RESEARCH **Open Access**

A guided network estimation approach using multi-omic information

Georgios Bartzis[1], Carel F. W. Peeters[1*], Wilco Ligterink[2] and Fred A. Van Eeuwijk[1]

*Correspondence:
Carel.Peeters@wur.nl

[1] Mathematical and Statistical Methods Group - Biometris, Wageningen University and Research, Wageningen, The Netherlands
[2] Laboratory of Plant Physiology, Wageningen University and Research, Wageningen, The Netherlands

Abstract

Intoduction: In systems biology, an organism is viewed as a system of interconnected molecular entities. To understand the functioning of organisms it is essential to integrate information about the variations in the concentrations of those molecular entities. This information can be structured as a set of networks with interconnections and with some hierarchical relations between them. Few methods exist for the reconstruction of integrative networks.

Objective: In this work, we propose an integrative network reconstruction method in which the network organization for a particular type of omics data is guided by the network structure of a related type of omics data upstream in the omic cascade. The structure of these guiding data can be either already known or be estimated from the guiding data themselves.

Methods: The method consists of three steps. First a network structure for the guiding data should be provided. Next, responses in the target set are regressed on the full set of predictors in the guiding data with a Lasso penalty to reduce the number of predictors and an L2 penalty on the differences between coefficients for predictors that share edges in the network for the guiding data. Finally, a network is reconstructed on the fitted target responses as functions of the predictors in the guiding data. This way we condition the target network on the network of the guiding data.

Conclusions: We illustrate our approach on two examples in Arabidopsis. The method detects groups of metabolites that have a similar genetic or transcriptomic basis.

Keywords: Multi-omics, Network reconstruction, Network integration

Introduction

Advances in high-throughput technology have enabled the massive collective quantification of molecular entities, such as messenger RNA, proteins, and metabolites. This age of omics has revolutionalized the field of systems biology, enabling biological systems to be studied using mathematical and computational modeling on high-dimensional omics data. In systems biology, an organism is viewed as a complex web of interacting molecular entities [19] studied in order to outline how cells, organs, and tissues behave at a molecular level [21].

A commonly used tool for analyzing omics data is network analysis. In network analysis, each omics level is assumed to have a network representation where complex associations are visualized by graphical structures. SNPs, genes, metabolites, and/or traits are typically represented by nodes in a graph and their associations (physical, genetic, or functional) by edges connecting them. Extracted patterns are then used to help elucidate biological mechanisms underlying traits of interest.

Methods for omic data integration

A key question in systems biology is how to model omics data at a systems level (integrative analysis), instead of each omics source separately [12]. Several approaches have been developed in the context of integrated analysis, see [5, 11]. One such approach for two sets of omics data is canonical correlation analysis (CCA) [8]. In order to solve CCA, the inverse of two covariance matrices needs to be computed which is problematic when the number of variables exceeds the number of samples, therefore penalization techniques can be implemented. Similarly, penalized partial least squares (PLS) regression [13] variants (sPLS; sparse Partial Least Squares) have been proposed in order to remove noisy variables resulting in variable selection for both sets of omics data [14].

An extension of sPLS is the sparse multi block partial least squares regression (sMB-PLS) [16] in which several genomic data are measured on the same samples. One dataset is considered the response data, while the rest acts as guiding sets. In an application using a dataset containing gene expression (response data), copy number variation, DNA methylation, and micro RNA expression, Li and et al. [16] identified combinations of multiple types of genomic markers that jointly impacted the expression of a set of genes. The covariance between the data blocks and the response block is maximized so that multidimensional modules are discovered associating the guiding with the response data.

Finally, network-based integration methods have also been proposed. The integration may be vertical (across omic-levels) or horizontal (one omic platform through time). The vertical approaches aim to provide a mechanistic understanding of molecular (de) regulation across the omic cascade. An overview of such methods can be found in [1]. In this work, we propose an integrative network reconstruction method where the network topology of one type of omics data is conditioned on the network topology of another omics source that is upstream in the omics cascade [4].

Aim

The question answered in this work is how to integrate information across multiple omics levels. To answer and better understand relationships between different biological functional levels, we need to combine a systems view (requiring network modeling) and a multimodal view (requiring data integration).

In this work, we study whether network reconstruction of a particular omics source can benefit from information from the network organization of another omics source. For example, is metabolite network reconstruction helped by using DNA information? Or does information on a gene expression network aid recovery of the metabolites' network organization? Under our setting, for N samples, there are two sets of omics data; the P-dimensional target dataset (denoted by $\boldsymbol{Y}_{N \times P}$ from hereon) for which the underlying network organization needs to be recovered by using a Q-dimensional guiding dataset (denoted by $\boldsymbol{X}_{N\times Q}$ from hereon) and information on its network structure which is represented by a $Q \times Q$ matrix.

For estimating the network organization of the target data using the *guiding* data set and its network organization, a guided network estimation approach is considered. First, the network organization of the guiding data is estimated. We then regress the target on the guiding data and keep the fitted values on which we estimate a network structure. Alternatively, a guiding network can be used that is available already.

Overview

The rest of the paper is organized as follows. In Sect. 2, we review some basic network concepts, and propose a guided approach for estimating the network organization of an omics source using information from another omic dataset. In Sect. 3, we demonstrate our approach on metabolite data coming from the *Arabidopsis thaliana* population.

In the first application, the metabolic network estimation is guided by utilizing SNP information. SNP data and their spatial organization are used as input (DNA structure can be seen as a linear network, with edge intensity analogous to the distance between the markers on the chromosome) [2]. We then identify and retain the part of metabolic variation related to SNP information and its structure and use it for estimating networks of metabolites. In the second data example, we guide the estimation of the metabolite networks by using information coming from gene-expression data and their network organization. Pairs of metabolites will share edges if they are associated to similar gene sets. Here, the data come from the Wageningen Seed Lab and contain SNP, transcriptomic, and metabolic information [10]. We consider this to be a standard dataset and demonstrate our integrative network approach. Our aim is to understand the metabolites from a SNP and gene level. Using network analysis we detect groups of metabolites having similar genetic or transcriptomic basis. We conclude the article with some discussion in Sect. 4.

Methods

Network analysis is a multivariate type of analysis aimed at recovering the underlying network structure of the data. We consider networks a representation of the pairwise (conditional) (in)dependencies between random variables. The nodes then represent metabolites or other molecular features and the edges represent pairwise dependency. An undirected network is typically encoded into a symmetric matrix $\boldsymbol{W}$ (intensity matrix). The element w_{ij} can be any type of association measure, e.g., the (absolute) partial correlation or (absolute) marginal correlation coefficient. The row- or column-sum of $\boldsymbol{W}$ is called strength and measures the total intensity of the connections of node i: $s(\boldsymbol{X})_i = \sum_j \boldsymbol{W}(\boldsymbol{X})_{ij}$.

