Computers in Biology and Medicine 114 (2019) 103417

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

An Integrated Gaussian Graphical Model to evaluate the impact of exposures on metabolic networks

Jai Woo Lee [a], Erika L. Moen [b], Tracy Punshon [c], Anne G. Hoen [b,d], Delisha Stewart [e], Hongzhe Li [f], Margaret R. Karagas [d], Jiang Gui [b,*]

[a] Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, USA
[b] Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH, USA
[c] Department of Biological Sciences, Dartmouth College, Hanover, NH, USA
[d] Department of Epidemiology, Geisel School of Medicine, Lebanon, NH, USA
[e] Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[f] Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadephia, PA, USA

ARTICLE INFO

Keywords:
Data integration
Gaussian graphical model
Lasso
Trace element exposures
Metabolic network

ABSTRACT

Examining the effects of exogenous exposures on complex metabolic processes poses the unique challenge of identifying interactions among a large number of metabolites. Recent progress in the quantification of the metabolome through mass spectrometry (MS) and nuclear magnetic resonance (NMR) has given rise to high-dimensional biomedical data of specific metabolites that can be leveraged to study their effects in humans. These metabolic interactions can be evaluated using probabilistic graphical models (PGMs), which define conditional dependence and independence between components within and between heterogeneous biomedical datasets. This method allows for the detection and recovery of valuable but latent information that cannot be easily detected by other currently existing methods. Here, we develop a PGM method, referred to as an "Integrated Gaussian Graphical Model (IGGM)", to incorporate exposure concentrations of seven trace elements—arsenic (As), lead (Pb), mercury (Hg), cadmium (Cd), zinc (Zn), selenium (Se) and copper (Cu—into metabolic networks. We first conducted a simulation study demonstrating that the integration of trace elements into metabolomics data can improve the accuracy of detecting latent interactions of metabolites impacted by exposure in the network. We tested parameters such as sample size and the number of neighboring metabolites of a chosen trace element for their impact on the accuracy of detecting metabolite interactions. We then applied this method to measurements of cord blood plasma metabolites and placental trace elements collected from newborns in the New Hampshire Birth Cohort Study (NHBCS). We found that our approach can identify latent interactions among metabolites that are related to trace element concentrations. Application to similarly structured data may contribute to our understanding of the complex interplay between exposure-related metabolic interactions that are important for human health.

1. Introduction

Metabolomics is the study of all low molecular weight molecules present in biological fluids and tissues and may be the most promising of the "omics" technologies used in exposome research. Applying metabolic profiling to the examination of normal or complicated pregnancies has emerged as an innovative unsupervised approach for exploring potential biomarkers and biological mechanisms of reproductive outcomes. Pregnancy is a dynamic period consisting of a series of minute physiologic fetal adjustments over time that affect the metabolism of nutrients in an effort to facilitate fetal development [1]. Human pregnancy and development are also susceptible to the toxic effect of metals, which may stunt infant growth and cause preterm delivery [2]. Therefore, it is critical to include environmental exposures such as essential nutrients or potentially toxic heavy metals to the study of the metabolic interaction network.

Previous work has focused on developing statistical methods in Graphical Models integrating elements from heterogeneous datasets. Recently, a joint Gaussian Graphical Model (jGGM) was applied to analyze significantly interacting genes in genomics data of common

* Corresponding author.
E-mail address: jiang.gui@dartmouth.edu (J. Gui).

https://doi.org/10.1016/j.compbiomed.2019.103417
Received 3 May 2019; Received in revised form 25 August 2019; Accepted 26 August 2019
Available online 31 August 2019

features over studies of independent samples [3]. While this method adjusts penalization in joint GGM using *a priori* pathway information validated in the existing biology literature, it does not consider impacts from external data. Additionally, while the Ising model considers the impact of subject-specific external variables, its applications are limited to multivariate binary genomic data [4]. As such, there is a need for a GGM that can handle continuous omics data beyond genomics and consider the impact of external variables.

In this paper, we present a method, an Integrated Gaussian Graphical Model (IGGM), that integrates metabolomics and trace element data to infer a metabolic network outside the realm of genomics. Our proposed method will allow us to conduct an integrative analysis of how trace elements affect metabolites and how metabolites interact with each other. We first use a simulation to demonstrate that this integrated approach is more powerful in estimating latent interactions of metabolites impacted by exposure variables than GGM, which estimates the network based only on metabolomics data. We then examine the optimal set of parameters, such as sample size and the number of strongly correlated neighbor metabolites of each trace element. Finally, we assess the estimated metabolic pathway consisting of the most statistically significant metabolic interactions detected by our proposed method and discuss these newly detected interactions of metabolites in the context of known associations in the literature.

2. Methods

2.1. Gaussian Graphical Model (GGM) and least absolute shrinkage and selection operator (Lasso) implementation

We start with GGM. We assume that the data are randomly sampled observational data from a multivariate normal distribution. Specifically, let X be a random normal p-dimensional vector and $X_1, X_2, \ldots X_p$ denote the p features. Let $X^{(k)}$ be the vector of feature values for the kth sample. We assume that $X \tilde{} N_p\ (0, \Sigma)$ and Σ represent a positive definite covariance matrix. Let $\Omega = w_{ij}$ be the precision matrix, which is defined as the inverse of the covariance matrix Σ. Let S be the empirical covariance matrix, and the penalized log-likelihood function can be written as

$$l = log(det(\Omega)) - tr(S\Omega) - \lambda\|\Omega\|_1 \quad (1)$$

here, tr denotes the trace, and $\|\Omega\|_1$ is the sum of absolute values of all elements in Ω. We then implement lasso to fit this model with linear regression with feature shrinkage and selection based on the validation of models defining interactions between two features in the data. $x_1, x_2, \ldots, x_N$ denote the vectors of values for each feature assigned in all samples. V = {1,2, …, p-1, p} is the set of nodes. To define the interactions among the variables, the conditional dependency in Gaussian Graphical Models can be implemented by penalized regression [5]. The penalized log-likelihood is defined as follows, and we aim to find the β coefficient that minimizes the formula below:

$$\left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right] + \lambda\sum_{j=1}^{p}|\beta_j| \quad (2)$$

β indicates the estimated coefficient, and y indicates a column vector in the data matrix. We utilize the lasso penalty in the form of an absolute value of β for a sparse network estimation.

