# A Comparative Study of Algorithms for Learning Causal Genotype-Phenotype Networks


Adèle H. Ribeiro, Júlia M. P. Soler, and André Fujita

adele@ime.usp.br, pavan@ime.usp.br, and fujita@ime.usp.br

Institute of Mathematics and Statistics - University of São Paulo


## I) Abstract


A challenging task in biomedical research is to understand precisely the **complex network of causal associations among phenotypes and outcomes**. Experimental studies such as clinical trials are the most trustworthy method of causality assessment. However, it may be unfeasible to carry out randomized experiments to discover all possible causal relationships when the number of variables is large.

In systems genetics, causal inference is supported by **Mendelian randomization**, which provides a natural randomization process where genotypes, rather than treatments, are randomly allocated to individuals. Furthermore, genetic variants robustly associated with phenotypes can be seen as **instrumental variables**, allowing inferences on the causal relation between phenotypes and outcomes [1].

In this work, we made a comparative study among four recent algorithms that use genetic variants as instrumental variables for learning the structure of a genotype-phenotype network, namely, (i) **QTL-directed Dependency Graph (QDG)** [2], (ii) **QTL-driven phenotype network (QTLnet)** [3], (iii) **Sparsity-aware Maximum Likelihood (SML)** [4], and (iv) **QTL+Phenotype Supervised Orientation (QPSO)** [5].

These algorithms are similar in the sense that they use QTL information to determine the causal direction among phenotypes. However, they were designed under different assumptions and therefore some may be more suitable than others for a particular biological application.

By **simulation studies**, we investigated advantages and limitations of these methodologies, under different configurations. Finally, we applied the algorithms to real data involving **cardiovascular phenotypes of F2 rats** and compared the inferred causal networks.

The QDG, SML and QPSO algorithms require as input an association network among phenotypes, which can be estimated by the PC skeleton algorithm available in the R/pcalg package, and a genetic mapping, which can be performed by using the R/QTL package. The QTLnet algorithm infers phenotype and genotype network simultaneously. The SML algorithm is fully formulated as a Structural Equation Model (SEM), but the others are based on Probabilistic Graphical Models. Thus, except for the SML algorithm, the magnitude of the causal effects must be estimated by fitting a SEM with the inferred structure.

The QDG and QTLnet algorithms are implemented in the R/QTLnet package. The QPSO algorithm was provided by the authors, and the SML algorithm is available as supporting information in the online version of its paper. Both are implemented in Matlab.


## II) The Algorithms

Main differences among the algorithms, which are mainly related to the ability in discovering networks with the following properties: QTLs with pleiotropic effects, traits associated with multiple QTLs, acyclic or acyclic structure, and reciprocal interactions.

| | QDG | QTLnet | SML | QPSO |
|---|---|---|---|---|
| Input | | | | |
| An phenotype association network | ✓ | | | ✓ |
| A genetic map in which each phenotype | | | | |
| must be associated with a distinct genetic variant | ✓ | | ✓ | |
| can be associated with multiple genetic variants | ✓ | | ✓* | ✓ |
| can share a genetic variant with other phenotypes | ✓ | | ✓* | ✓ |
| Assumptions | | | | |
| Phenotypes are normally distributed | | ✓ | ✓ | ✓ |
| The phenotype network is acyclic | | ✓ | | ✓ |
| Features | | | | |
| Discovers acyclic structures | ✓ | ✓ | ✓ | ✓ |
| Discovers cyclic structures | ✓ | | ✓ | |
| Discovers reciprocal associations | | | ✓ | |
| Performs multiple genetic (QTL) mapping | | ✓ | | |
| Estimates the magnitude of the causal effects | | | ✓ | |
| Output | | | | |
| The most likely structure network | ✓ | ✓ | ✓ | ✓ |
| A list of the most likely solutions | ✓ | ✓ | | |
| A goodness-of-fit measure of the solution | ✓ | ✓ | | ✓ |

(✓) indicates the presence of the feature in the corresponding algorithm.
(*) indicates that the feature could not be verified by our simulation studies.

## III) Simulation Studies

In our simulation studies, we adopted the following SEM that represents the associations among $p$ phenotypes and $q$ genetic variants for $n$ individuals:

$$\mathbf{Y} = \mathbf{M} + \mathbf{P}\mathbf{Y} + \mathbf{Q}\mathbf{X} + \mathbf{E},$$

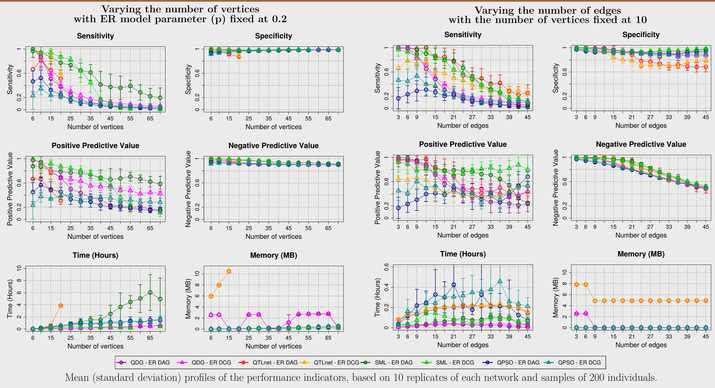
where,

- $\mathbf{Y}$ is a $p \times n$ matrix with $y_{ij}$ representing the $i$-th phenotype for the $j$-th individual;
- $\mathbf{M} = \mathbf{1}' \otimes \mu$ is a $p \times n$ matrix resulting from the Kronecker product between the transposed $n$-dimensional unity vector and the vector $\mu$ with the expected values of each phenotype;
- $\mathbf{Q}$ is a $p \times q$ matrix and $q_{ij}$ is the effect of the $j$-th genetic variant on the $i$-th phenotype;
- $\mathbf{X}$ is a $q \times n$ matrix and $x_{ij}$ is the genotype of the $i$-th genetic variant in the $j$-th individual;
- $\mathbf{P}$ is a $p \times p$ and $p_{ij}$ is the direct effect of the $j$-th phenotype on the $i$-th phenotype; and
- $\mathbf{E}$ is a $p \times n$ and $e_{ij}$ is the error term of the $i$-th phenotype for the $j$-th individual.

