies between random variables via a directed acyclic graph (DAG; a finite directed graph with no cycles).
Closeness (of a node): the sum of the length of the shortest paths between each node and all other nodes in the network is that node's closeness.
Degree: the number of edges connecting each node to the rest of the network is that node's degree. In a directed network, indegree of a node is the number of edges leading to that node and the outdegree of a node is the number of edges leading away from that node.
Diameter: the maximum of pairwise distances between every two nodes.
Edge (link): a link between each pair of nodes is called an edge.
Gaussian graphical model (GGM): an undirected probabilistic graphical model for estimation of partial correlations as a measure of conditional dependence of any two variables to infer direct interactions between them.
Markov random field (MRF): a probabilistic graphical model similar to Bayesian network that satisfies Markov properties (i.e., conditional probability distribution of future states of the process depends only on the present state). However, MRF is undirected and could be cyclic (i.e., nodes being connected in a closed chain).
Microbial synergism: when synchronized action of a group of microbes is greater than the sum of the individual species actions.
Mixed graphical model (MGM): a graphical model to represent heterogeneous (mixed) datasets including both continuous and discrete variables.
Neighborhood: the neighborhood of a node is made up of all the other nodes connected to that node by an edge.
Network (graph): a group of two or more interconnected objects (e.g., microbial taxa).
Node: each object in a network is called a node.
Path: a sequence of edges, which connects a pair of nodes by possibly walking through any number of other nodes.

random variables, decisional PGMs also consider decision and utility variables. Markov networks are examples of undirected graphical models satisfying the Markov property such that a variable is independent of all other variables given its neighbors. In general, however, undirected graphs are consisted of three main groups: (i) correlation graphs, (ii) partial correlation graphs, and (iii) conditional independence graphs. These models are described in detail in Box 2.

Probabilistic graphical model (PGM): a probabilistic model in which the conditional dependence between random variables are depicted as a graph. In such a graph, the edges correspond to direct probabilistic interactions between the variables.
Shortest path: a path between two nodes in a network such that the sum of the weights of its constituent edges is minimized.

Network Inference Methods Robust to Compositionality Bias

Microbiome data usually suffer from two problematic features that confound their analysis. Firstly, OTU data are compositional; meaning that microbial counts are interdependent due to the normalization of counts to the total number of counts in the sample. This interdependence can lead to spurious results when using traditional statistical methods such as Pearson's correlation. Secondly, the ratio of observations (samples) to the number of variables (OTUs) is small. Recently, there have been many efforts to develop network construction methods robust to these two issues. These methods are described in Box 3.

Detecting Biologically Important Clusters from Networks

Clusters (also known as modules) are elementary units of any biological network, and their identification and characterization provides us with more information about the local interaction patterns in the network and their contribution to the overall structure, connectivity, and function of the network (Figure 1, Key Figure, panel D). Clusters are biologically important when considered as isolated, taxonomic, evolutionary or functional modules. High modularity indicates that the network has dense connections within certain groups of nodes and sparse connections between these groups. Several approaches have been developed to detect clusters within networks with varying degrees of success, which is partly due to inherent clustering ambiguity of the real-world networks. Identifying clusters within a biological network (e.g., groups of coexisting or coevolving microbes contributing towards a disease, or groups of functionally related molecules) is a key issue in network biology and one that is likely to grow in importance in the near future. We can identify these clusters either by using network topology to reveal the modular structure of the data, or by taking advantage of supplementary data (such as 'omics data and biomedical metadata), along with network topology to find closely tied clusters of nodes within networks. Topological clustering methods including hierarchical top-down and bottom-up methods, multilevel and Markov clustering algorithm are described in detail in Box 4.

Box 1. Network Types

The vast majority of highly diverse networks found in both natural and anthropogenic systems can be assigned to a small number of network types based on their topology or architecture. The first network model described mathematically in the literature is the *random network* introduced in 1960 by Paul Erdŏ"s and Alfred Rényi [52]. This model assumes a network of randomly interconnected nodes, in which some nodes are sparsely connected, whereas others have many more links (Figure IB). Nodes' degrees, therefore, follows a Poisson distribution, and most nodes have a number of connections comparable to the network's average degree.

In natural or artificial networks, however, the distribution of nodes' degrees rarely, if ever, follows a Poisson distribution. Most networks show a power-law degree distribution, where a few nodes have a very large number of connections, while other nodes have no or few connections [53]. The highly connected nodes are called hubs, and networks following a power-law degree distribution are often called scale-free networks (Figure ID). Cellular networks, genetic regulatory networks, and protein–protein interaction networks are biological examples of scale-free networks [54,55]. S-metric developed by Li and colleagues [56] is a useful method for explaining the differences between networks that have identical degree sequence, especially if they are scaling (i.e., there are bivariate relationships of power-law types, by which one attribute relates to another attribute raised to a power, called power-law or scaling exponent).

