STATISTICAL APPROACHES FOR MATCHING THE COMPONENTS OF COMPLEX MICROBIAL COMMUNITIES

by

Chongci Tang

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
May 2016

Table of Contents

List of Tables

List of Figures

Abstract

A hierarchical Bayesian model called BiomeNet can be used to identify the functional units of metabolic reactions, subnetworks and community-level metabolic networks. The framework models metabolic structures by assuming each sample consists of many tightly connected subnetworks, which in turn are comprised of different reactions. When applying the method the number of subnetworks L must be pre-specified When L is set larger in BiomeNet, the inferred structures of the subnetwoks are expected to come out in a more trivial form. Three methods, LASSO, NNLS and a new method, MJSD, are applied to match a subnetwork in one analysis (say, when L=100) with several subnetworks when L is increased (say, L=200). RSS and JSD are applied as matching criteria to conduct multiple tests to judge the significance of the matches. From the results, I am able to identify those “predominant” subnetworks and give a reasonable conjecture that those “predominant” subnetworks always come out as unbroken blocks for any larger L values.

List of Abbreviations and Symbols Used

Symbols and Abbr.	Description
$\mathbf{X}$	2824 by 200 data matrix
$\mathbf{Y}$	2824 by 100 data matrix
Y_j	j_{th} column in $\mathbf{Y}$
X_j	j_{th} column in $\mathbf{X}$
y_i	i_{th} element in Y
$\boldsymbol{\beta}$	coefficients vector
JSD	Jensen Shannon divergence
RSS	Residual sum of squares
NNLS	Non-negative least square
LASSO	least absolute shrinkage and selection operator
LDA	Latent Dirichlet Allocation
S-R	substrate-product
LAR	least angle regression
df	degrees of freedom
MCMC	Markov chain Monte Carlo
PCA	Principle Component Analysis
IBD	Inflammatory bowel disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
MJSD	the new method based on minimizing JSD
FDR	False Discovery Rate
B-H procedure	Benjamin-Hochberg procedure

Acknowledgements

I would like to express my huge gratitude to my supervisor, Dr. Hong,Gu and co-supervisor Dr. Joseph Bielawski, for their guidance and encouragement for the two years. They spent a lot of time on instructing me and revise my thesis even though they are really busy. Without their support this thesis would not have been possible.

I would also like to thank my committee members, Dr. Toby Kenney and Dr. Bruce Smith for their kind comments and advice on my thesis. I would like to give my thanks to my friends and everyone else who helped me. My thanks also goes to my professors and classmates in Dalhousie University.

Chapter 1

Introduction

1.1 Background

From protists to humans, all plants and animals live in close association with microbial organisms. These microbes are believed to play a critical role in a wide variety of settings, from globally significant nutrient cycling to influencing human physiology [16]. The ecological community of commensal, symbiotic and pathogenic microorganisms that literally share the human body space have a profound effect on human health [11]. For this reason, in human microbiomics , research is focused on the relationship of the microbial communities to both health and disease status. In late 1990s, researchers found that the microbiome in the gut played a role in the development and maintenance of the human immune system, and the human microbiome is now implicated in a variety of autoimmune diseases like diabetes [10] [22]. A poor mix of microbes in the gut may also aggravate common conditions such as obesity [20]. The consensus opinion is that although the mammalian immune system seems to be designed to control microorganisms, it is in fact controlled by microorganisms [12].

Most microbes in most environments are not cultivatable, so studies of microbial communities must be made based on DNA sampled directly from the environment, although RNA, protein and metabolite based studies may also be performed [6]. There are two ways to use environmental DNA to study microbial communities. One is targeted amplicon studies (e.g., 16S SSU rDNA), which focuses on a known marker gene and is primarily informative about taxonomy [21]. The other that appeared more recently is shotgun metagenomics, which can be used to study the functional potential of the community [21]. Shotgun sequencing of total environmental DNA proceeds by randomly cutting the total DNA within a sample, and sequencing the many short fragments that are produced [14]. The resulting collection of DNA sequences represents the genomes of all the microbes in the sample, and is referred to as the "metagenome" of the sampled microbiota. The function of some of those sequences can be inferred via homology to reference sequence with known functions (e.g., through the KEGG database) [14]. However, only a fraction of the gene sequences can be assigned a function; the function of a very large number of sequences, will remain unknown [14]. Of particular interest are those sequences that encode a known enzyme, as they can be used to infer the community-level metabolic potential of the microbiota.

The complete set of metabolic and physical processes of a cell, called a *metabolic network*, determines the physiological and biochemical properties of the bacterium. The fundamental units of a metabolic network are its chemical reactions. The essential components of a reaction are the *substrates* (chemical compounds), the enzyme (a protein encoded by DNA) that will act upon the substrates, and the *products* that are produced by the action of the enzyme on its substrates (other chemical compounds). It is important to note that the products of one enzyme-mediated reaction can serve as the substrate for a different enzyme-mediated reaction. Such chemical reactions are organized into *subnetworks* (functional modules), which are themselves organized into the higher-level structures that biologists refer to as the metabolic network. In the field of microbiomics, the notion of the metabolic network is extended to include the full metabolic capacity, and the full set of interactions, carried out by a community of microbes.

In order to infer differential usage of metabolic networks among ecologically divergent microbial communities, a Bayesian modeling approach called BiomeNet was developed by Shafiei et al. [16]. The modeling framework differs from previous approaches [e.g., PCA and its variants] in attempting to capture the hierarchical nature of real metabolic networks [22]. Specifically, sets of reactions (related by exploiting a shared pool of substrates and products) are modeled as a metabolic subnetwork, and over-lapping mixtures of such metabolic subnetworks are used to model the full metabolic network [16]. When such networks are inferred from the model (as opposed to being defined solely according to biochemical expertise) they are called *metabosystems* [16]. Metabosystems are intended to represent hypothetical metabolic phenotypes, and samples are permitted to consist of overlapping mixtures of metabosystems.

Currently, the most widely used method to analyze community metabolic networks, and metagenomic variation in general, is Principle Component Analysis (PCA) and it variants [16]. BiomeNet is fundamentally different from such methods, in providing a framework to take advantage of dependencies between reactions encoded by metabolic networks without any transformation and reduction. This means that results obtained under BiomeNet offer biologists a direct functional interpretation. Thus, BiomeNet contributes a valuable addition to PCA [16].

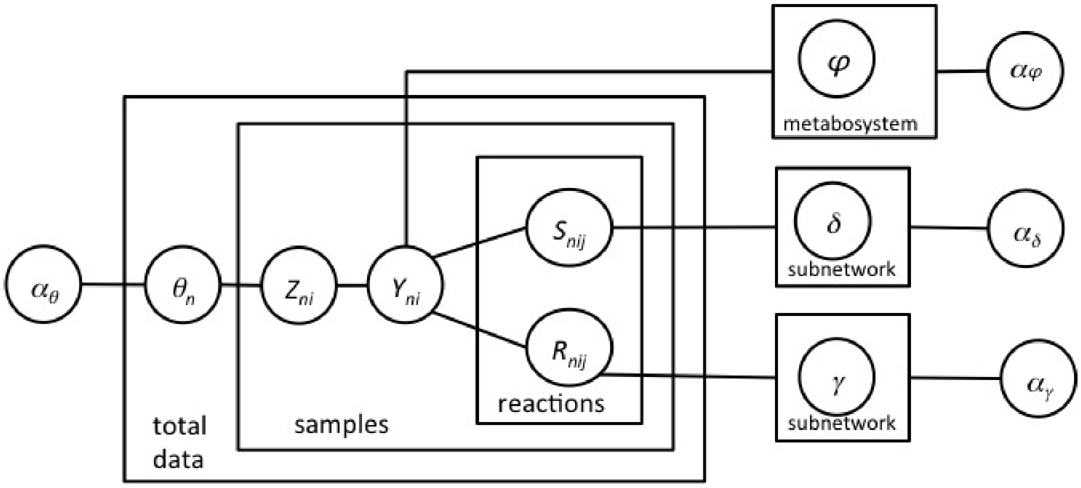
BiomeNet has been applied to a variety of datasets, including marine water samples, mammalian gut microbiomes and the gut microbiomes of inflammatory bowel disease (IBD) patients [16]. The case of IBD illustrates why a direct functional interpretation of the results is critical to microbiome researchers. IBD is a human disease characterized by chronic immune dysregulation in the gut. Using BiomeNet, Shafiei et al [16] found that IBD patients differed from healthy individuals according to the mixture of metabosystems found in their gut. Interestingly, the metabosystem having a greater prevalence among IBD patients was characterized by metabolic reactions associated with (i) close association with the human gut epithelium, (ii) resistance to dietary intervention, and (iii) interference with uptake of antioxidants connected to IBD [16]. IBD is a serious disease; it can cause severely disruptive pain, require surgery, and even cause increased mortality. Even more pressing is that it is on the increase worldwide, with Canada having among the highest rates. Thus, it is important that BiomeNet produces results that have a direct relationship to microbial metabolic capacity, and the posterior mixture weights have a straightforward interpretation. Inference under BiomeNet will be focus of this thesis, and the framework is described in more detail in the next section.

1.2 BiomeNet

BiomeNet is a hierarchical mixed-membership Bayesian model similar to Latent Dirichlet Allocation (LDA). LDA employs a mixture of Dirichlet priors to facilitate clustering, or classification. Standard LDA, however, cannot be used to resolve the underlying structure of a metabolic network in terms of easily interpretable parts such as metabolic pathways or subnetworks. In the metagenomic setting, the observed data (produced by shotgun sequencing) is the abundance of enzyme-encoding sequence reads in each sample; but, for input into BiomeNet the data must be converted to counts of substrate-product pairs. To do this, where possible, each enzyme-encoding sequence is assigned to its enzyme within the EC database. The EC database provides information about the biochemical reaction that the enzyme catalyzes, including all of the substrates and products for that reaction. Each reaction is decomposed into all possible substrates and products (this leads to a mini hyper-graph for each reaction), and all pairs of compounds are assigned the abundance of the associated enzyme-encoding sequence within the sample. As mammalian gut metagenomes can encode thousands of unique enzymes, the input data for BiomeNet is rich in information relevant to metabolic interactions.

BiomeNet is used to model metabolic interactions at the community-level. The model assumes that each microbiome sample is a mixture of K metabosystems, where K is assumed to be known and fixed. However, the contribution of the K metabosystems is permitted to differ between samples. Each metabosystem is itself viewed as a mixture of a fixed number (L) of metabolic subnetworks. Hence, metabosystems differ according to their particular mixture of subnetworks, with the subnetworks viewed as a mixture of metabolic reactions. Because reactions within a subnetwork are linked through shared chemical compounds, the subnetworks are actually modeled as a subset of compounds that are converted to another subset of compounds. This is why each enzymatic reaction must be decomposed into substrate-product pairs (S-R pairs), with each subnetwork having its own substrates and products. The lower levels of the hierarchy (e.g., reactions) can always contribute to any of the higher levels (e.g., to different subnetworks, metabosystem and samples) to different degrees [16]. The set of S-R pairs for each reaction in a sample is the only observable variable, with other variables (the subnetwork (Y) and metabosystem assignments (Z) of the reactions) being latent variables.

In order to capture the dependence among the many variables more concisely, a plate diagram is shown in Figure 1.1 with the mathematical definitions given in Table 1.1. The outer plate represents the total data, the middle plate represents samples, and the inner plate represents reactions within a sample. The relative contribution of each metabosystem to the n^{th} microbiome sample is modeled via the variable θ_n, which is a probability vector of K values that sum to one. Typically, a



$$\theta_n \sim \text{Dirichlet}(\alpha_\theta, K) \qquad Z_{ni} \sim \theta_n$$
$$\varphi_k \sim \text{Dirichlet}(\alpha_\varphi, L) \qquad Y_{ni} \sim \varphi_k$$
$$\delta_l \sim \text{Dirichlet}(\alpha_\delta, C) \qquad S_{nij} \sim \delta_l$$
$$\gamma_l \sim \text{Dirichlet}(\alpha_\gamma, C) \qquad R_{nij} \sim \gamma_l$$

Figure 1.1: Plate diagram for the BiomeNet model. This model specifies a generative process; coupling between substrate-product pairs is enforced by conditioning their generation on a single subnetwork membership [16].
Note: Adapted from *BiomeNet, A bayesian model for inference of metablic divergence among mibrobial communities* by Mahdi Shafei, Katherine A Dunn. *PLoS Comput Biol*, 10(11):e1003918, 2014.

sparse symmetric Dirichlet priori is placed on θ.

$$\theta \sim Dirichlet(\alpha_\theta)$$

The unique mixture of L subnetworks for each metabosystem is modeled with a probability vector, φ_k, of mixing probabilities that sum to one. Thus, given K metabosystems, a $K \times L$ matrix called φ is used to represent the relative contribution of the l^{th} subnetwork to the k^{th} metabosystem. Typically, a sparse symmetric Dirichlet prior is placed on φ.

$$\varphi_k \sim Dirichlet(\alpha_\varphi)$$

Because reactions are linked through shared chemical compounds, each subnetwork has its own substrate (S) and product (R) groups. Note that the products of one reaction can serve as the substrate of another reaction (i.e., the intermediary compounds in a metabolic pathway), and the membership of compounds to the S and R groups must be "soft" (probabilistic) rather than discrete. Thus, for L subnetworks, there will be L substrate and products groups. Each subnetwork has a vector of C compounds for its own substrate and products, denoted by δ_l and γ_l respectively, summing to one. For L subnetworks, there will be an $L \times C$ matrix, δ, for substrate compounds and another $L \times C$ matrix, γ, for product compounds. A value in row l and column c of one of these matrices gives the contribution of compound c to a subnetwork (as either a substrate or product, depending on the matrix). As above, a sparse symmetric Dirichlet prior is placed on δ_l and γ_l.

$$\delta_l \sim Dirichlet(\alpha_\delta)$$

$$\gamma_l \sim Dirichlet(\alpha_\gamma)$$

Variable	Type	Meaning
N	integer	number of microbiome sample(e.g 38)
K	integer	number of metabosystems (e.g 3)
L	integer	number of subnetworks (e.g 100)
C	integer	number of compounds (e.g 2713)
θ	probability	prior distribution of metabosystems in a sample
φ	probability	prior distribution of subnetworks in metabosystems
δ	probability	prior distribution of substrate compounds in subnetworks
γ	probability	prior distribution of product compounds in subnetworks
α	probability	concentration parameter of the Dirichlet distribution
n	integer	index of a sample
i	integer	index of a reaction
j	integer	index of a substrate-product pairs
Z_{ni}	vector	metabosystem assignment for the i^{th} reaction in the n^{th} sample
Y_{ni}	vector	subnetwork assignment for the i^{th} reaction in the n^{th} sample
S_{nij}	vector	substrate compounds assigned for the J_{in} S-R pairs
R_{nij}	vector	product compounds assigned for the J_{in} S-R pairs

Table 1.1: Mathematical definition of the plate diagram

While the prior probabilities are in the Dirichlet distributions, the posterior distributions are in a multinomial distribution. The relative contribution of each metabosystem and subnetwork to the sample n is modeled as

$$Z_{ni}|\theta_n \sim Multi(\theta_n)$$

$$Y_{ni}|Z_{ni} \sim Multi(\varphi_{z_{ni}})$$

Where Z_{ni} and Y_{ni} denote the metabosystem and subnetwork assignments for reaction i in sample n. Inference under BiomeNet involves sampling the posterior distribution of the subnetwork and metabosystem assignments for each reaction in the data and integrating out all other latent variables. Specifically, collapsed Gibb sampling is used to sample from the conditional distribution $P(Z, Y|R, S, \alpha_\theta, \alpha_\varphi, \alpha_\delta, \alpha_\gamma)$, with the latent variable θ, φ, δ, γ integrated out. The subnetwork and metabosystem assignments are sampled for a single reaction in a single sample, $P(Z, Y|S, R)$, given the set of subnetwork and metabosystem assignment for all other reactions in all other samples except only reaction i in sample n (Y_{-ni} and Z_{-ni} respectively). Each iteration of Gibb sampling not only provides a joint posterior distribution sampling point on $P(Z, Y|S, R)$, but also provide a sampling point from all marginal distribution $P(Z_{ni}, Y_{ni}|S, R)$. Based on the posterior distribution of metabosystems and subnetworks assignments, $P(Z|S, R)$ and $P(Y|S, R)$, the posterior distribution of θ and φ, and their means, can be directly inferred if sufficient iterations of Gibbs sampling have been carried out [16].

