List of Figures

List of Tables

Contents

1

Graphical Models in Genetics, Genomics and Metagenomics

Hongzhe Li

University of Pennsylvania

Jing Ma

University of Pennsylvania

CONTENTS

1.1 Introduction

High-throughput technologies have generated an enormous amount of tissue and cell-type specific genetic, genomic and metagenomic data. Measurements of gene expression at the single-cell level have also become possible and are promising data sources in studying brains, cancer and immunology [19]. The CRISPR (clustered regularly interspaced short palindromic repeats) screen has recently emerged as a powerful new approach in profiling gene essentiality at the genome scale and in facilitating the dissection of regulatory networks by gene editing [43]. These new data and technologies enable us to experimentally measure and define biomolecular interactions on a large scale. This chapter focuses on graphical models and network-based analysis in genetics, genomics and metagenomics, with an emphasis on incorporating biomolecular networks in answering fundamental biological questions.

1.1.1 The human interactome

As introduced in [chaper of Mukerjee and Bates], a biological network consists of a collection of biomolecules and their interactions that correspond to various cellular functional relationships, and is often represented as a graph with directed and/or undirected edges.

Throughout the chapter, the word 'interaction' is used to denote the presence of an edge between two nodes, which may be directed or undirected and defined experimentally or statistically depending on the context. Examples of important biological networks include gene regulatory networks, whose directed edges represent activation or repression relationships between genes; protein-protein interaction networks, whose nodes are proteins linked together by physical binding events; metabolic networks, whose nodes are metabolites and edges reflect the chemical reactions of metabolism. Other useful networks are gene co-expression networks [49], which are phenotypic networks in which genes are linked if they share similar co-expression patterns.

Using complex network theory, [3, 4] found that the topologies of biomolecular networks are far away from being random, but are in fact scale-free. In addition, these networks are often comprised of physically or functionally connected subnetworks, also called pathways, that work together to achieve certain biological functions. It is worth noting that both the nodes and the interactions described above can be tissue- and context-specific. An important goal of studying the human interactome is to elucidate the functional role of biological networks under selected tissues and contexts, so as to understand the mechanisms of disease onset and progression, and identify previously unknown genes and pathways associated with complex phenotypes.

1.1.2 Publicly available databases

The past few years have seen systematic efforts in collecting and storing biomolecular interactions, that are curated from literature and high-throughput experiments or estimated using statistical methods, in publicly available databases. Some of these databases span a wide range of data types, such as KEGG [25, 26, 27] that has structural information on genes, proteins and pathways, while others contain specific data types, such as iRefIndex [42] and STRING [45] that provide a critical assessment and integration of protein-protein interactions. These well-maintained and regularly updated databases allow us to answer important questions about the factors that control how signals pass through the biological network in response to external stimuli. Efficient and rigorous incorporation of known biological information about pathways and networks into analysis of multiple genomic data is a key component of integrative genomics.

However, an increasing body of evidence suggests that biomolecular interactions and canonical pathways in existing databases are incomplete and largely inaccurate. Complementary to existing knowledge, one can also computationally construct biological networks based on various types of molecular data and use the resulting networks/subnetworks in downstream analysis. Learning biological networks by integrating both perturbation experiments and observational data has been an active area of research. Interested readers are referred to [chapter of Mukerjee and Bates] for an overview of available methods. On the other hand, the validity of biological networks inferred from data-driven approaches may largely depend on the size of the study cohort which is often small compared to the number of genetic features, the quality of the data, and the tissue or context under which the data are collected. One expects that combining existing network information with data-driven approaches may facilitate better understanding of fundamental biological processes.

1.1.3 Genetic terminologies

Table 1.1 lists the key genetic terminologies and their definitions used throughout this chapter. More details about their biological contexts are available in [2] and [35].

TABLE 1.1
Definitions and abbreviations of genetic terminologies used in the paper.

SNP	single nucleotide polymorphism; DNA sequence variation occurring in which a single nucleotide differs among individual subjects
eQTL	expression quantitative trait loci; statistical associations between a SNP value and the expression level of a mRNA
Haplotype	alleles across different loci on the same chromosome
GWAS	genome-wide association study; statistical association between a SNP value and a trait (e.g. response to therapy) or disease
LD	linkage disequilibrium; non-random association of alleles (e.g. SNP values) at different genomic locations
Pathway	functionally related set of biomolecules (genes, proteins, metabolites)
Pleiotropy	one gene that influences two or more unrelated phenotypes
Polymorphism	genetic variations between individual subjects
PPI	protein-protein interactions

1.2 Network-based analysis in genetics

Human genetic research aims to identify the genetic variants that are associated with various complex phenotypes. A genetic variant may refer to (1) a single-nucleotide polymorphism (SNP), which is a common variant that occurs in at least 1% of a population, (2) a mutation, in a case where it is a rare variant, or (3) a copy-number variation/aberration (a CNV is change in copy number in germline cells, whereas a CNA is change in copy number arisen in somatic tissues). Graphical model and network-based methods play an important role in identifying biologically relevant variants by taking into considerations gene-gene interactions.

