HHS Public Access

Author manuscript

Comput Biol Chem. Author manuscript; available in PMC 2018 December 01.

Published in final edited form as:

Comput Biol Chem. 2017 December ; 71: 219–223. doi:10.1016/j.compbiolchem.2017.10.012.

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], and **Jiang Gui**[a,d]

[a]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH

[b]Department of Biological Sciences, Dartmouth College, Hanover, NH

[c]Department of Epidemiology, Geisel School of Medicine, Lebanon, NH

[d]Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH

[e]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Lebanon, NH

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.

Keywords

Penalized regression; Gaussian graphical model; Concentration matrix; Elemental network

*Corresponding author: jiang.gui@dartmouth.edu (Jiang Gui).

Disclosures and Ethics

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. The external blind peer reviewers report no conflicts of interest.

1. Background

The advancement of large-scale genomics, epigenomics, gene expression, and other biochemical assays has propelled the application of graphical models for analyzing 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, (2005);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, (2005). 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 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 analyzing 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, …, 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 Ω = 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:

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

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 hubs 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 matrices. We used Akaike information criterion (AIC), Bayesian information criterion (BIC), 5-fold and 10-fold cross validation to find tuning parameter which determine the optimized model. After obtaining the optimized model, we computed the true positive rate (TPR) and true negative rate (TNR). We defined accuracy as the average of TPR and TNR; The Area Under the curve (AUC) obtained by one run can be defined as balanced accuracy computed by the formula,

$$BalancedAccuracy=\frac{(sensitivity+specificity)}{2} \quad (2)$$

where sensitivity and specificity are known as true positive rate and true negative rate. We repeated the process 500 times and recorded the average accuracy. Sokolova et al., (2006)

Simulation of model selection compares accuracy with four different tuning parameter selection methods, AIC, BIC, and Cross-Validation (5-fold, and 10-fold). Figure 1 reports the accuracy against weights assigned to hubs. The results from AIC, BIC, 5-fold and 10-fold cross validation are plotted. As the weight increased from 1 to 5, accuracy sharply increased for all 4 methods. AIC outperformed BIC by a large margin. AIC is also more consistent than the cross-validation methods. Therefore, we focused on using AIC for tuning parameter selection in the application of high-dimensional data analysis.

3.2. Simulation of Weighted Nodes Selection

We tested the effects of hub selection on accuracy estimation. We simulated 4 different scenarios including assigning weights to all hubs, a mixture of hubs and non-hubs, and all non-hubs. We repeated each simulation 500 times and calculated the average accuracy for each scenario

Scenario 1: Weights were assigned to all 5 hubs

Scenario 2: Weights were assigned to 3 hubs and 2 non-hubs

Scenario3: Weights were assigned to 5 non-hubs

Scenario 4: Weights were assigned to 5 randomly selected nodes

Simulation II presents accuracy results by different scenarios of assigning weights to hubs and/or non-hubs. Figure 2 shows the effect of weighting when we have varying information on hubs. When we placed weights on all hubs (scenario 1), the accuracy increased sharply as the weights went up. If we only placed weights on a subset of hubs (scenario 2), the accuracy increased with a lesser slope. If we assigned weight on non-hubs or random nodes (scenarios 3 and 4), the accuracy would fluctuate without any significant improvement in accuracy. **The result is expected since true correlation values between non-hubs and other nodes are small in general. Hence, even after the weighting process, nodes with small true correlation values are rarely selected**

3.3. Application to placental metal concentration data

We collected 765 placental specimens from participants of the New Hampshire Birth Cohort and studied the metallome of human term placenta by measuring the bulk elemental composition of placental biopsies Punshon et al., (2015);Punshon et al., (2016). To estimate the network of metal elements in placental samples, we applied the Weighted Gaussian graphical model to the bulk elemental composition. Our goal is to better predict the covariance of each pair of metals and construct networks representing interactions among them. Network analysis is conducted by the estimation of the sparse precision matrix in

which each off-diagonal value indicates the conditional covariance between the two corresponding metals.

The placenta metal concentration dataset contained 27 metal elements and 763 samples. We excluded metal elements if more than 50% of their covariance values were missing. After pre-processing the dataset and excluding the metal elements V and U for measurement error, our final dataset included 25 metals.

