|  |
| --- |
| Inferring metabolite interactomes via Bayesian graphical model selection utilizing molecular structure informative priors |
|  |
| Patrick J. Trainora\*, Samantha M. Carlisleb, Shesh N. Raic§, Andrew P. DeFilippisd§ |
|  |
| aDepartment of Medicine, Division of Cardiovascular Medicine, University of Louisville  bDepartment of Pharmacology and Toxicology  cDepartment of Bioinformatics and Biostatistics  dDepartment of Medicine, Division of Cardiovascular Medicine, University of Louisville,  Jewish Hospital and Johns Hopkins University  §Equally contributing authors  \*Corresponding Author |
|  |

|  |  |  |
| --- | --- | --- |
| Keywords  Evolutionary Computation, Wisdom of Artificial Crowds,  Feature Selection,  Classification, Diagnostic Test,  Myocardial Infarction |  | Abstract  *Introduction*  *Materials and Methods*  *Results / Conclusions* |

1. Introduction

Untargeted profiling of the metabolome of an organism provides a view into the small molecule determinants of phenotype. While the genome of an organism may be conceptualized as blueprint for the composition and organization of an organism that is largely immutable (barring epigenetic modifications and genetic mutations) [citation], the metabolome of an organism is dynamic and variable [citation]. Sources of variation within the metabolome of a single organism include tissue-, cell-, and organelle-specific localization of metabolic processes (Shlomi, Cabili, Herrgård, Palsson, & Ruppin, 2008; Voet, Voet, & Pratt, 2013); environmental exposures (Southam et al., 2014); and host-microbe interactions and metabolite exchange (Moriya, Satomi, Murata, Sawada, & Kobayashi, 2017). While the generation of reference human genomes has facilitated the interrogation of gene-phenome associations (including human disease associations), the intrinsic variability in the metabolome of an organism likely precludes the generation of such a reference model. While a single reference model of the metabolome of an organism may not be sensible, significant efforts such as the Husermet project (Dunn et al., 2014) have been undertaken to quantify the repertoire of metabolites in specific biofluids for examining metabolite-metabolite and metabolite-phenotype associations. The learning of probabilistic models that describe the relationship between metabolites, henceforth “interactomes”, for a phenotype will allow for comparisons between phenotypes that while not global, are comprehensive given the sample media and analytical platform. A specific use case for such an interactome model is the generation of a plasma interactome model for stable heart disease to serve as a reference model for determining metabolic perturbations associated with acute disease events such as unstable angina and myocardial infarction (MI).

While correlation networks have been used to describe the relationships between metabolites in many metabolomics experiments (see for example [citation]), this approach is limited as the topology learned represents only the pairwise marginal associations between metabolites. Determining conditional relationships between metabolites allows for making inferences over how the abundance of a specific metabolites influences the abundance of other metabolites after conditioning on the abundance of other intermediates. In order to model such conditional probabilistic dependences between metabolite abundances, a Gaussian Graphical Model (GGM) approach may be employed as in the present work. GGMs provide a suitable framework for representing the joint probability distribution of metabolites that are detected in metabolomics experiments and for representing the probabilistic interactions between metabolites and have been employed in interaction modeling in other domains of molecular biology (see for example: ).

A significant challenge in evaluating the relationships between metabolites in an untargeted metabolomics experiment is that the dimension of metabolites may be greater than the number of samples. Even given a relatively high ratio of samples to metabolites detected, in the evaluation of pairwise conditional relationships between metabolites, the number of parameters to be estimated can be prohibitive. For example, if metabolites are detected, an evaluation of all pairwise conditional relationships would require the simultaneous estimation of 124,560 parameters. The use of regularization is a well-established approach for guaranteeing the existence of Gaussian Graphical Model parameters, amenable to the case that the sample size is less than (Banerjee, El Ghaoui, & d'Aspremont, 2008; Fan, Feng, & Wu, 2009; Friedman, Hastie, & Tibshirani, 2007; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007).