1.2.1 Network-assisted analysis in genome-wide association studies

Genome-wise association studies (GWAS) attemp to identify commonly occurring genetic variants that contribute to disease risk, and so far have identified thousands of SNPs that are associated with many human traits [5]. In its simpliest form, GWAS analysis is formulated as a sequence of logistic regressions where the disease status from all individuals serve as the response and each genotyped SNP is the covariate. The resulting p-value for each SNP is then corrected for multiple comparisons using e.g. the Bonferroni adjustment. Although this standard approach has the power of identifying common SNPs with strong effects on phenotypes, it ignores the possible synergistic effects of genetic variants on disease phenotypes. Therefore network-assisted methods have been proposed to prioritize the GWAS results and to identify subnetwork of genes that are associated with phenotypes. The rationale of such network-based methods is that topologically related genetic variants are more likely to produce similar phenotypic effects.

Among the available databases, the protein-protein interaction (PPI) networks from STRING [45], iRefIndex [42] and Reactome [10] are often used in network-assisted analysis (NAA). In addition, directed graphs such as the protein-DNA regulatory networks [28] have also been used to identify potential causal variants [30]. NAA starts with preprocessing the GWAS data to compute SNP- or gene-based statistical values such as p-values that measure the significance of associations between the tested SNPs and the phenotype. After overlaying these SNP- or gene-based p-values onto the network extracted from public databases,

NAA approaches search for subnetworks and assess the combined effects of multiple genes participating in the subnetworks through a gene set analysis. Depending on the null hypothesis tested, one may apply permutation tests that randomly swap case and control labels in the GWAS data or randomization tests that use randomly generated networks to estimate the null distribution, thereby evaluating the significance of the detected subnetworks [24]. Network information has also been explored for identification of the causal variant within a longer haplotype that is associated with the trait and for identification of causal variants among multiple genes within a pathway [30].

Motivated by network-based analysis of gene expression data in [47, 48], [9] and [32] proposed methods that incorporate the biological pathway information via a hidden Markov random field (HMRF) model for GWAS. These types of HMRF models have been further developed and applied to network-based analysis of rare variants (see Section 1.2.2). An important and closely related problem to GWAS is expression quantitative trait locus (eQTL) mapping where the phenotype of interest is gene expression (see detailed discussion in Section 1.3).

1.2.2 Co-expression network-based association analysis of rare variants

Advances in next-generation sequencing technologies have revolutionized biomedical research, including the ability to obtain the exome or whole genome sequencing of a large set of samples. The large number of single nucleotide variants (SNVs) uncovered in each single human genome or exome provide insights into the role of rare genetic variants in the risk of complex diseases, but also create computational challenges for genetic studies. Compared to a SNP which is a common variant, a SNV means a variation in a single nucleotide without any limitations of frequency. Analysis of rare variants from exome or whole genome sequencing has been a very active area of human genetic research. Given the very low minor allele frequencies of such rare variants, grouping the variants based on gene annotation or pathway information is the standard way of testing rare variant associations. However, such approaches have seen limited success because many rare variants are neutral and have no functional relevance.

Combining gene co-expression network with information on rare genetic variants, [34] developed a novel algorithm, DAWN, to model two types of data: rare variations from exome sequencing and gene co-expression in tissues that are related to disease risk. The algorithm is based on a HMRF model [32, 47, 48], whose graph structure is determined by gene co-expression. Specifically, for rare variants, gene-based tests for gene-disease associations are first applied to obtain the p-value for each gene [22]. To construct a more interpretable gene co-expression network, two important screening steps are applied. Step 1 first identifies a set of key genes, which are defined as those with relatively small p-values. In step 2, DAWN trims the set of key genes by excluding those that are not substantially co-expressed with any other measured genes. Further neighborhood selection is used to construct a sparse gene co-expression network.

Let n represent the total number of genes in the final network and Ω the corresponding $n \times n$ adjacency matrix. DAWN converts the gene-based p-values to normal Z-scores, $Z = (Z_1, \cdots, Z_n)$, to obtain a measure of the evidence of disease association for each gene. These Z-scores are assumed to have a Gaussian mixture distribution, where the mixture membership of Z_i is determined by the hidden state I_i indicating whether gene i is a risk gene ($I_i = 1$) or not ($I_i = 0$). The mixture model for Z_i can be formally expressed as

$$Z_i \sim P(I_i = 0)N(0, 1) + P(I_i = 1)N(\mu, \sigma^2),$$

where μ and σ^2 correspond, respectively, to the mean and variance of Z_i under the alternative ($I_i = 1$). An Ising model is used to model the conditional dependence structure of

the hidden states $(I_1, \ldots, I_n)$ with the probability mass

$$P(I = \eta) \propto \exp(b^T \eta + c\eta^T \Omega \eta) \text{ for all } \eta \in \{0, 1\}^n.$$

Using an iterative algorithm, one can estimate the parameters (b, c, μ, σ^2) and the posterior probability for each gene $P(I_i = 1 \mid Z)$, which can be used to select disease-associated genes.

It is worth pointing out that although the two screening steps used in DAWN significantly reduce the search space, they may also prematurely remove genes that are in fact risk genes from downstream analysis. This limitation may be addressed with improving quality of the sequencing data and larger study cohort.

