BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–6

Discovering main genetic effects with LABNet LAsso-Based Network Inference

Francesco Gadaleta [1,*], Co-Author [2] and Co-Author [2*]

[1]University of Liege, Montefiore Institute, Grande Traverse 10 Liege
[2]University of Liege, Montefiore Institute, Grande Traverse 10 Liege

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation: motivation
Results: results
Availability: availability of source code
Contact: francesco.gadaleta@ulg.ac.be

1 INTRODUCTION

Any biological system is characterised by interactions between components. The study of these interactions is essential to understanding the mechanisms that regulate complex diseases and to unravel the functional aspects of genetic compounds. In several fields of research, from social to telecommunication and biology, system interactions are more and more often represented by graphical models (26, 3, 15). Generally speaking, those are defined by a set of nodes and a set of edges. Each node usually represents a specific biological component that interacts with others to perform specific functions. Edges may have several meanings, depending on the type of interactions it represents, such as similarity, causality, distance, etc. In the field of network theory and genetics, the nodes of a graph usually represent genes and the edges represent the interactions among these genes. Consequently, a network graph of genetic interactions is a viable way to visualise clusters, modules or pathways, according to the purpose of the analysis. It is known that genes act in clusters and their individual effects tend to be characterised by a smaller magnitude within the system as a whole (Refs.) Graphical models facilitate the detection of the main genetic effects. Moreover, pathways of genes become more visible to the researcher who investigates the data, giving a more complete explanation of the biological function that the pathway itself performs. One viable way to represent the interactions of the nodes of a graph - and consequently the topology of the resulting network - might be represented by the adjacency matrix $\beta = \beta_{ij}$. The values of each cell (i, j) in the adjacency matrix represent the magnitude of the interaction between two nodes, whereas zeros are equivalent to absence of interaction between node i and node j. Specifically to the field of computational biology, the problem of learning the structure of a graph starting from the expression profile of a number of genes is a challenging task due to several reasons that are already known to the research community. The presence of noise in the measurements, the high dimensionality of genetic data and multicollinearity are just a few that are mentioned.

Despite active research in the field of high-density oligonucleotide arrays, noise still represents a consistent source of error. Any analysis subsequent to the measurement of a subset of genes should take into consideration the artifacts introduced by noise or by the computational method that mitigates it (23, 14).

It is very common to perform analyses over high-dimensional genetic data in which the number of genes p is much larger than the number of individuals n. Discovering the interactions between genes in such cases is extremely difficult. Methods that rely on the dispersion matrix whose elements (i, j) represent the covariance between the expressions of two genes are also affected by high-dimensionality. As a matter of fact, the maximum likelihood estimator cannot provide a reliable estimate of the covariance matrix for $n << p$ problems.

Moreover, gene expression profiles are affected by the presence of multicollinearity (19, 10), namely two or more genes or genetic compounds can be highly correlated. The presence of multicollinearity can influence the performance of regression-based models. The regression coefficient of a predictor variable's importance on the target variable has the tendency to lose precision with respect to the case in which the same genes were uncorrelated. From a biological perspective, it is broadly recognised that strong genetic correlations are utterly frequent in microarray data and that, in contrast, complete independence between any two gene expression measurements is rare (refs). Therefore, it is expected that functionally related genes are somehow co-expressed. This phenomenon can be expressed by assuming the presence of high correlation for a subset of genes in the dataset under study. Moreover, as the gene sets to be tested are usually chosen on the basis of functional annotation, it should be expected that many of the tested genes might be, in fact, correlated (11). Variable selection methods are even more sensitive to the presence of multicollinearity as they tend to select only one or few highly correlated variables.