Penalized estimation of GGM parameters provides a natural mechanism for integrating *a priori* knowledge regarding the molecular structure of metabolites with experimental metabolomics data. The integration of empirical data and scientific knowledge regarding metabolism is common in metabolomics studies. Typically, univariate and/or multivariate analyses first identify sets of metabolite features for which evidence of differences between experimental conditions or phenotypes are observed. After identifying interesting metabolite features, these sets can be tested for enrichment of specific metabolic pathways [citation] or biological processes greater than that expected by chance. A promising alternative to pathway analyses discussed in (Barupal & Fiehn, 2017) is to use structural similarity and chemical ontology to generate study-specific metabolite sets for contextualizing empirical results with *a priori* knowledge of metabolism. The current work is of a similar paradigm and predicated on the assumption that the individual biochemical reactions that result in statistical dependence between metabolic intermediates also generate statistical dependence in structural similarity between the same intermediates. However, rather than considering fixed sets of metabolites such as pathways or modules and subsequently quantifying enrichment of these sets in empirical results, we consider *a priori* knowledge of the relationships between metabolites as probabilistic statements about the relatedness of compounds. Thus, the *a priori* scientific knowledge is used to generate prior probability distributions that influence GGM model selection, so that posterior inference probabilistically combines empirical data and prior scientific knowledge to yield an updated model of the probabilistic interactions between metabolites. In the present work, we introduce a methodology for using 2D molecular structure similarity to generate prior distributions that control the degree of penalization in parameter estimation for learning a GGM metabolite interactome from metabolomics data. We evaluate the methodology using simulation studies that follow two different schema. Under the first schema, autoregressive processes were simulated for representing linear biological processes in which the correlation between simulated metabolites decreased in tandem with decreasing structural similarity. Under the second schema, metabolomics data collected previously was utilized with simulated metabolites “spiked-in”. The spiked-in simulated metabolites were simulated via a hierarchical model in which structural similarity was simulated first, followed by abundance distributions in which the correlation between abundances increased with structural similarity. Given both schema, we evaluated the ability of the proposed method to recover the true pairwise conditional correlations structures that were specified in advance.

1. Methods
   1. Gaussian Graphical Models (GGM)

We consider Markov Random Fields (MRFs) which are graphical models in which random variables are represented as vertices and edges in the edge set represent probabilistic interactions. Gaussian Graphical Models (GGMs) represent a special class of MRFs in which the underlying joint probability distribution represented by the graph is assumed to be multivariate Gaussian (Koller & Friedman, 2009). In addition to the joint distribution being multivariate Gaussian, the marginal distribution for each is Gaussian, as are the conditional distributions for . Given a multivariate Gaussian distribution , where is a vector of means and is the inverse of the covariance matrix , the entries of the concentration matrix are of particular importance as implies that and are conditionally independent and with respect to the graph topology, there does not exist an edge between and . Further from the entries of , the partial correlation coefficient between two random variables and can be computed as: .

* 1. GGM parameter estimation

It has been shown previously that if where represents the maximal clique size of the GGM then a maximum likelihood estimator does not exist (Buhl, 1993). Noting the likelihood function for the concentration matrix :

where is the sum of products matrix. As the log-likelihood function is not guaranteed to be convex, regularization of this likelihood has been proposed (Banerjee et al., 2008; Friedman et al., 2007; Meinshausen & Bühlmann, 2006) as a solution for estimating . Friedman et al. (2007) proposed a method, known as the graphical Lasso (Least Absolute Shrinkage and Selection Operator) for finding the maximum of the norm penalized log-likelihood:

via a coordinate descent algorithm.

A Bayesian approach has been proposed for the regularized estimation of (Wang, 2012) that provides a natural structure for integrating *a priori* scientific knowledge and high-throughput molecular biology data such as un-targeted metabolomics data. Wang (2012) introduced a hierarchical Bayesian representation of the regular graphical Lasso as well as the adaptive graphical Lasso (Fan et al., 2009). The frequentist adaptive graphical Lasso was devised to link the magnitude of the penalty parameter to the norm of individual concentration matrix entries and proposes the following penalized likelihood for :

with weights where and are estimates for the concentration matrix entries, such as regular graphical Lasso estimates.

In the above model formulation for the density of conditional on the realizations of ,  represents the double exponential, or Laplace distribution, and the exponential distribution with scale parameter . The space of positive definite matrices is represented by . Finally, represents the normalizing constant so that is a proper probability distribution. For the non-adaptive Bayesian graphical Lasso, for all and , in other words the shrinkage parameter is not specific to each concentration matrix entry. Wang (2012) chooses a gamma prior for , that is , where and are hyperparameters and develops a data-augmented block Gibbs sampler for sampling from the posterior distribution of . Further, it is shown that the conditional distribution of the shrinkage parameter is then . Wang (2012) proposes to link the scale hyperparameter for the shrinkage to the norm of the current MCMC iteration estimate for each . We propose that instead, [LOH].

* 1. Generating informative priors from molecular structure

To generate informative shrinkage priors for the adaptive Bayesian graphical Lasso, atom pair descriptors are determined for each metabolite feature for which a compound identification has made to a fixed level of confidence such as MSI (Metabolomics Standards Initiative) level 1. Atom pair descriptors were determined as in Chen and Reynolds (2002) and originally described in Carhart, Smith, and Venkataraghavan (1985). An atom pair is defined by the properties of each atom in a pair as well as the length of the shortest bond path between the atoms. The atom properties considered are the identity of each atom, the number of *π* electrons associated with each atom and the number of other atoms that are bonded to each atom (excluding hydrogens). For metabolites with PubChem annotation, SDF files were accessed and used to generate an atom pair (AP) library. The AP library and AP representation of identified metabolites were generated using *ChemmineR* version 2.28(Cao, Charisi, Cheng, Jiang, & Girke, 2008). To determine structural similarity between compound using AP representation, the Tanimoto coefficient between pairs of metabolites was computed, which is defined as (Chen & Reynolds, 2002):

where represents the count of unique atom pairs indexed by for molecule . The Tanimoto dissimilarity is then .