1.3 Network-based eQTL and integrative genomic analysis

Expression quantitative trait locus (eQTL) refers to the genomic regions that carry one or more sequence variants that affect the expression of a gene, typically measured by microarrays or high-throughput RNA sequencing. Such variation may suggest mechanisms under which phenotypic differences arise. Thus eQTL analysis has emerged as a key tool for elucidating the causal effects of regulatory variants on gene expressions and the clinical traits, where tissue-specific gene expressions can serve as possible mediators of the genetic variants.

If the eQTLs are located close to the genes they influence, they are called local eQTLs. Local eQTLs can act in *cis* by directly affecting only the expression of the gene that is on the same physical chromosome with it, as well as in *trans*, owing to changes in the function of a mediator [2]. In contrast, distant eQTLs refer to those that are located further away from the genes they influence and usually act in *trans* [37] (see Figure 1.1). As whole-genome sequencing data from different tissues/cell types become more accessible, there is a growing interest in integrative eQTL studies including joint analysis of eQTL mapping across multiple tissues for improved power [14] and Bayesian methods that combine external functional annotations with genetic association data for prioritizing causal variants in genome-wide association studies [16, 29].

Network-guided methods have also proven useful for identification of regulatory key driver genes associated with coronary heart disease [23], detection of *trans* acting eQTLs by modeling local gene networks [41] and eQTL mapping with mixed graphical Markov models [46]. Below we discuss in detail several novel approaches that incorporate network information for detecting eQTLs, and integration of GWAS and eQTL analysis. In such applications of causal inference in genetics, directed graphs provide an effective way of modeling various causal relationships.

1.3.1 Detection of trans-acting genetic effects

Recent studies suggest that a substantial proportion of the heritability in human gene expression cannot be explained by *cis* variants [20, 40], indicating the important contribution from trans acting genetic variants. Compared to *cis* eQTLs, detection of trans acting genetic effects remains a major challenge due to the relatively small effects of *trans* eQTLs and their specificity to the particular tissues and contexts [2, 12]. The high-dimensional multivariate nature of gene expression traits imposes a severe multiple testing burden, which further complicates *trans* eQTL mapping.

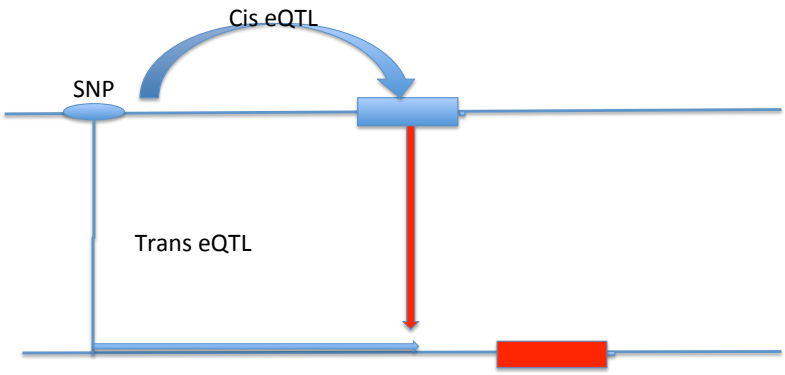

FIGURE 1.1
Schematic of eQTLs, and the two mechanisms *cis* and *trans* that a genetic variant can regulate gene expressions [37]. The boxes (in blue and red) represent protein-coding genes influenced by the SNPs, as indicated by the arrows. Local eQTLs that are located close to the genes they regulate can act both in *cis* and in *trans*, whereas distant eQTLs located further away from the genes they regulate usually act in *trans*.

To enhance the power for *trans* eQTL mapping, a simple but important principle is to account for as many competing sources of variation as possible. Consider the rationale in Figure 1.2. Here SNP A regulates gene A in *cis*. The directed edge between gene B and gene C, when unaccounted for, may reduce the signal and hence the power for detecting the trans association between SNP A and gene C. Gene B in Figure 1.2 is also called the *exogenous factor*, which is defined as any gene that (i) has a causal effect on gene C, and (ii) is independent of the genetic variant SNP A. By identifying and conditioning on all exogenous genes B, one is hopeful to increase the power for mapping *trans* eQTLs. This is also the main objective underlying GNet-LMM [41] which detects *trans*-acting genetic variants by modeling local gene regulatory networks. To be more specific, for each SNP A - gene C pair to be tested, GNet-LMM evaluates the following statistical dependencies to detect the V structure gene A $\rightarrow$ gene C $\leftarrow$ gene B (Figure 1.2):

$$\begin{array}{lll} \text{dep}(X_A, X_C), & \text{dep}(X_B, X_C), & \text{ind}(X_A, X_B), \\ \text{dep}(X_A, X_B \mid X_C), & \text{dep}(Z_A, X_A), & \text{ind}(Z_A, X_B). \end{array} \tag{1.1}$$

Here $\text{dep}(\cdot,\cdot)$ and $\text{ind}(\cdot,\cdot)$ denote, respectively, a statistical dependence and independence criterion (see [41] for detailed definitions of these criteria). Given the gene expression data, a standard correlation test is employed to assess the dependency between two genes. To evaluate the dependency between a SNP Z and a gene X, [41] used a linear mixed effects model of the form

$$X \sim N(Z\beta, \sigma_g^2 K_g + \sigma_n^2 \mathbf{I}).$$

Here K_g denotes the random effects covariance defined based on the genotype similarity with σ_g^2 being the variance of the random effects, σ_n^2 is the variance of the noise and $\mathbf{I}$ the identify matrix.