A computational approach that can deal with the aforementioned issues affecting genetic interactions belongs to the family of penalised linear regression. The core idea consists in reducing the number of meaningful interactions with each gene, in order to build sparser networks. Penalised linear regression has been investigated in seminal work reported in (25, 18, 8, 17), in which each variable is considered response and the remaining ones are independent covariates. In the aforementioned work, bootstrapping

[*] to whom correspondence should be addressed

has been extensively used to improve the stability of the predicted interactions. Unfortunately, the nature of genetic data and the presence of highly correlated variables play a detrimental role that can affect the overall reliability of these methods. Moreover, Lasso-based regression procedures are known to deal poorly with highly correlated variables since only one in a group of multi correlated covariates is selected. Bootstrapping does not seem to mitigate such a troublesome condition.

In this paper, we consider the use of Lasso penalised regression as a starting point. We subsequently rely on a permutation-based approach in order to increase the significance of the discovered interactions.

In Section 2, we describe the method in detail. In Section 3, we measure the performance of our approach on simulated genetic networks of different size. Conclusion and future developments are drawn in Section 5.

2 APPROACH

Microarray data are usually represented by the matrix $X = x_{ij}$ of the expression profiles of i genes and j individuals or sample tissues. The main goal of the approach described in the current section is to infer the network topology that regulates the main interactions of the genes under investigation. Network modelling has proven to be an effective approach in computational biology due to the straightforward representation of conditional dependency between variables (4, 24). Generally speaking, a network model is formed by a set of vertices G, representing the genes in our specific case, and a set of edges E representing the pairwise interactions. The existence of edge (i, j) represents the conditional dependency between gene i and gene j. If such an edge is not present, the two genes are considered conditionally independent, namely $(G_i \perp G_j)|G_k, \forall k = 1...i$. In the specific application described in this paper, we aim at finding the best set of neighbours associated to each gene. We translate the biological meaning of association into the terms specified by regression analysis. The expression value of a gene is related to one or more independent variables, represented by other genes of the dataset.

Regardless the number of mathematical models that have been considered for inferring the association between variables in genetics, linear regression is a type of analysis that has found large consensus in the field of computational biology due to its simplicity of modelling (1, 6). One limitation of linear regression methods prevails in assuming a linear dependency between variables, a hypothesis that does not always apply in biology. One stratagem to overcome such a limitation consists of splitting the problem of learning the topology of the entire network of genes into a number of smaller linear problems. This can be achieved by regressing each covariate against all the remaining ones. Such a strategy, which has been used first in the work reported in (18) makes the assumption of linearity more suitable to the analysis of biological data. Assuming the presence of linearities on a local scale is a much more convincing and appropriate conjecture that might find an application to data from genomics and proteomics. Another limitation that researchers have to take into account appears in the case of high-dimensional data. In such a scenario, the number of genes is usually some orders of magnitude larger than the number of the individuals. Without loss of generality, the problem of inferring the conditional independence between variables is equivalent to the problem of computing the sample covariance matrix of the interactions among variables. In the case of high dimensional data, as well as in a more relaxed case in which the number of individuals has a similar order of magnitude as the number of genes, the inverse of the sample covariance matrix does not exist (5). Therefore, all approaches that are based on the maximum likelihood estimator do not provide a good approximation of the covariance matrix, making the solution of the interaction problem numerically unstable and the discovered interactions unreliable. (FIXME to intro)

Penalised regression has been considered as a way to circumvent such limitation due to the presence of a penalty factor that encourages sparsity of the final network. Specifically, lasso is one such regression method that converts the problem of estimating the covariance matrix into an optimisation problem in which a convex function, applied to each variable, is minimised.