Graphical LASSO

A popular approach for obtaining the underlying structure of the data from a set of P correlated variables (measured in N samples) is the Graphical LASSO (GL). In GL, the observational vectors of the data $\boldsymbol{Z}_{N\times P}$, where $\boldsymbol{Z}$ denotes a general dataset (either guiding or target data), are assumed to follow a P-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. GL is based on the conditional independence of pairwise relationships, meaning that the precision matrix ($\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$) is estimated. When the element θ_{ij} is equal to zero, variables i and j are conditionally independent given all other variables. The penalized log likelihood using a LASSO penalty [6, 7] is:

$$\ell_{\lambda}(\boldsymbol{\Theta}) \propto \log|\boldsymbol{\Theta}| - \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1, \tag{1}$$

where $|| \bullet ||_1$ is the $L1$-norm and λ is a non-negative tuning parameter governing the sparsity of the estimated precision matrix $\hat{\boldsymbol{\Theta}}$. The tuning parameter λ can be chosen based on subsampling. Here, we use the Stability Approach to Regularization Selection (StARS) [17] to estimate a set of stable edges. When using StARS, sparse networks are estimated based on multiple overlapping subsamples of the data, for different λ values on a grid. For an optimal λ (resulting in a sparse and stable network under random subsampling) selected by StARS, the absolute estimated precision matrix (similar to [26]) $|\hat{\boldsymbol{\Theta}}|$ will be used here as the intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})$.

Visual representation

To visually represent the sparse precision matrix as a network, the P variables are represented as a set of P nodes/vertices, which are connected by a set of edges, dictated by the non-zero entries of $\boldsymbol{W}(\boldsymbol{Z})$. The intensity of the connections between variables can be visualized by edge thickness with wider edges representing stronger connections. By taking the optimal λ selected by StARS as fixed, the intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})_t = |\hat{\boldsymbol{\Theta}}_t|$ can be computed based on different subsamples $t = \{1, \ldots, T\}$. The edge-wise standard deviation computed over all $\hat{\boldsymbol{W}}(\boldsymbol{Z})_t$ can be an indicator of the edge's uncertainty. Since a network is a visual representation of an intensity matrix, we will be using both terms interchangeably and denote a network by its estimated intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{Z})$.

From guiding to target data

Let $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P\}$ be the $N \times P$ target omics data matrix. Further, assume that $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q\}$ is the $N \times Q$ guiding omics data matrix. If $\boldsymbol{Y}$ contains the concentration levels of P metabolites on N samples, $\boldsymbol{X}$ could contain, for those same samples, data from Q SNPs or the expression levels of Q genes.

To incorporate information from the guiding omics data into the analysis of the target data we work in a regression framework. Prior to any type of multivariate analysis, e.g. network analysis, each of the P variables of $\boldsymbol{Y}$ is regressed on all Q variables of $\boldsymbol{X}$. Subsequently, the fitted values $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ are obtained, e.g. using penalized regression [23]. Note that the OLS coefficient estimates cannot be obtained in the high dimensional case $Q > N$. LASSO regression has some attractive properties by performing variable selection, i.e. leading to zero coefficients for some of the variables. On the other hand, no information on the dependencies between $\boldsymbol{X}$ variables ($\hat{\boldsymbol{W}}(\boldsymbol{X})$) is used.

This drawback can be alleviated by using network-constrained regularization (NCR), as proposed by [15], where the underlying network organization of $\boldsymbol{X}$ is explicitly modeled when regressing each of the P variables of $\boldsymbol{Y}$ on $\boldsymbol{X}$ [2].

Network constrained regularization

We first assume that the data $\boldsymbol{X}$ have an underlying estimable network organization $\hat{\boldsymbol{W}}(\boldsymbol{X})$. For the p^{th} response ($\boldsymbol{y}_p$), the estimated regression coefficients $\hat{\boldsymbol{\beta}}_p \in \boldsymbol{R}^{Q\times 1}$ are obtained as:

$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}_p}{\arg\min} \left\{ (\boldsymbol{y}_p - \boldsymbol{X}\boldsymbol{\beta}_p)^\top (\boldsymbol{y}_p - \boldsymbol{X}\boldsymbol{\beta}_p) + \lambda_1 ||\boldsymbol{\beta}_p||_1 + \lambda_2 \sum_{i\sim j} \left(\frac{\beta_{p_i}}{\sqrt{s(\boldsymbol{X})_i}} - \frac{\beta_{p_j}}{\sqrt{s(\boldsymbol{X})_j}} \right)^2 \hat{\boldsymbol{W}}(\boldsymbol{X})_{ij} \right\}, \tag{2}$$

where $\sum_{i\sim j}$ denotes the sum over all adjacent unordered ij pairs, $s(\boldsymbol{X})_i$, $s(\boldsymbol{X})_j$ are the strengths of nodes i and j, and the term $\lambda_1 || \bullet ||_1$ is a LASSO-type penalty inducing a sparse solution in which not all Q predictors enter the model. For selecting the penalty parameters, cross-validation (CV) can be used for estimating the prediction error for set values of λ_1 and λ_2. The chosen penalties are the ones giving the lowest CV error (in our applications we used 5-fold CV). Note that (2) can be seen as a generalization of the elastic net [15].

The part accounting for the network structure of $\boldsymbol{X}$ when estimating $\hat{\boldsymbol{\beta}}_p$ in (2) is:

$$\lambda_2 \sum_{i\sim j} \left(\frac{\beta_{p_i}}{\sqrt{s(X)_i}} - \frac{\beta_{p_j}}{\sqrt{s(\boldsymbol{X})_j}} \right)^2 \hat{\boldsymbol{W}}(\boldsymbol{X})_{ij}. \tag{3}$$

The regression coefficients $\boldsymbol{\beta}_p$ are smoothed by penalizing the sum of weighted squares of the differences between β_{p_i} and β_{p_j}. Therefore, when nodes i and j share an edge with some weight ($w(\boldsymbol{X})_{ij} \neq 0$) in the network of $\boldsymbol{X}$, they will tend to have similar association to $\boldsymbol{y}_p$. This can be biologically justified since it is expected that connected nodes (in the case of SNPs/genes/metabolites) have similar function [15] and subsequently their coefficients should be shrunken towards each other. In expression (3) it can be seen that the regression coefficients are scaled, since it is expected that nodes with higher strength are more important and therefore have larger coefficients.

The linear model using the NCR criterion, unlike LASSO, preserves the grouping property, meaning that groups of connected variables (predictors linked in $\hat{\boldsymbol{W}}(\boldsymbol{X})$) will enter the model together. This result is shown in [15].

We then fit values of the target responses on the guiding predictors $\boldsymbol{X}$:

$$\hat{\boldsymbol{y}}_p(\boldsymbol{X}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_p, \tag{4}$$

for each p. These are used for network reconstruction on $\hat{\boldsymbol{Y}}$.

Three-step approach for network reconstruction

For recovering the network structure of the target omics data, i.e. $\boldsymbol{Y}$ using a guiding omics source $\boldsymbol{X}$, we thus have a general 3-step approach:

1. Represent the guiding structure with an estimated or a priori known intensity matrix, i.e. $\hat{\boldsymbol{W}}(\boldsymbol{X})$ using GL;

2. Evaluate expression (2) with $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\hat{\boldsymbol{W}}(\boldsymbol{X})_{ij}$ and retain the fitted data matrix $\hat{\boldsymbol{Y}}(\boldsymbol{X})$;
3. Use $\hat{\boldsymbol{Y}}(\boldsymbol{X})$ to reconstruct the target intensity matrix $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}(\boldsymbol{X}))$ using GL.