2.2. Integrated Gaussian Graphical Model (IGGM) and lasso implementation

GGM is expanded to incorporate a set of external variables in the network. In IGGM, X represents a random continuous $(p + k)$-dimensional matrix in which $X_1, X_2, \ldots X_p$ denotes the p features but $X_{p+1}, X_{p+2}, \ldots X_{p+k}$ denotes the k external variables. With additional k external variables, the set of nodes is {1, 2, …, p, p+1. p+2, …p + k}. To define the interactions among the variables, IGGM, similar to GGM, implements penalized regression for continuous data by minimizing the negative log-likelihood with penalty to find the β coefficient. We apply the linear regression formula [5–7]:

$$\left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right] + \lambda\sum_{j=1}^{p+k}f_j|\beta_j| \quad (3)$$

Each column of data X becomes y in each step. f_j denotes the penalty factor for the j^{th} element, with f_j equaling 0 for p+1 ≤ j ≤ p + k to not penalize external variables, and f_j equaling 1 for 1 ≤ j ≤ p to penalize features. X indicates a data matrix consisting of features and external variables. This process is illustrated in the diagram (Fig. 1).

As shown in Fig. 1, two data representing features and external variables are merged. Each column of X is chosen as y. The newly created N-by-(p-1+k) matrix X and y are plugged into the penalized regression; vector y is estimated using the combination of (p-1+k) columns in X. This process continues for all features, that is, p times in total. For all p iterations, features are penalized, and external variables are nonpenalized. After implementing all of these iterations, we obtain a resulting p-by-p matrix representing interactions of the p features. The features that are regulated by the external variable form a densely connected subnetwork (Fig. 2).

For example, as seen in Fig. 2, external variables named T1, T2 and T3 are strongly related to five features named M1-5, M6-10 or M11-15, respectively. After we apply IGGM, we observe that the subnetworks of the five features strongly relate to each external variable while all other features maintain the same number of interactions or noninteractions. In this way, two different types of data can be integrated. The matrix shown on the right represents interactions of features in the network after applying IGGM. To validate the effectiveness of the proposed model, we use the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to determine the tuning parameters that determine the optimized model. After obtaining the optimized model, we compute the true positive rate (TPR), which predicts true subnetworks of the data adjusted by external variables, for example, a subnetwork consisting of M1-5, which includes ten interactions, M1-M2, M1-M3, M1-M4, M1-M5, M2-M3, M2-M4, M2-M5, M3-M4, M3-M5, and M4-M5, and the true negative rate (TNR), which predicts nonselected edges in subnetworks.

2.3. Alternative methods: Gaussian Graphical Model with penalized external variable (GGM-PE), Ising models, and integrated Ising models (I-Ising)

We explore GGM-PE as an alternative GGM approach that estimates an interaction matrix of features in the combined feature and external variable matrices with penalties on both features and external variables. We also consider binary Ising models that evaluate the same variables and parameters, including p features, k external variables, and the penalty factor, f, as described in Sections 2.1 and 2.2. To this end, we generate a median-cutoff dichotomized version of an N-by-p data matrix X as U, and a response vector Y of length N as V. Values in the data are dichotomized into two values, 0 or 1. The set of values in the data are represented as R = {1,0}, and in the response vector, V, any data values greater than the median are defined as 1; $v_i = I(r_i{=}1)$. We apply the logistic regression formula [5–7]:

$$Pr(R = 1|U = u) = \frac{e^{(\beta_0 + \beta^T u)}}{1 + e^{(\beta_0 + \beta^T u)}} \quad (4)$$

By using a log-odds transformation, the above model is modified as follows:

$$Log\left(\frac{Pr(R = 1|U = u)}{Pr(R = 0|U = u)}\right) = \beta_0 + \beta^T u \quad (5)$$

We aim to minimize the penalized logistic regression function using

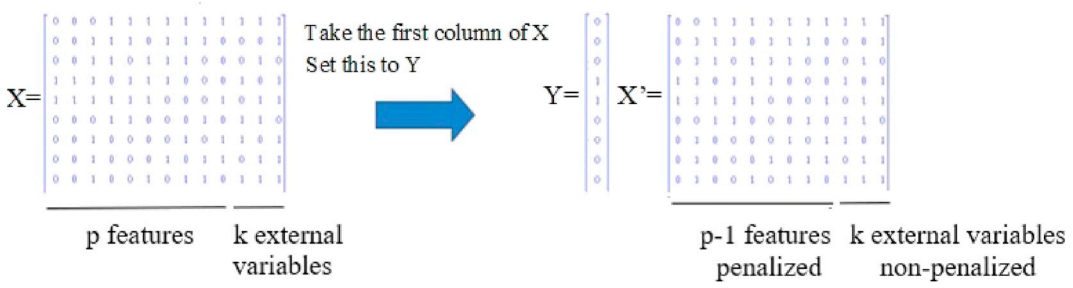

Fig. 1. Main process in integrated Gaussian Graphical Models.

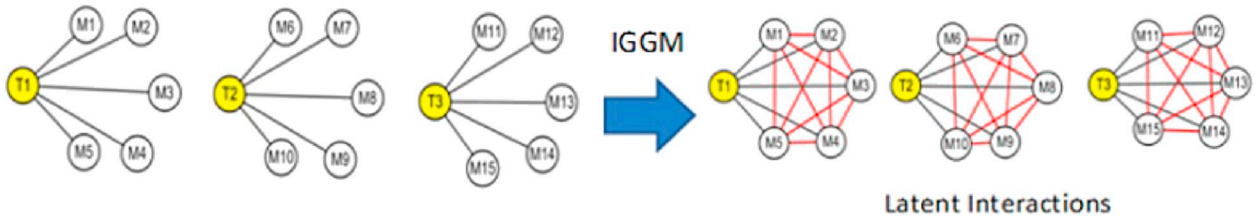

Fig. 2. How integrated Gaussian Graphical Models find latent interactions of features affected by external variables. A yellow T node indicates an external variable, and a white M node indicates a feature, each representing a trace element and a metabolite in the real data application. Black edges represent strong correlations between an external variable and its neighbors, and red edges represent latent interactions after considering the impact of an external variable on the network of features. In this example, three sets of five features impacted by each external variable do not overlap.. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

negative binomial log-likelihood. Our goal is to find coefficients that minimize the function, as shown below.

$$\left[\frac{1}{N}\sum_{i=1}^{N}\left(v_i*(\beta_0+u_i^T\beta)-log\left(1+e^{(\beta_0+u_i^T\beta)}\right)\right)\right]+\lambda\sum_{j=1}^{p+k}|\beta_j| \quad (6)$$

Extending this model to an I-Ising model considering external variables, we again minimize the function to find the optimal β coefficients.

$$\left[\frac{1}{N}\sum_{i=1}^{N}\left(v_i*(\beta_0+u_i^T\beta)-log\left(1+e^{(\beta_0+u_i^T\beta)}\right)\right)\right]+\lambda\sum_{j=1}^{p+k}f_j|\beta_j| \quad (7)$$

We, therefore, have five penalized regression functions corresponding to GGM, GGM-PE, IGGM, Ising model, and I-Ising.