Small-world networks, by contrast, describe a model in which most nodes are accessible to every other node through a relatively short **path** (Figure IC). In other words, a small-world network is a network in which **diameter** increases proportionally to the logarithm of the number of nodes [57]. In this regard, small-world networks might resemble random networks. However, unlike random networks they show high local clustering among their components, which makes them more similar to regular networks (i.e., highly ordered nonrandom networks where all the nodes have exactly the same degree; Figure IA). This intermediate status of small-world networks has allowed Humphries and Gurney [58] to develop a testable measure of small-world-ness based on the compromise between high local clustering and **shortest path** length of the networks. Many real-life networks including technological, biological, social, and information networks have been characterized as small-world networks [57].

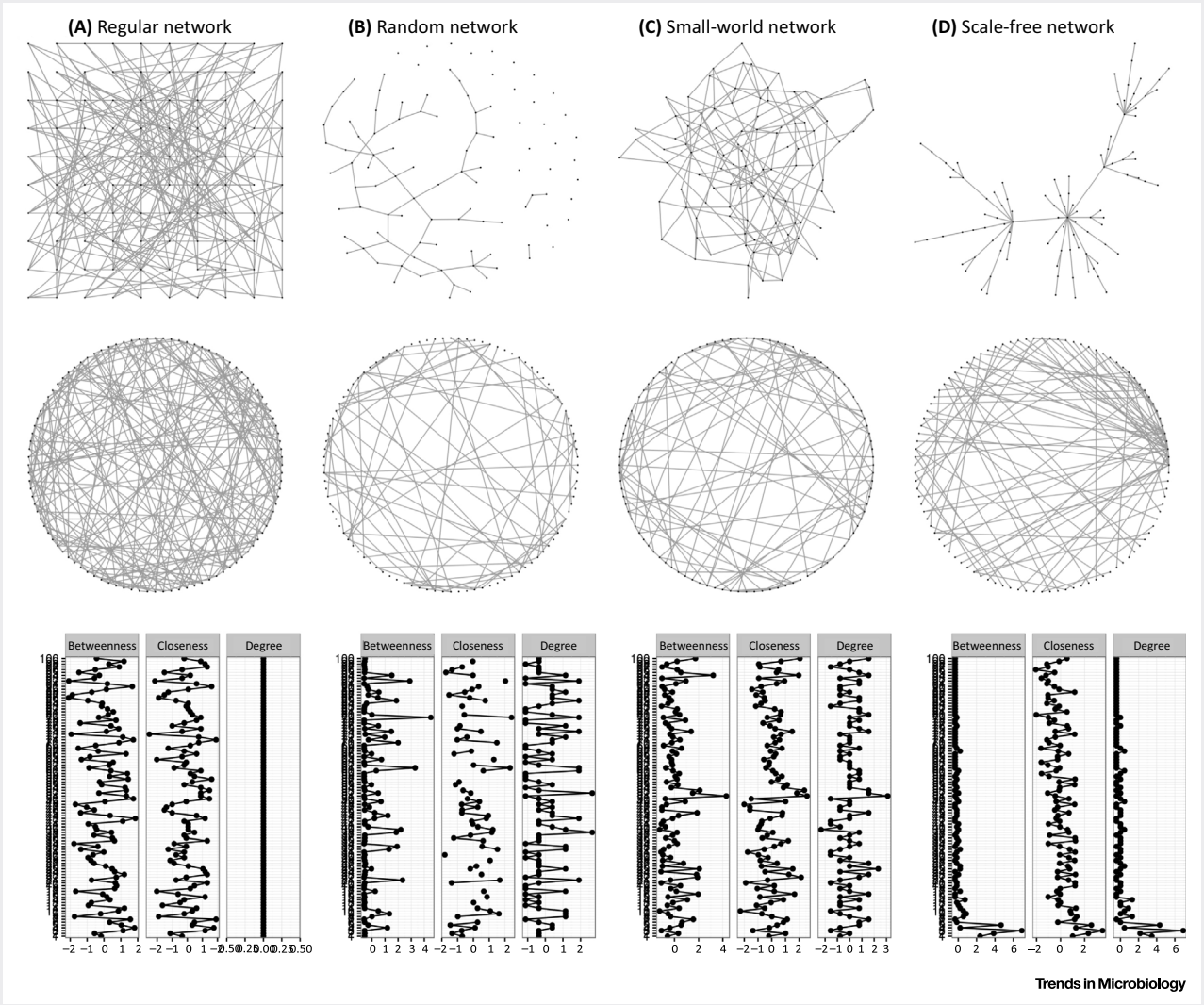

Figure I. Four Main Types of Network. The top panel shows the network representations of the four main types, each consisting of 100 nodes. The middle panel shows the same networks in a circular layout to accentuate the differences between the network types. The bottom panel shows plots of two node centrality measures as well as the degree distribution for each of the above networks (all measurements are standardized). (A) Regular network: each node has exactly the same number of links. (B) Random network: nodes are randomly connected to each other. (C) Small-world network: most nodes can be reached from any other node through a short path. (D) Scale-free network: the degree distribution of nodes follows a power law.