Given X_i the expression of gene i and the expression profiles of the remaining genes (referred to as X, for simplicity), the lasso-based estimate consists of providing a solution for Equation 1

$$\hat{\Theta}^{a,\lambda} = \underset{\text{s. t. } \Theta:\Theta_a=0}{\operatorname{argmin}} \left(\frac{1}{n}\|X_i - X\Theta\|_2^2 + \lambda\|\Theta\|_1\right) \qquad (1)$$

The vector of regression coefficients Θ determines the conditional independence structure between variables. The l_1-norm of the coefficient vector tends to shrink the coefficients of some variables to zero, removing them from the set of selected variables associated to the response, as extensively explained in (25). The right choice of the shrinkage factor λ is crucial to controlling the rate of false positives and false negatives. Regardless the number of approaches to approximate the optimal λ, reported in (27, 7, 13), a reliable estimate that is widely used in practice is provided by cross-validation (9). We use a 3-fold cross validation approach and estimate $\hat{\lambda_{cv}}$ from a subset of the data. Cross-validation can be a time consuming task especially when applied to datasets with a high number of covariates. Therefore, we estimate the shrinkage factor that minimises the expected generalisation error, for a grid of λ values, on the 10% of the total number of genes. The R package *glmnet* has been used to provide such an estimate.

The method we describe in this paper is a two-step approach that recursively performs the regression of Equation 1 of each gene, considered as response, with respect to all remaining genes, considered as independent variables. The response gene is not included in the set of independent variables. Regardless biological evidence that supports the existence of self interactions and positive/negative feedback loops within regulatory networks (12, 2, 21), those are not considered here, in order to avoid complex interactions and simplify as much as possible the inferred network topology.

In step 1, the set S of variables associated with the current response gene is selected. We use a lasso method that does not fit the intercept. As explained, the choice of the optimal λ occurs prior to this stage.

In step 2, we use a permutation-based approach to assess the significance of the associated edges detected in step 1. The values of response variable are permuted a number of times. For each permutation we count how many times each variable within the set S of selected genes has been selected again. At the end of the permutation test, the variables with the smallest counter are selected as the best candidate variables associated with the current response gene. This approach is supported by the fact that after permuting the response variable, the genes selected at step 1 should be no longer associated and therefore should be considered as selected by chance.

Algorithm 1 Variable selection and permutation-based stability test

```
1: procedure LASSO2NET($X_i, X, B, fanout, best$)
2:     $fit \leftarrow lasso.cv(X_i, X)$
3:     $\lambda_{cv} \leftarrow fit.lambda$
4:     $S \leftarrow fit.coeffs$
5:     $S \leftarrow sort(S, decreasing)[1 : best]$
6:     while $r < B$ do
7:         $X_i^{perm} \leftarrow permute(X_i)$
8:         $permfit \leftarrow lasso(X_i^{perm}, X, \lambda_{cv})$
9:         $update(counter[S])$ update counters of selected variables
10:        $r \leftarrow r + 1$
11:    end while
12:    $sel \leftarrow sort(counter[S], increase)[1 : fanout]$ order and select first fanout
13:    return $sel$
14: end procedure
```

The procedure described above is summarised in Algorithm 1. It selects the $best$ number of genes associated to the current response. Namely, the vector of the associated genes is sorted in decreasing order and the first $best$ are selected (*line 5*). The parameter $best$ can be tuned in order to select a variable number of strong genetic effects according to the type of disease under investigation and the dataset at the researcher's disposal which, in turn, might determine the amount of significant genetic compounds to be considered for further analysis. At each permutation, the counters of the selected variables are updated (*line 9*) and after B permutations the first $fanout$ genes are selected. These variables represent the most stable genes associated with the response variable (*line 12*).

The algorithm described above finds a solution of Equation 1 for each response variable. Subsequently, it finds the most stable non-zero regression coefficients associated to each gene. When the described procedure is performed on the entire set of genes, an adjacency matrix can be built directly from the counters of selected variables. The aforementioned adjacency matrix represents the network topology of the inferred network of interactions. Since we are interested in discovering interactions we convert the non-zero values to 1 in the adjacency matrix, in order to denote the presence of an edge in the graph. As one would expect, the method does not guarantee the adjacency matrix to be symmetric. A symmetrisation procedure would be required before further analysis or visualisation of the predicted network. We symmetrise the adjacency matrix by the OR rule which considers two variables as associated if only one of the two variables is associated with the other. Namely, $gene_i \leftrightarrow gene_k \iff gene_i \rightarrow gene_k \vee gene_i \leftarrow gene_k$.