2.4. Datasets and additional preprocessing and validation steps for real data application

The New Hampshire Birth Cohort Study (NHBCS) assembled data on placental elements, including metals, metalloids and nutrient elements, along with metabolomic data from newborn cord blood. This presents an opportunity to apply network methods to analyze complex interactions of metabolites and their association with placental trace elements. We applied our simulated method to analyze interactions of cord blood metabolites for 381 women enrolled in NHBCS. Placental bulk elemental concentrations include potentially toxic heavy metals, such as arsenic (As), lead (Pb), mercury (Hg), and cadmium (Cd), and essential nutrient elements, such as zinc (Zn), selenium (Se), and copper (Cu). The NHBCS enrolled pregnant women, 18–45 years old, receiving prenatal care in one of the study clinics, used a private well that served ¡15 households or 25 individuals and were not planning to move prior to delivery. All samples were analyzed by inductively coupled plasma mass spectrometry (7700x Agilent, Santa Clara, CA), with analysis following the quality control procedures described in EPA 6020a [8]. Cord blood metabolomics data originally consisting of 188 metabolites were acquired by quantitative/semiquantitative targeted LC-MS/MS on the Biocrates 'p180' platform [9–11]. In the real data application, cord blood metabolomics data include 220 measured metabolite features.

Both trace elements and metabolomics data were imputed with the classification and regression tree method in the "mice" R package and log-transformed. After the trace element and metabolomics data were preprocessed, IGGM was applied to these data with 100 bootstrap replications. Edges in the network are considered to be significant when they are chosen more than 90 times out of 100 bootstrap replications.

2.5. Validation study for simulation

To validate the effectiveness and efficiency of IGGM, we tested whether IGGM correctly identified interactions of features impacted by external variables and consistently detected original interactions of features in the data. First, we needed to prove that IGGM detected additional statistically significant interactions of features highly correlated with each external variable compared to other Graphical Model methods. Second, we needed to demonstrate that IGGM successfully detected a set of original feature interactions with similar or better accuracy compared to GGM, which does not consider the impact of external variables on features. We compared the accuracy of the estimated network using the following five models: IGGM, GGM-PE, GGM, Ising Model and I-Ising Model. We assessed the GGM-PE method that implements GGM by penalizing both features and external variables. The Ising model dichotomizes the continuous data and applies the logistic regression to estimate the neighbors of each feature. The I-Ising model implements a similar process as the binary Ising model but also considers the impact of external variables.

We simulated models on features and external variables with different sample sizes and numbers of features strongly correlated with each external variable. We used {p, d, c, and n} to denote the parameters of our simulation, where p is the number of features, d is the number of features that are dependent on the external variables, c is the number of external variables, and n is the sample size. We treated the features as continuous variables in GGM. Based on the four main parameters described above, d neighbors of c external variables have strength of interactions ranging from 0.8 to 1 with their external variable and strength of interactions of d neighbors ranging from −1 to −0.8. The correlation matrix based on this precision matrix represented

interactions of features and external variables with correlations ranging from 0.45 to 0.55 for strong correlations and 0.01 to 0.10 for weak correlations and original interactions of features with two cliques of size 5 and 10 plus two hubs with 5 and 10 neighbors. We plugged this correlation matrix into the mvrnorm function in the R package to generate an *n*-by-*p* continuous matrix, giving samples under N(0, Σ). We used a median cutoff to transform features of this continuous matrix data to binary matrix data and used the Ising model to estimate the binary network.

This simulation measured two different dependencies between features and external variables: dependency of features impacted by their external variables and the original interactions of features. Regarding the dependency of features driven by external variables, we hypothesized that the features that are dependent on an external variable should form a clique (complete subnetwork) after we accommodate the external variable in the network analysis. The accuracy of detecting external variable-driven interactions (latent interactions) of features impacted by external variables was measured as follows:

$$(TPR1 + TNR1) / 2*100\% \quad (8)$$

This accuracy was measured among external variables and features strongly correlated with external variables. TPR1 describes a true positive rate indicating correctly detected dependencies of all possible interactions in the cliques of features strongly affected by external variables. We assumed that *d* features are "highly correlated" with one external variable. In this case, if the method correctly identified a complete subnetwork of d features defined as $\binom{d}{2}$ edges, 100% accuracy of detection was achieved. TNR1 describes a true negative rate indicating correctly detected noninteractions between features impacted by weakly correlated external variables. To measure the accuracy of detecting noninteractions of features highly correlated with any external variables, we assumed an interaction matrix of (*c*d*) features highly correlated with any external variables. If the method correctly identified all noninteractions of features related to weakly correlated external variables defined as $\binom{c*d}{2}$ - $\binom{d}{2}$ edges, 100% accuracy of detection was obtained.

Regarding the original interactions of features, we measured the accuracy by the average of TPR2, true positive rate indicating the correctly identified interactions of features, and TNR2, true negative rate indicating the correctly identified noninteractions of features, as shown below.

$$(TPR2 + TNR2) / 2*100\% \quad (9)$$

From our simulations, we hypothesized that IGGM would outperform the other four methods in detecting latent interactions of features driven by external variables. We further hypothesized that IGGM would produce a similar set of original interactions of features as GGM and outperform the binary Ising model and the I-Ising model using median-cutoff dichotomization.

3. Results

3.1. Comparisons of methods over different parameters and conditions for simulated data

We evaluated the performance of the proposed IGGM, GGM-PE, GGM, Integrated Ising and Ising methods in edge recovery with simulation settings varying by sample size and the number of neighbors of external variables, assuming strong correlations between each external variable and its selected features. We first considered the level of accuracy of recovering edges considering different sample sizes and the number of neighbors of external variables. We then evaluated how the different numbers of neighbors of external variables impact the correlation strength between external variables and their neighbors.

a) The numbers of samples from 110 to 260 with 100 features, 3 external variables with the number of neighbors of each external variable ranging from 4 to 6

Increasing the sample size had a modest positive impact on model performance. The accuracy in detecting latent and existing interactions of features for all methods increased by 3–5% points as the sample size increased from 110 to 260. In this simulation, GGM methods outperformed Ising methods in detecting existing and latent interactions of features. Among GGM methods, IGGM better detected a higher proportion of original interactions than did GGM. IGGM significantly outperformed GGM in detecting latent interactions of features when an external variable was strongly correlated with more than four features. While IGGM and GGM-PE performed similarly in detecting original interactions of features, IGGM outperformed GGM-PE in detecting latent interactions of features driven by external variables. In computing the true negative rate, we found that all models achieved an almost perfect true negative rate in detecting latent interactions, demonstrating that all methods except IGGM did not efficiently estimate the majority of true latent interactions in each simulation. For all five graphical models, a larger sample size was associated with higher accuracy in detecting original interactions of features, which indicated that a larger sample size helps to recover the edges of both latent and existing interactions of features given these three conditions: the fixed number of neighbors of an external variable, the fixed number of features, and strong correlations between features and their neighbor external variable. This empirical result supports the theoretical findings of Ravikumar et al., implying that a large sample size enables GGM models to recover a higher number of existing and latent interactions of features [12].