By using the proposed multi-step approach, the two omics sources are no longer treated independently. The resulting estimated network of the target data is conditioned on the network organization of the guiding data.

Application to data

We now use the proposed methods for estimating metabolite networks while using information from other omics sources that have a network organization of their own.

We use a Recombinant Inbred Line (RIL) population of a cross between two Arabidopsis accessions, i.e. Bayreuth (Bay-0) and Shahdara (Sha). In this population we want to study the metabolite similarities subject to variation coming from lower leveled omics sources (SNPs or Genes). In our first example we utilize SNP data and their spatial relationship to estimate metabolite networks. Metabolites will be connected if they have the same genetic basis (similar QTLs). In the second example, we use gene expression data and their underlying network organization information when we estimate metabolite networks. Therefore, we identify metabolites with similar transcriptomic basis.

Data

Seeds from 164 lines of the Arabidopsis Bay-0×Sha RIL population were divided into four sub-populations (41 lines each) representing four important developmental stages of seed germination; (1) freshly harvested primary dormant dry seeds (PD), (2) after-ripened non-dormant dry seeds, (3) seeds imbibed for 6 h (6 H), and (4) seeds at radical protrusion (RP).

For determining the metabolite concentrations, all 164 lines were subjected to gas chromatography time of flight mass spectrometry giving 7537 peaks, representing 161 metabolites based on retention time and correlation structure [10, 24]. In total, $P = 64$ metabolites were annotated and were further used in our analysis. Gene expression analysis was performed using the Affymetrix AtSNPtile microarray on the same sub-populations and developmental stages as the metabolites, where the expression levels of 29304 genes were extracted. The top 10% most varying genes ($Q1 = 2931$ genes) were retained for further analysis. Concentration levels of the metabolites and gene expression levels were log transformed and adjusted for the four developmental seed stages by subtracting the mean levels from each group. Finally, information on $Q2 = 1059$ markers (5 chromosomes) was available. More information on the study design and data can be found in [10] and [9]. For the rest of the paper, since metabolites will be the target dataset, we will denote their $N \times P$ data matrix as $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P\}$. The $N \times Q1$ gene expression data and the $N \times Q2$ SNP data matrix will be used as guiding dataset and will be denoted as $\boldsymbol{X}^G = \{\boldsymbol{x}_1^G, \ldots, \boldsymbol{x}_{Q1}^G\}$ and $\boldsymbol{X}^S = \{\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_{Q2}^S\}$, respectively.

From SNPs to metabolites

Step 1: The SNP network representation

By having map information known, we represent the SNP data as the simplest type of network, i.e. a one-dimensional linear 'network'. We represent with $\alpha_{(1)}, \ldots, \alpha_{(p)}$ the ordered (in ascending order) genetic/physical position of the markers on the chromosome. The intensity of the connections between neighboring nodes is the relative (genetic/physical) marker proximity is calculated as:

$$\hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ij} = \hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ji} = 1 - \frac{\alpha_{(j)} - \alpha_{(i)}}{\alpha_{(p)} - \alpha_{(1)}}, \text{ where } i = 1, \dots, p-1 \text{ and } j = i+1, \tag{5}$$

where $\hat{\boldsymbol{W}}(\boldsymbol{X}^S)_{ij} = 0$ for all other cases and for markers belonging to different chromosomes.

Step 2: Estimating the metabolite part related to genetic variation

In order to use $\boldsymbol{X}^S$, and $\boldsymbol{W}(\boldsymbol{X}^S)$ for estimating $\boldsymbol{Y}^M(\boldsymbol{X}^S)$, we work with the NCR as described in Sect. 2.3. Sets of SNPs that relate to each metabolite are identified. For metabolite p, the vector of coefficients $\boldsymbol{\beta}_p^S$ is estimated and used for obtaining the metabolite fitted values as:

$$\hat{\boldsymbol{y}}_p^M(\boldsymbol{X}^S) = \boldsymbol{X}^S \hat{\boldsymbol{\beta}}_p^S.$$

Step 3: Estimating metabolite network related to genetic variation

By using GL coupled with StARS on $\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S)$, the metabolite network using SNP information $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ was estimated and is visualized in Fig. 1. The optimal regularization parameter λ^S equalled 0.651, resulting in 98 edges between the metabolites. In the same figure, the network using the original metabolite values ($\boldsymbol{Y}^M$), i.e. $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ is depicted. In order to compare the two networks, we controlled the sparsity of $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$: select the regularization parameter giving the same number of edges (98 out of 2016 possible edges resulting in sparsity of 0.049). Therefore, the tuning parameter governing the network sparsity in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ was selected to be 0.554.

Results and comparison

By examining Fig. 1, we first see that the uncertainty of the edges is lower in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ compared to $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. The top connected (hub-nodes) metabolites in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ are *Proline, Valine, Threonine, Xylose*, and *Serine* with 16, 13, 12, 12, and 10 edges respectively. On the other hand, when we see the network of metabolites with respect to SNP variation, the top connected metabolites are *Serine, (2-Hydroxyethyl)-methanamine, Isoleucine*, and *Proline* with 10, 9, 9, and 7 edges, respectively.

Here, we highlight the major differences between the networks by comparing them. Differences between the networks are visualized in Fig. 2. Edges are colored with green if they only appear in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$, red if they only appear in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$, and grey if they appear in both.

Interestingly, the metabolite losing the most edges by conditioning on SNP information (12) is *Xylose*, showing that the similarity with other metabolites was due to the non-genetic variation. Other metabolites losing multiple edges when we use SNP information are: *Proline* (11), *Valine* (11), *Asparagine* (7), and *Glucose* (7).

On the other hand, the metabolites that gained multiple edges by conditioning on SNP information are *Glutarate* (with 7) and *2-Oxoglutarate, Benzoate, Digalactosylglycerol,*