Revealing the modular structure of a microbiome co-occurrence network based on its topology will provide us with invaluable insights into complex polymicrobial interactions and co-occurrence patterns. However, topological approaches reveal no information on the underlying mechanisms shaping the modularity. A more insightful analysis will use other types of available data in combination with network topology to infer biologically significant modules. Such integrative algorithms can expand on topological approaches by evaluating both network connectivity and the correlations within biological and medical profiles across multiple samples or conditions. We refer the reader to a recent review of integrative approaches for finding modules in biological networks through merging topological data with 'omics data [27]. Many of these methods are potentially applicable to the integration of topological structure with biomedical profiles for the inference of modules in microbiome co-occurrence network. Integrative methods, for example, are being used to unify gene expression profiles with the topology of protein–protein interaction networks for the computation of joint membership probability of

Box 2. Network Construction Using Probabilistic Graphical Models (PGMs)

Correlation Graphs

In a correlation PGM, Pearson's correlation or Kendall's tau (which is more robust to outliers) can be used as the measure of associations in conjunction with a permutation test or a false discovery rate (FDR) calculator to control for multitesting. However, classical tests relying on Pearson's or Kendall's correlation coefficients are not guaranteed to detect all modes of dependence between the random variables.

Partial Correlation Graphs

Partial correlation measures the degree of association between two random variables excluding the effect of a set of controlling random variables. Unlike Pearson correlation, partial correlation of two OTUs calculates the amount of correlation left after eliminating the influence of other OTUs. Graphical Gaussian models (GGMs) are one example of methods using partial correlation to calculate relationships within data. GGMs are shown to be very effective in inferring conditional dependency and modeling interactions among genes [59]. Partial correlation has also been used to infer bacterial–bacterial, viral–bacterial and viral–viral associations in microbial communities in the upper respiratory tract of healthy children [26]. The main drawback of this method is that it cannot reliably estimate the partial correlation in high dimensional datasets, where the number of samples is smaller than the number of dimensions. This is due to the sample covariance of the data being noninvertible, which is required to calculate partial correlations.

Conditional Independence Graphs

Two events *A* and *B* are conditionally independent given a third event *C*, precisely if the knowledge of whether *A* occurs provides no information on the likelihood of *B* occurring and vice versa. In the special case of normality (i.e., Gaussian distribution), conditional independence graphs are equivalent to partial correlation graphs. Graphs constructed based on conditional independence are the most accurate and informative undirected graphs, but they are also the most difficult to construct due to the larger ratio of observations to the number of variables required for accurate estimation of conditional relationships. Most high-throughput datasets lack this property. To resolve this issue, low-order conditional independence graphs have been proposed to simplify the graphical model. This approach is based on the assumption that for sparse graphical models, the zero- and first-order conditional independencies are reasonably accurate estimation of the full conditional independencies within data. Low-order conditional independence graphs have previously been used to find dependencies between gene expression profiles and construct genetic networks [60]. Given the sparsity of microbiome data, we recommend using low-order conditional independence graphs to infer microbiome networks.

coregulated modules. A similar approach can be adopted for polymicrobial infections to infer shifts in microbiome composition in response to various antibiotics, by merging antibiotic treatment profiles with microbiome network topology.

Detecting Hub (Keystone) Species from Networks

One of the advantages of network analysis is that it can identify the most important nodes or hubs in a given network (Figure 1, panel C). In microbiome analysis, the importance of nodes can take a variety of meanings depending on the context and application. For example, the most important node may be the most influential member in the microbial community, the most essential microbe for community stability, the etiological agent of disease, or the organism responsible for disease transmission. Methods used to identify network hubs fall into one of three general categories: (i) centrality indices, (ii) node influence metrics, and (iii) link analysis methods. Node centrality indices identify which nodes are more central in a given network. A more central node is expected to have more links, reach all the other nodes more quickly, and control the flow between the other nodes. Examples of node centrality indices are **degree** centrality [28], **node**- and **edge**-betweenness [28,29], **closeness** [28] and Eigen-centrality [30]. Centrality indices suffer from two general drawbacks: (i) they only rank nodes, but do not measure the difference between them [31], and (ii) they underestimate the power of non-hub nodes due to heterogeneous topology of complex networks [32]. Berry and Widder [33] investigated the applicability of centrality measures on co-occurrence networks to find keystone species in microbial communities.

Node influence metrics, by contrast, are global metrics devised to measure the influence of all the nodes in a network. Accessibility [34] and Expected Force [35] metrics are two well known

Box 3. Network Construction Methods Robust to Compositionality

SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) is a novel statistical method trying to combine data transformations developed for compositional data with a sparse graphical model inference framework. SPIEC-EASI builds microbiome networks using sparse **neighborhood** and inverse covariance selection algorithms. SPIEC-EASI has been used to predict previously unknown microbial associations using data from the American Gut project [61].