1.3 The challenge of applying BiomeNet to real data

The strength of BiomeNet is that it provides a method to learn the community-level structure of the data in terms of subnetworks and metabosystems, rather than according to curated pathways, or some other biology-centered definitions. Model-derived structures help researchers to investigate the latent metabolic structure of any microbial community, and to understand microbial community ecology, without having to assume that metabolic pathways have been previously defined and annotated in a way that is suitable to the communities under study. However, BiomeNet does have limitations, and further development of the analytical framework is warranted. The number of metabosystems K in each sample, and the number of subnetworks L

must be pre-specified [16]. Without strong biological criteria of selecting the values of K and L, the user of BiomeNet must either (i) try different number of K and L and compare the discrepancy to choose a reasonable values, or (ii) choose arbitrarily high values for K and L and set the concentration parameter of the Dirichlet priors close to zero as a means of pushing BiomeNet to characterize metabosystems by a relatively few "predominant" subnetworks and reactions. The latter is a strategy for minimizing variance and maximizing interpretability in the face of high values for K and L. The problem with both strategies is that there is no objective means of assessing which, if any, subnetworks within a metabosystem, or reactions within a subnetwork, warrant "predominant" status and further biological interpretation.

A metagenomic dataset comprised of 38 mammalian gut microbiomes [16] provides both a good illustration of the challenges, as well as a good test case of future method development. This dataset is rich in metabolic information, having 2824 unique reactions involving 2713 compounds. In the original study, Shafiei et al [16], set $K=3$ metabosystems, to match the number of dietary niches (carnivore, omnivore and herbivore) represented by the sampled mammals. However, the authors found that they could only separate the carnivores from the herbivores, and concluded that the there was strong signal for 2 metabosystems, and only weak signal for a third. Shafiei et al. [16] had no *a priori* basis to select a good value for L, so they experimented with different numbers of subnetworks (i.e., L=50, 100, 150, 200) and assessed the robustness of the reaction composition of the metabosystems to L. [14]. Although their results suggested that a good value of L was likely somewhere between 100 and 150, they provided no means of objectively assessing which were the most important subnetworks.

When L is larger than needed, there will be redundant subnetworks in BiomeNet. The redundant subnetworks will carry very small weight for many of the metabosystems and their reactions will have only a trivial impact in composition of each metabosystem. Even when the concentration parameter of the Dirichlet priors are close to zero, and BiomeNet has been encouraged to use relatively few predominant subnetworks and reactions, it remains unclear how to decide at what point the mixture weight of a given metabosystem should be considered "trivial". Thus the motivation of my research is to address this challenge.

I employed the mammal dataset described above to develop and investigate several alternative methods for discriminating the so-called "predominant" subnetworks from those making only trivial contributions to a metabosystem. The logic that underpins my approach is that the informative subnetworks should make consistent contributions to each metabosystem across alternative values of L (as long as L is not too small), and that matching the components (i.e., reactions) of subnetworks can be used to identify them. However, comparison of results across alternative values of L is far from straightforward. First, subnetworks are unlikely to be labeled in the same way. For example, subnetwork 10 under L=100 might correspond to subnetwork 33 under L=150. Second, the signal for a subnetwork in one analysis (say, when L=100) could be split among several subnetworks when the value for L is increased (say, when $L = 200$). Thus, I investigated model-based methods for matching subnetworks from one case as linear combinations of subnetworks derived from another case. This strategy overcomes the problem of potentially complex relationships between the components of subnetworks derived from different analyses, and it has no requirements about labeling of subnetworks among the different analyses.

To achieve this, I summarize the information about the components of each subnetwork in a $N \times L$ reaction matrix. Here, N is the number of unique reaction in the dataset, which is 2824 for the 38 mammalian gut metagenomes. Each column in the $N \times L$ matrix represents a reaction's posterior mixing probability to the corresponding subnetwork [16] (detailed desciption of the reaction matrix in Chapter 2). In the thesis, I first evaluate two classical methods, LASSO (least absolute shrinkage and selection operator) and NNLS (non-negative least squares) regression, for "matching" results structured as described above. I then describe a new method similar to forward selection for matching the components of subnetworks. The effectiveness of these three methods are compared, and I show that the new method is more suitable in the case of the 38 mammalian gut metagenomes.

1.4 Structure of the thesis

This thesis is comprised of four chapters. The next two chapters (2 and 3) are devoted to method development and application to real data, In Chapter 2, I introduce two classical methods (LASSO and NNLS) and the new method, and describe how they can be applied to the problem of matching the metabolic components of complex microbial communities. As an example, I match one subnetwork from the analysis of L=100 to the subnetworks from that of L=200 using the three methods. In Chapter 3, these three methods are applied to all the 100 subnetworks in the $L = 100$ run of BiomeNet to match the subnetworks in the L= 200 run. I use RSS and JSD as criteria to rank the good matches and give examples of a good and a bad match according to those criteria. I then discriminate those well matched subnetworks by using permutation test and present the estimated coefficient matrices to find the features of the good matches. Chapter 4 concludes with the strengths and weaknesses of all three methods and provides directions for future research.

Chapter 2

Methods for matching the metabolic components of complex microbial communities

As introduced in Chapter 1, the strength of BiomeNet is that it provides a method to learn the community-level structure of the data in terms of subnetworks and metabosystems. One way to find if BiomeNet output is consistent with the information on the true community-level structure is to match the subnetworks to find whether the informative subnetworks always can be identified across alternative values of L when K is fixed. Here K is fixed at 3 as this was the value employed in the original study of Shafiei.et.al [16]. If BiomeNet works well, the signal for a subnetwork in one analysis (say, when L=100) could be either the same or split among several subnetworks when the value for L is increased (say, L=200). Because subnetworks are unlikely to be labeled in the same way, some method is needed to identify the relationships between subnetworks derived from different applications of BiomeNet to the same data. The relationships could be either a one to one correspondence, or a subnetwork from one case being split into several subnetworks in another case.

Each subnetwork is represented by a column of the posterior mixing probabilities in the $N \times L$ reaction matrix, where N=2824 is the number of unique reactions for the 38 mammalian gut metagenomic data and L is the number of subnetworks. Denote the $N \times L_1$ reaction matrix from one run of BiomeNet as Y, and the $N \times L_2$ reaction matrix from different run of BiomeNet as X, the matching problem that one subnetwork in $\mathbf{Y}$ corresponds to one or more subnetworks in $\mathbf{X}$ can be represented mathematically as follows:

$$Y = \mathbf{X}\boldsymbol{\beta} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{L_2} X_{L_2}$$

Thus the matching problem naturally can be formulated as a regression without intercept, in which case minimization of the sum of squared residuals can be used as the criterion to achieve the best fit. Matching subnetworks from one case as linear combinations of another case should perform well because the expectation is that a large subnetwork in L=100 will be split into several smaller (and largely non-overlapping) subnetworks when L=200. Thus the linear regression model is chosen to represent the relationship between subnetworks. Usually, the least square linear regression coefficient vector is not sparse. However, in this application one subnetwork is expected to be matched by either one or only a few subnetworks, i.e. the consistency of the subnetwork outputs from BiomeNet is associated to the sparsity and the quality of the matches.

So, a regularized version of the least squares solution may be preferable to satisfy the sparsity of $\boldsymbol{\beta}$. Naturally, LASSO regression is a good candidate method since it can result in a good regression model with a smaller subset of predictors for LASSO fitting. A modified LARS algorithm [11] can be used to return a full LASSO path, and a sparse solution can be achieved with a proper model selection criterion.

If a subnetwork in Y is split into several subnetworks in X, then ideally the regression coefficients should be non-negative which means $\beta \geq 0$. LASSO regression only satisfies the requirement of sparsity of β, but not the non-negativity of β. Thus another good candidate method is the NNLS regression which assures the non-negativity of the coefficients. The NNLS solution is not only non-negative, it can also be sparse to some extent, and further sparsity can be achieved by a proper thresholding algorithm.

Furthermore, since each column in matrices $\mathbf{X}$ and $\mathbf{Y}$ represents the posterior mixing probability over all reactions, the sum of all elements in a column should be 1. If one column can be represented by a linear combination of several other columns, i.e., Y=$\mathbf{X}\boldsymbol{\beta}$, thus $1 = \boldsymbol{1^T}Y = \boldsymbol{1^T}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{1^T}\boldsymbol{\beta}$. This implies another constraint on $\boldsymbol{\beta}$. Both LASSO and NNLS don't necessarily output the solution that strictly satisfy this third constraint for β.

As a measure of the quality of matching of two multinomial probability distributions, residual sum of squares is not the most proper criterion. A natural criterion for this purpose is the Jensen-Shannon Divergence. Thus a new method is developed by directly optimizing the Jensen-Shannon divergence between the target vector Y and a sparse non-negative linear combination of X such that the coefficients sum to 1.

In the rest of this chapter, I first review the methods of LASSO and NNLS and then describe our newly developed method which is more suitable for this application. Following the review of each method, I show an example of its performance on matching a subnetwork from $L = 100$, denoted as Y_1, with subnetworks from L=200 obtained from the mammalian metagenomics data by using BiomeNet.

2.1 LASSO Regression

2.1.1 Review of the method

Given a set of input measurements $X_1, X_2, \ldots, X_L$ and an outcome measurement Y, the LASSO [4] regression fits a linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_L X_L = \mathbf{X}\boldsymbol{\beta}$$

with a constraint on $|\boldsymbol{\beta}|$, the L_1 norm of the parameter vector, not greater than a given value.

The criterion is:

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{argmin}} && \sum_{i=1}^{N}(y_i - \sum_{j=1}^{L} x_{ij}\beta_j)^2 \\ &\text{subject to} && \sum_{j=1}^{L} |\beta_j| \leq s \end{aligned}$$

The bound s is a tuning parameter [2]. Of course when s is large enough, the constraint has no effect, and the solution of LASSO regression is just the same as the least square linear regression.

The problem can be equivalently represented as an unconstrained minimization problem with the L_1 norm penalty $\lambda|\boldsymbol{\beta}|$ added as following:

$$\text{argmin}\frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{L} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{L}|\beta_j|$$

An alternative regularized version of least squares is Ridge regression [4], which uses the constraint that $||\boldsymbol{\beta}||^2$, the L_2-norm of the parameter vector, is no greater than a given value. Here, LASSO regression is preferred to Ridge regression, since with the penalty λ increasing, all parameters of Ridge regression are shrunk towards 0, but not exactly equal to 0; while in LASSO more parameters are shrunk to zero which results in a sparse solution [2]. Since our purpose is to find several subnetworks that exhibit the strongest linear relation with the target subnetwork, sparsity is desirable.

Because the LASSO loss function is not quadratic, the optimization problem can not be solved using general convex optimization methods. Here the entire LASSO path is calculated by a modified form of the LARS (least angle regression) algorithm [2]. The R package "lars" is used to solve the LASSO fitting. By default in the package, each variable is standardized to have unit L_2-norm. In this application, the variables are multinomial probabilities that sum to 1; so, it's not proper to standardize the variables.

The solution path is piecewise linear – there are a finite number of the points at which the regression projection changes its direction [2]. The sparsity of the solution is controlled by the regularization parameter λ, which in general is chosen by a cross-validation procedure on training data. In our application, it is not proper to perform the cross-validation, because both responses and predictors are multinomial distribution probabilities. The sparsity of these vectors makes the variance of the cross-validation results very large and it is not proper to ignore the property that the sum of these vectors are all 1's. An alternative method to cross-validation for choosing the tuning parameter is to use Mallow's Cp as the model selection criterion [2]. A smaller Cp value indicates a better model. The Cp criterion is defined as.

$$\mathrm{Cp} = \frac{\mathrm{SSE}_k}{\mathrm{MSE}_p} - n + 2k$$

where

$\mathrm{SSE}_k = \sum_{i=1}^{n}(y_i - \hat{y}_{i(k)})^2$ is the sum of squares for the model with k predictors.

$\hat{y}_{i(k)}$ is the predicted value of the ith observation y_i from the model with k regressors.

MSE_p is the mean square error on the complete set of p regressors. n is the sample size.