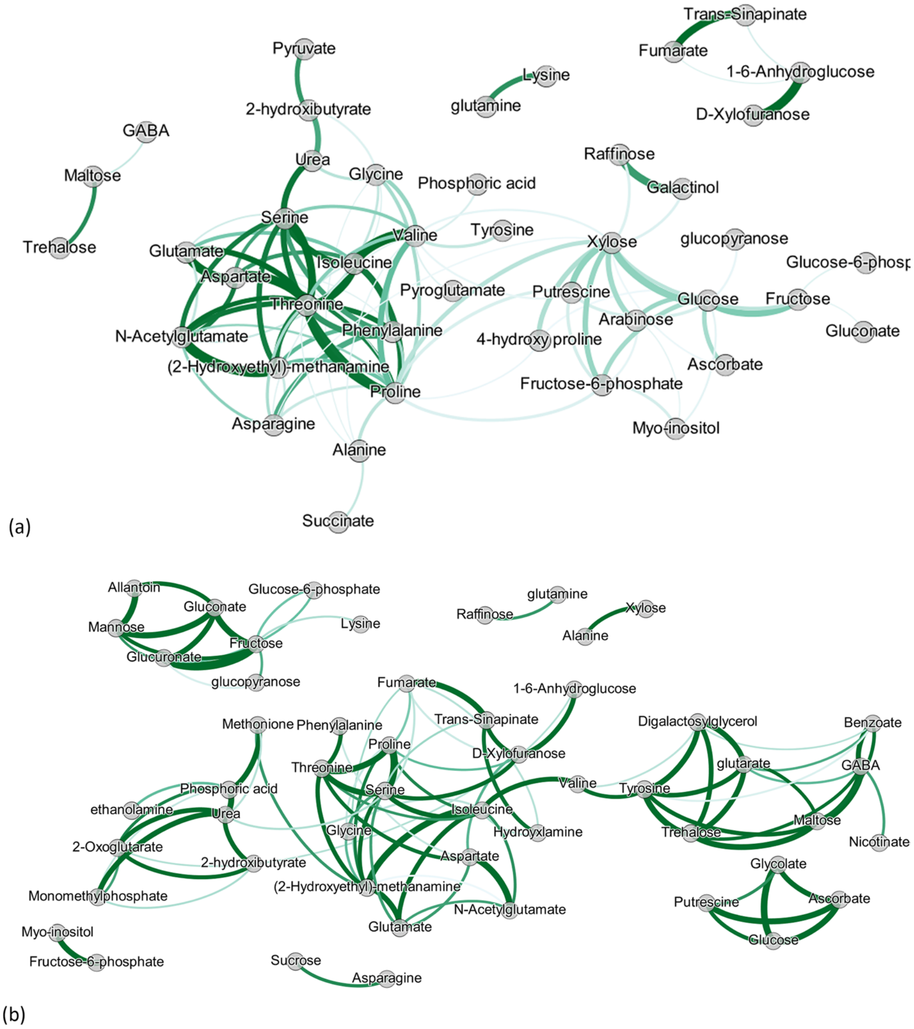

Fig. 1 Estimated metabolite networks when: (**a**) using the original metabolite data ($\boldsymbol{W}(\boldsymbol{Y}^{M})$), and (**b**) using information on SNPs and their network structure ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{S}))$). Edges' width denotes the intensity of the connection between two nodes, while edges' opacity indicates the uncertainty as measured by the edges' standard deviation

D-Xylofuranose, Fumarate, Glucuronate, Phosphoric acid, and *Tyrosine* (with 5 edges each), showing that their genetic similarity with other metabolites was stronger but it was concealed by their non-genetic variation part.

Finally, the top metabolites retaining many edges are: *Isoleucine* (8), *Serine* (6), *Threonine* (6), *(2-Hydroxyethyl)-methanamine* (5), and *Proline* (5) showing that their genetic similarity with other metabolites was stronger than the non-genetic.

Connection between QTLs and metabolite network

The vector of estimated SNP coefficients can also be used to detect QTLs. Regions where we find SNPs with non-zero coefficients should be highlighted as possible QTL regions. In Figs. 3, 4, 5 and 6 we provide some results of the correspondence between *Composite Interval Mapping* (CIM; *qtl* R-package) [27] and QTL detection using NCR while in the

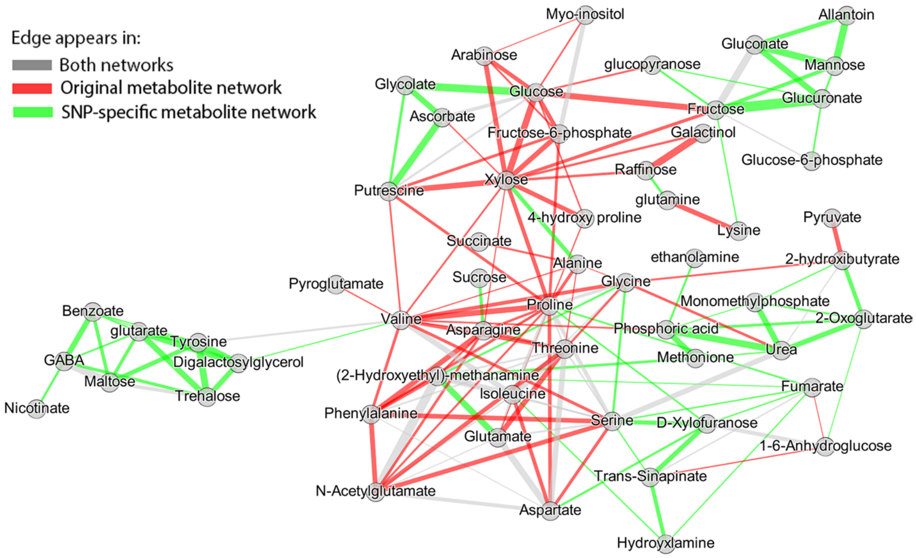

Fig. 2 Difference between network based on the original metabolite values ($\boldsymbol{W}(\boldsymbol{Y}^M)$) and network reconstructed when SNP information is used ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$). Green edges denote the unique edges that appear in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$. Red denote the unique edges appearing in $\boldsymbol{W}(\boldsymbol{Y}^M)$. Grey edges are the common edges between $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$. The width of the edges denotes the difference between the connections' intensity of $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$

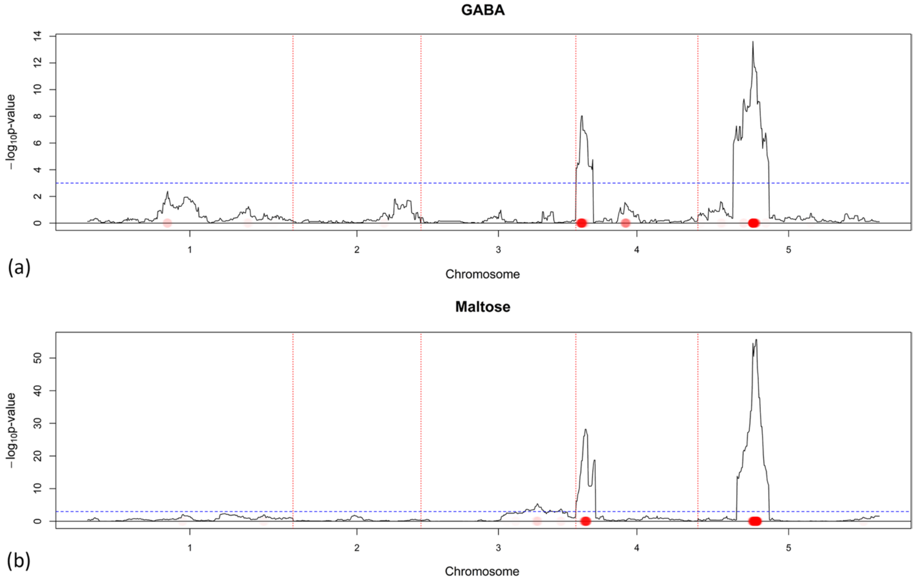

Fig. 3 QTL detection for GABA (**a**) and Maltose (**b**) using CIM and NCR when the guiding dataset is SNP data. GABA and Maltose share an edge in both $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$ which can be justified by their QTL profile. The $-\log_{10}$ *p*-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ *p*-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red dots on the x-axis are in most areas where the $-\log_{10}$ *p*-value score has high values