We used the following approach on real data: first, we log-normalized and scaled the placenta metal concentration data. Then, we applied AIC to find tuning parameters to determine the optimized model, because AIC give us the most consistent estimate in accuracy from simulation. To ensure robust results, we ran 200 bootstraps for each estimated network and plotted the edges that were selected at least 190 times.

We targeted three different elements of interest; zinc (Zn), cadmium (Cd) and arsenic (As) as possible hubs in the network. We chose Zn as a result of observing spatial localization of zinc in the syncytiotrophoblast, the outer membranes of the fetal chorionic villi in a previous study Punshon et al., (2016), and because of its multiple essential intracellular roles. We chose Cd as a possible hub because it is a common non-essential metal contaminant that can be detected in the placenta (from exposure to cigarette smoke, as well as dietary exposure) whose competition with Zn is well-supported in the scientific literature Wang et al., (2016);Da Silva and Williams, (2001);Bridges and Zalups, (2005);Ballatori, (2002). We also chose As as a further non-essential element of interest because groundwater in New Hampshire contains elevated concentrations of arsenic, and up to 40% of residents in the region obtain their household drinking water from an underground well. Both of these non-essential elements cross the human placenta, with As crossing particularly efficiently and Cd showing evidence of placental retention Punshon et al., (2016).

For applied results, first, we estimated the unweighted element network generated by QUIC. In the resulting network (Figure 3), Zn, magnesium (Mg), strontium (Sr), barium (Ba) and phosphorus (P) are highly connected. Next, we assigned weights of 3 to As Cd and Zn (Figure 4). Since males and females tend to have different Zn concentrations, we repeated the weighted analysis on males and females separately (Figure 5 and 6).

The network from the unweighted model (Figure 3), as well as weighted (Figure 4) and sex-specific models (Figure 5 and 6) identified fundamental elemental associations consistent with known chemical and biological roles of the elements. The common features of these models are, (1) a separate network containing only Ca and K; (2) an interconnected sub-network of Mg, P, Ba and Sr, and (3) the appearance of Zn as a hub of multiple elemental associations.

The connection between Ca and K distinct from the larger network speaks to the many fundamental and regulatory cellular processes in which these two elements participate. In addition to a major role in metabolic control, Ca provides mechanical stability as a component of cell walls and membranes, is involved in cell division and hormonal activities. Calcium is also a signalling molecule, where short bursts of Ca movement across membranes are triggered either by electrical impulses or hormonal signals. Likewise, K is

involved in osmotic control, electrolytic equilibria across cell membranes, stability of polyelectrolytes and DNA, while also being connected to the chemical uptake of organic metabolites. Calcium and K are more abundant in the placenta by several orders of magnitude than other elements measured in this study Da Silva and Williams, (2001).

Another common feature of the resulting networks is the appearance of a smaller sub-network consisting of Mg, P, Sr and Ba. The close association between Mg and P is well known: Mg strongly binds to pyrophosphates such as ATP (adenosine triphosphate), widely known as the molecular unit of currency in intracellular energy transfer. By contrast, Ba and Sr are not essential elements, but their chemical similarity to Ca and Mg, (also in group IIA) may explain their connections with Mg in this network. Both of these elements are ubiquitous environmental contaminants, the former from coal combustion, but they are both also present in a diverse range of industrial products from ceramics, lubricants, paint pigments, fluorescent lights to medicine. With the exception of group IA and IIA metals, Zn is the most common catalytic metal ion in the cell cytoplasm; it is involved in a great variety of enzyme reactions, and its abundance in placenta is comparable to that of K and Ca. Zn is a component of insulin, S-100, phospholipase C, carbonic anhydrase, superoxide dismutase, alcohol dehydrogenase and metallothionein, among others. In all models, Zn is connected with multiple metals, in particular with Mn, Se, Pb and Cd. Zinc is associated directly with Cd in the weighted models. Also, a Lewis acid, Cd^{2+} is a known to substitute for Zn^{2+} in many Zn-containing enzymes. In the placenta, in particular, Cd has been noted to impair Zn transport Wang et al., (2016).