Running the testing procedure in (1.1) over all possible combinations of (SNP A, gene A, gene B, gene C) is a daunting task. To reduce the search space, one possible solution is to require the presence of *cis* or *trans* association between SNP A and gene A. Conditioning on

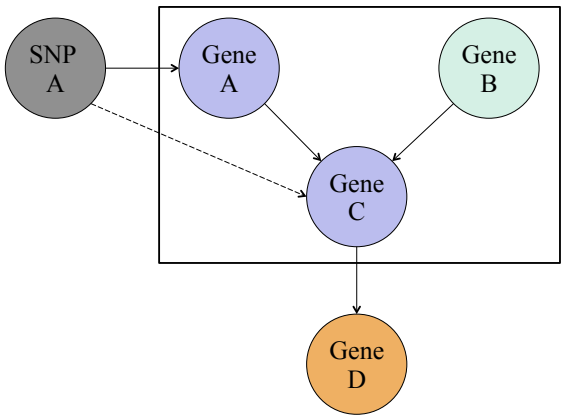

FIGURE 1.2
Graphical model illustration of the GNet-LMM algorithm [41]. The principle to improve power for detecting *trans* association between SNP A and gene C is to identify and condition on all exogenous genes with incoming edges (gene B in green). Exogenous genes represents either cofounding sources of variation or regulatory effects between genes, and are defined by testing for V-structures gene A $\rightarrow$ gene C $\leftarrow$ gene B (black box) that are linked to SNP A via gene A.

the expressions of all identified exogenous genes B, an extended linear mixed effects model can be applied to detect whether there is a trans association between SNP A and gene C [41].

1.3.2 A causal mediation framework for integration of GWAS and eQTL studies

Mapping eQTLs is also frequently used for unraveling the causal mechanism leading from genotype to phenotype, a crucial step for developing effective treatments of diseases. The increasing ability and power to map not only *cis* acting eQTLs but also *trans* acting eQTLs has greatly facilitated investigations into putative causal intermediates between genotype and the phenotype (node A, B in Figure 1.3). It has been recognized that gene expressions, albeit associated with genetic variants, may not be causal for the phenotype as the associations could be the result of responses to the phenotype (node E, F in Figure 1.3) or side effects (node C, G in Figure 1.3). Thus novel statistical methods beyond association analysis are needed to prioritize causal mediating genes.

When both gene expressions and genetic variants are measured on the same set of individuals with different phenotypes, [33] developed a sparse instrumental variable regression approach in order to identify the phenotype-associated genes whose expressions are controlled by genetic variants, where the genome-wide genetic variants served as instrumental variables. PrediXcan [18] is a gene-based association method that estimates gene expressions determined by an individual's genetic profile and subsequently correlates imputed gene expression with the phenotype under investigation to identify genes involved in the etiology of the phenotype.

There has been a recent interest in integrating summary data from GWAS and eQTL

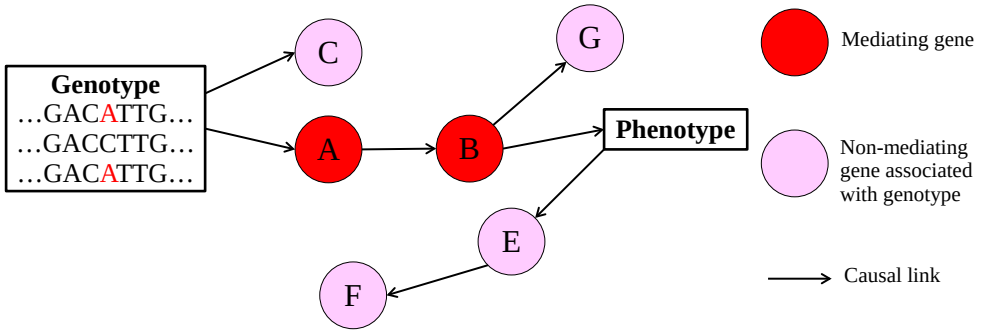

FIGURE 1.3
Distinguishing causal intermediate genes between genetic variation and phenotype (modified based on [17]). Genes can mediate the effect of genetic variation (genotype) on phenotype (nodes A and B), but gene expressions may also be affected by genetic variants irrespective of the phenotype (nodes C and G) or as a consequence of the phenotype (nodes E and F). A major challenge is to go beyond association analysis and identify mediating genes like A and B, which are valuable intervention points for understanding the molecular chain of causality.

studies in order to identify genes whose expression levels are associated with complex trait because of pleiotropy. One advantage of such approaches is that the summary-level data from GWAS and eQTL studies can come from two completely different sets of individuals, thereby effectively increasing the sample size for association analysis. The rationale for such an integrative analysis is articulated in [50] and is illustrated in Figure 1.4. If the phenotypic difference is caused by a genetic variant mediated by gene expression or transcription, then we should expect simultaneous association between phenotype, gene expression and the genetic variant (see Figure 1.4 (a) and also [50]). Such a simultaneous association can be due to causality with gene expression as mediator, pleiotropy where the same causal variant is associated with both phenotype and gene expression, or due to linkage where the shared association is because of linkage disequilibrium (LD) with two distinct causal variants, one affecting gene expression and another affecting phenotype (see Figure 1.4 (b)).