Compared to the other four approaches, the computational approach taken by IGGM was more consistent and accurate over different simulation settings, specifically regarding the number of neighbors of an external variable. IGGM consistently detects latent interactions of features perfectly, as shown in Table 1. The changes in the number of neighbors of the external variable affect results only in the case of GGM-PE. The detection accuracy of latent interactions produced by GGM-PE drastically decreased as the number of neighbors of the external variable increased from 4 to 6, which implies that given a fixed sample size, a higher number of neighbors of external variables was associated with lower accuracy in detecting latent interactions by GGM-PE. GGM, I-Ising and Ising consistently failed to identify latent interactions of features.

b) The different numbers of neighbors of external variables and varied correlations between external variables and their neighbors with IGGM

The results provided in Fig. 3 were generated by IGGM in the simulation setting of 100 features, 200 samples, and 3 external variables with varied numbers of neighbors ranging from 4 to 6. In the left-most matrix in Fig. 3, each group of 6 features strongly and commonly correlated with each of the three external variables forms a complete subnetwork with weak interactions. In the middle matrix, three groups of features strongly correlated with the common external variable form three different complete subnetworks with moderately strong correlations. In the right-most matrix, each group of four features strongly correlated with each external variable form complete subnetworks with strong correlations.

As shown in the simulations with the different numbers of neighbors, increasing the number of neighbors decreased the strength of estimated latent interactions of features. The correlations between external variables and their selected neighbors were uniformly strong across simulation settings. We observed that the estimates of latent interactions with an external variable weakened with a greater number of neighbors. Thus, the IGGM provided more interpretable results than GGM-PE in

Table 1
Detection Accuracy Results by different methods, sample size, and the number of neighbors of external variables.

sample size	The number of neighbors of an external variable	6		5		4	
	Methods	Interaction Types					
		Original feature interactions	External variable –driven (latent) feature interactions	Original feature interactions	External variable –driven (latent) feature interactions	Original feature interactions	External variable driven –(latent) feature interactions
110	IGGM	81.9%	100.0%	81.7%	100.0%	81.7%	100.0%
	GGM-PE	80.6%	61.1%	80.6%	76.7%	80.6%	100.0%
	GGM	76.4%	50.0%	76.4%	50.0%	77.8%	50.0%
	I-Ising	66.6%	50.0%	66.8%	50.0%	69.3%	50.0%
	Ising	67.3%	50.0%	68.0%	49.3%	68.0%	50.0%
140	IGGM	82.4%	100.0%	82.2%	100.0%	83.0%	100.0%
	GGM-PE	80.9%	64.4%	82.5%	83.3%	81.7%	100.0%
	GGM	74.9%	50.0%	76.5%	51.7%	77.2%	52.8%
	I-Ising	73.9%	50.0%	68.1%	50.0%	69.6%	50.0%
	Ising	74.6%	49.5%	68.0%	49.7%	70.0%	50.0%
170	IGGM	82.3%	100.0%	82.2%	100.0%	81.9%	100.0%
	GGM-PE	81.7%	62.8%	81.7%	80.0%	81.7%	100.0%
	GGM	80.3%	50.0%	81.8%	53.3%	81.8%	50.0%
	I-Ising	71.4%	49.8%	71.7%	48.7%	72.9%	55.6%
	Ising	72.8%	49.5%	72.2%	48.7%	73.7%	50.0%
200	IGGM	81.7%	100.0%	81.4%	100.0%	81.4%	100.0%
	GGM-PE	81.8%	63.3%	81.8%	83.3%	81.8%	100.0%
	GGM	76.0%	51.1%	76.0%	50.0%	78.2%	58.3%
	I-Ising	67.0%	52.0%	68.7%	50.0%	67.2%	50.0%
	Ising	66.5%	50.0%	69.3%	51.0%	68.6%	49.0%
230	IGGM	83.5%	100.0%	82.5%	100.0%	82.5%	100.0%
	GGM-PE	82.8%	56.7%	82.8%	79.2%	82.8%	100.0%
	GGM	79.4%	61.1%	79.4%	58.3%	78.7%	50.0%
	I-Ising	66.8%	50.0%	70.6%	52.7%	67.7%	50.0%
	Ising	67.0%	48.1%	71.5%	50.0%	69.0%	49.5%
260	IGGM	85.8%	99.5%	84.6%	100.0%	83.9%	99.5%
	GGM-PE	84.0%	60.0%	83.5%	80.0%	85.2%	99.0%
	GGM	81.8%	61.1%	82.1%	56.7%	82.1%	55.6%
	I-Ising	69.1%	50.0%	70.8%	48.7%	71.3%	54.2%
	Ising	68.9%	48.6%	72.0%	47.3%	70.7%	49.0%

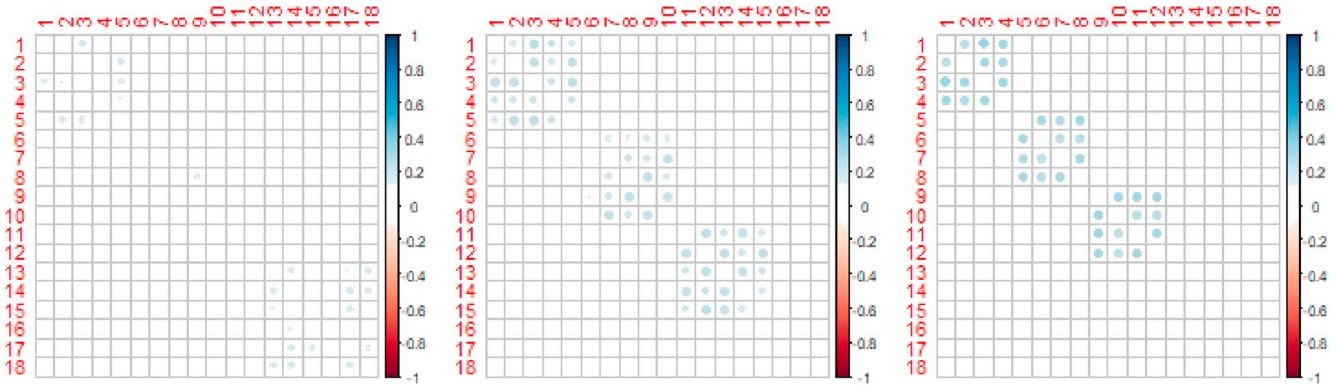

Fig. 3. IGGM estimation of latent interactions by different numbers of neighbors of external variables. Each matrix neighbor represents latent interactions of 18 features with 3 external variables that have 6 neighbors (left-most), 15 features with 3 external variables that have 5 neighbors (middle), and 12 features of 3 external variables that have 4 neighbors (right-most).

that GGM-PE could not detect latent interactions with a higher number of neighbors while IGGM did.