The complexity of the models in the LASSO path can be represented by their degrees of freedom. The number of degrees of freedom for a linear model is defined as the true dimension of the linear subspace in which the predictors of the model lie. However defining the number of predictors in the LASSO model is only an approximation because each dimension shouldn't be counted as a full degree of freedom due to the shrinkage applied [2]. The approximation works due to a useful conclusion in [19] that when fitting a linear model via LASSO stopping at some number of steps $k < p$, the df of the modified LAR procedure at any stage, is approximately equal to the number of predictors in the model. Thus if the selected model includes k non-zero coefficients, I define the degrees of freedom (d.f.) as k.

2.1.2 An example of LASSO fitting

As an example, I use one subnetwork from the $L = 100$ case, denoted as Y_1, to illustrate the LASSO fitting. Note that the same Y_1 variable is also used as illustration example for the NNLS and our newly developed method.

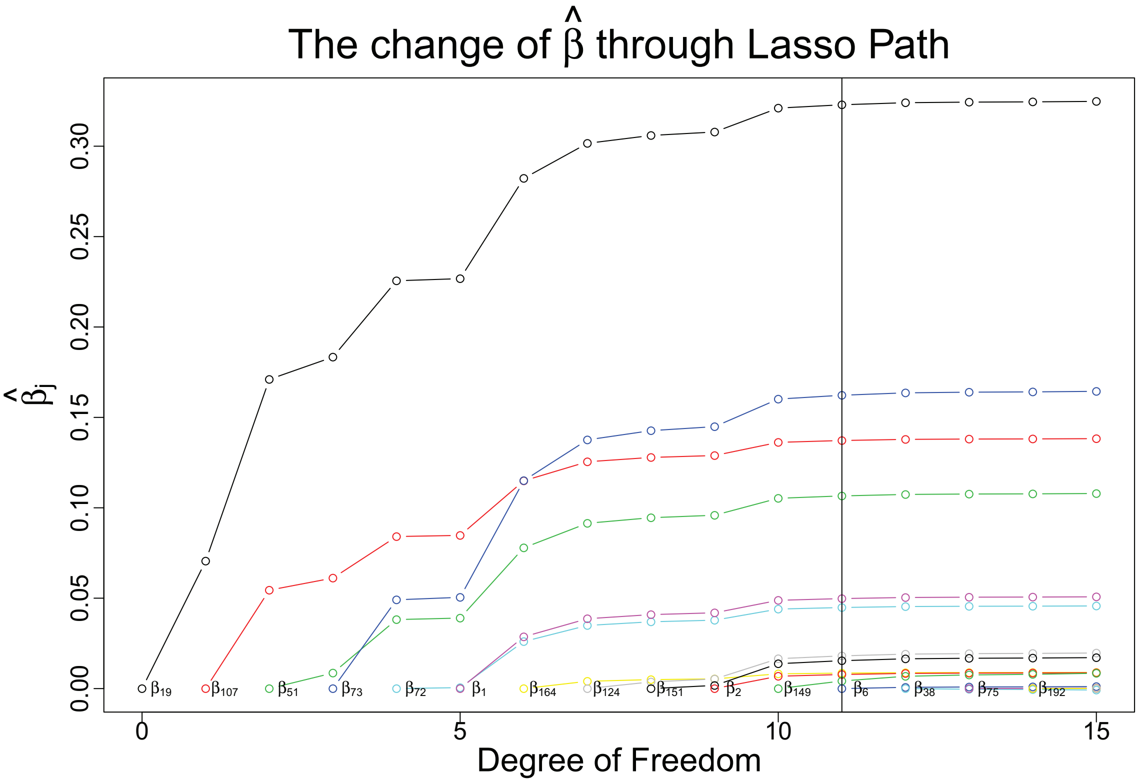


Figure 2.1: Profiles of LASSO coefficients. A vertical line is drawn at df=11, the value chosen by smallest Cp criterion. The profiles are piece-wise linear, and so are computed only at the first 15 steps of LASSO path. β_j is the coefficients of the predictor X_j . With the increase of df, the predictor is added into the model one by one.

In this particular LASSO regression, there are 153 steps in the whole LASSO path which means that 153 variables are added into the model at the end of LASSO. Df is used to denote the number of variables that are already selected into the model at each step.

df	index	Cp	RSS	λ
1	19	441.7240109	0.0513347	0.0899638
2	107	331.2993961	0.0495664	0.0519204
3	51	232.4238521	0.0479797	0.0425626
4	73	7.6186055	0.0444123	0.0357959
5	72	-1.3932404	0.0442391	0.0196929
6	1	-119.2749048	0.0423535	0.0187911
7	164	-120.7537535	0.0422988	0.0061926
8	124	-119.9081055	0.0422807	0.0055229
9	151	-126.6824536	0.0421426	0.0053128
10	2	-130.3282597	0.0420538	0.0035695
11	149	-130.6962969	0.0420166	0.0019435
12	6	-128.8005140	0.0420150	0.0006002
13	38	-126.9147353	0.0420132	0.0004717
14	75	-124.9477789	0.0420126	0.0002889
15	192	-122.9586742	0.0420125	0.0002141
16	17	-120.9757184	0.0420122	0.0001849
17	109	-118.9774784	0.0420122	0.0001306
18	26	-116.9820573	0.0420121	0.0001240
19	179	-114.9826678	0.0420121	0.0001062
20	93	-112.9835537	0.0420121	0.0001037
21	70	-110.9861158	0.0420120	0.0001002
22	101	-108.9870484	0.0420120	0.0000899
23	99	-106.9874792	0.0420120	0.0000860
24	31	-104.9894883	0.0420120	0.0000842
25	65	-102.9904490	0.0420120	0.0000759
26	47	-100.9914000	0.0420120	0.0000717
27	48	-98.9939734	0.0420119	0.0000674
28	184	-96.9943649	0.0420119	0.0000549
29	178	-94.9949087	0.0420119	0.0000528
30	177	-92.9981477	0.0420119	0.0000499

Table 2.1: The first 30 steps of the LASSO path on Y_1. Df denotes degrees of freedom of the model. Index indicates which predictor is included in the model at each step.

Table 2.1 shows the first 30 steps of the LASSO path. From Table 2.1, the smallest Cp value corresponds to df=11, which means 11 variables are selected by this model. The LASSO path is dispalyed in Figure 2.1, where a vertical line is drawn at df=11 which is the best model selected by Mallow's Cp criterion.

The coefficients of the selected model are plotted in Figure 2.2. From the plot, X_{19} ranks highest in matching Y and is followed by X_{107} and X_{51}. Meanwhile, X_2, X_{149} and X_{124} play a trivial role in matching Y since the corresponding coefficients are relatively small.

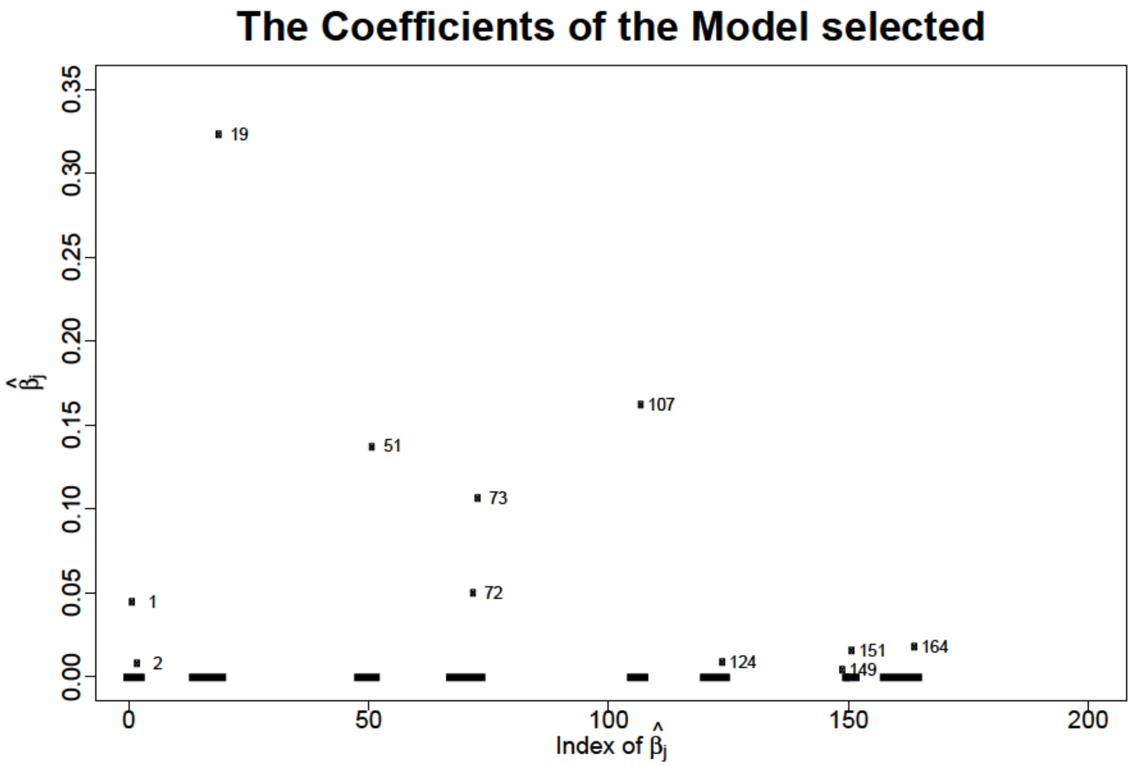


Figure 2.2: Profiles of LASSO coefficients of the selected best model on Y_1. Among the 200 estimated coefficients, 11 variables are selected into model, all with positive coefficients. The index numbers of the predictors are marked next to each point.

2.2 NNLS Regression

2.2.1 Review of the method

Given X_1, X_2, ..., X_L in $\mathbf{X}_{N\times L}$ matrix as the predictor variables and a column vector Y as the response variable, the criterion of NNLS regression is [18]:

$$\begin{array}{ll} \underset{\beta}{\operatorname{argmin}} & ||Y - \mathbf{X}\boldsymbol{\beta}||_2 \\ \text{subject to} & \beta_j \geq 0,\ j = 1, \ldots, L. \end{array}$$

The NNLS optimization problem is a quadratic programming problem [18],

$$\begin{aligned} \underset{\boldsymbol{\beta}\geq 0}{\operatorname{argmin}}||Y - \mathbf{X}\boldsymbol{\beta}||_2 &= \underset{\boldsymbol{\beta}\geq 0}{\operatorname{argmin}}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - Y^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X}^TY + Y^TY) \\ &= \underset{\boldsymbol{\beta}\geq 0}{\operatorname{argmin}}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\beta - 2Y^T\mathbf{X}\boldsymbol{\beta} + Y^TY) \\ &= \underset{\boldsymbol{\beta}\geq 0}{\operatorname{argmin}}(\frac{1}{2}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - Y^T\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

There are many similar algorithms to solve the NNLS optimization problem, of which the first widely used algorithm is an active set method published by Lawson and Hanson in their 1974 book [18]. The R package “nnls” is used to solve the NNLS fitting in this study.

The NNLS regression could result in many positive coefficients in the model, which is not convenient for selecting the most important predictors. In our example most of the estimated coefficients are extremely small and can be neglected without much increase in RSS. Therefore, thresholding these small positive coefficients has little influence on the precision of the matching.

Here, hard-thresholding is used in model selection of NNLS. The hard-thresholding works by setting the threshold limit t equal to a sequence of increasing numbers. To calculate the corresponding RSS, all the coefficients that are less than or equal to the threshold t are set to zero. Hence I obtain a sequence of RSS corresponding to the sequence of thresholds. Our thresholding criterion is to choose the maximum threshold such that the difference in RSS before and after thresholding is less than 10^{-5}.

2.2.2 An example of NNLS fitting

As an example, I again use Y_1 to illustrate the NNLS fitting. Among the 200 predictor variables NNLS analysis resulted in 71 positive coefficients with the rest of the coefficients all zero coefficients.

Figure 2.3 shows the change in the values of RSS versus threshold. The threshold t is set from 0 to 0.05 with an increment of 0.001. The vertical line is the model indicating that the threshold is 0.009. The coefficients smaller than 0.009 are thresholded to zero. Before applying for any thresholding, the RSS is 0.04201208, and after thresholding at 0.009 the RSS is 0.04201267 with 11 nonzero coefficients left. The difference before and after thresholding is no more than 10^{-5}. Given such a very small difference in RSS it is reasonable to ignore the loss of precision.

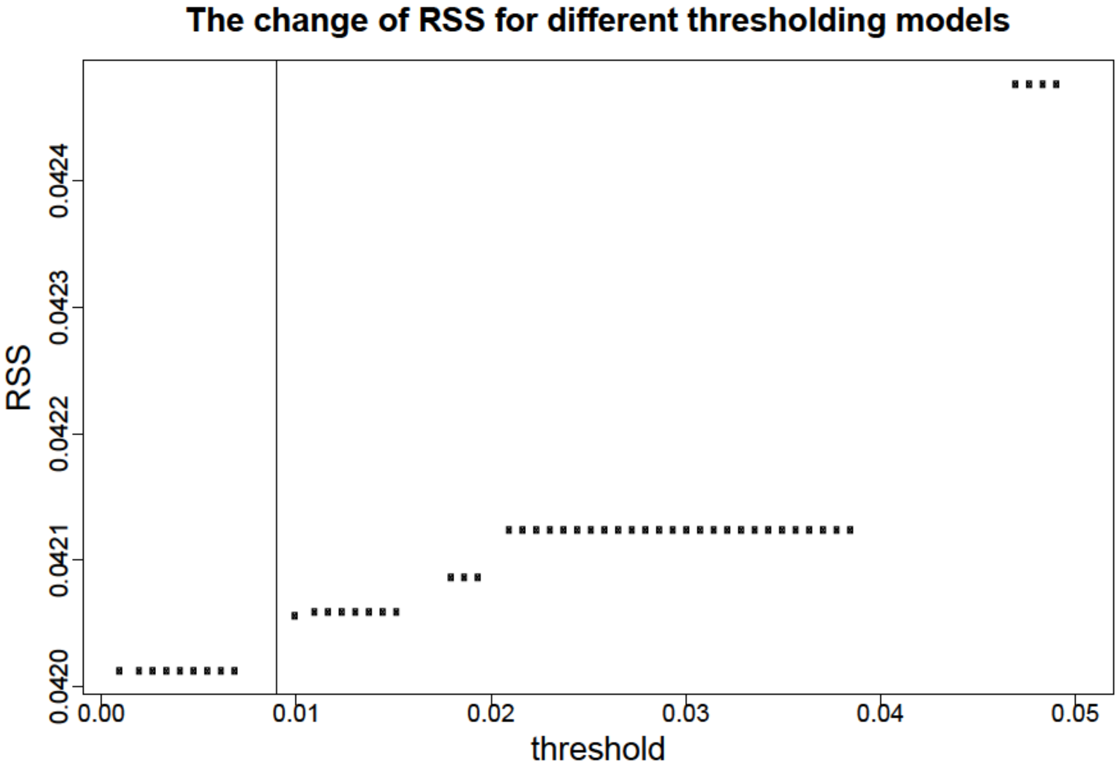


Figure 2.3: Change of RSS with increment of threshold in NNLS regression of Y_1. The values of RSS are plotted verses the values of the thresholds. The vertical line is the model chosen when threshold is 0.009.