A summary data-based mediation test was proposed in [50] to identify gene expressions that are associated with complex traits. For each of the GWAS identified SNPs, [50] performed a mediation test with each of the gene expressions that has at least one *cis*-eQTL at a p-value $< 5 \times 10^{-8}$. Specifically, let Z be a genetic variant (e.g. a SNP), X the expression level of a gene and Y the trait. Using the mediation framework illustrated in Figure 1.5, the two-step least squares estimate of the effect of X on Y is

$$\hat{b}_{XY} = \hat{b}_{ZY}/\hat{b}_{ZX},$$

where $\hat{b}_{ZY}$ and $\hat{b}_{ZX}$ are the least squares estimates of Y and X on Z, respectively. One can interpret b_{XY} as the effect size of X on Y free of confounding from non-genetic factors. In addition, the variance of $\hat{b}_{XY}$ can also be estimated from the GWAS and eQTL summary statistics. One can thus use the test statistic $\hat{b}^2_{XY}/\operatorname{Var}(\hat{b}_{XY})$ to test whether gene X is significantly associated with the trait Y. To differentiate pleiotropy from the less interesting

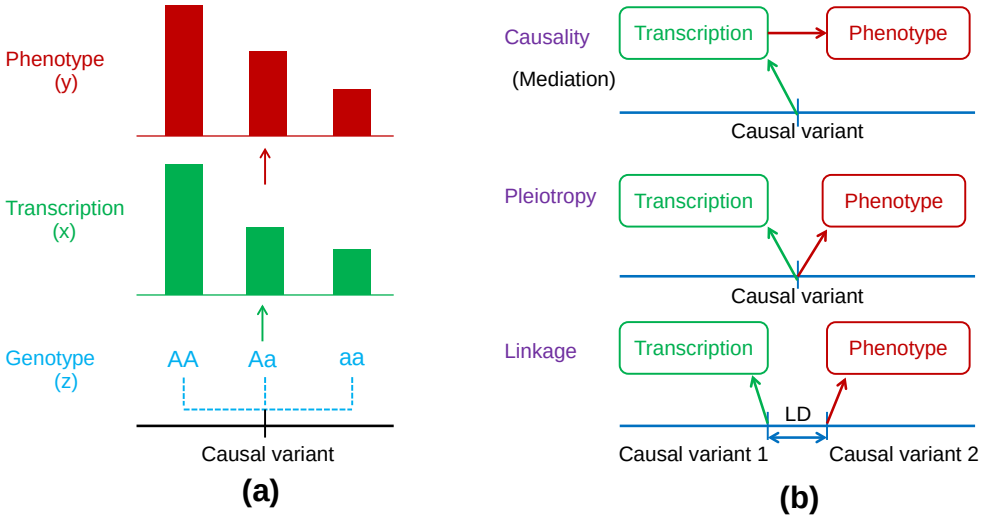

FIGURE 1.4
Association between gene expression and phenotype through shared genotypes (modified based on [50]). (a) Simultaneous associations caused by shared causal variants; (b) three possible causes of simultaneous association, causality, pleiotropy and linkage [21].

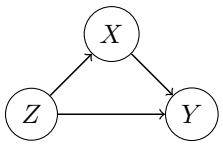

FIGURE 1.5
A simple mediation framework to link genetic variant Z to the gene expression X and the trait of interest Y.

case of linkage, [50] tested against the null hypothesis that there is a single causal variant, or equivalently the absence of heterogeneity in the b_{XY} values estimated for the SNPs in the *cis*-eQTL region.

1.4 Network models in metagenomics

Metagenomics has emerged as a powerful tool for learning microbial communities by directly extracting genetic materials from environmental samples. Microorganisms such as bacteria and archaea do not exist in isolation but form complex ecological interaction networks. These microorganisms are naturally assembled into interacting communities, and these community structures are directly linked to microbial processes. Therefore, the identification of key players in a taxonomically complex sample and understanding these complex interdependency are necessary to understand the ecology of a particular habitat.

1.4.1 Covariance based on compositional data

One challenge in microbial network construction is that we cannot measure the true abundances of the microbes. Instead, the current sequencing technologies such as 16S rRNA sequencing or shotgun metagenomic sequencing only provides information on the relative abundances of the microbial taxa. Such relative abundances are often given in terms of proportions with a unit sum. In other words, the data are compositional. Another feature of the compositional data is the presence of many zeros because many taxa are absent from the sample or their abundances are below the detection level due to insufficient sequening depths. Recent attempts in microbial network analysis focused on estimating the covariance matrix of compositional data [15]. However, one caveat with learning directly from the compositional data is that the unit sum constraint can lead to large spurious correlations, as illustrated in Figure 6 of [31].