SparCC (Sparse Correlations for Compositional data), by contrast, infers associations in compositional data by estimating the linear Pearson's correlations between the log-transformed components [62]. SparCC co-occurrence network analysis was employed to find a putative symbiotic relationship between *Chlorella vulgaris* (a high lipid-producing alkaliphilic alga) and *Pseudomonas* sp. in an outdoor, open pond used to produce algal biofuels [63]. However, log-transformation based methodologies should be used with caution, because assigning statistical significance to associations inferred by these methods have been shown to be problematic. Log transformation cannot be applied to zeros, which are frequent in microbiome data. To address this issue, zeros are usually substituted with a small value, known as a pseudocount. However, the choice of pseudocount values can influence the results drastically [64].

Faust *et al.* [19] developed a permutation-renormalization bootstrap method (ReBoot) to evaluate the significance of Pearson's correlation coefficients (as well as other similarity, dissimilarity and correlation measures; packaged in a tool called CoNet [65]) estimated from compositional data while mitigating the compositionality bias. ReBoot was applied on 20 different 16S rDNA sequencing data sets to investigate how co-occurrence networks differ across biomes and what factors influence their properties [66]. The main finding of this study was that count matrix properties, such as sequencing depth and evenness, are potential confounding factors that might influence network construction and should be taken into account while interpreting or comparing microbiome networks.

REBACCA (Regularized Estimation of the BAsis Covariance based on Compositional dAta), a newer method in this category, tries to identify significant co-occurrence patterns by finding sparse solutions to a system with a deficient rank [67]. To be specific, REBACCA estimates the correlations between pairs of basis abundance using the log ratio transformation of count or proportional data. Application of REBACCA on a metagenomic dataset of mouse skin microbiota showed that the microbial correlation patterns in immunized samples are different from the nonimmunized ones due to the response of a group of Bacteroidetes and clostridia (associated with anaerobic infections) to immunization [67].

CCLasso (Correlation inference for Compositional data through Lasso) is yet another method developed to infer correlations from compositional data [68]. CCLasso uses least squares with L1 penalty after log ratio transformation for raw compositional data to infer the correlations among microbes through a latent variable model. $\ell 1$ penalized estimation methods are usually used to prevent overfitting due to either collinearity of the covariates or high-dimensionality.

MENAP (Molecular Ecological Network Analysis Pipeline) is a Random Matrix Theory (RMT)-based method that is developed to address the issue of arbitrary choice of threshold used to include or exclude interaction from ecological networks [69]. This method automatically identifies a threshold and defines an adjacency matrix based on that. Finally, an undirected network graph will be constructed from the adjacency matrix.

MInt (Microbial Interaction) is a Poisson-multivariate normal hierarchical model to find taxon-taxon interactions from metagenomic count data by controlling for confounding predictors at the Poisson layer, and capturing direct microbial interactions at the multivariate normal layer, using an $\ell 1$ penalized precision matrix [70]. MInt was shown to outperform SparCC and graphical lasso methods in both synthetic and experimental experiments.

examples of node influence metrics. The former uses the diversity of random walks to measure how accessible the rest of the network is from a given start node, whereas the latter measures node influence from an epidemiological perspective by computing the spreading power of all network nodes using the combinatorics inherent in local topology. Despite being fundamentally different, both of these measures can be meaningfully computed from the structure of the network alone. Finally, link analysis algorithms are iterative and interactive data-analysis techniques used to evaluate connections between nodes as well as identify, analyze, and visualize patterns in data. The PageRank algorithm [36] is a well known example (developed and popularized by Google) from this category that works by counting the number and quality of links to a node to determine its importance and assign a numerical weight to it. The underlying assumption of PageRank is that hubs are likely to be more connected to other nodes when