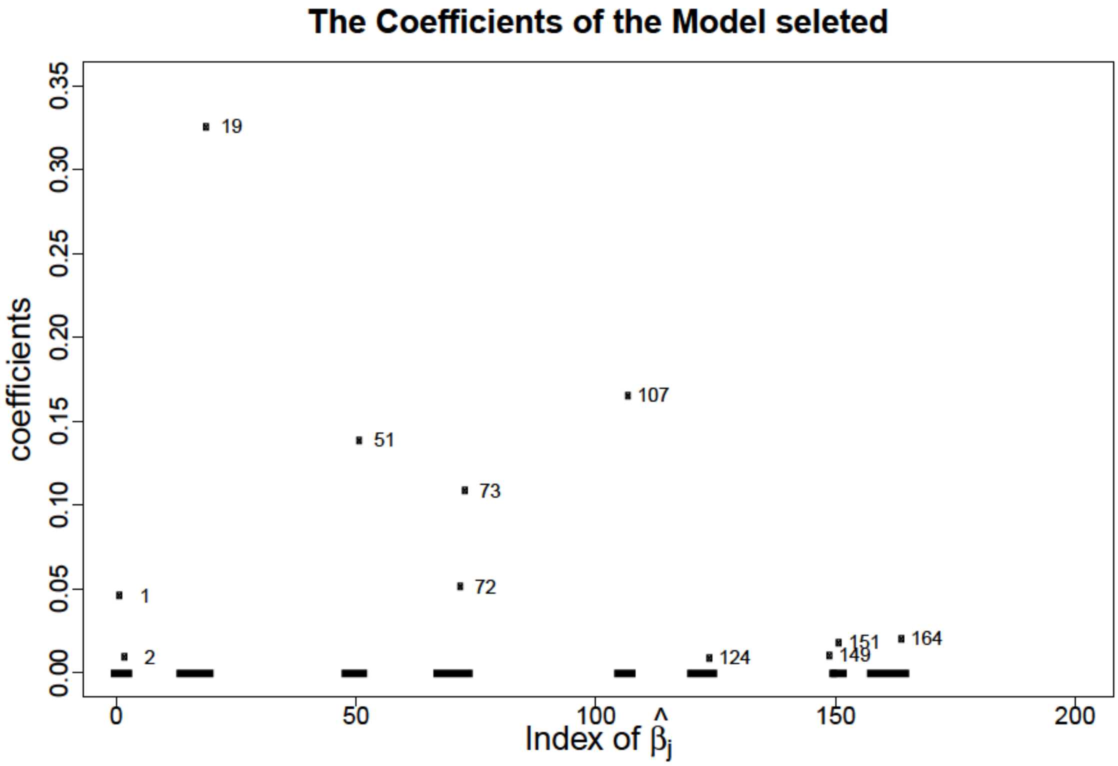


Figure 2.4: The coefficients of the NNLS model of Y_1 after thresholding. Among the 200 estimated coefficients, 11 variables remain in the model. The index numbers of nonzero β_j are marked next to the points.

The best model after thresholding is plotted in Figure 2.4. From the plot, X_{19} ranks highest in matching Y and is followed by X_{107} and X_{51}. Meanwhile, X_2, X_{149} and X_{124} play a trivial role in matching Y since the corresponding coefficients are very small. The results are very similar to the LASSO profile (by comparing Figure 2.4 and Figure 2.2).

2.3 A method based on minimizing JSD

Different from LASSO and NNLS regression, our proposed method uses Jensen-Shannon Divergence (JSD) instead of RSS as the optimization criterion. JSD is a widely used method of measuring the similarity between two probability distributions, and it is widely applied in bioinformatics and genome comparisons [9].

The Jensen-Shannon Divergence between two multinomial probability vectors $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ is defined as [9]:

$$\text{JSD}(x||y) = \frac{1}{2}(\sum_{i=1}^{n} x_i \log \frac{x_i}{h_i} + \sum_{i=1}^{n} y_i \log \frac{y_i}{h_i}) \tag{2.1}$$

where $h = \frac{x+y}{2}$ is the mean vector of x and y.

From the definition, it is clear that JSD is a modified version of Kullback-Leibler divergence. The direct application of Kullback-Leibler divergence is not possible to measure two multinomial distributions with many zero probability categories.

Often the logarithm in the JSD definition uses 2 as its base, in which case JSD has the property: $0 \leq \text{JSD} \leq 1$. Thus when the logarithm uses e as its base, it has the property: $0 \leq \text{JSD} \leq \ln(2)$.

In this thesis, e is used as the logarithm base. Jensen-Shannon Distance is defined as the square root of Jensen-Shannon Divergence [7]. As a distance measure, Jensen-Shannon Distance satisfies all three properties required of a distance [9].

2.3.1 The method

Given a set of multinomial probability vectors $X_1, X_2, ..., X_p$ in the matrix $\mathbf{X}$ and a target multinomial probability vector Y selected from the matrix $\mathbf{Y}$, the aim is to find a linear combination of several X variables which has smallest JSD with the variable Y. i.e.

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_L X_L$$

with the coefficients satisfying $\sum_{j=1}^{L} \beta_j = 1,\ \beta_j \in [0,1]$. This guarantees that $\hat{Y}$ is a multinomial probability vector too.

The objective function for finding the β is

$$\min_{\beta} \text{JSD}(Y||\hat{Y})$$

Directly optimizing this function is challenging. At first several optimization algorithms such as gradient descent and Newton's methods with constraints [13] were tried to solve the problem directly. But the Jacobian or Hessian is unavailable analytically and too expensive to compute numerically at every iteration. Then I considered an alternative method to Newton's methods (i.e., Quasi-Newton methods) without requiring the Jacobian in order to search for zeros [13] [3]. However, Quasi-Newton methods often find a local maximum and minimum of a function and the solution is highly dependent on the initial values [3]. Since different sets of initial values can result in different local maximum and minimum, the need to set the 200 initial values became a big problem. The output of 200 coefficients are not sparse at all if the input numbers are not sparse. The challenge posed by directly optimizing this function leads us to consider other methods.

A heuristic searching strategy which is similar in nature to the forward variable selection procedure is proposed. First, I calculate the JSD between Y and each X_j variable, JSD($Y||X_1$), JSD($Y||X_2$),. . . , JSD($Y||X_L$), which are then ordered in an increasing order. Thus, the X variables are ordered as $X_{[1]}, \ldots, X_{[L]}$, so that the variables in X which are more similar to Y will be considered first. I first assign the best match of Y as $X_{[1]}$ and calculate the JSD, then I consider whether $X_{[2]}$ should be added into the model. If the linear combination of $X_{[1]}$ and $X_{[2]}$ lead to a smaller JSD with Y variable, then $X_{[2]}$ is added into the model, otherwise the procedure is stopped. Adding X variables according to this procedure ensures that the linear combination satisfy the constraints on the coefficients while the JSD to Y is decreased at each step. The optimization criterion at first step is following:

$$\begin{array}{ll} \min\limits_{\beta} & \mathrm{JSD}(Y||(\beta_1 X_{[1]} + \beta_2 X_{[2]})) \\ \text{subject to} & \beta_1 + \beta_2 = 1,\ \beta_1 \in [0,1],\ \beta_2 \in [0,1] \end{array}$$

The R package named “General-purpose Optimization” is used to solve the optimization problem with constraint.

The procedure stops when adding another variable will no longer reduce the JSD (i.e., the coefficient for the newly added variable is zero). The details of this forward variable selection procedure are summarized in the following algorithm:

Note that MJSD might not find the optimum result for all the coefficients. This could happen when the selected subnetworks overlap each other. However, when one subnetwork is split into almost non-overlapping several smaller subnetworks (as we expect will be a common case), the MJSD method will provide nearly optimum results. In the next section I give an example showing that the selected X variables are non-overlapping subnetworks.

Algorithm 1: A forward selection procedure to sequentially minimize JSD

```
1. Initialization
    (a) Calculate the JSD between $Y$ to each $X$ variable, order $X$ variables according to their JSD from smallest to largest.
    (b) Set an set $P = \varnothing$, set $R = [X_{[1]}, \ldots, X_{[p]}]$. Move $X_{[1]}$ from set R to set P, set $\hat{Y} = X_{[1]}$, $\beta^* = 1$.
2. At the $i$th step (i=2,... $p$):
    (a) Move $X_{[i]}$ variable from set R to set P: calculate $\beta^*$ that minimize JSD($Y||\beta^*\hat{Y} + (1 - \beta^*)X_{[i]}$).
    (b) If $1 - \beta^* > 0$, update $\hat{Y} = \beta^*\hat{Y} + (1 - \beta^*)X_{[i]}$. Go back to step 2(a).
        Else if $1 - \beta^* = 0$, the recursive procedure stops, output $\hat{Y}$.
```

2.3.2 An example of JSD minimization method

I again use Y_1 as an example to illustrate the method's ability in finding the best matching set of variables in X.

The JSD of each of the 200 X variables with Y_1 are shown in Figure 2.5. Here the smallest one is JSD($Y_1||X_{19}$), the second smallest one is JSD($Y_1||X_{73}$). In Figure 2.5, most of the 200 X variables have JSD near the upper bound of ln(2). The X variable with smallest JSD values will be selected by the model at first.

When β^* converges to 1, it shows that the program adds all the 11 smallest X_j variables to the linear combination resulting in the final optimal JSD as 0.3511393. Table 2.2 shows the updated parameters of β^* and JSD at each step. X_j is the variable added into the model in each step. At the beginning the initial β^* is equal to 1, and then it becomes a value between 0 and 1 with more variables selected into the model. The β^* stops at 0.9957691 as shown in Table 2.2. β^* is 1 at the 12 step.

	X_j	β^*	$\text{JSD}(Y\|\|\hat{Y})$
1	X_{19}	1.0000000	0.4572171
2	X_{73}	0.7081007	0.4019348
3	X_{51}	0.7652943	0.3733276
4	X_{107}	0.9293715	0.3700184
5	X_1	0.8874797	0.3557269
6	X_{72}	0.9648953	0.3524371
7	X_{151}	0.9987738	0.3524122
8	X_{164}	0.9980228	0.3523677
9	X_2	0.9857460	0.3513407
10	X_{149}	0.9984250	0.3513016
11	X_{124}	0.9957691	0.3511393

Table 2.2: The updated parameters β^* for MJSD method on Y_1

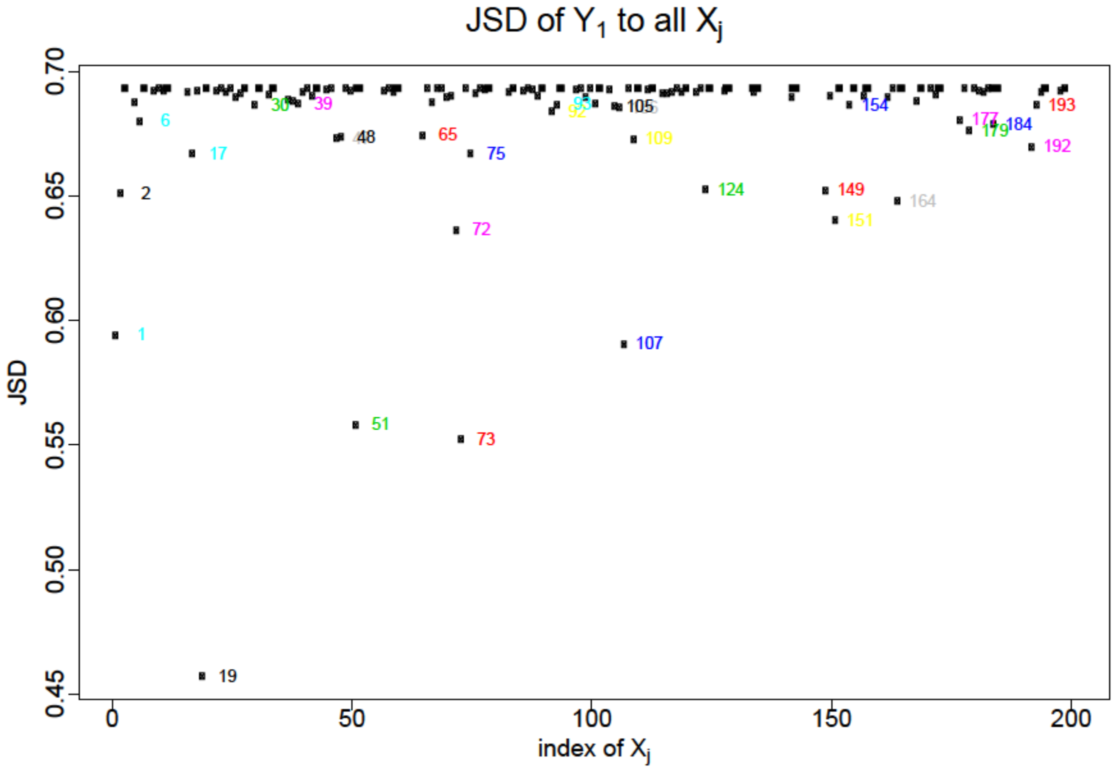


Figure 2.5: Profile of JSD of Y_1 to all 200 X variables. JSD values are plotted verses the index of 200 X variables. The index number of some X variables are also marked next to the points.

$\hat{\beta}_j$ are plotted in Figure 2.6. $\hat{\beta}_j$ stands for the coefficient of X_j that has been selected into the model after the whole procedure stops. $\sum \hat{\beta}_j = 1$. X_{19} still ranks highest in matching Y, but the weight calculated by this method is even larger than LASSO and NNLS methods. The same set of variables are selected by all three methods, which include X_{19}, $X_{51}, X_{73}, X_{107}, X_1, X_{72}, X_2, X_{124}, X_{149}, X_{124}, X_{164}$, however the order that these variables enter the model, and the coefficients, are different.