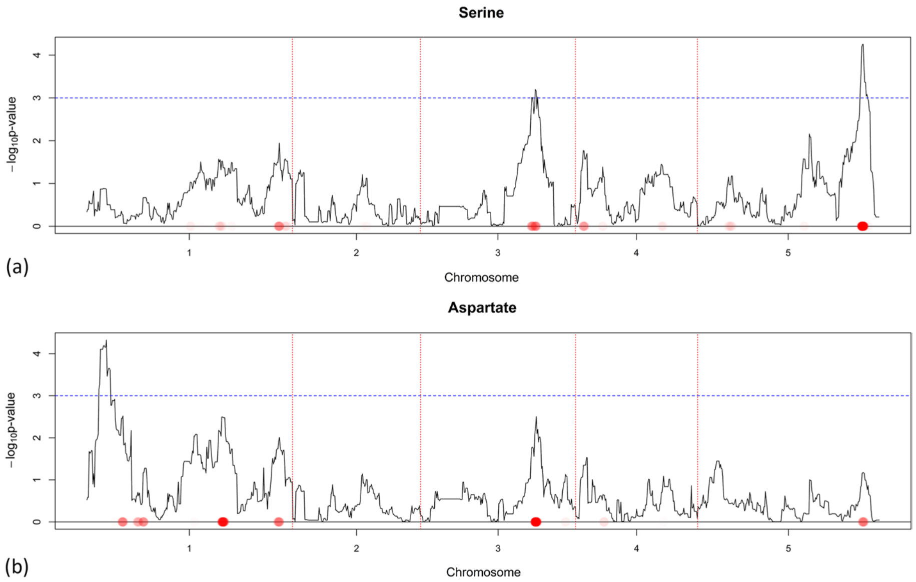

Fig. 4 QTL detection for Serine (**a**) and Aspartate (**b**) using CIM and NCR when the guiding dataset is SNP data. They only share an edge in $\boldsymbol{W}(\boldsymbol{Y}^M)$, but do not share an edge in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$, which can indicate that the unique QTLs (for a pair of metabolites) can neutralize correlation induced by common QTLs. In this case unique QTLs are responsible for a bigger part of the metabolic variation. The $-\log_{10} p$-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10} p$-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red dots on the x-axis are in most areas where the $-\log_{10} p$-value score has high values

Additional file 1 we present all metabolites. By closely inspecting Figs. 3, 4, 5 and 6, it is evident that by using NCR we find positions on the chromosome with high CIM test statistic and subsequently possible QTL regions.

The GABA/Maltose pair (Fig. 3) clearly had similar QTLs and thus share an edge. The Serine/Aspartate (Fig. 4) and the GABA/Glucose-6-phosphate (Figs. 3, 5) pairs had an overlap in their QTL profiles but share no edge, which might indicate that the non common potentially identified QTLs are responsible for a big part of the metabolic variation. Finally, the Fructose/Glucose-6-phosphate (Fig. 5) and GABA-Glycolate (Figs. 3, 6) pairs do not have overlap in QTLs justifying why there are no edges in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$.

Summarizing, when two metabolites are connected, we usually observe a similarity in QTLs. On the other hand, when metabolites do not share an edge, this is generally due to dissimilar QTLs. Still, there can be situations where metabolites with similar QTLs are not connected, because of either measurement noise, or because non-overlapping QTLs account for a big part of the metabolic variation.

Multigraph representation

An informative representation can be obtained by visualizing a network that combines all data used here. In Fig. 7, $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$ and all markers have been visualized; the nodes have been colored so that visual inspection is easier. The 5 chromosomes have been

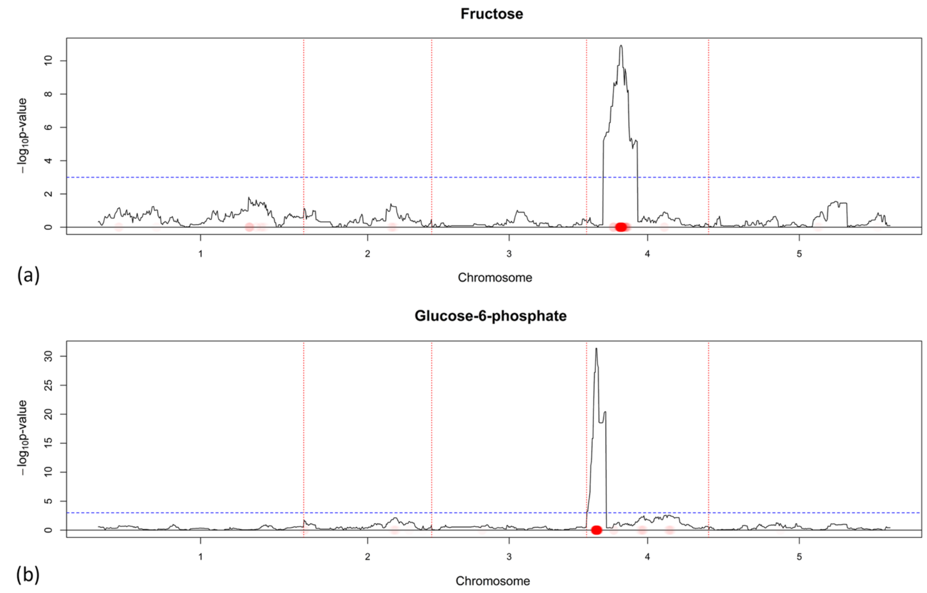

Fig. 5 Fructose (**a**) and Glucose-6-phosphate (**b**) do not have similar QTLs and therefore do not share an edge in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^S))$. The $-\log_{10}$ *p*-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ *p*-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red dots on the x-axis are in most areas where the $-\log_{10}$ *p*-value score has high values

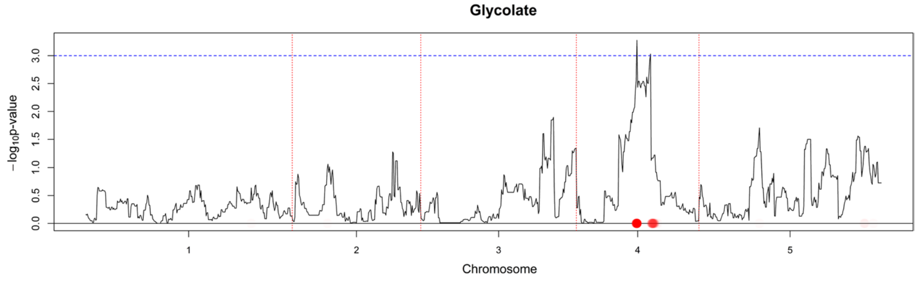

Fig. 6 Glycolate's QTL profile using CIM and NCR. The $-\log_{10}$ *p*-value score for every marker is plotted when CIM is used. The red dotted vertical lines are plotted as a visual separation between the 5 chromosomes and the chromosome number is indicated on the x-axis below each segment. The dotted horizontal blue line marks the $-\log_{10}$ *p*-value score of 3. Red dots on the x-axis are placed on marker positions for which NCR estimated non-zero coefficients. The color transparency indicates the magnitude of the regularized estimated coefficient. The correspondence between CIM and NCR can be seen by noticing that red dots on the x-axis are in most areas where the $-\log_{10}$ *p*-value score has high values

depicted as circular with the start and end being at the topmost point (one moves clockwise from start to end). Edges between a metabolite p and SNPs denote the non-zero estimated coefficients $\hat{\boldsymbol{\beta}}_p^S$.