The main goal of the work described here is to detect the structure of the network of the main genetic interactions, passing over the causality interactions. Detecting an interaction between two variables is sufficient to build the overall structure.

3 RESULTS

In order to evaluate the performance of the method described in Section 2, we need to compare the predicted network to the real network that generated the data. In real biological applications this procedure is usually not possible, due to the fact that the real network is, in fact, unknown. In the specific case described thus far, we take advantage of synthetic data that make such a performance evaluation practical.

Since our algorithm is designed to analyse gene expression profiles, we generate synthetic microarray data with the Gene Net Weaver software package (GNW) (20). The aim of GNW is to generate in-silico networks extracting modules from biological networks. These networks are simulated to produce gene expression data (steady states or time series) (16). The aforementioned framework can be used to evaluate the performance of our inference method by comparing the predicted network with the golden standard network that generated the dataset.

We perform the approach described above on synthetic microarray data generated from simulated networks of 50 and 200 nodes. The parameters used in our experiments are summarised in Table 1.

cross-validation	3-fold on 10% genes
best	80% genes
fanout	1
B	0-500

Table 1. Parameters of LABnet for both 50-node and 200-node networks

A set of networks has been inferred with an increasing number of permutations. One important characteristic that arises from our experiments consists in the fact that by increasing the number of permutations, the connectivity of the network is increased proportionally (Figure 3). Within the same figure it is shown that the number of false positives is limited regardless the number of predicted edges and permutations. We measure the connectivity of the network by counting the number of the predicted edges.

In Figure 6, the false positive rate, usually referred to as accuracy, is not affected by the number of permutations but by the number of predicted edges which increases accordingly (as shown in Figure 4).

In Figure 4 the reader can also notice that by increasing the number of permutations the false negatives, or missed edges, tend to decrease. Since higher connected networks are usually affected by an increasing number of false negatives, we consider the method described above a promising approach with potential benefits to the analysis of large genetic networks.

Another outcome that is worth mentioning regards the true positive rate which follows the same trend of the number of permutations (Figure 5). Within the same figure the Matthew Correlation Coefficient (MCC) is also reported. The MCC is a correlation coefficient between the observed and the predicted classification (presence or absence of edge) and takes into account both true and false negatives of the overall predicted network.