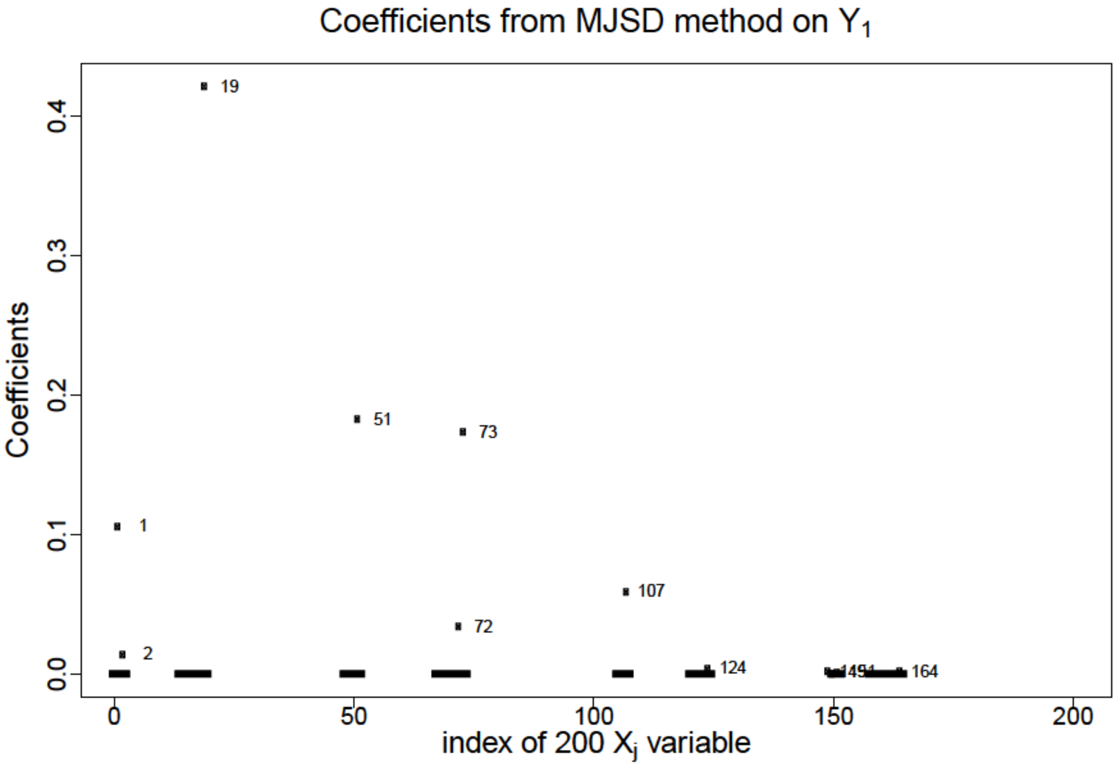


Figure 2.6: Coefficients of the variables resulted from the MJSD method. Among the 200 estimated coefficients, 11 X variables are selected into model with corresponding positive coefficients. The index numbers of the predictors are marked next to the points.

Next, I investigated the overlap in the reaction composition of the 11 subnetworks with positive coefficients that were added into the model by MJSD. Figure 2.7 shows all the nonzero reactions in the 11 selected Xvariables. For this plot, reactions were reordered as follows. First, the nonzero reactions of X_{19} were ordered from largest to smallest. Then, any remaining nonzero reactions of X_{51} were ordered from largest to smallest and added to the sequence derived from X_{19}. This process was repeated until all the nonzero reactions of all the 11 X variables were reordered. The distribution of reactions in Figure 2.7 shows that the selected subnetworks are largely non-overlapping at the reaction level (i.e., they are comprised of different reactions). Thus the the MJSD method should provide very similar results to those that would be obtained by directly optimizing the linear combination of those 11 X variables.

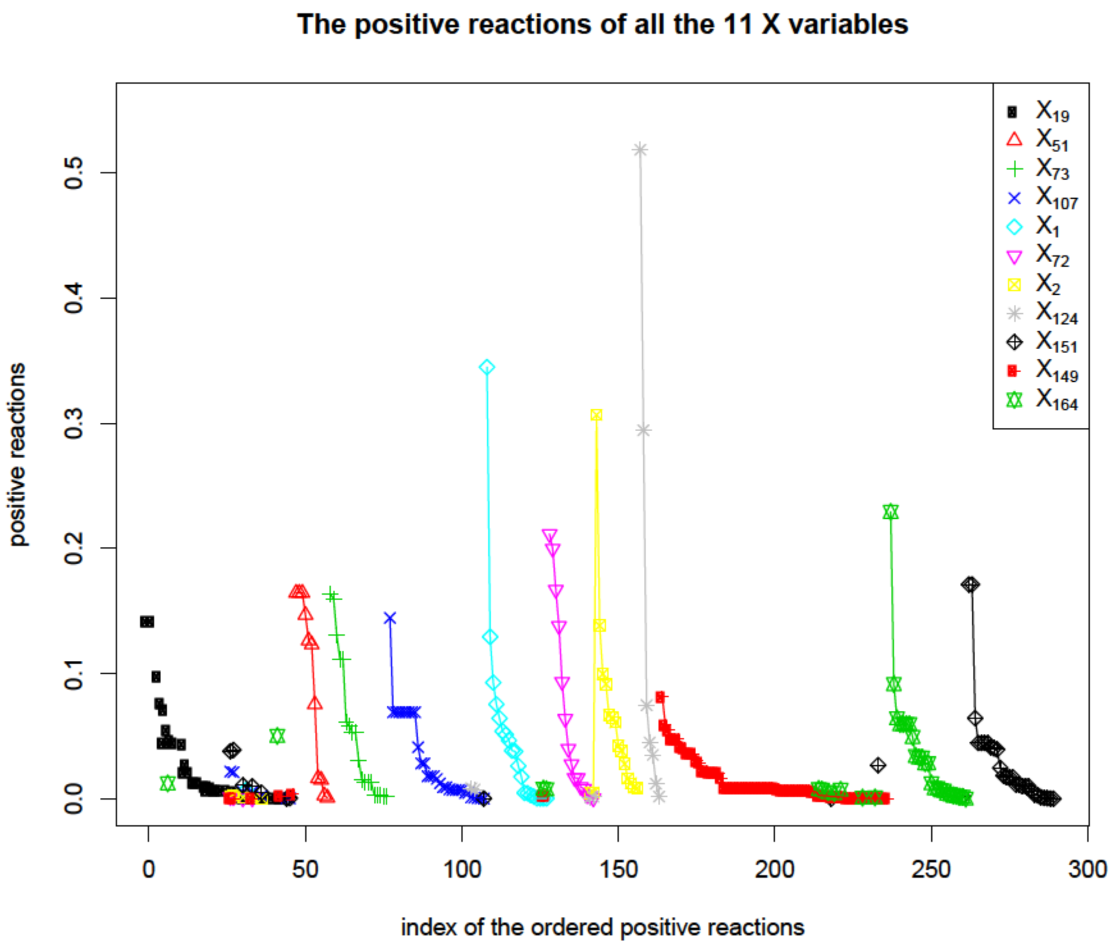


Figure 2.7: The positive reactions of all the 11 X variables. The legend indicates the index of the 11 X variables and their corresponding point types.

2.4 Summary

LASSO regression results satisfy the requirement of sparsity of $\hat{\beta}$, through a proper model selection criterion to select the best model from a full LASSO path. In our example, the best model already satisfies the non-negativity requirement of the coefficients. The non-negativity was naturally satisfied in 93% of all the 100 subnetwork matchings when using Cp-statistics for model selection. For the remaining 7% of subnetworks, some very small negative coefficients result from LASSO fitting among their matching X variables. By changing these small negative coefficients to zero, the RSS is not much changed.

NNLS regression results naturally satisfy the non-negativity constraints of the coefficients. The NNLS solution is not only non-negative, but also sparse to some extent. However, there are still too many very small positive coefficients in the NNLS regression results. Hence, further sparsity has to be achieved through an appropriate thresholding algorithm without much increase in the final RSS.

The new method, named as MJSD, uses a forward variable selection procedure to simplify the optimization problem. It is based on directly minimizing JSD to obtain the positive and sparse coefficients that sum to 1, without the model selection step in LASSO or the hard-thresholding in NNLS. Note that since LASSO and NNLS don't necessarily output a solution satisfying $\hat{\beta}\mathbf{1}$=1, the resulting $\hat{Y}$ from LASSO and NNLS are not probability vectors.

In order to compare the results, I listed all the final coefficients from these three methods in Table 2.3. The three methods have selected the same subset of variables in $\mathbf{X}$ to match Y_1, but the coefficients are different. The LASSO and NNLS results are more similar in this case. To compare the fitting of these three methods, both RSS and JSD on Y_1 are used as the criteria. RSS for the three methods can be directly calculated from RSS $= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. However, before calculating the JSD for LASSO and NNLS matches, I need to normalize the vector $\hat{Y}$ to sum to 1. The normalized vector is denoted as $\hat{Y}' = [\hat{y}'_1, \ldots, \hat{y}'_N]^T$, where $\hat{y}'_i$=$\hat{y}_i / \sum_{i=1}^{N} \hat{y}_i$. Then the JSD($Y||\hat{Y}'$) is used as criterion for LASSO and NNLS.

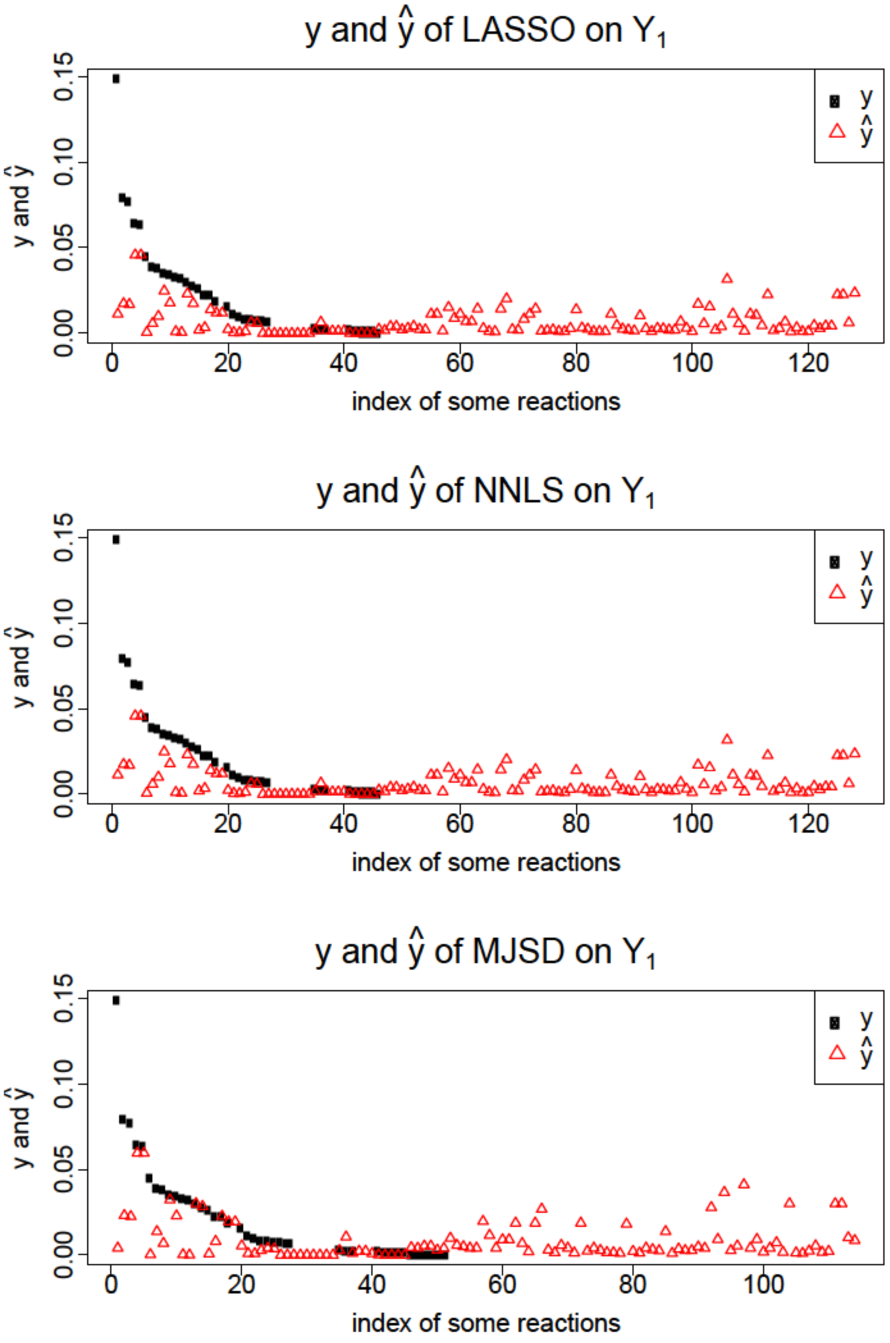
JSD and RSS for the three methods on Y_1 are also shown in Table 2.3. When using JSD as the criterion, the matching of MJSD is better than LASSO and NNLS, and NNLS is a little better than LASSO. However, when using RSS as criterion, the matching of LASSO and NNLS are better than MJSD, and NNLS is still a little better than LASSO. This outcome is reasonable because LASSO and NNLS use RSS as optimization criterion, while MJSD use JSD as optimization criterion.

	LASSO	NNLS	MJSD
$\hat{\beta}_{19}$	0.3235519	0.3256803	0.4213065
$\hat{\beta}_{107}$	0.1631789	0.1654160	0.0590834
$\hat{\beta}_{51}$	0.1371404	0.1387444	0.1824729
$\hat{\beta}_{73}$	0.1067218	0.1085419	0.1736745
$\hat{\beta}_{72}$	0.0496812	0.0512407	0.0496812
$\hat{\beta}_{1}$	0.0446626	0.0461363	0.1060616
$\hat{\beta}_{164}$	0.0184734	0.0204525	0.0184734
$\hat{\beta}_{151}$	0.0159389	0.0179019	0.0159389
$\hat{\beta}_{124}$	0.0081506	0.0091444	0.0081506
$\hat{\beta}_{2}$	0.0076767	0.0092262	0.0076767
$\hat{\beta}_{149}$	0.0072906	0.0105463	0.0072906
$\sum \hat{\beta}_j$	0.8824671	0.9030308	1
JSD	0.3636888	0.3650386	0.3511393
RSS	0.04201496	0.04201208	0.04487792

Table 2.3: $\hat{\beta}_j$, RSS, JSD of three methods on Y_1

To view the fitting of these three methods more intuitively, I plot the matching of some y's and $\hat{y}$'s in Figure 2.8. All the 2824 reactions are too many to plot, with most of them being zeros. Thus only those reactions which have either positive y's or positive $\hat{y}$'s are plotted in the Figure 2.8. The index of reactions are ordered from largest y's to smallest y's. On the whole the matching results of these three methods are similar, except for some small differences. It can be seen that large reactions are matched slightly better in MJSD, while small reactions are matched slightly better in LASSO and NNLS.

