Computational Biology and Chemistry 71 (2017) 219–223

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Penalized estimation of sparse concentration matrices based on prior knowledge with applications to placenta elemental data

Jai Woo Lee[a], Tracy Punshon[b], Erika L. Moen[c,e], Margaret R. Karagas[c], Jiang Gui[a,d,*]

[a] Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, United States
[b] Department of Biological Sciences, Dartmouth College, Hanover, NH, United States
[c] Department of Epidemiology, Geisel School of Medicine, Lebanon, NH, United States
[d] Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH, United States
[e] The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Lebanon, NH, United States

ARTICLE INFO

Article history:
Received 8 August 2017
Received in revised form 29 October 2017
Accepted 30 October 2017
Available online 4 November 2017

Keywords:
Penalized regression
Gaussian graphical model
Concentration matrix
Elemental network

ABSTRACT

Identifying patterns of association or dependency among high-dimensional biological datasets with sparse precision matrices remains a challenge. In this paper, we introduce a weighted sparse Gaussian graphical model that can incorporate prior knowledge to infer the structure of the network of trace element concentrations, including essential elements as well as toxic metals and metaloids measured in the human placentas. We present the weighted L1 penalized regularization procedure for estimating the sparse precision matrix in the setting of Gaussian graphical models. First, we use simulation models to demonstrate that the proposed method yields a better estimate of the precision matrix than the procedures that fail to account for the prior knowledge of the network structure. Then, we apply this method to estimate sparse element concentration matrices of placental biopsies from the New Hampshire Birth Cohort Study. The chemical architecture for elements is complex; thus, the method proposed herein was applied to infer the dependency structures of the elements using prior knowledge of their biological roles.

1. Background

The advancement of large-scale genomics, epigenomics, gene expression, and other biochemical assays has propelled the application of graphical models for analysing high-dimensional biological datasets (Damaraju et al., 2014; Pierson et al., 2015; Vinciotti et al., 2016). Graphical models visually represent dependencies, or relationships, among stochastic variables (i.e., genes, proteins, or chemical elements). The Gaussian graphical model assumes that the set of variables follows a multivariate normal distribution, and a precision matrix is used to represent the inverse of the covariance matrix. Gaussian graphical models often assume a sparse precision matrix, characterized by many zeros, which occur when the variation for a given variable is predicted by a small subset of other variables in the matrix (Dempster, 1972). There have been several efforts to address the computational challenges encountered when using graphical models to infer complex relationships in biological datasets (Meinshausen and Buhlmann, 2004; Schafer and Strimmer, 2004; Dobra et al., 2004). For example, the precision matrix from a large-scale gene association network was estimated using a novel shrinkage covariance estimator, but this method failed to account for the sparsity of the precision matrix (Schafer and Strimmer, 2004). Penalized approaches (Li and Gui, 2006; Friedman et al., 2008; Hsieh et al., 2011) addressed the sparse structure of high-dimensional gene expression datasets by using threshold gradient descent and lasso (least absolute shrinkage and selection operator) penalty. Powered by coordinate decent procedure, penalized approaches are remarkably computationally efficient and facilitate the application to large data sets with thousands of parameters.

Some of the challenges to identifying patterns of association between biological entities via modelling sparse precision matrices can in part be alleviated by incorporating prior knowledge in the Gaussian graphical model estimation. In this paper, we propose to use prior knowledge of chemical associations to improve the estimation of their dependency structures in bulk elemental concentrations measured in placental tissue samples. Essential elements are homeostatically regulated some more tightly than others due to their potential for cellular damage. Selective

* Corresponding author at: Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH, United States.
E-mail address: jiang.gui@dartmouth.edu (J. Gui).

https://doi.org/10.1016/j.compbiolchem.2017.10.012
1476-9271/

transport of essential metal and non-metal elements across biological membranes uses chemical properties such as ionic radius and charge to discriminate between elements; a fallible process which allows entry of potentially toxic elements. These shared transport routes serve as prior information for associations between elements in our study. Shared transport routes have proven robust in other biological systems, for instance, the transport of cadmium, zinc and manganese by iron transporters in plants (Korshunova et al., 1999), suppression of iron, manganese and zinc uptake by cadmium (Eide et al., 1996); co-transport of calcium and strontium (Twardock, 1963), co-transport of arsenate and phosphate (Rosen and Liu, 2009), as well as co-transport of lead and calcium (Simons, 1988). Identification of orthologous genes between species has been critical to understanding their function, which are strongly conserved in membrane transport proteins. In the human placenta, membrane transport is pivotal to function due to the intense bidirectional traffic of nutrients, respiratory gases, waste products and hormones between the maternal and fetal blood supplies. We hypothesize that analysing elemental associations in the placenta may therefore be particularly informative.