The MCC has been calculated as

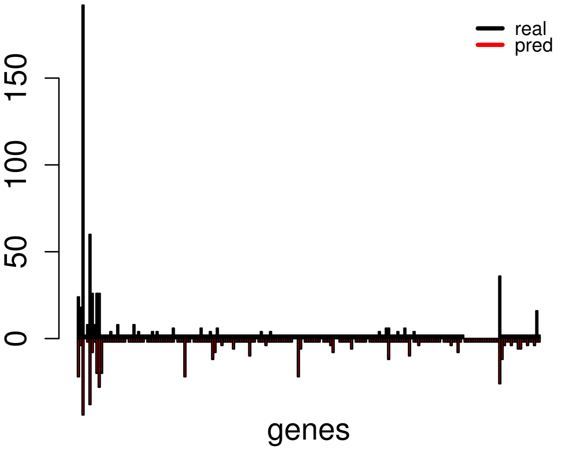

Fig. 1: Degree correlation across real and predicted nodes

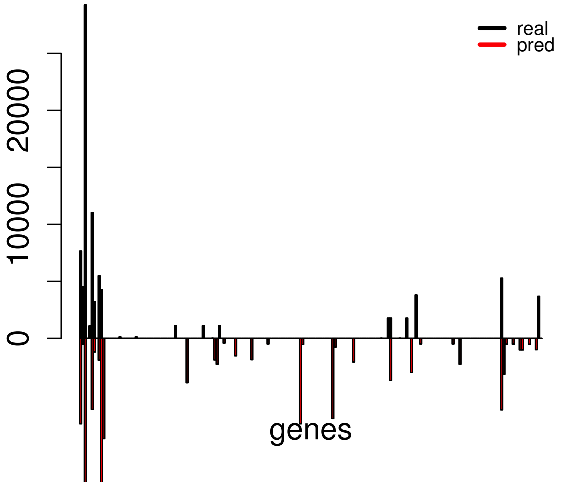

Fig. 2: Betweenness correlation across real and predicted nodes

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{\begin{array}{l}(TP + FP) \times (TP + FN) \\ \quad \times (TN + FP) \times (TN + FN)\end{array}}} \tag{2}$$

In order to compare the predicted network to the golden standard using a measure that takes into account the global structures of the graphs, two global measures have been provided, such as the degree correlation DC and the betweenness correlation BC.

DC is the correlation between the vector of the degrees of all genes in the real network and those of the predicted network. It is calculated as

$$DC = cor(\bar{d}_{gold}, \bar{d}_{pred}) \tag{3}$$

where d is the i-dimensional vector containing the degree of each gene.

Similarly, the betweenness correlation BC is the correlation between the same two vectors where the degree has been replaced by the betweenness centrality measure.

BC is calculated as

$$BC = cor(\bar{b}_{gold}, \bar{b}_{pred}) \tag{4}$$

where $b = b(i) = \sum_{q \neq i \neq r} \frac{\sigma_{qr}(i)}{\sigma_{qr}}, \forall i$, σ_{qr} is the total number of shortest paths from node q to node r and $\sigma_{qr}(i)$ is the number of shortest paths from q to r that pass through gene i.

Betweenness centrality is, in our opinion, more helpful than simple connectivity. This measure is a direct indicator of how connected the node is and its importance with respect to the global network topology.

As it can be seen in Figure 1 and Figure 2 there is a strong degree correlation (0.83) and betweenness correlation (0.86) between the nodes of the predicted and real networks. The two measures and the aforementioned strong correlations support the evidence that the topology of the real network is conserved within the predicted network, following the same power law degree distribution of the original network that generated the data. Due to the fact that GNW generates network from real life templates, we expect similar results in real biological data.

4 DISCUSSION

Despite the encouraging results of the method we described in the previous sections and performed on simulated genetic networks, we address some limitations we intend to investigate in the near future.

As already stated, genetic data are usually affected by measurement noise and a high number of variables collected from different datasets such as gene expression profiles, SNPs, methylation and clinical data.

The curse of dimensionality can set a limit on the number of permutations to perform. Due to the fact that our method relies on permuting each response variable, in order to increase the stability of the discovered interactions, the overall performance is directly affected by the total number of genes in the dataset. We are investigating possible solutions to mitigate the curse of dimensionality by limiting the discovery of interactions to highly connected genes. This strategy would detect the local structure around genes usually referred to as network hubs. We do not interpret this fact as a limitation since biological networks usually manifest a scale free topology, in which only few nodes are highly connected to the rest of the graph (3, 22).

The variable selection procedure consistently depends on the value of the shrinkage factor λ, estimated on a subset of the covariates. Obviously, it might occur a prior exclusion of significant genes from further analyses in the case of a too restrictive shrinkage factor. An alleviation to this risk (which can directly determine the