Sex-dependent functional differences in the placenta may underlie the observed sex-specificity of responses to in utero exposure to metal contaminants, as well as to non-communicable chronic disease Kippler et al., (2013). These differences may be a result of genetic and hormonal differences. The sex-specificity of the metal networks was explored by separating placental specimens on the basis of whether they had supported a male or a female pregnancy (Figures 5 and 6), producing subtly different networks. In particular, we observed differences in the connections with Zn and Cd; in females, Mn was co-connected with Zn and Cd, and in males, none of elements was co-connected with Zn and Cd.

4. Conclusion

In this paper, we presented a novel procedure for estimating the sparse precision matrix in the setting of Gaussian graphical models. This paper makes two unique contributions: first, we offer a computationally efficient algorithm that incorporates network structure for better accuracy estimation in a precision matrix. We have developed the functions for this procedure in R, which are available upon request. Second, we carefully compared three popular tuning parameter selection methods (cross validation, AIC and BIC) and found that AIC yields the best results in most cases.

Despite the advantages stated above, one limitation of this method is that the weight selection is ad-hoc. In future work, we plan to identify a method to optimize the weight selection using either publicly available datasets or via splitting the data into training and

test sets. We believe that an optimal weight to maximize the accuracy can be derived based on the known network structures.

In summary, we demonstrate in simulation and real data that our proposed method is efficient and can accurately estimate the precision matrix. We believe that it will play an important role as part of a research strategy to understand chemical interactions that embraces the complexity of the network structure.

Acknowledgments

The study is funded by the following: R01LM012012 from National Library of Medicine, P20GM104416 National Institute of General Medical Sciences, P42007373 and P01ES022832 National Institute of Environmental Health Sciences, RD83544201 Environmental Protection Agency and R25CA134286 National Cancer Institute.

References

Damaraju E, Allen E, Belger A, Ford J, McEwen S, Mathalon D, et al. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. NeuroImage: Clinical. 2014; 5:298–308. [PubMed: 25161896]

Pierson E, Koller D, Battle A, Mostafavi S. GTEx Consortium; Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput Biol. 2015; 11(5):e1004220. [PubMed: 25970446]

Vinciotti V, Wit EC, Jansen R, de Geus EJ, Penninx BW, Boomsma DI, et al. Consistency of biological networks inferred from microarray and sequencing data. BMC Bioinformatics. 2016; 17(1):254. [PubMed: 27342572]

Dempster AP. Covariance selection. Biometrics. 1972:157–75.

Meinshausen, N., Buhlmann, P. Seminar fr Statistik, Eidgenssische Technische Hochschule (ETH). Zrich: 2004. Consistent neighbourhood selection for sparse high-dimensional graphs with the Lasso.

Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005 Mar.216:754–64.

Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90(1):196–212.

Li H, Gui J. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics. 2006 Apr.72:302–17.

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008 Jul.93:432–41.

Hsieh CJ, Dhillon I, Ravikumar P, Sustik M. Sparse inverse covariance matrix estimation using quadratic approximation. Paper presented at the Advances in Neural Information Processing Systems. 2011:2330–2338.

Korshunova YO, Eide D, Clark WG, Guerinot ML, Pakrasi HB. The IRT1 protein from Arabidopsis thaliana is a metal transporter with a broad substrate range. Plant Mol Biol. 1999; 40(1):37–44. [PubMed: 10394943]

Eide D, Broderius M, Fett J, Guerinot ML. A novel iron-regulated metal transporter from plants identified by functional expression in yeast. Proc Natl Acad Sci U S A. 1996 May 28; 93(11): 5624–8. [PubMed: 8643627]

Twardock, A. The Transfer of Calcium and Strontium Across Biological Membranes. Academic Press, Inc; New York: 1963. Studies on the movement of calcium and strontium across the bovine mammary gland; p. 327-39.

Rosen BP, Liu Z. Transport pathways for arsenic and selenium: a minireview. Environ Int. 2009; 35(3): 512–5. [PubMed: 18789529]

Simons T. Active transport of lead by the calcium pump in human red cell ghosts. J Physiol (Lond). 1988; 405(1):105–13. [PubMed: 2855638]

Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006; 93(3):509–24.