We use the covariance between metal elements across placental tissues to define the precision matrix. Some metals will be highly connected in the precision matrix and thus fit the characterization of a hub. We propose to add weight to the network hubs, identified using prior biological knowledge, to increase the accuracy of the estimated graph. We introduce a weighted lasso regularization procedure for penalized estimation of a sparse precision matrix in the setting of Gaussian graphical models. Then, we demonstrate its application to identifying networks based on metal element concentrations in human placenta biopsy specimens. Such a regularization procedure aims to account for the sparsity of the precision matrix in the estimation stage while taking advantage of prior biological knowledge in the weighting scheme. After obtaining the estimate of the precision matrix, we applied a bootstrap procedure to identify the edges of the graph. Through simulations and application to real data sets, we demonstrate that this procedure is computationally feasible for both large and small sample cases and provides biologically meaningful results. Our method would be applicable to any biological dataset with prior knowledge of variables which are likely to be hubs in the network.

2. Methods

2.1. Gaussian graphical model

In the following estimations using the Gaussian graphical model, we assume that the data are randomly sampled observational or experimental data from a multivariate normal probability model. Specifically, let X be a random normal p-dimensional vector and $X_1, X_2, \ldots, X_p$ denote the p elements, where p is the number of elements. Let V={1, 2, ..., p} be the set of vertices, or nodes, and $x^{(k)}$ be the vector of elemental concentration levels for the kth sample. We assume that $X \sim N_p(0, \Sigma)$ and Σ is 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 E be the set of edges connecting the set of V vertices, and the Gaussian graphical model G = (V, E) represents an undirected graph and satisfies the linear restrictions: $e_{ij} = 0 \Rightarrow w_{ij} = 0$. Here e_{ij} is an indicator variable for the existence of an edge between vertex i and j. The Gaussian graphical model is also called a covariance selection model (Dempster, 1972) or a Gaussian concentration graph model. L1 (Lasso) regularization is well suited to compensate for the sparse nature of the true network in real data estimated with sparse Gaussian graphical models. Let S be the empirical covariance matrix, the penalized log-likelihood function can be written as:

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

Here tr denote the trace and $\| \Omega \|_1$ is the sum of absolute values of all elements in Ω. We applied the R statistical software library QUIC developed by Hsieh et al. (2011) to coordinate descent procedure to estimate the penalized coefficients.

2.2. Weighted Gaussian graphical model

We proposed a weighted Gaussian graphical model to account for the structure of the underlying true network. Using prior knowledge about the underlying biology, we can assign weights to likely hub candidates. Assigning a weight to a hub effectively assigns that weight to all of the hub's adjacent edges. If we assign weights only to the strongest candidates (i.e., those with evidence from the literature), we limit our ability to make novel discoveries. If we assign weights evenly to all nodes, we do not provide adequate weight to nodes with some support. Consequently, we aimed to develop functional networks, which integrate publicly available biological data to provide both the opportunity to make new discoveries while also allowing us to focus on candidates that are more promising than randomly selected nodes.

Unlike the weighted False Discovery Rate (Genovese et al., 2006; Gui et al., 2015), where weights have to satisfy certain restrictions, we can freely up-weight and down-weight nodes. We propose the weighted procedure as follows:

We assume that $W = diag\{w_1, w_2, \ldots, w_p\}$ is a p-by-p diagonal matrix and X is an n-by-p data matrix.
1. Identify a subset of nodes (hubs) and the corresponding weights $w_1, w_2, \ldots, w_p$. Here $w_i = 1$, when $i \notin$ hub set, and $w_i > 1$, when $i \in$ hub set.
2. Transform data X to $X^w = XW$.
3. Apply QUIC to X^w to estimate the network.

The QUIC is a forward selection procedure which will include the edges with larger corresponding covariance first. Therefore, adding a weight to each hub will increase its connectivity in the estimated network. This approach is equivalent to apply Lasso to a set of weighted variables in linear regression setting. The weighted variables will be less penalized than the rest and therefore have a better chance to get non-zero coefficients after penalization. We will use simulations to demonstrate this point and determine the optimal weight.

3. Results and discussion

To demonstrate the strength of the proposed method, we designed two simulations: Simulation I focused on comparing the tuning parameter selection methods. Simulation II aimed to estimate the accuracy of our proposed method under different settings.

3.1. Simulation of model selection

We simulated input covariance matrices with 40 nodes and 120 samples. Each matrix had 5 hubs with 8 edges on average and 35 non-hub nodes with 1–2 edges on average. First, we assigned weights to the hubs ranging from 1 (that is, no different from non-hubs) to 16 (that is, the adjacent edges to each hub have a weight of 16). Then, we applied existing penalized regression methods to the input covariance matrices in order to produce estimated precision