The above is just an example of one matching. The other 99 column vectors in the 2824×100 reaction matrix can be matched by the same methods that were applied to Y_1. In fact, this example is not a relatively good match among all the 100 matches because $\hat{Y}_1$ has quite large deviation from Y_1, as shown in Figure 2.8. The results for all the 100 matchings of three methods are shown in Chapter 3, including several more examples of both good and bad matches.



y and $\hat{y}$ of LASSO on Y_1

y and $\hat{y}$ of NNLS on Y_1

y and $\hat{y}$ of MJSD on Y_1

Figure 2.8: Comparison of some y_i's and $\hat{y}_i$'s of three methods. Black dots denote y's and red dots denote $\hat{y}$'s. Of all 2824 reactions, those which have either positive y's or positive $\hat{y}$'s are plotted in the figure. The index of reactions are ordered according to the decreasing values for y's.

Chapter 3

Results for matching the metabolic components of complex microbial communities

In Chapter 2, a single subnetwork was used as an example to show the performance of three methods. I compared the results of the three methods and found that all of them can achieve the goal of matching the subnetworks to some extent.

In this chapter, these three methods are applied on all the 100 subnetworks in the $L = 100$ run of BiomeNet to match the subnetworks in the $L = 200$ run. First I calculate RSS and JSD for all the 100 matches on three methods as criteria to rank the good matches and give examples of a good and a bad match according to those criteria. Then I test all the matches using all of the three methods, with the null distributions generated by permuting each variable in the $\mathbf{X}$ matrix a thousand times to discriminate those significantly well matched Y variables. Finally I present the estimated coefficient matrices of the three methods to find the features of the good matches.

3.1 RSS and JSD of all the matches

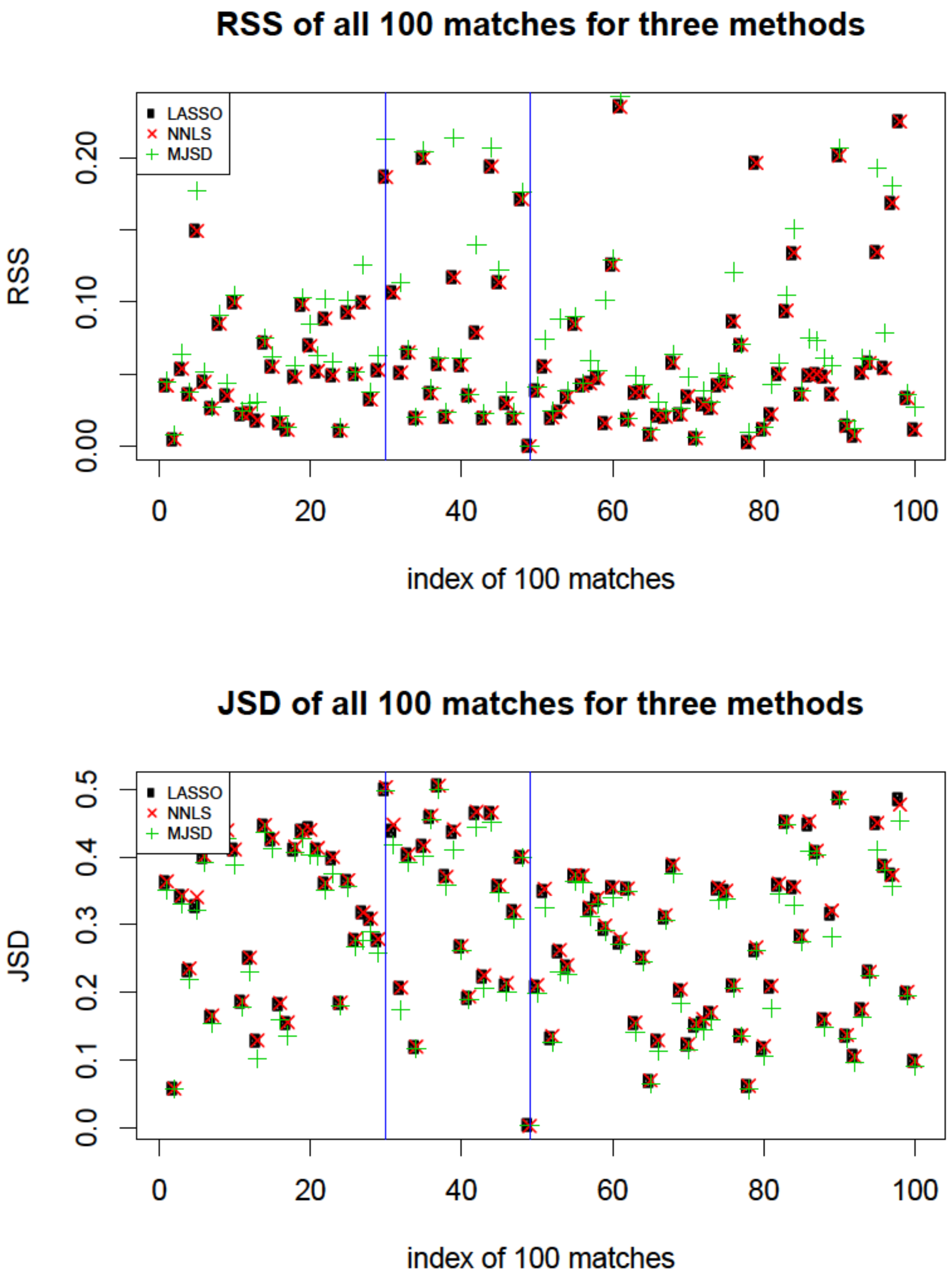
Among all the 100 matches for each method, I still use RSS and JSD as criteria to rank matches between subnetworks.

The RSS and JSD of all 100 matches by each method are plotted in Figure 3.1. When using RSS as criterion, most of the black and red points are overlapping with each other and lower than the green points, which shows that LASSO and NNLS perform equally well on most of the 100 matches, and a little better than MJSD. When using JSD as criterion, most of the green points are lower than the black and red points which shows that MJSD performs better than LASSO and NNLS on most of the 100 matches. Between NNLS and LASSO, which is better depends on the particular matches.

The scatter distributions of RSS and JSD show that not all 100 Y variables are matched equally well by the same criterion. Some subnetworks are matched very well from all the three methods while some subnetworks can not be matched well by any of the three methods. Some subnetworks can be matched a little better by one or two methods.

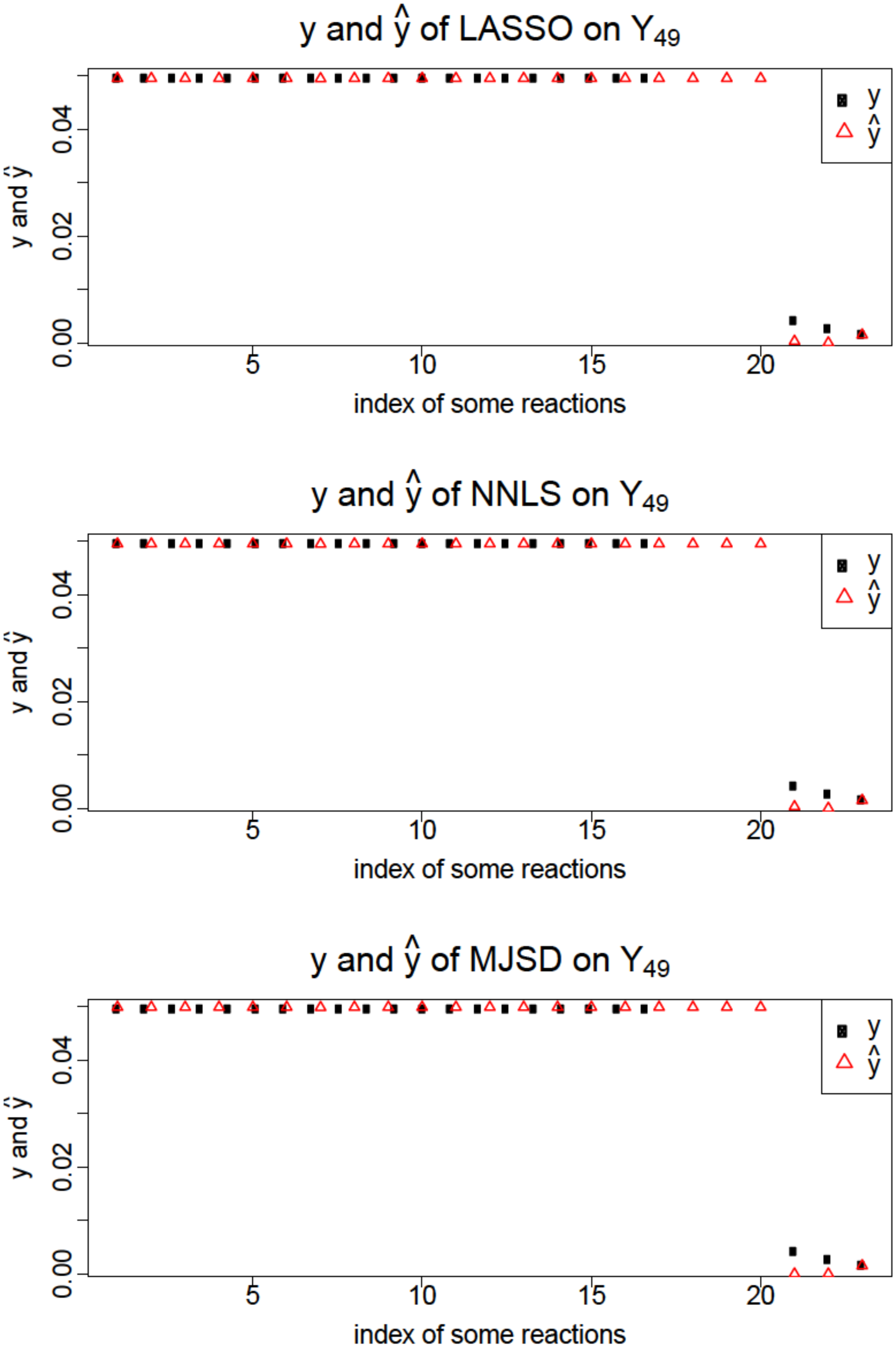
I give an example of a good match on Y_{49} in Figure 3.2, and an example of a bad match on Y_{30} in Figure 3.3. Y_{49} has both smallest RSS and JSD among 100 Y variables. All the large y_i's are matched well by $\hat{y}_i$'s, and those zero y_i's are matched by zero $\hat{y}_i$'s as shown in Figure 3.2. y and $\hat{y}$ of Y_{30} are shown in Figure 3.3. Although the match of large y_i's are better in MJSD than LASSO and NNLS, the deviation on the zero weight elements are larger by MJSD as well. In general, Y_{30} is not matched well by any of the three methods.

The quality of the matches ranked by RSS and JSD are not exactly the same, but not far from each other. For example, one variable which ranks as 30th in JSD might rank 35th by RSS.


RSS of all 100 matches for three methods

JSD of all 100 matches for three methods

Figure 3.1: RSS and JSD of all 100 matches for three methods. Black circle points indicate values of LASSO, red cross points indicate values of NNLS, and green plus points indicate values of MJSD. Y_{49} and Y_{30} are marked by blue vertical lines.



y and $\hat{y}$ of LASSO on Y_{49}

y and $\hat{y}$ of NNLS on Y_{49}

y and $\hat{y}$ of MJSD on Y_{49}

Figure 3.2: Comparison of some y_i's and $\hat{y}_i$'s of three methods on Y_{49}. Black dots denote y's and red dots denote $\hat{y}$'s. Of all 2824 reactions, those which have either positive y's or positive $\hat{y}$'s are plotted in the figure. The index of reactions are ordered according to the decreasing values for y's.