Gui J, Greene CS, Sullivan C, Taylor W, Moore JH, Kim C. Testing multiple hypotheses through IMP weighted FDR based on a genetic functional network with application to a new zebrafish transcriptome study. BioData Min. 2015 Jun 17.8 17,015-0050-8. eCollection 2015.

Sokolova M, Japkowicz N, Szpakowicz S. Beyond Accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. Proc Advances in Artificial Intelligence (AI 2006). 2006:10151021.

Punshon T, Davis MA, Marsit CJ, Theiler SK, Baker ER, Jackson BP, et al. Placental arsenic concentrations in relation to both maternal and infant biomarkers of exposure in a US cohort. Journal of Exposure Science and Environmental Epidemiology. 2015; 25(6):599–603. [PubMed: 25805251]

Punshon T, Li Z, Marsit CJ, Jackson BP, Baker ER, Karagas MR. Placental Metal Concentrations in Relation to Maternal and Infant Toenails in a US Cohort. Environ Sci Technol. 2016; 50(3):1587–94. [PubMed: 26727403]

Wang H, Wang Y, Bo Q, Ji Y, Liu L, Hu Y, et al. Maternal cadmium exposure reduces placental zinc transport and induces fetal growth restriction in mice. Reproductive Toxicology. 2016; 63:174–82. [PubMed: 27319394]

Da Silva, JF., Williams, RJP. The biological chemistry of the elements: the inorganic chemistry of life. Oxford University Press; 2001.

Bridges CC, Zalups RK. Molecular and ionic mimicry and the transport of toxic metals. Toxicol Appl Pharmacol. 2005; 204(3):274–308. [PubMed: 15845419]

Ballatori N. Transport of toxic metals by molecular mimicry. Environ Health Perspect. 2002 Oct; 110(Suppl 5):689–94.

Kippler M, Engstrm K, Mlakar SJ, Bottai M, Ahmed S, Hossain MB, et al. Sex-specific effects of early life cadmium exposure on DNA methylation and implications for birth weight. Epigenetics. 2013; 8(5):494–503. [PubMed: 23644563]

When weights are placed only on hubs, the accuracy increases sharply as the weights go up.

When weights are assigned on random nodes, accuracy fluctuates without clear improvement.

There is a connection between Ca and K distinct from the larger network regardless of weights.

There exists a smaller sub-network consisting of Mg, P, Sr and Ba regardless of weights