Key Figure

Microbiome Network Analysis

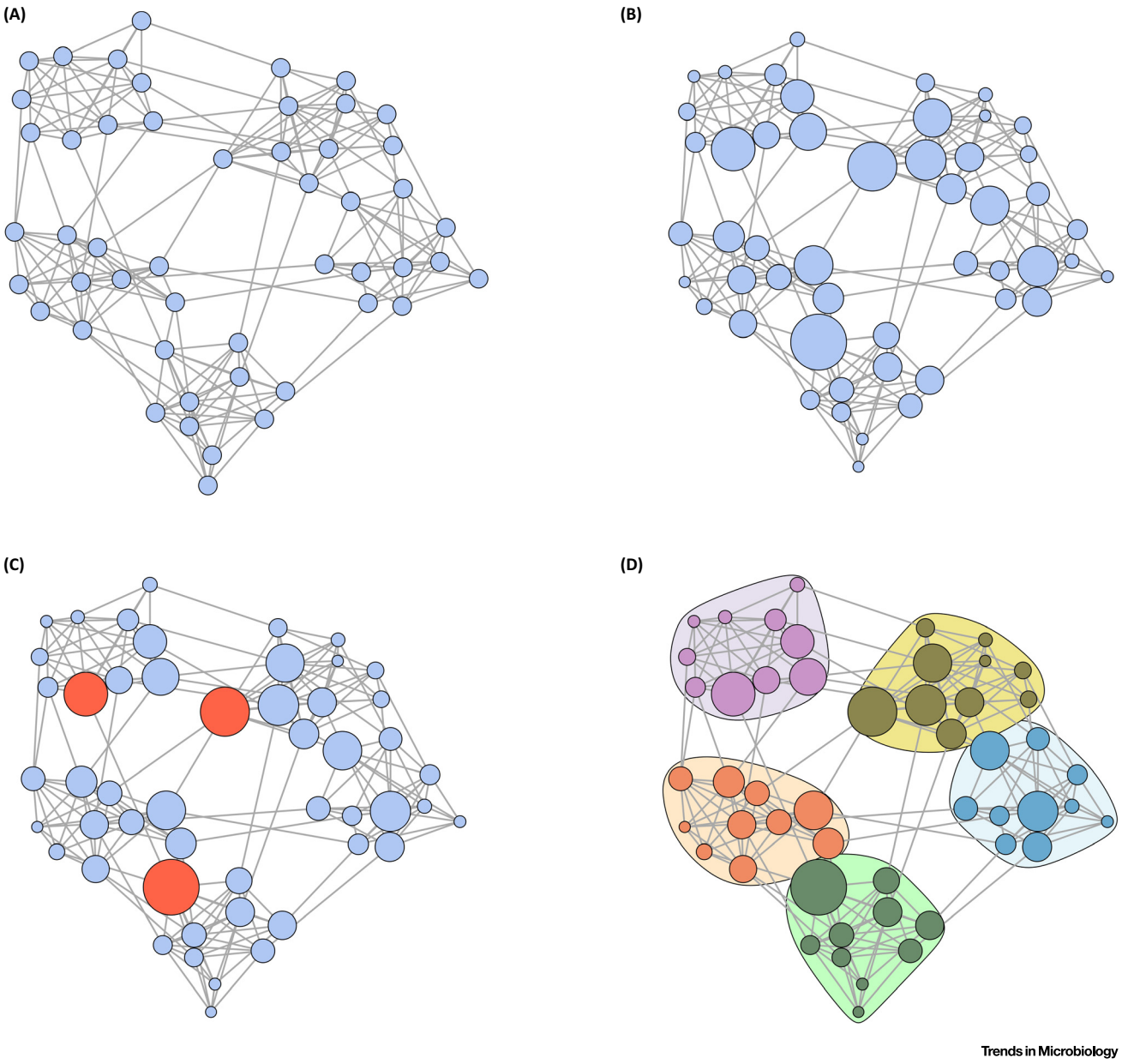

Figure 1. (A) A microbiome network built from an OTU table. Each blue node represents a microbe from the microbiome, and each gray link represents a pairwise co-occurrence or interaction. (B) The same microbiome network with nodes' sizes proportionate to HITS scores computed for all the microbes. (C) The same microbiome network with hub (keystone) species highlighted in red. (D) The same microbiome network with microbes clustered into five distinct groups.

Box 4. Topological Clustering Methods for Partitioning Data into Biologically Meaningful Subgroups

Edge betweenness is a popular method in which links are removed in the decreasing order of their betweenness scores [71]. The betweenness score of any given edge is the number of shortest paths that pass through that edge.

Leading eigenvector tries to optimize a quality function called modularity [72]. In each step, the network is split into two parts in a way that the separation itself yields a significant increase in the modularity. One drawback of this method is that it might not work on degenerate (or sparse) networks.

Fast greedy tries to optimize a modularity function too, but in a greedy manner. Initially, every node belongs to a separate cluster, and clusters are merged iteratively if the merge is locally optimal until it is not possible to increase the modularity any more. The algorithm is fast and is the method of choice as a first approximation test due to its lack of tunable parameters [72].

Walktrap algorithm is based on the general idea that, if one performs random walks on a network, then the walks are more likely to stay within the same cluster due to the assumed higher level of interconnectedness within clusters. Walktrap is slightly slower than fast greedy approach, but it is shown to be more accurate [72].

Multilevel clustering (Louvain), in every step, reassigns the nodes to clusters in a local, greedy manner to find the cluster in which it achieves the highest contribution to modularity [73]. Then, all the clusters are collapsed into single nodes and the process continues. The stopping condition is met when there is only one single node left or when the modularity cannot be increased any more in a step.