By looking at Figs. 2 and 7 we identify three interesting metabolite groups. The first consists of: *Nicotinate, GABA, Benzoate, Glutarate, Tyrosine, Digalactosylglycerol, Valine, Trehalose*, and *Maltose*. In Fig. 2, the edges between the metabolites are green meaning that they are grouped together after using SNP information. Some of them are

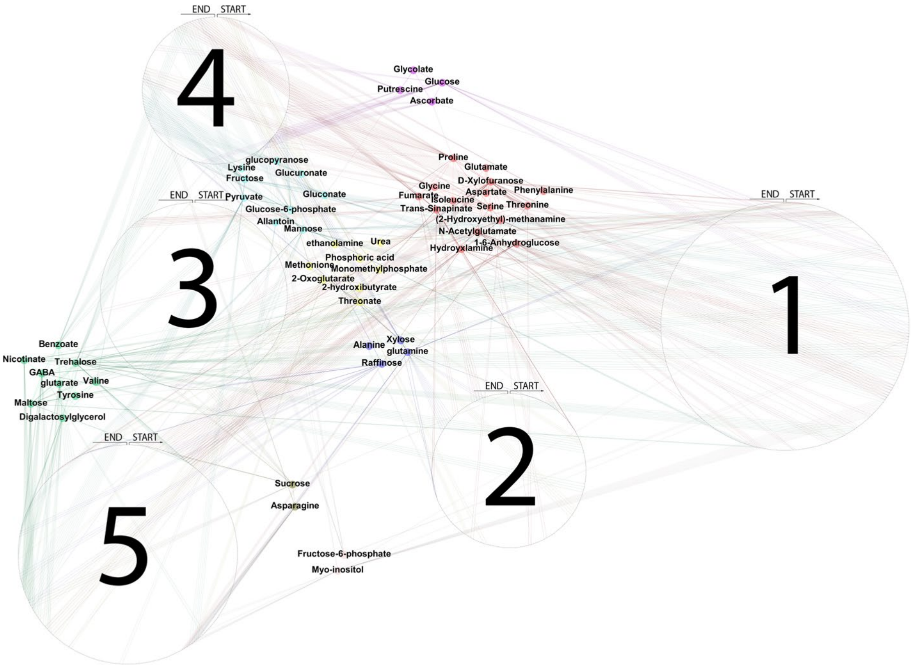

Fig. 7 Combined network of metabolites and SNPs. $\boldsymbol{W}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{S}))$ is visualized together with the five chromosomes which have been folded to be represented by five circular structures. The start and end of each chromosome is at the topmost part (moving clockwise for proceeding from start to end). Non-zero SNP coefficients for every individual model have been visualized as edges connecting metabolites and SNPs. Metabolites have been colored to ease visual inspection

connected to the biosynthesis of alkaloids derived from shikimate pathway. In Fig. 7 they have been represented as the dark green colored cluster sharing many edges with chromosome 5.

The second interesting metabolite group consists of: *Allantoin, Gluconate, Fructose, Glucuronate, Mannose, Glucopyranose*, and *Glucose-6-phosphate* and has been colored as ciel blue in Fig. 7. These metabolites did not share any connections with any metabolites in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^{M})$ but formed a cluster when SNP information was used ($\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^{M}(\boldsymbol{X}^{S}))$). Those metabolites are involved in sucrose metabolism, glycolysis and are either sugars or closely related to sugars.

Finally, the most interesting metabolite group contains: *Phenylalanine, Proline, Isoleucine, Aspartate, N-Acetylglutamate, Glutamate, (2-Hydroxyethyl)-methanamine, Glycine, Serine,* and *Threonine*. This metabolite group is the one with most grey edges in Fig. 2, showing that it retained most of its edges when we include non-genetic SNP variation. All metabolites in this group are contained in the biosynthesis of amino-acids. They have been colored red in Fig. 7 and show strong association with chromosomes 1, 4, and 5.

From genes to metabolites

In the first example we used SNP data, where the network structure was simple. Nevertheless, in many applications, e.g. gene expression data, the underlying network structure is far more complicated than a linear distance-based network and not known *a priori*. In this second example we recover metabolite networks by utilizing gene information.

Step 1: Reconstruction of the gene expression network

In order to reconstruct the gene expression network, we use GL coupled with StARS on $\boldsymbol{X}^G$. The selected regularization parameter based on subsampling was equal to 0.82, resulting in 3347 edges (sparsity of 0.00078). Our strict selection was based on the intention to minimize edges between metabolites due to false positives in gene expression data. The sparse gene expression intensity matrix $\hat{\boldsymbol{W}}(\boldsymbol{X}^G)$ contains the absolute values of the resulting inverse covariance matrix.

Step 2: Estimating the metabolite part related to transcriptional variation

We use expression (2) with $\boldsymbol{Y}^M$ as response and the gene expression data $\boldsymbol{X}^G$ as predictors, having an estimated network structure $\hat{\boldsymbol{W}}(\boldsymbol{X}^G)$. The p-th metabolite is regressed on all genes. The vector of estimated coefficients $\hat{\boldsymbol{\beta}}_p^G$ related to the $Q1$ genes is used for recovering the fitted metabolite values related to transcriptional variation ($\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G)$) as:

$$\hat{\boldsymbol{y}}_p^M(\boldsymbol{X}^G) = \boldsymbol{X}^G \hat{\boldsymbol{\beta}}_p^G \tag{6}$$

Step 3: Metabolite networks related to gene variation

To estimate metabolite networks, we use GL on the fitted metabolite values related to transcriptional variation, i.e. $\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G)$. For comparing with $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$, the regularization parameter λ^G was selected equal to 0.69, resulting in 98 edges for the metabolite network related to gene variation ($\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$). The resulting network has been visualized in Fig. 8 together with $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. The edges' width in both figures, denotes the intensity of the connection between the metabolites. The opacity represents the uncertainty for the edge intensity and has been computed based on resampling as in example 1. In Fig. 8, we see that the uncertainty of the edges is lower in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$ compared to $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$. By examining $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$, we see that the top connected metabolites are *Arabinose, Xylose, Glucose, Raffinose, Fructose-6-phosphate*, and *Monomethylphosphate* with 13, 13, 12, 12, 11, and 11 edges, respectively.