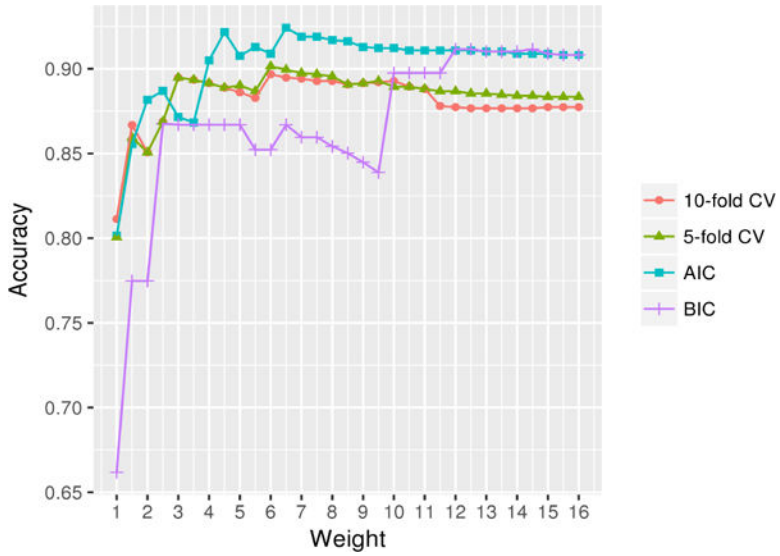

Figure 1.
Accuracy by different weights assigned to hubs

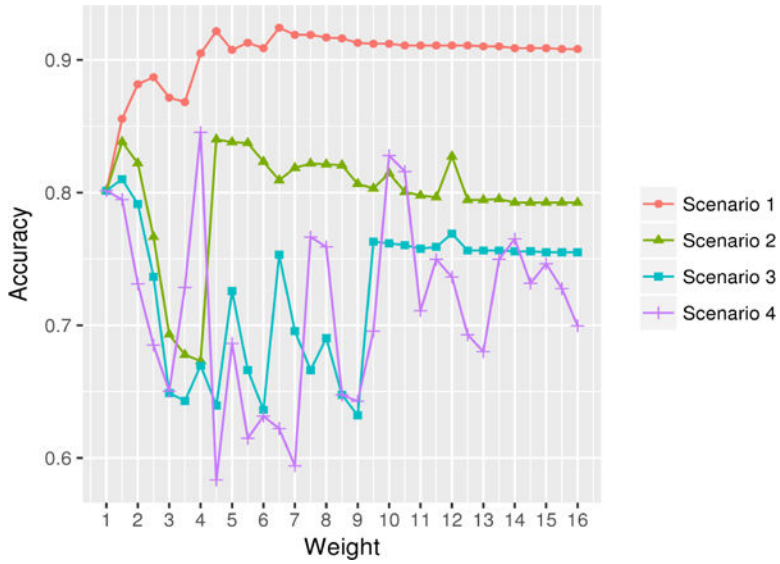

Figure 2.
Accuracy by different weighting strategies

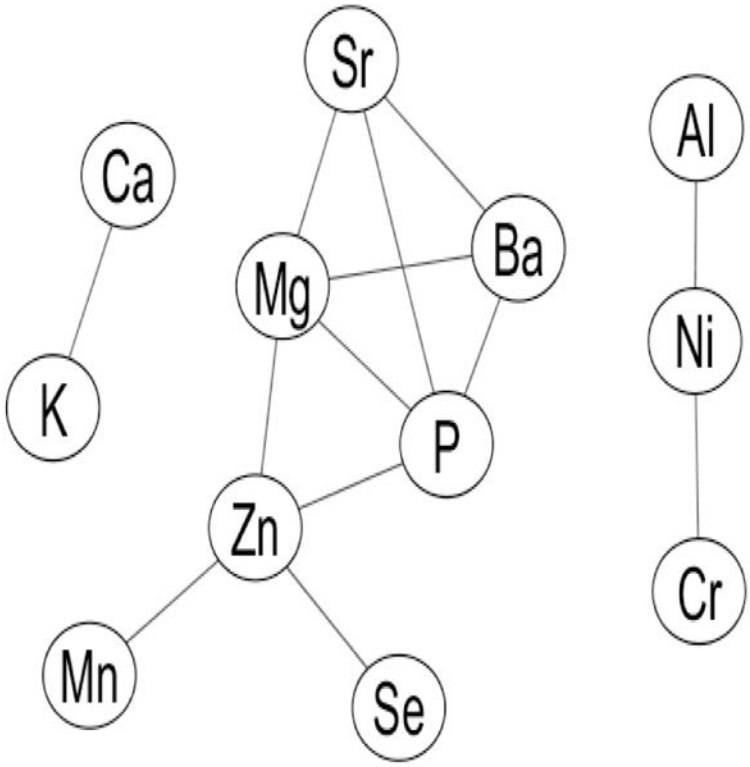

Figure 3.
Network from unweighted model