After determining the Tanimoto dissimilarity between each pair of metabolites, the gamma distribution hyperpriors for the shrinkage parameter can be determined as

* 1. Posterior inference of model parameters

[LOH]

* 1. Efficacy analysis via simulation studies

To evaluate the efficacy of the proposed method, we employed extensive simulation studies. Two simulation study schema were employed. Under the first schema, a simple AR(1) process (autoregressive process of order 1) was simulated for representing a linear biological process in which the correlation between simulated metabolites decreased in tandem with decreasing structural similarity. Under the second schema, metabolomics data collected previously was utilized with simulated metabolites “spiked-in”. The spiked-in simulated metabolites were simulated via a hierarchical model in which structural similarity was simulated first, followed by abundance distributions in which the correlation between abundances increased with structural similarity.

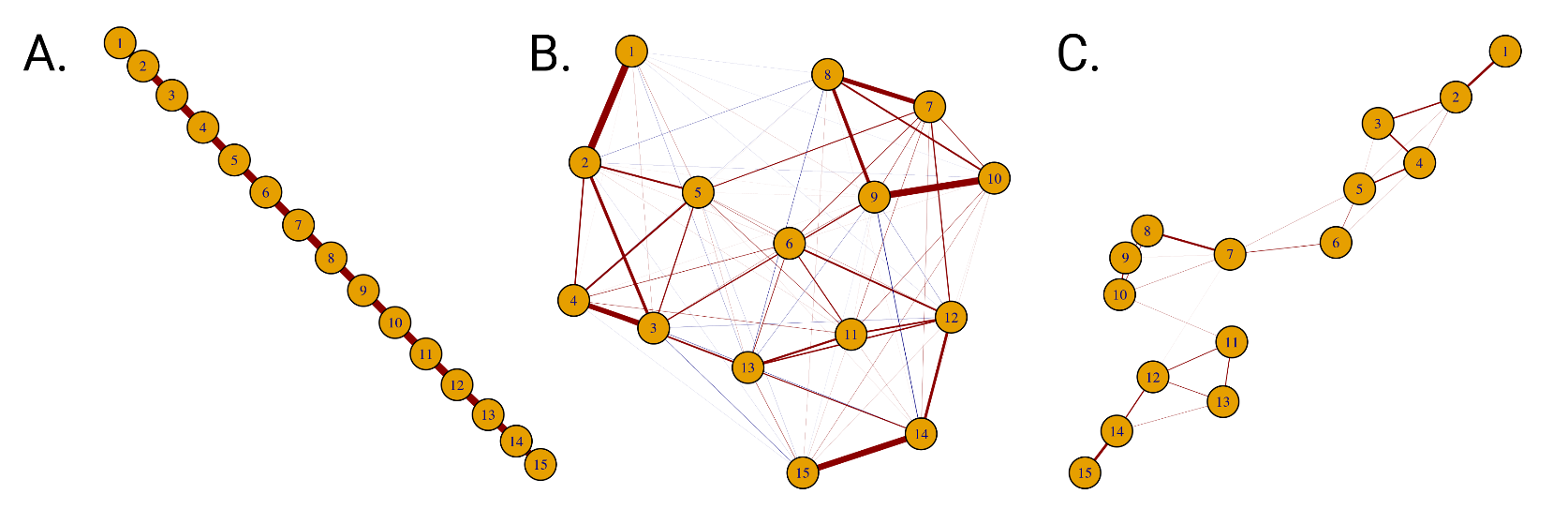
* 1. A plasma interactome for stable heart disease

Heart disease is the most prevalent cause of death globally (Benjamin et al., 2017). As a disease, heart disease does not represent a uniform condition, but rather a collection of diseases of varying etiologies (Kasper, 2015). Of particular interest in the study of coronary artery disease (CAD) is the elucidation of the precipitants of acute disease events such as myocardial infarction (Arbab-Zadeh & Fuster, 2015) or unstable angina [citation], as well as their downstream metabolic consequences. In order to determine metabolic perturbations associated with acute events, a reference model of the metabolic footprint of plasma from humans with stable heart disease is desirable.