3.2. Application to placental trace element and cord blood plasma metabolomics data in NHBCS

In the real data application, we integrated one trace element with 220 metabolites. We hypothesized that the correlations between a trace element (external variable) and a metabolite (feature) would be low in our real data application and that each trace element would correlate with a sparse subnetwork of metabolites. We identified latent and existing interactions of metabolites affected by potentially toxic heavy metal elements and essential nutrient trace elements. The latent interaction is an underlying biological interaction of metabolites that can be detected only when the impact of the trace element on specific metabolites is considered in the model, and existing interaction of metabolites can be detected with or without considering the impact of the trace element on metabolites. We considered a metabolic interaction as existing if those interactions were detected by GGM, and a metabolic interaction was latent if those edges were not detected by GGM but detected by IGGM. Perfectly distinguishing latent interactions of metabolites from existing interactions of metabolites was not possible

because in each estimation of interactions, detected latent and existing interactions could be mutually inclusive. As shown in Fig. 4 and Fig. 5, an edge between a metabolite and a metabolite unrelated to a trace element could indicate, for example, that they share substrates or enzymes, and an edge between a metabolite and a metabolite associated with a trace element suggests existing but undetected regulation or signaling pathways of metabolites by the same trace element. Our model focuses on the latter whereby we seek to identify latent interactions of metabolites that can be detected only by integrating a specific trace element into our model. It was also assumed that both existing and latent interactions are biologically meaningful. All latent interactions detected by IGGM with the impact of trace elements are listed in Table S1 in Appendix A.

In our real data, the most commonly detected latent interaction was between proline (Pro) and phenylalanine (Phe), which in the metabolic network is driven by As, Pb, Hg, Zn or Se. Another latent interaction between sphingolipids sm_C18:1 and lysophosphatidylcholine lysoPC_a_C20:4 was found in the metabolic network driven by As, Se, and Cu. Latent interactions detected in the metabolic network driven by As, Se and Cu result in subnetworks of more than two metabolites. The metabolic network driven by As contains a subnetwork of three metabolites, sm_C18:1, lysoPC_a_C20:4, and PC_aa_C36:0, where both sm_C18:1 and PC_aa_C36:0 were connected to lysoPC_a_C20:4. The metabolic network driven by Se contained a subnetwork of four metabolites consisting of phosphatidylcholine (2x O-acyl) PC_aa_C36:0, lysophosphatidylcholine lysoPC_a_C20:4, sphingolipids sm_C18:1, and phosphatidylcholine (2x O-acyl) PC_aa_C38:4. The metabolic network driven by Cu contained two independent subnetworks, one consisting of sphingolipid sm_C24:1 and phosphatidylcholine (1x O-acyl, 1x O-alkyl) PC_ae_C40:1 and the other consisting of phosphatidylcholine (2x O-acyl) PC_aa_C36:0, lysophosphatidylcholine lysoPC_a_C20:4, and sphingolipids sm_C18:1. The metabolic network driven by Cd contained a latent interaction between lysophosphatidylcholine lysoPC_a_C18:2 and an amino acid, histidine (His). The metabolic network driven by Cu contained a latent interaction between an amino acid, lysine (Lys) and α-aminoadipic acid (α-AAA). The names of metabolites are listed in Table S2 in Appendix A.

We next assessed evidence in the literature supporting our network estimation results. In Hg-driven and As-driven metabolic networks, the toxic effect of Hg or As on changes in proline concentration was found in a plant, Spinacia oleracea L, in the process of adaptation against Hg or As stress [13,14]. Many peptides include the NPF motif consisting of asparagine, proline and phenylalanine [15]. Thus, the structures of Hg-proline-phenylalanine or As-proline-phenylalanine are plausible. However, for Zn-driven and Se-driven metabolic networks, we found evidence that Zn increases the accumulation of proline and that the concentration of proline increased with high selenate treatment [16]. The Asn-Pro-Phe (NPF) motif consisting of asparagine, proline, and phenylalanine supported our findings on the pathways involving Zn-proline-phenylalanine and Se-proline-phenylalanine. With respect to the Pb metabolomic network, Pb has been positively associated with an increase in serum creatinine concentrations [17], and an increase in symmetric dimethylarginine (SDMA) was found to affect changes in creatinine concentrations [18]. For the Cu-driven metabolic network, a correlation between the concentrations of dietary Cu from the copper