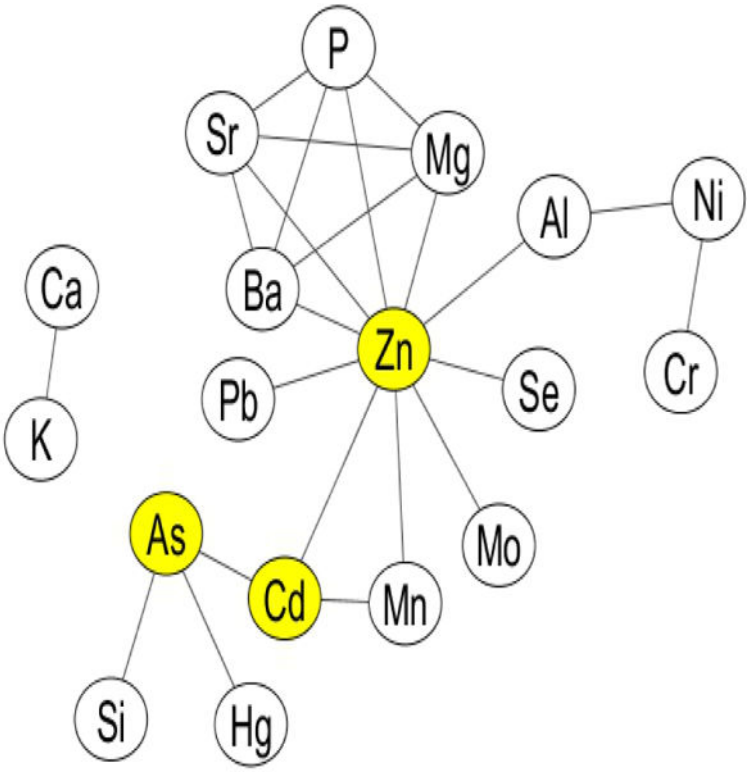

Figure 4.
Network from weighted model

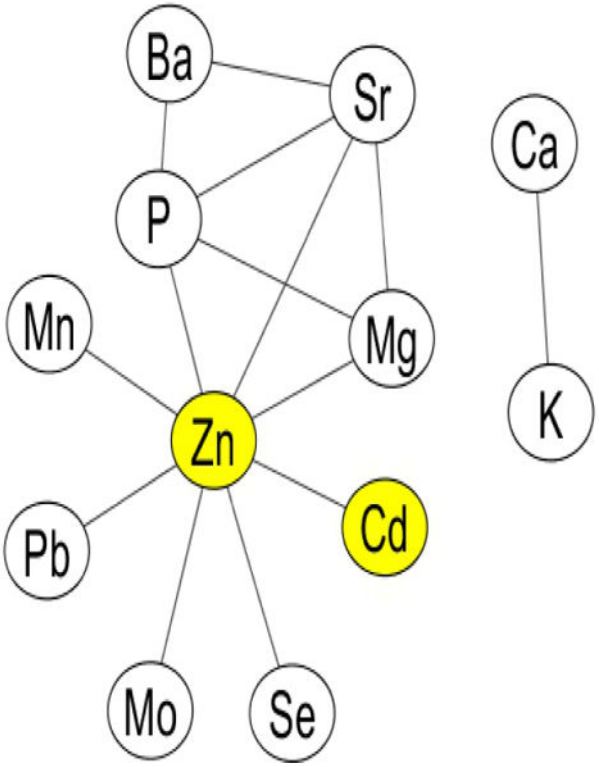

Figure 5.
Network from weighted model for males only

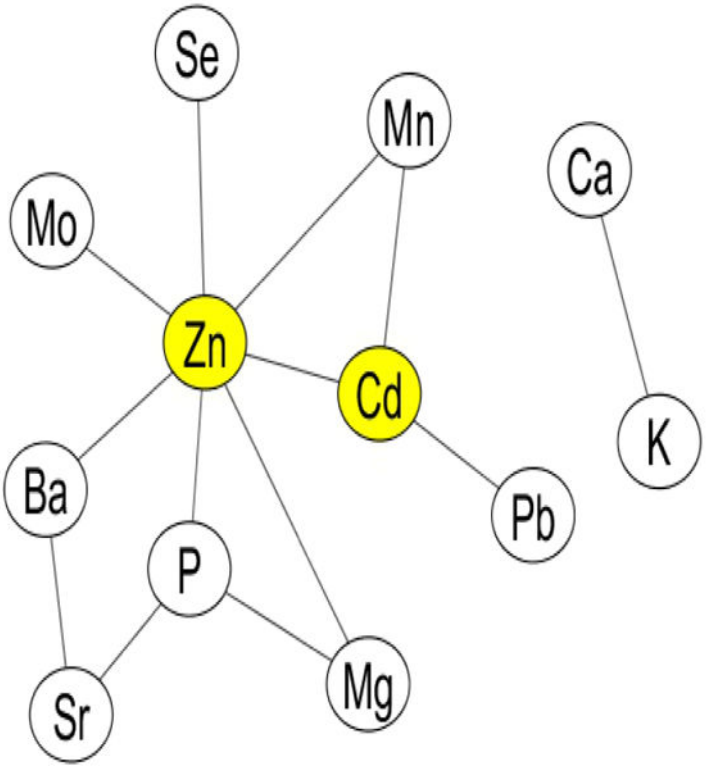

Figure 6.
Network from weighted model for females only