It is instructive to first examine the quantity that is estimable based on compositional data. Let $W = (W_1, \ldots, W_p)^T$ with $W_j > 0$ for all j be a vector of latent variables, called the *basis* counts (e.g., true bacterial counts), that generate the observed compositional data via the normalization

$$X_j = \frac{W_j}{\sum_{i=1}^p W_i}, \quad j = 1, \ldots, p.$$

Estimating the covariance structure of W based on X has traditionally been considered infeasible owing to the apparent lack of identifiability. Nonetheless, [6] showed that the *basis covariance matrix* Ω_0 is approximately identifiable as long as it belongs to a class of large sparse covariance matrices, where $\mathbf{\Omega}_0 = (\omega_{ij}^0)_{p\times p}$ is defined by

$$\omega_{ij}^0 = \mathrm{Cov}(Y_i, Y_j), \quad Y_j = \log W_j.$$

To see this, recall one of the matrix specifications of compositional covariance structures introduced by [1] is the *variation matrix* $\mathbf{T}_0 = (\tau_{ij}^0)_{p\times p}$ defined by

$$\tau_{ij}^0 = \mathrm{Var}(\log(X_i/X_j)) = \mathrm{Var}(\log W_i - \log W_j) = \omega_{ii}^0 + \omega_{jj}^0 - 2\omega_{ij}^0,$$

or in matrix form,

$$\mathbf{T}_0 = \boldsymbol{\omega}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\omega}_0^T - 2\mathbf{\Omega}_0, \tag{1.2}$$

where $\boldsymbol{\omega}_0 = (\omega_{11}^0, \ldots, \omega_{pp}^0)^T$ and $\mathbf{1} = (1, \ldots, 1)^T$. One can see from the decomposition (1.2) that $\mathbf{\Omega}_0$ is unidentifiable, since $\boldsymbol{\omega}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\omega}_0^T$ and $\mathbf{\Omega}_0$ are in general not orthogonal to each other (with respect to the usual Euclidean inner product).

On the other hand, one can similarly define the *centered log-ratio covariance matrix* $\Gamma_0 = (\gamma_{ij}^0)_{p\times p}$ by

$$\gamma_{ij}^0 = \mathrm{Cov}\{\log(X_i/g(\mathbf{X})), \log(X_j/g(\mathbf{X}))\},$$

where $g(\mathbf{x}) = (\prod_{j=1}^p x_j)^{1/p}$ is the geometric mean of a vector $\mathbf{x} = (x_1, \ldots, x_p)^T$. Letting $\boldsymbol{\gamma}_0 = (\gamma_{11}^0, \ldots, \gamma_{pp}^0)^T$, [6] shows that

$$\mathbf{T}_0 = \boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T - 2\mathbf{\Gamma}_0. \tag{1.3}$$

Unlike (1.2), the following proposition shows that (1.3) is an orthogonal decomposition and hence the components $\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T$ and $\mathbf{\Gamma}_0$ are identifiable. In addition, by comparing the decompositions (1.2) and (1.3), one can bound the difference between $\mathbf{\Omega}_0$ and its identifiable counterpart $\mathbf{\Gamma}_0$ as follows.

Proposition 1 *The components $\boldsymbol{\gamma}_0 \mathbf{1}^T + \mathbf{1}\boldsymbol{\gamma}_0^T$ and $\mathbf{\Gamma}_0$ in the decomposition* (1.3) *are orthogonal to each other. Moreover, for the covariance parameters $\mathbf{\Omega}_0$ and $\mathbf{\Gamma}_0$ in the decompositions* (1.2) *and* (1.3),

$$\|\mathbf{\Omega}_0 - \mathbf{\Gamma}_0\|_{\max} \le 3p^{-1}\|\mathbf{\Omega}_0\|_1.$$

Proposition 1 implies that the covariance parameter $\mathbf{\Omega}_0$ is *approximately* identifiable as long as $\|\mathbf{\Omega}_0\|_1 = o(p)$. Consequently one can use $\mathbf{\Gamma}_0$ as a proxy for $\mathbf{\Omega}_0$, which greatly facilitates the development of new methodology and associated theory. [6] developed a composition-adjusted thresholding (COAT) method under the assumption that the basis covariance matrix is sparse, and showed that the resulting procedure can be viewed as thresholding the sample centered log-ratio covariance matrix and hence is scalable for large covariance matrices.

1.4.2 Microbial community dynamics

The microbial communities are highly dynamic, and are constantly responding to perturbations in the environment. Several large-scale time-series microbiome data have been generated to gain insights into the dynamics of gut microbiome over time. The human microbiota time series study in [7] covers two individuals at four body sites over 396 time points, including gut, tongue, left palm and right palm. [11] reported coupled longitudinal datasets of human lifestyle and microbiota by tracking two healthy male volunteers and their commensal microbial communities each day over the course of a year. These studies not only show overall stability of the microbial communities, but also marked community disturbance following changes in the environment. However, the small number of subjects involved in the above two studies proves to be inadequate for valid inference of time-varying microbial networks.