Metabolites are mainly connected because they are associated to similar (or connected) genes. On the other hand, metabolites that are not connected are usually associated with different sets of genes.

Network of differences between $W(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$ and $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$

To highlight the major differences between the networks we visualize their differences in Fig. 9. Edges are colored with green if they only appear in $\hat{\boldsymbol{W}}(\hat{\boldsymbol{Y}}^M(\boldsymbol{X}^G))$, red if they only appear in $\hat{\boldsymbol{W}}(\boldsymbol{Y}^M)$, and grey if they appear in both. By examining the differences between the networks, we make the following observations.

The metabolites losing most of their edges are *Proline* (13) and *Valine* (12) showing that their similarity with other metabolites is not due to the transcriptional part of variation. Those metabolites lost many edges in the SNP example as well, showing that their correlation with other metabolites is driven by other sources of variation. Other metabolites losing multiple edges when we use only gene information are: *Aspartate* (7), *Threonine*(7), and *Glutamate* (6).

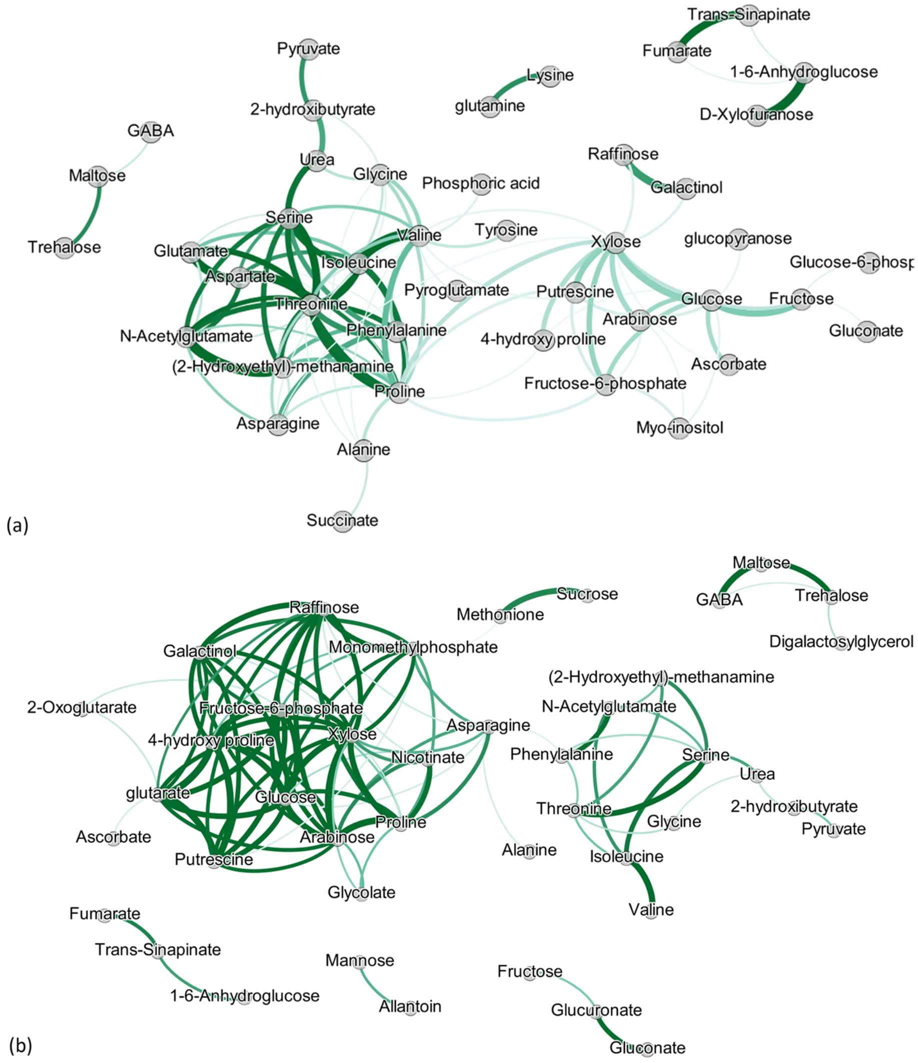

Fig. 8 Estimated metabolite networks when: (**a**) using the original metabolite data ($\boldsymbol{W}(\boldsymbol{Y}^{M})$), and (**b**) using information on SNPs and their network structure ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^{M}(X^{G}))$). Edges' width denotes the intensity of the association between two nodes, while edges' opacity indicates the uncertainty as measured by the edges' standard deviation

On the other hand, the metabolites gaining edges when gene variation is used are *Monomethylphosphate* (11), *Arabinose* (10), *Glutarate* (10), *Raffinose* (10), and *Galactinol* (8). Finally, the metabolites keeping most of their edges are: *Xylose* (9), *Serine* (6), *Fructose-6-phosphate* (5), *Glucose* (5), and *Threonine* (5).

Another finding standing out when looking at Fig. 9 is the two metabolite clusters. One consists of the following amino-acids: *Proline*, *Phenylalanine*, *t Threonine*, *Isoleucine*, *Valine*, *Glycine* and *Serine*. Lastly, the cluster containing many green edges is composed of several metabolites that are related to abiotic stress responses in plants like those related to the Raffinose family of oligosaccharides [22].

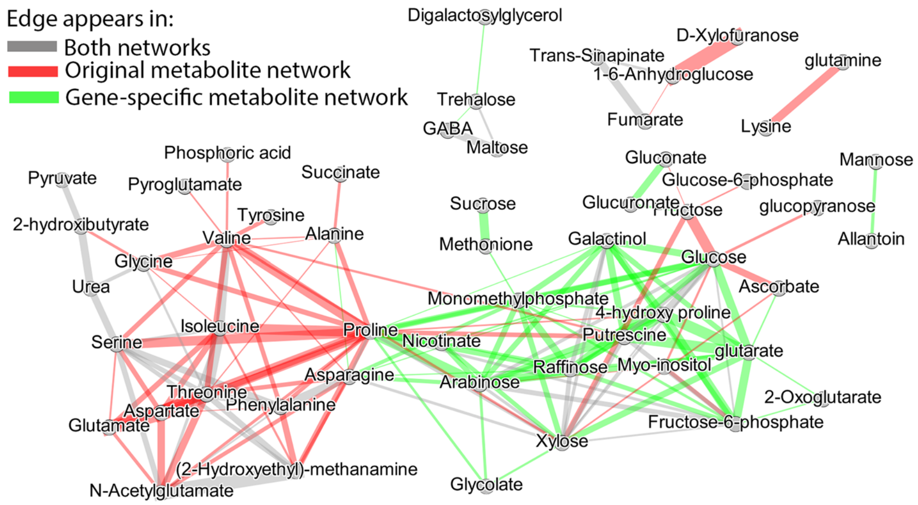

Fig. 9 Difference between network based on the original metabolite values ($\boldsymbol{W}(\boldsymbol{Y}^M)$) and network reconstructed when gene expression is used ($\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$). Green edges denote the unique edges that appear in $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$. Red denote the unique edges appearing in $\boldsymbol{W}(\boldsymbol{Y}^M)$. Grey edges are the common edges between $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$. The width of the edges denotes the difference between the connections' intensity of $\boldsymbol{W}(\hat{\boldsymbol{Y}}^M(X^G))$ and $\boldsymbol{W}(\boldsymbol{Y}^M)$

Discussion

In this work, we studied whether estimating the network structure of a particular omics level can be supported by using information coming from the network organization of another omics level. We proposed a three-step approach (Sect. 2.5) based on regularized regression that was demonstrated in two applications. Using this approach, in both applications the recovered networks contained edges with lower uncertainty compared to the original data.