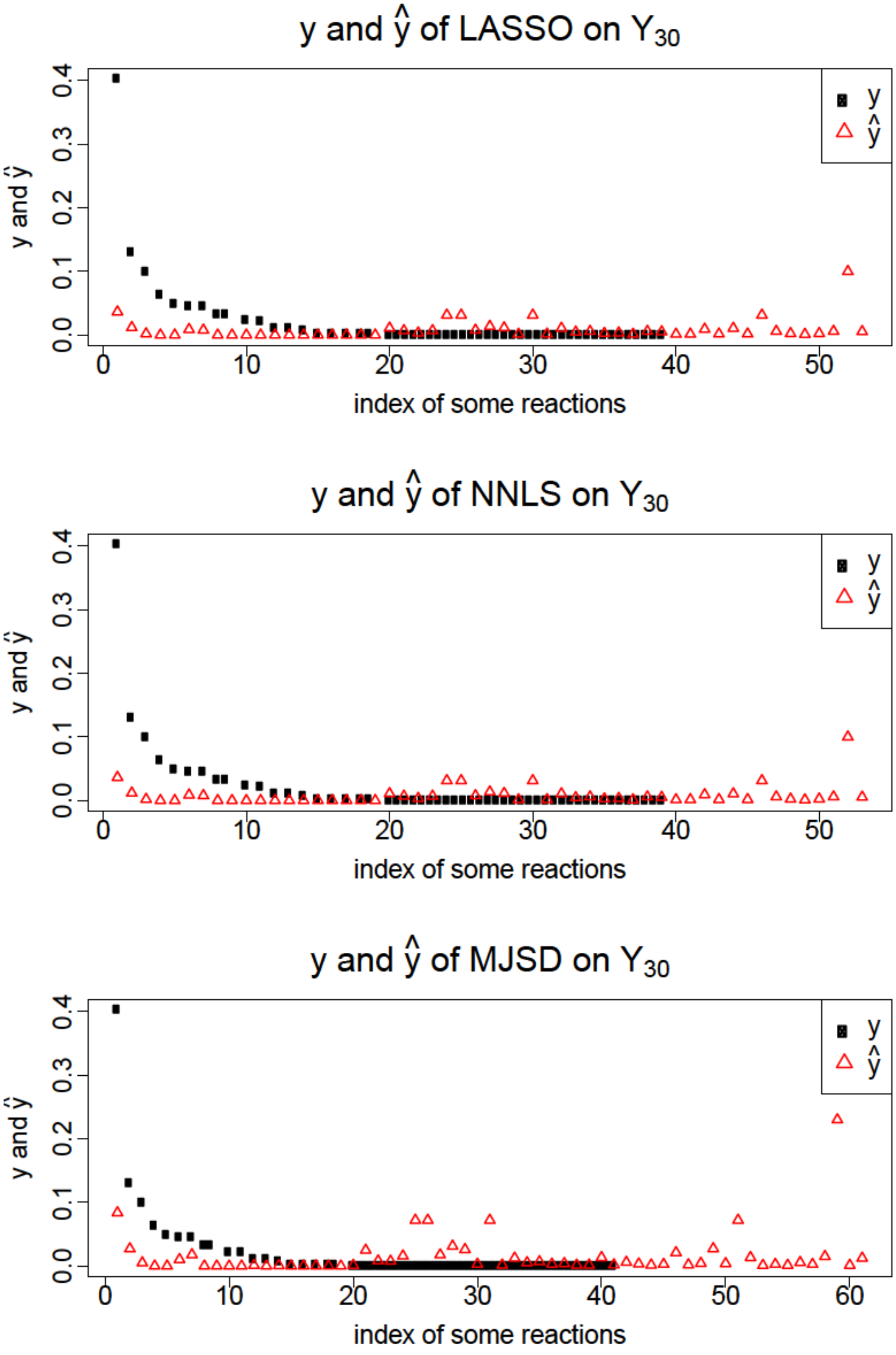
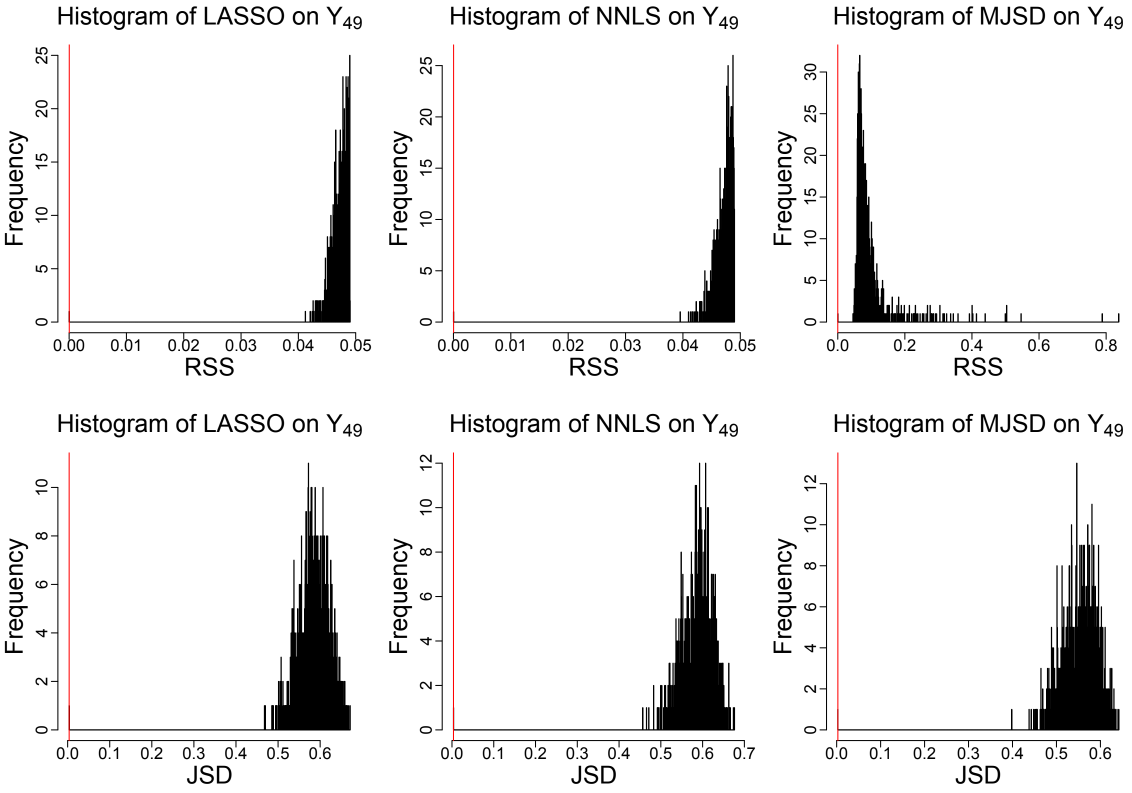


Figure 3.3: Comparison of some y_i's and $\hat{y}_i$ by three methods on Y_{30}. Black dots denote y's and red dots denote $\hat{y}$'s. Of all 2824 reactions, those which have either positive y's or positive $\hat{y}$'s are plotted in the figure. The index of reactions are ordered according to the decreasing values for y's.

3.2 Quality of the matches

The quality of the matches can be judged by whether they are much better than random matches and how sparse the matches are. In this section I first describe a permutation test and apply it to each match to find those which are significantly better than the random matches, then I compare the sparsity of the matches.

First, the elements of each column in the **X** matrix are permuted to obtain a permuted **X** matrix of which the columns are still probability vectors that sum to 1 but the weights for reactions in a column are different from the original **X** matrix. Then all the three methods are applied to match the permuted **X** matrix with each column in the **Y** matrix, and calculate the JSD and RSS of the new 100 matches. I repeat the permutation and matching process 1000 times, and obtain 1000 RSS and 1000 JSD values from each of the 100 matches for each of the three methods.



Histogram of LASSO on Y_{49} | Histogram of NNLS on Y_{49} | Histogram of MJSD on Y_{49}

Histogram of LASSO on Y_{49} | Histogram of NNLS on Y_{49} | Histogram of MJSD on Y_{49}

Figure 3.4: RSS and JSD permutation histogram of Y_{49} for three methods. The red vertical lines indicate the JSD values of the original matches.

Here I still use the Y_{49} and Y_{30} as examples. As shown in Figure 3.4 and Figure 3.5, there are six permutation histograms for each Y variable, for two criteria and three methods. Y_{49} is a good match thus the RSS and JSD values of original matches seem significantly smaller than permuted ones, while that of Y_{30} are not so significant. For all the 100 matches, P-values are used to decide how many matches are significantly better than random matches.

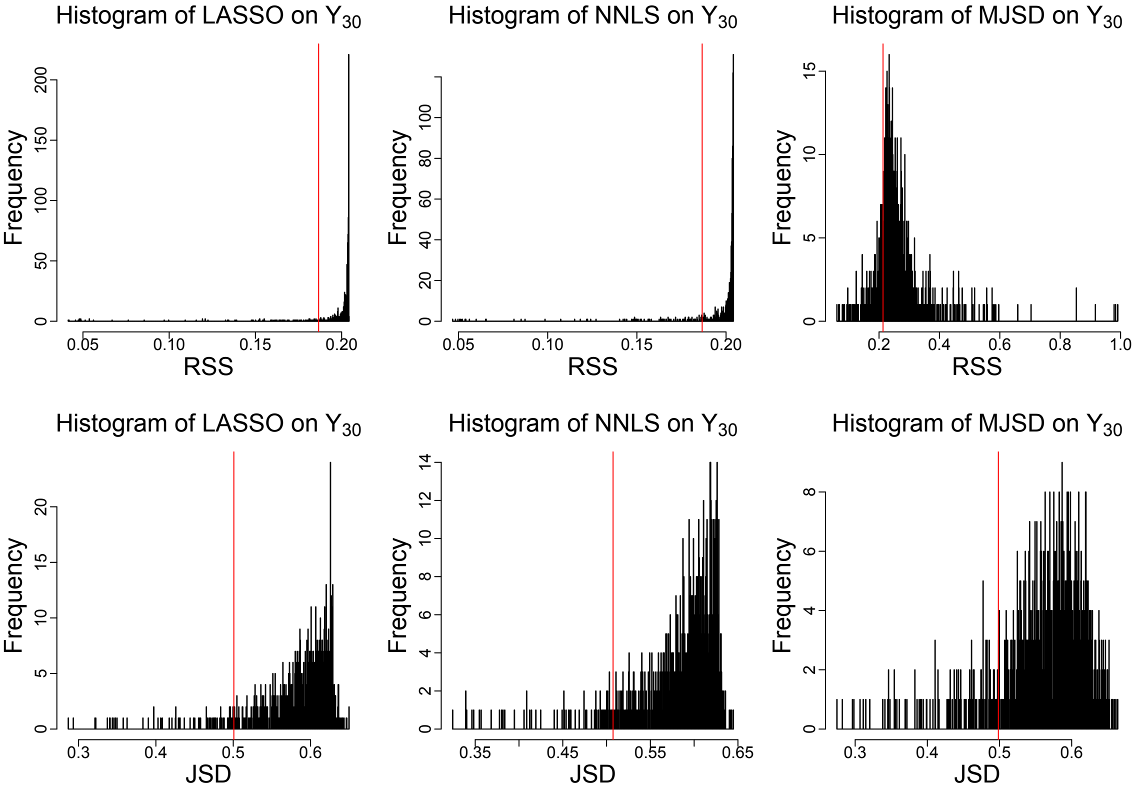


Figure 3.5: RSS and JSD permutation histogram of Y_{30} for three methods. The red vertical lines indicate the RSS and JSD values of the original matches.

The rank percentage is used as the nonparametric P-value for each test. For example, if the rank of the original JSD among all 1000 JSD values from the randomly permuted data is i, then the P-value is $\frac{i}{1000}$.

Usually when the P-value is smaller than 0.05, the match is judged as significantly better than random matches. But when conducting multiple comparisons, a more proper procedure in this situation is to control the FDR (false discovery rate) [1]. The BH (Benjamin-Hochberg) procedure is applied to control the FDR as follows:

1. Rank all the 100 P-values from smallest to largest as $P_{(1)}$, $P_{(2)}$,…, $P_{(100)}$;

2. For a given α (0.05), find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$. Here m=100 which is the number of independent tests.

3. Reject the null hypothesis for all $H_{(i)}$ for $i = 1, \ldots, k$.

Thus k stands for the number of the significant matches among all the 100 after controlling the FDR. The k values for the three methods under two criteria are shown in Table 3.1.

k	RSS	JSD
LASSO	56	66
NNLS	56	72
MJSD	89	95

Table 3.1: The number of the significant matches among all the 100 matches

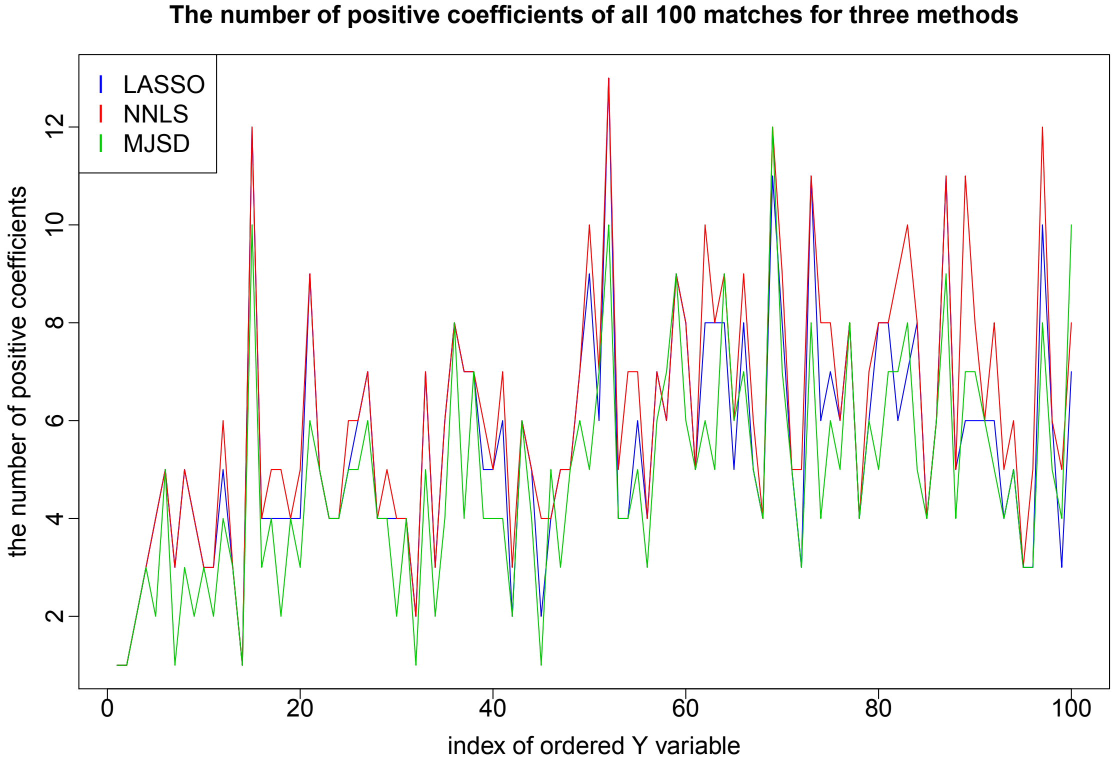


Figure 3.6: The number of positive coefficients of all 100 matches for three methods. Blue lines indicate the number of positive coefficients of all 100 matches for LASSO, red lines indicate that for NNLS and green lines indicate that for MJSD.

Table 3.1 shows that more significant matches are found by MJSD than LASSO and NNLS. This is because MJSD gives different estimates from that of LASSO and NNLS (although only slightly).

In spite of the permutation test, the sparsity is another important property to judge if a method is good in achieving the matching purpose. Figure 3.6 shows the number of positive coefficients of all 100 matches for the three methods. The green lines are obviously lower than the red and blue lines, indicating that that the most of matches from MJSD are more sparse than the matches from LASSO and NNLS. The indices of Y variables are ordered in the same sequence as that in Figure 3.7.

Based on the results of the MJSD method and the JSD criterion that 95 matches are judged significantly better than random matches, and the sparsity of MJSD is better than LASSO and NNLS, I will use MJSD and JSD to find the features of those best matches in the next section.