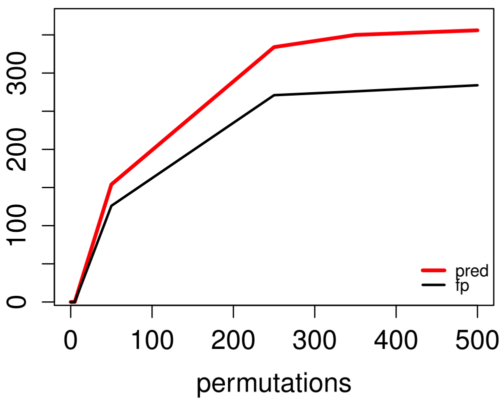

Fig. 3: Number of predicted edges and false positives vs. number of permutations

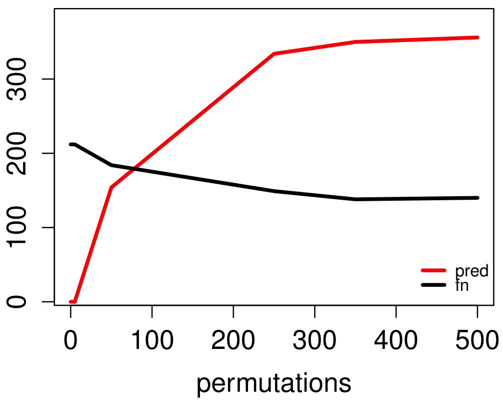

Fig. 4: Number of predicted edges and false negatives vs. number of permutations

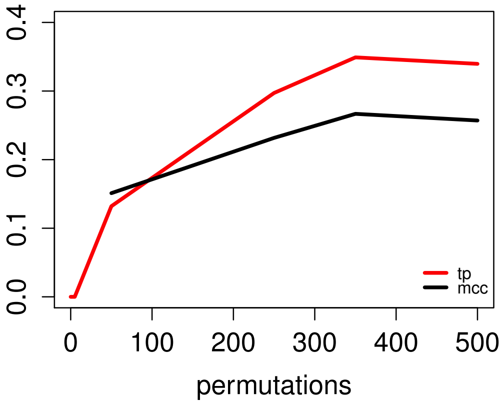

Fig. 5: True positives and Matthew Correlation Coefficient vs. number of permutations

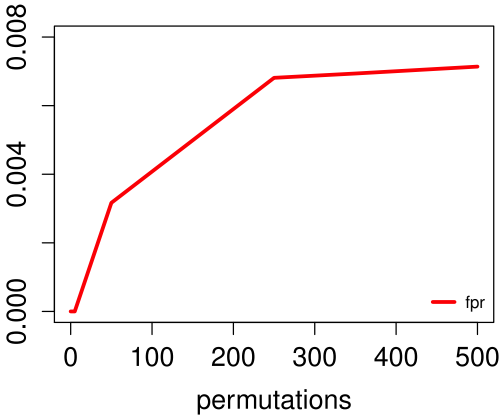

Fig. 6: False positive rate vs. number of permutations

false negative rate) consists in replacing the pure lasso penalty with an elastic net procedure of the type

$$\hat{\Theta}^{a,\lambda} = \underset{\text{s. t. }\Theta:\Theta_a=0}{\text{argmin}} \; (\frac{1}{n}||X_i - X\Theta||_2^2 + \alpha||\Theta||_1 + (1-\alpha)||\Theta||^2) \quad (5)$$

In that case it would be necessary to estimate an additional parameter α. To the other extreme, a pure ridge-regression procedure would not benefit from the permutation-based stability test, due to the fact that our method ignores the regression coefficients and selects the subset of genes with the best permutation score.

Another aspect we intend to probe regards the direction of the interactions. In our analysis we ignore the direction of each edge in the graph. A relaxation of the problem of learning the network topology consists in considering the interaction $i \rightarrow j$ equivalent to the interaction $j \rightarrow i$. Although this simplification makes the construction of the overall network consistently easier, it might lead to inconsistencies from a biological perspective. As a matter of fact, gene regulations are known to have a direction, usually referred to as activation and inhibition (FIXME refs). Learning the directionality of network edges represents an additional complexity that is plausible to deal with in the presence of a large number of samples or by integrating complementary data sources of known interactions. Therefore, the need for integrating different data sources is twofold: data integration can increase the stability of all discovered interactions and their direction and, specifically to our method, it can reduce the number of required permutations per gene. We believe that data integration can consistently improve the overall performance of the described approach.