For addressing missingness within guiding and target datasets, an expectation-maximization (EM) algorithm adapted for penalized network estimation offers a theoretical solution, but may escalate computational complexity. Alternatively, matrix completion techniques provide a pragmatic preprocessing step to impute missing data (e.g., [18]), thereby preparing the dataset for our network-based approach.

A natural extension of our three-step method is by using more than two datasets, e.g. SNPs, genes, and metabolites. To estimate such networks, we work sequentially from one omics source to the next. We start from SNP data and their linear structure and work our way to estimate gene expression data subject to SNP variation. Then we use the fitted gene expression values and their estimated network organization to estimate metabolite networks. Even though the rationale of such application is intuitive (propagate information from one omics level to the next), the interpretation is challenging.

By taking a step forward, since metabolites determine many quality traits (nutritional value, drought tolerance, etc) [25] and are closely related to the phenotypes [3], we could also study phenotypic associations using network analysis. By using our three-step approach for modeling phenotypic associations, we would be able to identify metabolites, genes, and DNA regions responsible for these traits. Using this approach, in plant genetics, plant breeders and physiologists can improve adaptation to environmental stress, food quality, and crop yield [20].

Finally, an interesting point of discussion is the choice of NCR in step 2 over other candidate methods, e.g., the LASSO or elastic net. In [15], these three methods have been compared in different scenarios with respect to sensitivity (true positives), specificity (true negatives) and prediction mean squared error (PMSE). The NCR procedure resulted in better PMSE making it a principal candidate for our multi-step approach. Another alternative candidate method to relate the guiding and the target datasets would be to use L2 regularization instead of L1 in (2) making it a Ridge-NCR procedure. The solution of the Ridge-NCR problem with application in genomic prediction may be more interesting, as the L1 regularization tends to drop collinear variables from the model that can potentially carry relevant information. We have presented the results of a Ridge-NCR analysis elsewhere [2]. Lastly, a hybrid between L1 and L2 penalties, aka elastic net-NCR can also be considered. Similar to LASSO, this alternative can produce reduced models by estimating zero-valued coefficients. In addition, not all collinear variables are eliminated, potentially retaining relevant information (similar to Ridge).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05778-7.

Additional file 1. QTL concordance.

Author contributions

Georgios Bartzis, Carel F.W. Peeters and Fred A.v. Eeuwijk conceptualized and wrote the main manuscript, prepared figures and tables. Wilco Ligterink provided insights on the results. All authors reviewed the manuscript.

Funding

Not applicable

Availability of data and materials

The Arabidopsis thaliana data are available upon resonable request from the Authors of Joosen et al. [10]. R-code can be found through the following publicly available GitHub repository: https://github.com/bartzis-georgios/guided-network-multi-omic.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no conflict of interest.

Received: 24 October 2023 Accepted: 11 April 2024
Published online: 30 May 2024

References

1. Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, Ederveen TH, Mulder N, Chimusa ER, t Hoen PA. Computational approaches for network-based integrative multi-omics analysis. Front Mol Biosci. 2022;9:1214.
2. Bartzis G, Peeters CFW, Eeuwijk FV. psblup: incorporating marker proximity for improving genomic prediction accuracy. Euphytica. 2022;218(5):1–14.
3. Beisken S, Eiden M, Salek RM. Getting the right answers: understanding metabolomics challenges. Expert Rev Mol Diagn. 2015;15(1):97–109.
4. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev. 2007;26(1):51–78.

5. Fabres PJ, Collins C, Cavagnaro TR, Rodríguez López CM. A concise review on multi-omics data integration for terroir analysis in vitis vinifera. Front Plant Sci. 2017;8:1065.
6. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41.
7. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, vol. 2. New York: Springer; 2009.
8. Jendoubi T, Strimmer K. A whitening approach to probabilistic canonical correlation analysis for omics data integration. BMC Bioinform. 2019;20(1):1–13.
9. Joosen RVL. Imaging genetics of seed performance. Wageningen: Wageningen University and Research; 2013.
10. Joosen RVL, Arends D, Li Y, Willems LA, Keurentjes JJ, Ligterink W, Jansen RC, Hilhorst HW. Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. Plant Physiol. 2013;162(2):553–66.
11. Joyce AR, Palsson BØ. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol. 2006;7(3):198–210.
12. Kitano H. Systems biology: a brief overview. Science. 2002;295(5560):1662–4.
13. Lê Cao KA, Le Gall C. Integration and variable selection of 'omics' data sets with pls: a survey. J Société Française de Statistique. 2011;152(2):77–96.
14. Lê Cao KA, González I, Déjean S. integromics: an r package to unravel relationships between two omics datasets. Bioinformatics. 2009;25(21):2855–6.
15. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24(9):1175–82.
16. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012;28(19):2458–66.
17. Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. In: Advances in neural information processing systems, pp 1432–1440
18. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res. 2010;11:2287–322.
19. Nielsen J, Jewett MC. The role of metabolomics in systems biology. In: Metabolomics. Berlin: Springer; 2007. p. 1–10.
20. Okazaki Y, Saito K. Recent advances of metabolomics in plant biotechnology. Plant Biotechnol Rep. 2012;6(1):1–15.
21. Raja K, Patrick M, Gao Y, Madu D, Yang Y, Tsoi LC (2017) A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries. Int J Genom. 2017.
22. Sengupta S, Mukherjee S, Basak P, Majumder AL. Significance of galactinol and raffinose family oligosaccharide synthesis in plants. Front Plant Sci. 2015;6:656.
23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodological). 1996;58:267–88.
24. Tikunov Y, Laptenok S, Hall R, Bovy A, De Vos R. Msclust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics. 2012;8(4):714–8.
25. Wang H, Paulo J, Kruijer W, Boer M, Jansen H, Tikunov Y, Usadel B, Van Heusden S, Bovy A, Van Eeuwijk F. Genotype-phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and qtls in ripe tomatoes. Mol BioSyst. 2015;11(11):3101–10.
26. Weber M, Striaukas J, Schumacher M, Binder H. Regularized regression when covariates are linked on a network: the 3cose algorithm. J Appl Stat. 2023;50(3):535–54.
27. Zeng ZB, A composite interval mapping method for locating multiple qtls. In: Proceedings, 5th World Congress on Genetics Applied to Livestock Production, University of Guelph, Guelph, Ontario, Canada, vol 7. 1994.

Publisher's Note