Markov clustering algorithm (MCL) tries to simulate a stochastic flow within the network structure, strengthening the flow where nodes are highly interconnected and weakening it in other regions until the flow process stabilizes [74]. MCL has found great popularity in network biology and has been used to characterize protein families within protein networks [75], detect orthologous and homologous groups [76], predict protein complexes from protein interaction networks [77], find gene clusters base on expression profiles [78], and to find clusters of putative pathogens and growth-promoting bacteria in phyllosphere microbiome [23].

compared to non-hub nodes. Hyperlink-Induced Topic Search (HITS) [37] is another example of link analysis algorithms that estimates an authority score and a hub score for each node in a network. A higher authority score for a node means that node receives connections from many other nodes, whereas a high hub score means that node is pointing to many other nodes. In undirected microbiome networks, however, adjacency matrix is symmetric and the hub scores are the same as authority scores.

Capturing Microbiome Dynamics

Under steady conditions, microbial communities can remain stable for long periods of time [38], but they can also change abruptly in response to small perturbations, such as antibiotic treatment [39] or change in diet [40]. Hence, dedicated time-series analysis tools must be used to take full advantage of temporal data, reveal periodic patterns, build predictive models, or quantify irregularities that make community behavior unpredictable. Local similarity analysis (LSA) [41–43] is a popular method developed to study temporal changes in the composition of microbiota through inferring significant associations among OTUs as well as between OTUs and their host without requiring substantial data reduction. LSA has been successfully employed to relate the taxonomic groups to various seasonal events using temporal dynamics data, revealing both contemporaneous and time-lagged correlation patterns among populations in the bacterioplankton community members and their associations to environmental variables [44].

An alternative to LSA is using Bayesian network models. There are two types of Bayesian network model for dynamic processes: dynamic Bayesian networks (DBNs), and temporal event networks (TENs). A DBN consists of a series of time slices (or snapshots), each representing the state of all the variables at a certain time. In contrast, a node in a TEN represents the time of occurrence of an event or a change in its state. TENs are simpler, more efficient, and thus more suitable for problems involving only few state changes in the temporal range of interest. Application of DBN model on infant gut microbial ecosystem has resulted in capturing specific

relationships and general trends, such as increasing amounts of clostridia, residual amounts of bacilli, and increasing amounts of Gammaproteobacteria that later receded in favor of clostridia [45]. Moreover, Faust and colleagues introduced a time-varying network construction method to infer temporal variation of microbial interactions. They suggested that this method could be combined with DBN to build a time-varying dynamic Bayesian network method [46].

Integrative Network Analysis

Microbiome data are usually accompanied by various types of metadata such as age, body-mass index, diet, antibiotic regimens, and performance measures of the affected organs. Integrating significant microbial associations detected from OTU tables with metadata measurements of interest not only will provide us with valuable insights into the dynamics of the interactions between external factors and the microbial community, but also can help us understand how the detected relationships might change when additional variables are taken into account. Nevertheless, determining the conditional independence relationships of all variables within a microbial community is a daunting challenge. For example, a well known probabilistic approach to represent dependencies within data is using an MRF in the form of an undirected graphical model. However, estimating the MRF underlying the multivariate distribution over all variables is extremely difficult when variables are coming from different domains (e.g., continuous, ordinal, categorical and count-valued).

Mixed graphical models (**MGM**) have recently been proposed to estimate the MRF underlying a joint distribution over mixed variables [47]. More advanced methods, such as data fusion approaches, are capable of mining heterogeneous datasets and directly exploit associated data without the need to transform data types into a common data space. Data fusion approaches have been successfully applied to predict gene function [48], mine disease–disease associations [49], to predict drug toxicity [50], and infer gene networks [51]. Mixed graph models have a huge yet untapped potential for finding relationships between microbial associations (e.g., in the form of network hubs or clusters) and external factors or conditions that are assumed to influence or be under influence of microbial community. For example, these models can help tie a change in disease severity to alterations in microbial interactions, or associate a modification in antibiotic regimen to shifts in microbiome composition.

Concluding Remarks

Despite significant advances in the past few years, the use of network biology is still in its infancy, mainly due to its complexity in terms of both concept and implementation. While we can see formidable potential for network-based analytic approaches, significant work must still be done before we approach a full systems-level understanding of microbe–microbe and host–microbe interactions using network theory. Understanding, diagnosing, and therapeutically treating dysbioses can undergo tremendous progress through the application of advanced network theory approaches, ranging from characterization of network topologies and cluster detection, to the integrative dynamic analyses that are able to characterize the interplay between microbiomes and external stimuli. Additionally, most current work concentrates on snapshots of activity in a few selected environments and in an abstract space. Moreover, the vast majority of network-based studies so far have focused on microbial co-occurrence or co-exclusion networks, which usually lead to partial or inadequate interpretation of the state and properties of the microbial community. However, a more insightful understanding of these complex interactions will require the analysis of data collection as a whole using advanced network theory. Integrative analyses will enable us to look at all types of interaction (e.g., metabolic, regulatory, spatial, etc.) within the microbiome, and can offer further insights into how the microbiome affects the host and vice versa. The result of such an integrative analysis will be a network of networks, which will lead to a more comprehensive understanding of the complex interplay of internal and external interactions involved in microbiome behavior, or the role of the microbiome in host health and

Outstanding Questions

Are there keystone hub species (OTUs) that have an outsized influence on microbiome and host function, and can we identify these via network theory? What roles do these hub species play in ecological community stability, and how do they influence host and environmental health?