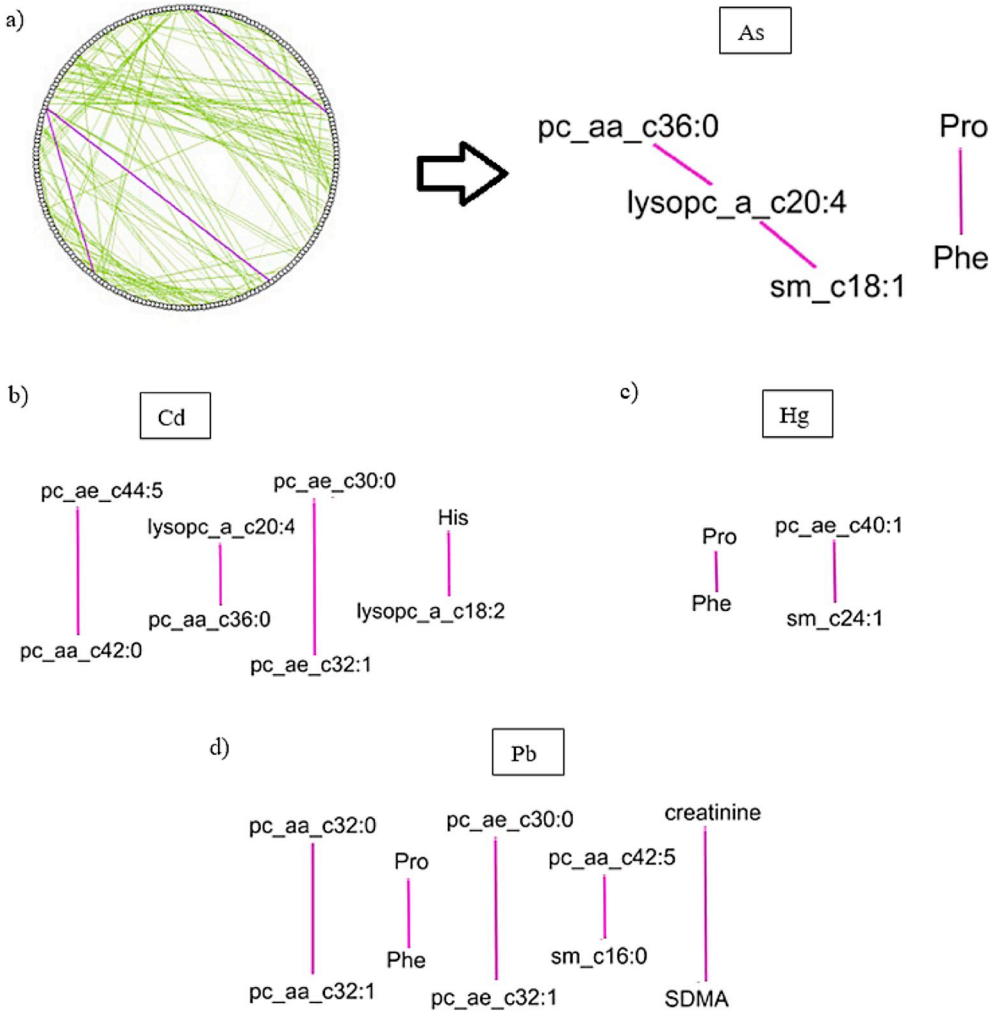

Fig. 4. Network visualizations of metabolite interactions for potentially toxic heavy metals; a) As, b) Cd, c) Hg and d) Pb. In the interaction network of metabolites (upper-left), yellow-green edges indicate interactions of metabolites detected by both GGM and IGGM, and purple edges indicate interactions metabolites impacted by trace elements detected by only IGGM. Latent interactions of metabolites impacted by each of As, Cd, Hg and Pb are represented by purple edges. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

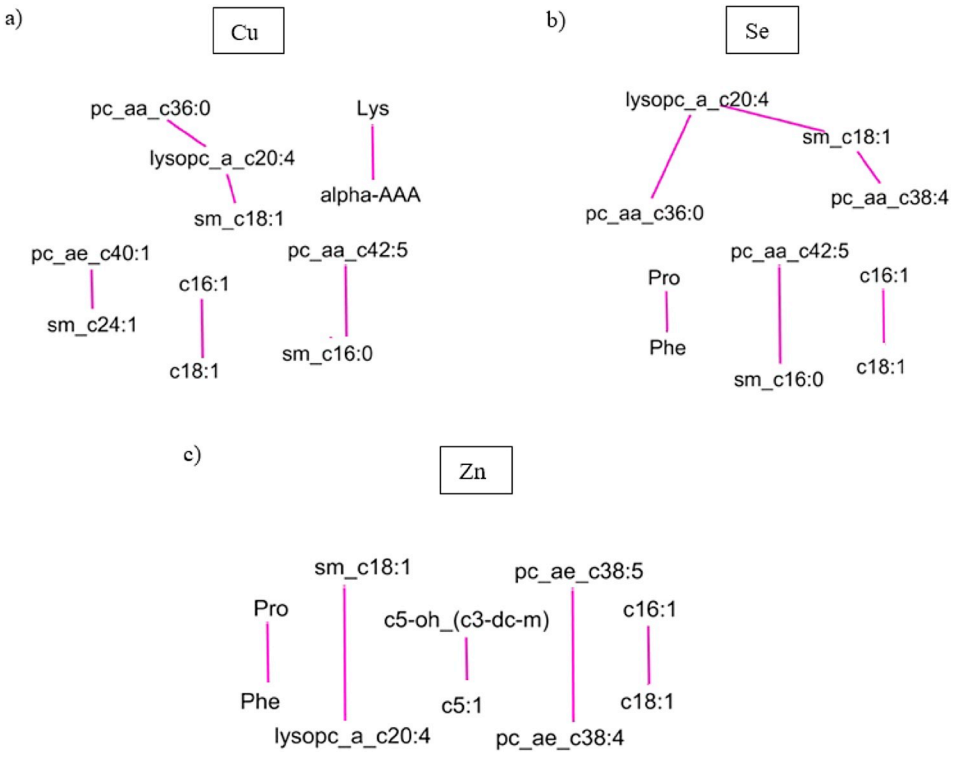

Fig. 5. Network visualizations of latent interactions of metabolites for essential nutrient metals; a) Cu, b) Se and c) Zn. Latent interactions of metabolites impacted by each of Cu, Se and Zn are represented by purple edges. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

lysine complex (CuLys) and that of serum copper has been identified previously [19]. The mammalian degradation of lysine ultimately converges at the level of α-aminoadipic semialdehyde (α-AASA), and α-AASA is converted to α-AAA. Therefore, the structure of Cu-Lysine α-AAA is possible [20]. It has been shown that α-AAA is a metabolite of the lysine metabolism pathway, possibly indicating oxidative stress in human plasma samples [21]. The toxic effect of Cd on changes in histidine concentration in response to Cd stress [22].

4. Discussion

Investigating the complex structures occurring in biological systems remains an emerging area of methodologic research. Recent progress in high-dimensional biomedical data analysis approaches has enabled quantification of metabolomic data from broad-spectrum metabolomics from mass spectrometry (MS) [23,24] and nuclear magnetic resonance (NMR) [25–27]. Advances in statistical analysis [28], network analysis [29] and software development [30] have resulted in a better understanding of associations between biological pathway members and the drivers of these pathways.

Using the methodological tools of network science, researchers have employed correlation-based techniques to uncover associations between biological entities, including metabolites. However, these approaches can detect abundant spurious correlations of low-concentration components of the data [31]. To obtain a more filtered set of statistically significant correlations between variables, probabilistic graphical models such as Markov random fields, which detect sets of conditional independence and dependence between components of the data, have been introduced [32]. The probabilistic graphical models evaluate conditional dependence and independence between components of two heterogeneous biomedical datasets as well as between components within the same biomedical dataset. While probabilistic graphical models of "omics" data have been applied to Gaussian Graphical Models, such methods have not incorporated continuous covariates.

There is a growing literature on the impact (e.g., regulation, enhancement or disruption) of trace elements, including nutrient and toxicant metal or metalloid elements (referred to as trace elements) on metabolites or metabolomic pathways. In particular, metals such as Pb, Cd, Hg and As can regulate CYP1A1 at various levels of the aryl hydrocarbon receptor signaling pathway. These toxic heavy metals can affect CYP1A1 detoxification and drug metabolism mechanisms [33]. Importantly, studies have also shown that metabolites are not independent but highly correlated. Metabolism is involved in the regulation of cellular processes, enzyme reactions and transport [34,35]. Therefore, latent relationships between metabolites and trace elements may reflect metabolic perturbations occurring between metabolites and may be responsible for either maintenance of health or disease occurrence or progression [36]. A limitation to prior computational approaches [37] is that they modeled the impact of an exogenous factor, specifically a trace element, on metabolomic profiles without considering the complex network structure or latent interactions between the metabolites. For instance, amino acids/metabolites interacting with common substrates are expected to be observed but sometimes cannot be detected [34] by the current statistical methods.