3.3 Heatmaps of the estimated coefficient matrix

In order to find some features of those best matched subnetworks based on the results from MJSD, I present the estimated coefficients in the heatmap. The estimated coefficients can be displayed in a form of 200×100 matrix since each of the 200 X variables has an estimated coefficient in matching each of the 100 Y variables. However, the heatmap of the estimated coefficient matrix in its original order is irregular and uninformative. Thus I order the labels of Y variables according to their quality of the matching so that the better matched variables ranked first, and the labels of X are also ordered so that lower ranks of the X variables are mostly corresponding to the lower ranks of the Y variables.

The indexes of Y and X variables are ordered as follows:

1. Initialization:

   (a) Order the 100 Y variables according to their JSD from the corresponding $\hat{Y}$ in an increasing order. Then $Y_1, Y_2, \ldots, Y_{100}$ are in a new sequence $Y_{[1]}, Y_{[2]}, \ldots, Y_{[100]}$.

   (b) Order 200 X variables according to the sequence of Y variables one by one. Select those X variables which have positive coefficients in matching

$Y_{[1]}$. Then order the selected X variables according to their estimated coefficients from largest to smallest.

2. At the ith step (i=2, . . . , 100):

   (a) Select from the remaining X variables which have positive coefficients in matching $Y_{[i]}$.

   (b) Order the selected X variables according to their coefficients from largest to smallest. Add the sequence of ordered X variables for $Y_{[i]}$ behind the sequence of the ordered X variables for $Y_{[i-1]}$.

The Heatmaps of the reordered estimated coefficient matrix of the MJSD method is shown in Figure 3.7. Because the heatmaps of the NNLS and LASSO are extremely similar to Figure 3.7, they are not shown.

The labels of 100 Y variables are ordered by JSD, so the best matched Y variables are on the left side. The blue background indicates that most estimated coefficients are zeros. The red lumps represent large coefficients.

The yellow lines indicate the matches that are not significantly better than random matches. All the insignificant matches are also the ones with large JSD values. From the heatmap it also can be found that the best matches (on the left side) are one to one matches.

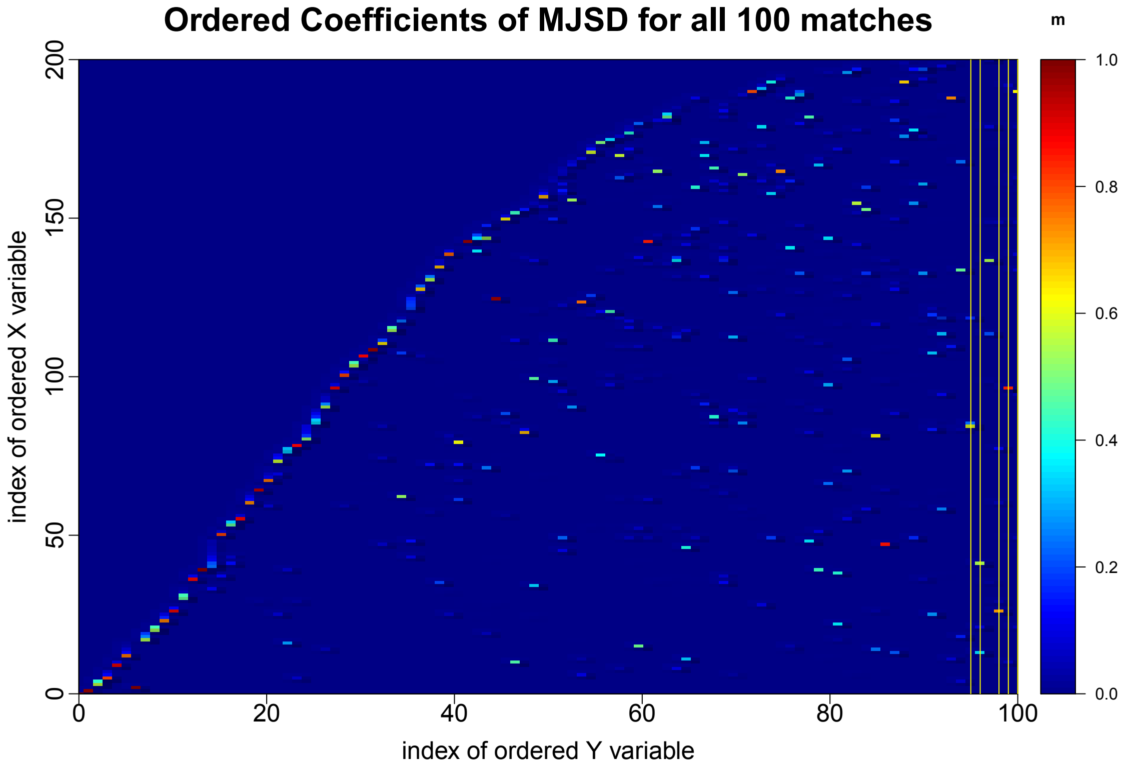


Figure 3.7: Heatmap of the ordered coefficients for the MJSD. The m is the color key, indicating the value and its corresponding color. Yellow lines indicate the matches that are not significantly better than random matches. The indexes of Y variables are ordered in the same sequence as that in Figure 3.7.

Chapter 4

Discussion and suggestions for future work

The BiomeNet analytical framework has many strengths, and chief among them is that its model-derived structures have a direct interpretation in terms of community metabolic function [16]. This makes BiomeNet an important addition to the exiting tools, such as PCA, that are widely used in metagenomics [22]. However, BiomeNet also has limitations. The number of community-level metabolic structures (the K metabosystems and L subnetworks within the model) must be fixed beforehand. The typical approach is to select very large values for K and L, and employ sparse and symmetric Dirichlet priors to encourage BiomeNet to concentrate the signal into relatively a few structures. However, there is no framework for judging which structures, if any, obtained in this way are better than random, and thereby warrant further biological interpretation. The purpose of my thesis was to address this issue.

To meet this challenge, I investigated three methods (LASSO, NNLS, and a new method called MJSD) to match the subnetworks obtained under one case for L as a linear combination of subnetworks derived from a different case. The basic idea is that those subnetworks that represent the real signal within a complex microbial community should make consistent contributions to the overall structure of the community. I then developed a permutation-based method to infer which structures (at the subnetwork level) are better than random. In this way, I have extended the value of the BiomeNet framework by providing biologists with the tools to infer (i) how much structure to include within the BiomeNet model for community-level metabolism (i.e., how to determine the proper choice for L for the data in hand), and (ii) which subnetworks are significant and thus warrant further biological attention.

All three methods identified the same subnetworks via the criterion that they can be matched across analyses as a linear combination. The LASSO and NNLS methods required model selection and thresholding, respectively, to satisfy the sparsity and non-negativity requirements. Alternatively, the MJSD method directly results in positive and sparse coefficients that sum to 1. Using JSD as criterion in MJSD, instead of RSS in LASSO or NNLS, improved the match of the larger weight subnetworks to some extent; however, it also tended to increase the deviation on matching the zero weight elements. Thus, the RSS of MJSD is larger than that of LASSO and NNLS for some matches.

Usually R^2 is used to estimate how well data fits a linear model. However R^2 is not a good criterion in our case because matching larger elements in a Y variable is our focus instead of the total variance in Y being explained by the regression. Measuring the similarity of two probability vectors Y and $\hat{Y}$ is our purpose. Thus, RSS and JSD were chosen as criterion instead of R^2. After multiple tests, it was shown that JSD is a more proper criterion than RSS to assess the similarity of two probability distributions, and MJSD is a more suitable method in this application than LASSO and NNLS.

The benefits to biologists are illustrated by my application of these methods to the 38 Mammalian gut metagenomes introduced in Chapter 1. The results for these data show that the predominant subnetworks (i.e., those that have the best matches) may not be much separated with larger L. The reason may be that a well-matched subnetwork consists of reaction groups that really do function together as a unit. Although I have not examined more than two values for L (100 and 200), my results suggest that the best value should be 100 or slightly less than 100 by the permutation test. Additional analyses of these, and other data, are required to determine if it is generally the case that predominant subnetworks are not much separated by larger L (or if datasets with less signal are more sensitive to the value of L). Regardless, Shafiei et al. [16] were largely unaware of this property of the mammal dataset, as they had no way to match the results they had obtained under different values of L.

Shafiei et al. [16] suggested that a single subnetwork (subnetwork 49 when L=100) was critical to the divergence between the gut communities found in carnivores and herbivores. It is noteworthy this subnetwork was the same one that I inferred as having the strongest signal within the data. However, Shafiei et al. [16] had no means of assessing if their results were any better than chance. This is important, because Shafiei et al. [16] uncovered a potentially informative metabolic signal within this subnetwork; it is prevalent in carnivores, and it is rich in reactions related to importation of extracellular saccharides (N-acetylmuramic acid, N-acetylglucosamine, fucose, glucose and mannose). This metabolic capacity gives carnivores greater ability exploit endogenous carbohydrate sources (the cell walls of the gut bacteria, whose outer membranes are composed of N-acetylmuramic acid and N-acetylglucosamine, and the fucosylated mucins secreted by the hosts' large intestine). My results are important because they revealed that the association of these reactions within a single subnetwork is highly significant. Furthermore, my results indicate that other subnetworks which differ between carnivores and herbivores (e.g., subnetwork 17 and subnetwork 72) also are significant, and warrant further biological investigation, even though they were not investigated by Shafiei et al. [16].

Although the above examples reveal that the well matched subnetworks are most possibly biologically meaningful reaction groups, further investigation is needed to demonstrate that these matched subnetworks are not correlated noise grouped together by two times of BiomeNet running. An important future work will be to develop the permutation test by permuting the input data directly and then run the BiomeNet on the permuted data. For example our input data include the E.C. abundance counts for 38 mammalian gut samples, which is in a form of $C \times N$ matrix, where C is the number of E.C. abundance counts and N is the number of samples. By randomly permuting the elements of each column, the biological connections between different reactions will be lost. The obtained matrix includes mainly random noises. Running BiomeNet on such permuted data twice and comparing the resulted subnetworks by the methods proposed in this thesis, we expect to find very few or no significantly matched subnetworks. Such results will demonstrate that the matched subnetworks from real data are biologically meaningful to some extend.

There are several other areas in which further work is warranted. The methods described in this thesis should be extended to the level of metabosystems within BiomeNet. Second, further analysis of other real datasets is required to determine if additional improvement of the methods might be needed. Third, there is a critical need to be able to build predictive models based on the metabolic structures (subnetworks and metabosystem) uncovered by BiomeNet. Recall the example of IBD that was introduced in Chapter 1; tools are needed to infer which metabolic structures might be predictive a particular IBD disease phenotype, or which might be predictive of response to certain IBD therapies. One possibility would be to use the methods described in this thesis to select significant subsystems as features to be input into classic methods for supervised training of predictive models [22]. Fourth, the methods can be extended to the problem of matching the structures obtained from a similar modeling framework called BioMiCo [15] designed for the taxonomic components (rather than the metabolic components) of microbial community structure.

The methods developed in this thesis can be either directly applied or extended in many more general matching problems. For example, matching the subnetworks derived from two or more different data sets can help to reveal the common metabolic structures for different groups of samples; matching the components that are derived on similar subjects but from different labs or locations can help to remove the unwanted variance and pool the resource for the same study. With the accumulation of large microbial or metagenomics data from different resources, the ability to pool data together for more informative inference is fundamentally important. In addition, the methods can also be extended for matching other types of vectors than sparse multinomial probabilities. Since PCA is still widely applied on all kinds of high dimensional data reduction, one possibility is to apply the developed methods in this thesis to match the PCA components. In different applications, which of the three methods can lead to a better result will depend on the nature of the problem, it is possible that LASSO or NNLS can work better than MJSD when matching PCA components.

Bibliography

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[3] Robby Haelterman. *Analytical study of the least squares quasi-Newton method for interaction problems*. PhD thesis, Ghent University, 2009.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[5] Dazhi Jiao, Yuzhen Ye, and Haixu Tang. Probabilistic inference of biochemical reactions in microbial communities from metagenomic sequences. *PLoS Comput Biol*, 9(3):e1002981, 2013.

[6] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, José C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, 2012.

[7] Jianhua Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[8] Brian D Muegge, Justin Kuczynski, Dan Knights, Jose C Clemente, Antonio González, Luigi Fontana, Bernard Henrissat, Rob Knight, and Jeffrey I Gordon. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, 332(6032):970–974, 2011.

[9] Frank Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[10] John Penders, Ellen E Stobberingh, Piet A van den Brandt, and Carel Thijs. The role of the intestinal microbiota in the development of atopic disorders. *Allergy*, 62(11):1223–1236, 2007.

[11] Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

[12] June L Round and Sarkis K Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323, 2009.

[13] Victor S Ryaben'kii and Semyon V Tsynkov. *A theoretical introduction to numerical analysis*. CRC Press, 2006.

[14] Nicola Segata, Daniela Boernigen, Timothy L Tickle, Xochitl C Morgan, Wendy S Garrett, and Curtis Huttenhower. Computational meta'omics for microbial community studies. *Molecular systems biology*, 9(1):666, 2013.

[15] Mahdi Shafiei, Katherine A Dunn, Eva Boon, Shelley M MacDonald, David A Walsh, Hong Gu, and Joseph P Bielawski. Biomico: a supervised bayesian model for inference of microbial community structure. *Microbiome*, 3(1):1, 2015.

[16] Mahdi Shafiei, Katherine A Dunn, Hugh Chipman, Hong Gu, and Joseph P Bielawski. Biomenet: A bayesian model for inference of metabolic divergence among microbial communities. *PLoS Comput Biol*, 10(11):e1003918, 2014.

[17] Christian Sigg, Bernd Fischer, Björn Ommer, Volker Roth, and Joachim Buhmann. Nonnegative cca for audiovisual source separation. In *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, pages 253–258. IEEE, 2007.

[18] Martin Slawski, Matthias Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

[19] Ryan J Tibshirani, Jonathan Taylor, et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

[20] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, 444(7122):1027–131, 2006.

[21] Hege Vestheim and Simon N Jarman. Blocking primers to enhance pcr amplification of rare sequences in mixed samples–a case study on prey dna in antarctic krill stomachs. *Frontiers in Zoology*, 5(1):1, 2008.

[22] Shaoguang Wu, Ki-Jong Rhee, Emilia Albesiano, Shervin Rabizadeh, Xinqun Wu, Hung-Rong Yen, David L Huso, Frederick L Brancati, Elizabeth Wick, Florencia McAllister, et al. A human colonic commensal promotes colon tumorigenesis via activation of t helper type 17 t cell responses. *Nature medicine*, 15(9):1016–1022, 2009.