In order to determine changes in the plasma metabolome associated with myocardial infarction (MI) characterized by thrombotic etiology versus non-thrombotic etiology, DeFilippis and colleagues recruited a human cohort as previously described (DeFilippis et al., 2016; DeFilippis et al., 2017; Trainor et al., 2017). [LOH].

1. Results
   1. Simulation studies

[LOH]



Blah blah blah.

* 1. Plasma interactome for stable heart disease

[LOH]

|  |  |
| --- | --- |
| A) Full heatmap | B) Zoomed-in perspective of two clusters    Figure: Heatmap showing structural similarity for the N compounds detected by chromatography coupled-mass spectrometry in the heart disease data with PubChem entries available. |

1. Discussion

References

Arbab-Zadeh, A., & Fuster, V. (2015). The Myth of the “Vulnerable Plaque”. *Journal of the American College of Cardiology, 65*(8), 846-855. doi:10.1016/j.jacc.2014.11.041

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *J. Mach. Learn. Res., 9*, 485-516.

Barupal, D. K., & Fiehn, O. (2017). Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports, 7*(1). doi:10.1038/s41598-017-15231-w

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., . . . Muntner, P. (2017). Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association. *Circulation, 135*(10), e146-e603. doi:10.1161/cir.0000000000000485

Buhl, S. L. (1993). On the Existence of Maximum Likelihood Estimators for Graphical Gaussian Models. *Scandinavian Journal of Statistics, 20*(3), 263-270.

Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., & Girke, T. (2008). ChemmineR: a compound mining framework for R. *Bioinformatics, 24*(15), 1733-1734. doi:10.1093/bioinformatics/btn307

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Modeling, 25*(2), 64-73. doi:10.1021/ci00046a002

Chen, X., & Reynolds, C. H. (2002). Performance of Similarity Measures in 2D Fragment-Based Similarity Searching:  Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Computer Sciences, 42*(6), 1407-1414. doi:10.1021/ci025531g

DeFilippis, A. P., Chernyavskiy, I., Amraotkar, A. R., Trainor, P. J., Kothari, S., Ismail, I., . . . Bhatnagar, A. (2016). Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction. *J Thromb Thrombolysis, 42*(1), 61-76. doi:10.1007/s11239-015-1292-5

DeFilippis, A. P., Trainor, P. J., Hill, B. G., Amraotkar, A. R., Rai, S. N., Hirsch, G. A., . . . Bhatnagar, A. (2017). Identification of a plasma metabolomic signature of thrombotic myocardial infarction that is distinct from non-thrombotic myocardial infarction and stable coronary artery disease. *PLoS One, 12*(4), e0175591. doi:10.1371/journal.pone.0175591

Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., . . . Kell, D. B. (2014). Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics, 11*(1), 9-26. doi:10.1007/s11306-014-0707-1

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics, 3*(2), 521-541. doi:10.1214/08-aoas215

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441. doi:10.1093/biostatistics/kxm045

Kasper, D. L. (2015). *Harrison's principles of internal medicine* (19th edition / editors, Dennis L. Kasper, MD, William Ellery Channing, Professor of Medicine, Professor of Microbiology, Department of Microbiology and Immunobiology, Harvard Medical School, Division of Infectious Diseases, Brigham and Women's Hospital, Boston, Massachusetts and five others . ed.). New York: McGraw Hill Education.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. Cambridge, MA: MIT Press.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics, 34*(3), 1436-1462. doi:10.1214/009053606000000281

Moriya, T., Satomi, Y., Murata, S., Sawada, H., & Kobayashi, H. (2017). Effect of gut microbiota on host whole metabolome. *Metabolomics, 13*(9). doi:10.1007/s11306-017-1240-9

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology, 26*(9), 1003-1010. doi:10.1038/nbt.1487

Southam, A. D., Lange, A., Al-Salhi, R., Hill, E. M., Tyler, C. R., & Viant, M. R. (2014). Distinguishing between the metabolome and xenobiotic exposome in environmental field samples analysed by direct-infusion mass spectrometry based metabolomics and lipidomics. *Metabolomics, 10*(6), 1050-1058. doi:10.1007/s11306-014-0693-3

Trainor, P. J., Hill, B. G., Carlisle, S. M., Rouchka, E. C., Rai, S. N., Bhatnagar, A., & DeFilippis, A. P. (2017). Systems characterization of differential plasma metabolome perturbations following thrombotic and non-thrombotic myocardial infarction. *Journal of Proteomics*. doi:10.1016/j.jprot.2017.03.014

Voet, D., Voet, J. G., & Pratt, C. W. (2013). *Fundamentals of biochemistry : life at the molecular level* (4th ed.). Hoboken, NJ: Wiley.

Wang, H. (2012). Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis, 7*(4), 867-886. doi:10.1214/12-ba729

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika, 94*(1), 19-35. doi:10.1093/biomet/asm018