Various unsupervised multidimensional data integration approaches [38] using matrix factorization Bayesian methods or network-based methods have been developed. However, within complex biological systems, there are hidden biological signals or regularities [39], which cannot be found with currently existing data integration approaches. Furthermore, multiomics data integration approaches have rarely been applied to metabolomics and metallomics [40] datasets. An integration method over multiple heterogeneous datasets that can handle variables from distinct "omics" data rather than from original data with prior information is needed. Although some regularized methods such as group lasso can include or exclude a group of variables based on prior information, it fails to include external variables such as environmental exposures in the model. Furthermore, group lasso would lock all coefficients within a group to be either all zero or all nonzero. This fixed

structure gives little flexibility to the estimated network. We utilized the Gaussian Graphical Model (GGM) to analyze continuous "omics" datasets outside genomics, which efficiently investigates the impact of external variables as a different data integration approach from the Ising model for genomics data with relevant covariate information. In this paper, we introduced a novel multiomics data integration network-based method that can effectively identify latent interactions of entities by integrating metabolomics and metallomics data representing dietary constituents or environmental exposures that elude existing network analysis methods. We specifically developed a statistical method that simultaneously identified (1) latent intrametabolite associations related to a specific trace element and (2) existing intrametabolite associations in a high-dimensional setting.

In this study, we show that integrating trace elements with metabolites in GGM identifies novel interactions between metabolites that are regulated by trace elements. While our method has adequate power to detect interactions when the data follow a multivariate normal distribution, when the underlying distribution is non-Gaussian, our method could suffer from lower power or a high false positive rate. A potential remedy for this is to apply a semiparametric version of GGM [41] or to transform the data into binary data and then apply the Ising model instead. We will investigate this in our future studies. In summary, our method is hypothesis-generating. With further validation, the identification of interactions by our method can increase knowledge of complex biological systems that are critical to children's health.

We compared IGGM with other network methods in simulation with various parameters, such as the sample size or the connectivity of variables in the network. IGGM has consistently outperformed other methods. In real data applications, IGGM detected latent interactions between metabolites, indicating potential indirect effects that trace element exposures may have on the metabolic network. While studies have shown links between concentrations of proline and those of Zn, As, Se, Hg, or Pb, the proline – phenylalanine edges that we identified using IGGM indicate that the metabolic network is further impacted through potential indirect effects of trace elements. It has been shown that the in vivo placental transport of phenylalanine and proline in human pregnancy exists in the human body. Additionally, in vivo phenylalanine placental transport is also affected in the process of intrauterine growth-restriction pregnancies [42].

Disclosures and ethics

As a requirement of publication, the author(s) have provided to the publisher signed a confirmation of compliance with legal and ethical obligations, including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section.

Conflicts of interest

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section.

Acknowledgments

This study is funded in part by the following grants: R01LM012012 and R01LM012723 from the National Library of Medicine, P20GM104416 from the National Institute of General Medical Sciences, P42007373 and P01ES022832 from the National Institute of Environmental Health Sciences, RD8354201 from the Environmental Protection Agency and R25CA134286 from the National Cancer Institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103417.

References

[1] Janet C. King, Physiology of pregnancy and nutrient metabolism, Am. J. Clin. Nutr. 71 (5) (May 2000) 1218S–1225S. https://doi.org/10.1093/ajcn/71.5.1218s.
[2] D.R. Mattison, Environmental exposures and development, Curr. Opin. Pediatr. 22 (2) (2010) 208–218. https://doi.org/10.1097/MOP.0b013e32833779bf.
[3] P. Danaher, P. Wang, D.M. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes, J. R. Stat. Ser. Soc. B Stat. Methodol. 76 (2) (2014) 373–397. https://doi.org/10.1111/rssb.12033.
[4] J. Cheng, E. Levina, P. Wang, J. Zhu, A sparse Ising model with covariates, Biometrics 70 (2014) 943–953. https://doi.org/10.1111/biom.12202.
[5] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22.
[6] Robert Tibshirani, Bien Jacob, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, J. Ryan, Tibshirani. Strong rules for discarding predictors in lasso-type problems, J. R. Stat. Soc. Ser. B 74 (2) (2010) 245–266. https://doi.org/10.1111/j.1467-9868.2011.01004.x.
[7] Glmnet Vignette, Trevor Hastie, Junyang Qian. https://web.stanford.edu/hastie/glmnet/glmnetalpha.html. (Accessed 26 June 2014).
[8] T. Punshon, Z. Li, C.J. Marsit, B.P. Jackson, E.R. Baker, M.R. Karagas, Placental metal concentrations in relation to maternal and infant toenails in a U.S. Cohort, Environ. Sci. Technol. 50 (3) (2016) 1587–1594, https://doi.org/10.1021/acs.est.5b05316.
[9] B.H. Walsh, D.I. Broadhurst, R. Mandal, D.S. Wishart, G.B. Boylan, et al., The metabolomic profile of umbilical cord blood in neonatal hypoxic ischaemic encephalopathy, PLoS One 7 (12) (2012), e50520, https://doi.org/10.1371/journal.pone.0050520.
[10] N.M. Denihan, B.H. Walsh, S.N. Reinke, B.D. Sykes, R. Mandal, D.S. Wishart, D. I. Broadhurst, G.B. Boylan, D.M. Murray, The effect of haemolysis on the metabolomic profile of umbilical cord blood, Clin. Biochem. 48 (7–8) (2015) 534–537. ISSN 0009-9120, https://doi.org/10.1016/j.clinbiochem.2015.02.004.
[11] B. Englich, et al., Maternal cytokine status may prime the metabolic profile and increase risk of obesity in children, Int. J. Obes. 41 (2017) 1440–1446. https://doi.org/10.1038/ijo.2017.113.
[12] P. Ravikumar, M.J. Wainwright, J.D. Lafferty, High-dimensional Ising model selection using l1-regularized logistic regression, Annu. Stat. 38 (3) (2010) 1287–1319. https://doi.org/10.1214/09-AOS691.
[13] M. Pavlik, et al., The effect of arsenic contamination on amino acids metabolism in Spinacia oleracea L, Ecotoxicol. Environ. Saf. 73 (6) (2010) 1309–1313. https://doi.org/10.1016/j.ecoenv.2010.07.008.
[14] P. Theriappan, A.K. Gupta, P. Dhasarathan, Accumulation of proline under salinity and heavy metal stress in cauliflower seedlings, J. Appl. Sci. Environ. Manag. 15 (2) (2011) 251–255. http://doi.org/10.4314/jasem.v15i2.68497.
[15] A.E. Salcini, et al., Binding specificity and in vivo targets of the EH domain, a novel protein-protein interaction module, Genes Dev. 11 (17) (1997) 2239–2249. http://doi.org/10.1101/gad.11.17.2239.
[16] K.V. Alia, S.K. Prasad, P. Pardha-Saradhi, Effect of zinc on free radicals and proline in Brassica and Cajanus, Phytochemistry 39 (1) (1995) 45–47. https://doi.org/10.1016/0031-9422(94)00919-K.
[17] A.Z. Pollack, et al., Kidney biomarkers associated with blood lead, mercury, and cadmium in premenopausal women: a prospective cohort study, J. Toxicol. Environ. Health 78 (2) (2015) 119–131. http://doi.org/10.1080/15287394.2014.944680.
[18] J. Braff, E. Obare, M. Yerramilli, J. Elliott, M. Yerramilli, Relationship between serum symmetric dimethylarginine concentration and glomerular filtration rate in cats, J. Vet. Intern. Med. 28 (2014) 1699–1701. https://doi.org/10.1111/jvim.12446.
[19] G.A. Apgar, E.T. Kornegay, M.D. Lindemann, D.R. Notter, Evaluation of copper sulfate and a copper lysine complex as growth promoters for weanling swine, J. Anim. Sci. 73 (9) (1995) 2640–2646. https://doi.org/10.2527/1995.7392640x.
[20] E.A. Struys, C. Jakobs, Metabolism of lysine in α-aminoadipic semialdehyde dehydrogenase-deficient fibroblasts: evidence for an alternative pathway of pipecolic acid formation, FEBS (Fed. Eur. Biochem. Soc.) Lett. 584 (1) (2010) 181–186. https://doi.org/10.1016/j.febslet.2009.11.055.