The most comprehensive study on the progression of the infant gut microbiota thus far examined 58 preterm infants in a neonatal intensive care unit, with repeated measurements taken every few days on all study subjects starting within the first days of life, and ending at approximately one month of age [36]. This set of densely sampled microbiome data provides the relative abundances of microbial taxa measured over time, revealing important information on ecological dynamics. One approach towards inference of time-varying microbial networks is based on multivariate functional data analysis. For each taxon, its abundance trajectory can be treated as functional data. One can then apply techniques developed for estimating the covariance structure of multivariate functional data [39] to reconstruct the microbial network at each time point. However, the compositional nature of the data and also the excessive zeros require appropriate modification to these existing methods.

See [13] for a brief review of dynamic network inference from metagenomic data.

1.5 Future directions and topics

As our knowledge of biological networks increases, incorporation of such networks in analysis of biological data proves invaluable in genetic association studies, and to some degree in analysis of genetical genomics or eQTL studies. Looking forward, single-cell measurements and gene editing tools such as CRISPR-Cas will lead to detailed understanding of the biological networks at the single-cell level. The data sets from such studies are large and require new statistical and computational methods.

Metagenomics is an emerging field that holds a great promise in biomedical research. Most of the published works are still at the level of establishing association between taxa composition and microbial gene abundances with various covariates or disease states. A major challenge is to go beyond association studies and elucidate causalities. Mathematical modeling of the human gut microbiome at a genome scale is a useful tool to decipher microbe-microbe, diet-microbe and microbe-host interactions [38]. Graphical models, espe-

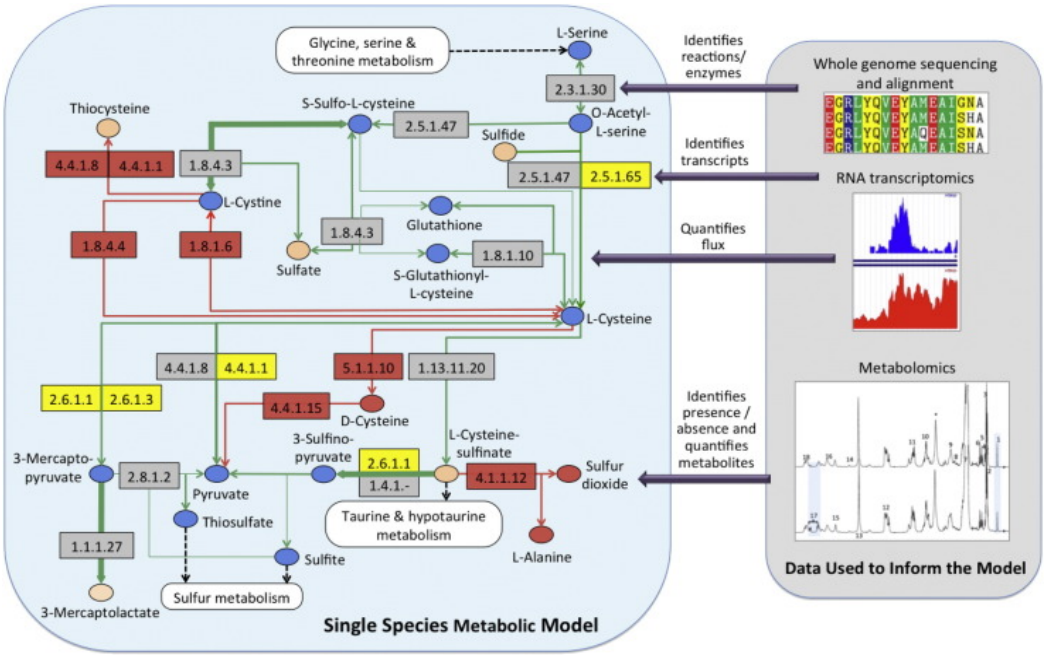

FIGURE 1.6
Subset of a microbial cysteine/methionine metabolic network for one bacterial species. The model is constructed based on the bacterial genome. Each box represents a reaction. The numbers within the boxes are KEGG Enzyme Commission (EC) number and code for specific enzymes present in each reaction. Gray boxes represent reactions that occur in this bacteria, as predicted by its genome. Red boxes denote reactions that are not predicted by the genome. Circles represent metabolites consumed and produced within the reaction network. Arrows represent reaction pathways that do (green) or do not (red) occur in this bacteria, as predicted by the model. Black dashed arrows indicate input or output from or to other metabolic networks. This figure is reproduced from [44] under a Creative Commons license. doi:10.1016/j.atg.2016.02.001.

cially causal graphical models, provide a natural and useful tool for elucidating such causal pathways. Parallel to advances in sequencing technologies, publicly available database of experimentally elucidated metabolic pathways from all domains of life, such as MetaCyc [8], is becoming more and more complete. Currently, MetaCyc contains more than 2400 pathways from 2788 different organisms, including those involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes and genes. This provides important resources for analysis of microbiome and metagenomic data. Figure 1.6 presents an example of such a metabolic network. How to incorporate the metabolic networks/pathways into analysis of metagenomic data is an important area for future research.