We endorse our approach to be deployed in a pipeline in order to 1) analyse different data sources 2) build the local network from each dataset 3) increase the stability of predicted interactions by permutation and 4) integrate each singular network into a more stable and complete graph.

5 CONCLUSION

LABNet is a lasso-based approach to detect main genetic interactions from gene expression profiles. Penalised regression in concert with the permutation-based approach determines whether the predicted interactions are stable across experiments. The higher number of permutations not only improves the sensitivity of the method by reducing the number of false negatives, but it also determines the number of predicted edges. This does not seem to affect consistently the false positive rate. Due to the features that we have described and the promising results on synthetic data, the approach is a good candidate to investigate further and expand for the analysis of heterogeneous genetic data.

ACKNOWLEDGEMENT

Thank you thank you thank you Thank you thank you thank you Thank you thank you thank you Thank you thank you thank you Thank you thank you thank you Thank you thank you thank you

Funding: Text Text Text Text Text Text Text Text.

REFERENCES

[1]Carl A Anderson, Allan F McRae, and Peter M Visscher.

[2]Roi Avraham and Yosef Yarden. Feedback regulation of egfr signalling: decision making by early and delayed loops. *Nature Reviews Molecular Cell Biology*, 12(2):104–117, 2011.

[3]Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Sci. Am.*, 288(5):50–59, 2003.

[4]Albert-László L. Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, January 2011.

[5]S Buhl. On the existence of maximum likelihood estimators for graphical gaussian models, 1993.

[6]Petra Bkov. Linear regression in genetic association studies. *PLoS ONE*, 8(2):e56976, 02 2013.

[7]Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 04 2004.

[8]Michael Finegold and Mathias Drton. Robust graphical modeling of gene networks using classical and alternative T-distributions. *The Annals of Applied Statistics*, 5(2A):1057–1080, August 2011.

[9]Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[10]Akhil Garg and Kang Tai. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control*, 18(4), 2013.

[11]J J Goeman and P Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, April 2007.

[12]Jennifer Hallinan and Paul T. Jackway. Network motifs, feedback loops and the dynamics of genetic regulatory networks. In *CIBCB*, pp. 90–96. IEEE, 2005.

[13]Kei Hirose, Shohei Tateishi, and Sadanori Konishi. Tuning parameter selection in sparse regression modeling. *Comput. Stat. Data Anal.*, 59:28–40, March 2013.

[14]Klebanov L and Yakovlev. How high is the level of technical noise in microarray data? 2007.

[15]Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, DREAM5 Consortium, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, August 2012.

[16]Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. *Journal of Computational Biology*, 16(2):229–239, 2009. WingX.

[17]Nicolai Meinshausen and Peter Bhlmann. Stability selection. Technical report.

[18]Nicolai Meinshausen and Peter Bhlmann. High dimensional graphs and variable selection with the lasso. *ANNALS OF STATISTICS*, 34(3):1436–1462, 2006.

[19]V M Roso, F S Schenkel, S P Miller, and L R Schaeffer. Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. *J Anim Sci*, 83(8):1788–800, 2005.

[20]Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011. wingx.

[21]Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(S-6), 2007.

[22]Sandy Shaw. Evidence of scale-free topology and dynamics in gene regulatory networks. In Antony Satyadas and Sergiu M. Dascalu, editors, *IASSE*, pp. 37–40. ISCA, 2003.

[23]E.G. Sifakis, A. Prentza, D. Koutsouris, and A.A. Chatziioannou. Evaluating the effect of various background correction methods regarding noise reduction, in two-channel microarray data. *Computers in Biology and Medicine*, 42(1):19 – 29, 2012.

[24]Edwin K. Silverman and Joseph Loscalzo. Network medicine approaches to the genetics of complex diseases. *Discovery medicine*, 14(75):143–152, August 2012.

[25]Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[26]Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome Networks and Human Disease. *Cell*, 144(6):986–998, March 2011.

[27]Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006.