[21] L. Leppik, K. Kriisa, K. Koido, K. Koch, K. Kajalaid, L. Haring, M. Zilmer, Profiling of amino acids and their derivatives biogenic amines before and after antipsychotic treatment in first-episode psychosis, Front. Psychiatry 9 (2018) 155. https://doi.org/10.3389/fpsyt.2018.00155.
[22] V. Zemanova, M. Pavlik, D. Pavlikova, P. Tlustos, The significance of methionine, histidine and tryptophan in plant responses and adaptation to cadmium stress, Plant Soil Environ. 60 (9) (2014) 426–432.
[23] J. Ren, A. Zhang, L. Kong, X. Wang, Advances in mass spectrometry-based metabolomics for investigation of metabolites, R. Soc. Chem. 8 (2018) 22335–22350. https://doi.org/10.1039/c8ra01574k.
[24] A.N. Halliday, et al., Recent developments in inductively coupled plasma magnetic sector multiple collector mass spectrometry, Int. J. Mass Spectrom. Ion Process. 146–147 (1995) 21–33. https://doi.org/10.1016/0168-1176(95)04200-5.
[25] A.H. Emwas, The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research, Methods Mol. Biol. 1277 (2015) 161–193, https://doi.org/10.1007/978-1-4939-2377-9_13.
[26] G.A. Nagana Gowda, D. Raftery, Recent advances in NMR-based metabolomics, Anal. Chem. 89 (1) (2017) 490–510. https://doi.org/10.1021/acs.analchem.6b04420.
[27] M. Halabalaki, K. Vougogiannopoulou, E. Mikros, A.L. Skaltsounis, Recent advances and new strategies in the NMR-based identification of natural products, Curr. Opin. Biotechnol. 25 (2014) 1–7. https://doi.org/10.1016/j.copbio.2013.08.005.
[28] A. Korman, A. Oh, A. Raskind, D. Banks, Statistical methods in metabolomics, Methods Mol. Biol. 856 (2012) 381–413. https://doi.org/10.1007/978-1-61779-585-5_16.
[29] S. Li, et al., Predicting network activity from high throughput metabolomics, PLoS Comput. Biol. 9 (7) (2013). http://doi.org/10.1371/journal.pcbi.1003123.
[30] J. Xia, I.V. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0–making metabolomics more meaningful, Nucleic Acids Res. 43 (W1) (2015) W251–W257. https://doi.org/10.1093/nar/gkv380.
[31] C.S. Calude, G. Longo, The deluge of spurious correlations in big data, Found. Sci. 22 (3) (2017) 595–612. https://doi.org/10.1007/s10699-016-9489-4.
[32] M. Layeghifard, D.M. Hwang, D.S. Guttman, Disentangling interactions in the microbiome: a network perspective, Trends Microbiol. 25 (3) (2017) 217–228. https://doi.org/10.1016/j.tim.2016.11.008.
[33] A. Anwar-Mohamed, R.H. Elbekai, A.O. El-Kadi, Regulation of CYP1A1 by heavy metals and consequences for drug metabolism, Expert Opin. Drug Metabol. Toxicol. 5 (5) (2009) 501–521. https://doi.org/10.1517/17425250902918302.
[34] M. Suarez-Diez, et al., Plasma and serum metabolite association networks: comparability within and between studies using NMR and MS profiling, J. Proteome Res. 16 (7) (2017) 2547–2559. https://doi.org/10.1021/acs.jproteome.7b00106.
[35] J. Krumsiek, J. Bartel, F.J. Theis, Computational approaches for systems metabolomics, Curr. Opin. Biotechnol. 39 (2016) 198–206. https://doi.org/10.1016/j.copbio.2016.04.009.
[36] P.A. Vorkas, et al., Metabolic phenotyping of atherosclerotic plaques reveals latent associations between free cholesterol and ceramide metabolism in atherogenesis, J. Proteome Res. 14 (3) (2015) 1389–1399. https://doi.org/10.1021/pr5009898.
[37] A. Ilyas, M.H. Shah, Multivariate statistical evaluation of trace metal levels in the blood of atherosclerosis patients in comparison with healthy subjects, Heliyon 2 (1) (2016). Article e00054, https://doi.org/10.1016/j.heliyon.2015.e00054.
[38] S. Huang, K. Chaudhary, L.X. Garmire, More is better: recent progress in multi-omics data integration methods, Front. Genet. 8 (2017) 84. https://doi.org/10.3389/fgene.2017.00084.
[39] Ebrahim Ali, et al., Multi-omic data integration enables discovery of hidden biological regularities, Nat. Commun. 7 (2016). https://doi.org/10.1038/ncomms13091.
[40] R. Haas, A. Zelezniak, J. Iacovacci, S. Kamrad, S. Townsend, M. Ralser, Designing and interpreting 'multi-omic' experiments that may change our understanding of biology, Curr. Opin. Struct. Biol. 6 (2017) 37–45. https://doi.org/10.1016/j.coisb.2017.08.009.
[41] H. Liu, F. Han, M. Yuan, et al., High-dimensional semiparametric Gaussian copula graphical models, Ann. Stat. 40 (2012) 2293–2326. https://doi.org/10.1214/12-aos1037.
[42] Cinzia L. Paolini, Anna Maria Marconi, Stefania Ronzoni, Michela Di Noio, Paul V. Fennessey, Giorgio Pardi, Frederick C. Battaglia, Placental transport of leucine, phenylalanine, Glycine, and proline in intrauterine growth-restricted pregnancies, J. Clin. Endocrinol. Metab. 86 (11) (1 November 2001) 5427–5432. https://doi.org/10.1210/jcem.86.11.8036.