What groups of co occurring or co evolving microbes are found in microbiomes? Do these groups function as symbiotic microbial consortia that influence host and environmental health? Is network theory the best analytical approach for identifying these co occurrences?

How can network theory be most effectively used to find the signatures of ecological interactions that result in microbiome stability or dysbiosis?

How can integrative, systems-level analyses of microbiomes be used to better understand commensal and pathogenic mechanisms of host–microbe interactions?

What are the impacts of ecological, evolutionary, and co evolutionary processes occurring within the microbiome on the host?

disease. Application of network biology will significantly enhance our understanding of human microbiome, in particular, and will ultimately have important implications for our understanding of diseases and the eventual targeted pharmaceutical modification of diseased modules. Enhanced knowledge of microbial interactions will also improve manipulation of microbiome composition and function, which can occur through the introduction of new members to the community (e.g., probiotics and fecal transplantation), or the removal of unwanted members (e.g., antibiotics and intestinal lavage).

Acknowledgments

This work was supported by the Canadian Institutes of Health Research (CIHR) Canadian Microbiome Initiative Emerging Team Grant (CMF108027) and Cystic Fibrosis Canada.

References

1. Hooper, L.V. *et al.* (2012) Interactions between the microbiota and the immune system. *Science* 336, 1268–1273
2. Thaiss, C.A. *et al.* (2016) The microbiome and innate immunity. *Nature* 535, 65–74
3. Hacquard, S. *et al.* (2015) Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* 17, 603–616
4. Pop, M. *et al.* (2014) Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol.* 15, R76
5. Gülden, E. *et al.* (2015) The gut microbiota and Type 1 Diabetes. *Clin. Immunol.* 159, 143–153
6. Yu, Y-N. and Fang, J-Y. (2015) Gut microbiota and colorectal cancer. *Gastrointest. Tumors* 2, 26–32
7. Morgan, X.C. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13, R79
8. Jalanka-Tuovinen, J. *et al.* (2014) Faecal microbiota composition and host–microbe cross-talk following gastroenteritis and in post-infectious irritable bowel syndrome. *Gut* 63, 1737–1745
9. Perry, R.J. *et al.* (2016) Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. *Nature* 534, 213–217
10. Stecher, B. *et al.* (2012) Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1269–1274
11. Rodrigues Hoffmann, A. *et al.* (2016) The microbiome: the trillions of microorganisms that maintain health and cause disease in humans and companion animals. *Vet. Pathol.* 53, 10–21
12. Hansen, T.H. *et al.* (2015) The gut microbiome in cardio-metabolic health. *Genome Med.* 7, 33
13. Rogers, G.B. *et al.* (2013) Interpreting infective microbiota: the importance of an ecological perspective. *Trends Microbiol.* 21, 271–276
14. Passos da Silva, D. *et al.* (2014) Bacterial multispecies studies and microbiome analysis of a plant disease. *Microbiol. Read. Engl.* 160, 556–566
15. Dalton, T. *et al.* (2011) An in vivo polymicrobial biofilm wound infection model to study interspecies interactions. *PloS One* 6, e27317
16. Murray, J.L. *et al.* (2014) Mechanisms of synergy in polymicrobial infections. *J. Microbiol. Seoul Korea* 52, 188–199
17. Proulx, S.R. *et al.* (2005) Network thinking in ecology and evolution. *Trends Ecol. Evol.* 20, 345–353
18. Faust, K. and Raes, J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* 10, 538–550
19. Faust, K. *et al.* (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8, e1002606
20. Lima-Mendez, G. *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science* 348, 1262073
21. Arumugam, M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473, 174–180
22. Barberán, A. *et al.* (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* 6, 343–351
23. Copeland, J.K. *et al.* (2015) Seasonal community succession of the phyllosphere microbiome. *Mol. Plant. Microbe Interact.* 28, 274–285
24. Chen, E.Z. and Li, H. (2016) A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32, 2611–2617
25. Weiss, S. *et al.* (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681
26. van den Bergh, M.R. *et al.* (2012) Associations between pathogens in the upper respiratory tract of young children: interplay between viruses and bacteria. *PloS One* 7, e47711
27. Mitra, K. *et al.* (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732
28. Freeman, L.C. (1978) Centrality in social networks conceptual clarification. *Soc. Netw.* 1, 215–239
29. Brandes, U. (2001) A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177
30. Bonacich, P. (1987) Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182
31. Bauer, F. and Lizier, J.T. (2012) Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: a walk counting approach. *EPL Europhys. Lett.* 99, 68007
32. Sikic, M. *et al.* (2013) Epidemic centrality - is there an underestimated epidemic impact of network peripheral nodes? *Eur. Phys. J. B* 86, 440
33. Berry, D. and Widder, S. (2014) Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Microb. Symbioses* 5, 219
34. Viana, M.P. *et al.* (2013) Accessibility in networks: A useful measure for understanding social insect nest architecture. *Chaos Solitons Fractals* 46, 38–45
35. Lawyer, G. (2014) Understanding the spreading power of all nodes in a network. *Sci. Rep.* 5, 8665
36. Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, pp. 107–117, Amsterdam
37. Kleinberg, J.M. (1999) Authoritative sources in a hyperlinked environment. *JACM* 46, 604–632
38. Faith, J.J. *et al.* (2013) The long-term stability of the human gut microbiota. *Science* 341, 1237439
39. Cho, I. *et al.* (2012) Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* 488, 621–626
40. David, L.A. *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563
41. Xia, L.C. *et al.* (2013) Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinforma. Oxf. Engl.* 29, 230–237
42. Xia, L.C. *et al.* (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol.* 5, 1–12