Bibliography

[1] John Aitchison. The statistical analysis of compositional data. *Journal of Royal Statistical Society, Series B*, 44(2):139–177, 1982.

[2] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[3] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[4] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[5] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, 2012.

[6] Yuanpei Cao, Wei Lin, and Hongzhe Li. Large covariance estimation for compositional data via composition-adjusted thresholding. *ArXiv e-prints*, 2016.

[7] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, 2011.

[8] Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[9] Min Chen, Judy Cho, and Hongyu Zhao. Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLoS Genetics*, 7(4):e1001353, 2011.

[10] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691–D697, 2011.

[11] Lawrence A David, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, 15:R89, 2014.

[12] Benjamin P Fairfax, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, and Julian C. Knight. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175):1246949, 2014.

[13] Karoline Faust, Leo Lahti, Didier Gonze, Willem M de Vos, and Jeroen Raes. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology*, 25:56–66, 2015.

[14] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS Genetics*, 9(5):e1003486, 2013.

[15] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8(9):e1002687, 2012.

[16] Sarah A Gagliano, Michael R Barnes, Michael E Weale, and Jo Knight. A bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. *PloS One*, 9(5):e98122, 2014.

[17] Julien Gagneur, Oliver Stegle, Chenchen Zhu, Petra Jakob, Manu M Tekkedil, Raeka S Aiyar, Ann-Kathrin Schuon, Dana Pe'er, and Lars M Steinmetz. Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS Genetics*, 9(9):e1003803, 2013.

[18] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[19] Charles Gawad, Winston Koh, and Stephen R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17:175–188, 2016.

[20] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, 2012.

[21] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48:245–252, 2016.

[22] Xin He, Stephan J Sanders, Li Liu, Silvia De Rubeis, Elaine T Lim, James S Sutcliffe, Gerard D Schellenberg, Richard A Gibbs, Mark J Daly, Joseph D Buxbaum, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8):e1003671, 2013.

[23] Tianxiao Huan, Bin Zhang, Zhi Wang, Roby Joehanes, Jun Zhu, Andrew D Johnson, Saixia Ying, Peter J Munson, Nalini Raghavachari, Richard Wang, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 33(6):1427–1434, 2013.

[24] Peilin Jia and Zhongming Zhao. Network-assisted analysis to prioritize gwas results: principles, methods and perspectives. *Human Genetics*, 133(2):125–138, 2014.

[25] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[26] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[27] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016.

[28] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Computational Biology*, 9(3):e1002886, 2013.

[29] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10(10):e1004722, 2014.

[30] Mark DM Leiserson, Jonathan V Eldridge, Sohini Ramachandran, and Benjamin J Raphael. Network analysis of gwas data. *Current Opinion in Genetics & Development*, 23(6):602–610, 2013.

[31] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[32] Hongzhe Li, Zhi Wei, and John Maris. A hidden markov random field model for genome-wide association studies. *Biostatistics*, 11(1):139–150, 2010.

[33] Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, 2015.

[34] Li Liu, Jing Lei, Stephan J Sanders, Arthur Jeremy Willsey, Yan Kou, Abdullah Ercument Cicek, Lambertus Klei, Cong Lu, Xin He, Mingfeng Li, et al. Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism*, 5(1):1, 2014.

[35] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[36] Michael J McGeachie, Joanne E Sordillo, Travis Gibson, George M Weinstock, Yang-Yu Liu, Diane R Gold, Scott T Weiss, and Augusto Litonjua. Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks. *Scientific Reports*, 6(20359), 2016.

[37] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B*, 368:201220262, 2013.

[38] Cecilia Noecker, Alexander Eng, Sujatha Srinivasan, Casey M. Theriot, Vincent B. Young, Janet K. Jansson, David N. Fredricks, and Elhanan Borenstein. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems*, 1(1), 2016.

[39] Alexander Petersen and Hans-Georg Müller. Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika*, 103(1):103–120, 2016.

[40] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics*, 7(2):e1001317, 2011.

[41] Barbara Rakitsch and Oliver Stegle. Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biology*, 17(1):1, 2016.

[42] Sabry Razick, George Magklaras, and Ian M Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):1, 2008.

[43] Jeffry D Sander and J Keith Joung. Crispr-cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32:347–355, 2014.

[44] Jaeyun Sunga, Vanessa Haleb, Annette C. Merkel, Pan-Jun Kima, and Nicholas Chiab. Metabolic modeling with big data and the gut microbiome. *Applied & Traslational Genomics*, in press, 2016.

[45] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[46] Inma Tur, Alberto Roverato, and Robert Castelo. Mapping eqtl networks with mixed graphical markov models. *Genetics*, 198(4):1377–1393, 2014.

[47] Zhi Wei and Hongzhe Li. A markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.

[48] Zhi Wei and Hongzhe Li. A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data. *Annals of Applied Statistics*, 2(1):408–429, 2008.

[49] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1128, 2005.

[50] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature Genetics*, 48:481–487, 2016.