43. Ruan, Q. *et al.* (2006) Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinforma. Oxf. Engl.* 22, 2532–2538
44. Eiler, A. *et al.* (2012) Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J.* 6, 330–342
45. McGeachie, M.J. *et al.* (2016) Longitudinal prediction of the infant gut microbiome with dynamic Bayesian networks. *Sci. Rep.* 6, 20359
46. Faust, K. *et al.* (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66
47. Murray, J.S. *et al.* (2013) Bayesian Gaussian copula factor models for mixed data. *J. Am. Stat. Assoc.* 108, 656–665
48. Žitnik, M. and Zupan, B. (2015) Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 41–53
49. Žitnik, M. *et al.* (2013) Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* 3, 3202
50. Žitnik, M. and Zupan, B. (2014) Matrix factorization-based data fusion for drug-induced liver injury prediction. *Syst. Biomed.* 2, 16–22
51. Žitnik, M. and Zupan, B. (2015) Gene network inference by fusing data from diverse distributions. *Bioinformatics* 31, i230–i239
52. Erdŏ"s, P. and Rényi, A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* 5, 17–61
53. Barabási, A-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* 286, 509–512
54. Nafis, S. *et al.* (2014) Apoptosis regulatory protein–protein interaction demonstrates hierarchical scale-free fractal network. *Brief. Bioinform.* 16, 675–699
55. Teschendorff, A.E. *et al.* (2015) Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. *Sci. Rep.* 5, 9646
56. Li, L. *et al.* (2005) Towards a theory of Scale-free graphs: definition, properties, and implications. *Internet Math.* 2, 431–523
57. Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of "small-world" networks. *Nature* 393, 440–442
58. Humphries, M.D. and Gurney, K. (2008) Network "small-world-ness": a quantitative method for determining canonical network equivalence. *PLoS One* 3, e2051
59. Ma, S. *et al.* (2007) An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Res.* 17, 1614–1625
60. Wille, A. and Bühlmann, P. (2006) Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* 5, 1
61. Kurtz, Z.D. *et al.* (2015) Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11, e1004226
62. Friedman, J. and Alm, E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8, e1002687
63. Bell, T.A.S. *et al.* (2016) A lipid-accumulating alga maintains growth in outdoor, alkaliphilic raceway pond with mixed microbial communities. *Microbiotechnology Ecotoxicol. Bioremediation* 6, 1480
64. Costea, P.I. *et al.* (2014) A fair comparison. *Nat. Methods* 11, 359-3489
65. Faust, K. and Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape. *F1000Research* 5, 1519
66. Faust, K. *et al.* (2015) Cross-biome comparison of microbial association networks. *Syst. Microbiol.* 6, 1200
67. Ban, Y. *et al.* (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* 31, 3322–3329
68. Fang, H. *et al.* (2015) CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* 31, 3172–3180
69. Deng, Y. *et al.* (2012) Molecular ecological network analyses. *BMC Bioinformatics* 13, 113
70. Biswas, S. *et al.* (2015) Learning microbial interaction networks from metagenomic count data. In *Research in Computational Molecular Biology* (Przytycka, T.M., ed.), pp. 32–43, Springer
71. Girvan, M. and Newman, M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826
72. Pons, P. and Latapy, M. (2005) Computing communities in large networks using random walks. In *Proceedings of the 20th International Conference on Computer and Information Sciences*, pp. 284–293, Berlin, Heidelberg
73. Blondel, V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008, P10008
74. Van Dongen, S. (2008) Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 121–141
75. Hua, Z. *et al.* (2013) Epigenomic programming contributes to the genomic drift evolution of the F-Box protein superfamily in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16927–16932
76. Richards, V.P. *et al.* (2014) Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biol. Evol.* 6, 741–753
77. Lei, X. *et al.* (2016) Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. *Inf. Sci.* 329, 303–316
78. Wang, J. *et al.* (2013) Construction and application of dynamic protein interaction network based on time course gene expression data. *PROTEOMICS* 13, 301–312