The phenotypes in $\mathbf{Y}$ were simulated from a zero $\mu$ vector, error terms $e_{ij} \sim \mathcal{N}(0, 1)$, a genetic architecture $\mathbf{Q}$ with additive effects of 0.8, and a matrix $\mathbf{P}$ with all causal effects among phenotypes equal to 0.8. The QTL genotypes in $\mathbf{X}$ were generated from an F2 intercross using the R/QTL package, so that QTLs of each simulated network are unlinked and in linkage equilibrium.

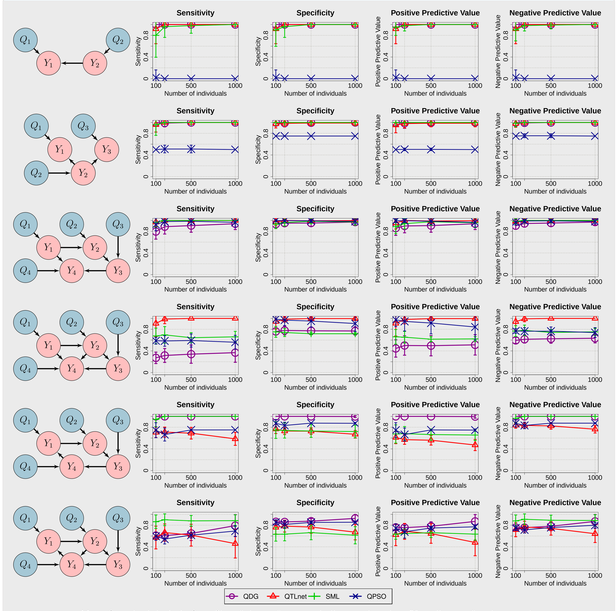
The phenotype network structure is specified according to the study. By using the R/igraph package, we simulated phenotype networks with structures corresponding to **Erdős-Rényi (ER) random graphs**. The performance and accuracy of the algorithms were evaluated as the number of phenotypes increases (by varying the number of vertices from 5 to 70 and fixing the probability of an edge existing between two phenotypes at 0.2), and as the structure becomes less sparse (by varying the number of edges from 3 to 45 and fixing the number of phenotypes at 10). In these cases, every phenotype is affected by an unique and distinct QTL. Also, **specific phenotype network structures and genetic architectures** were simulated to emphasize problematic scenarios.

## IV) Comparing Erdős-Rényi (ER) Random Phenotype Networks



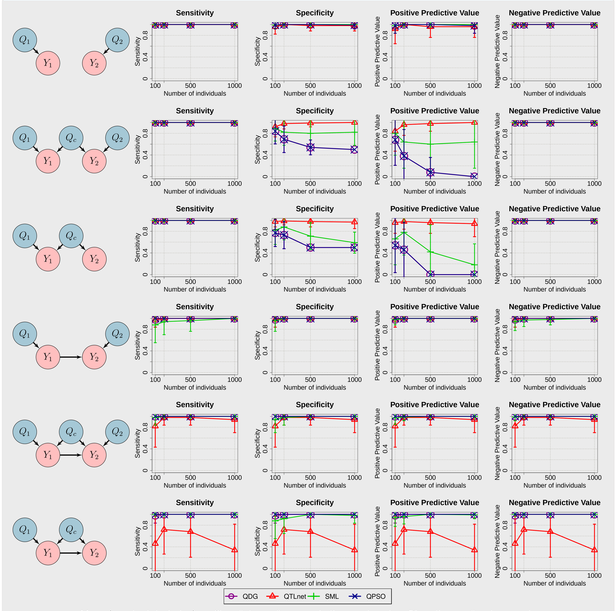

Mean (standard deviation) profiles of the performance indicators, based on 10 replicates of each network and samples of 200 individuals.

## V) Comparing Specific Phenotype Network Structures




Mean (standard deviation) profiles of the performance indicators, based on 50 replicates of each network.

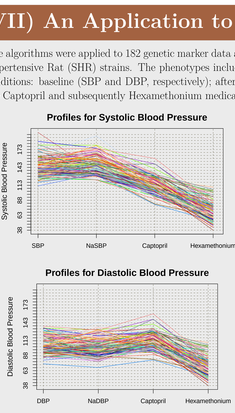
## VI) Comparing Specific Genetic Architectures




Mean (standard deviation) profiles of the performance indicators, based on 50 replicates of each network.

## VII) An Application to F2 Rat Data

The algorithms were applied to 182 genetic marker data and 7 phenotypic data of 156 F2 rats from a cross between Brown-Norway (BN) and Spontaneously Hypertensive Rat (SHR) strains. The phenotypes include the weight in grams, and the systolic and diastolic blood pressure (in mmHg) in the following conditions: baseline (SBP and DBP, respectively); after 13 days into a high sodium diet (NaSBP and NaDBP, respectively); and after administration of the Captopril and subsequently Hexamethonium medication [6].



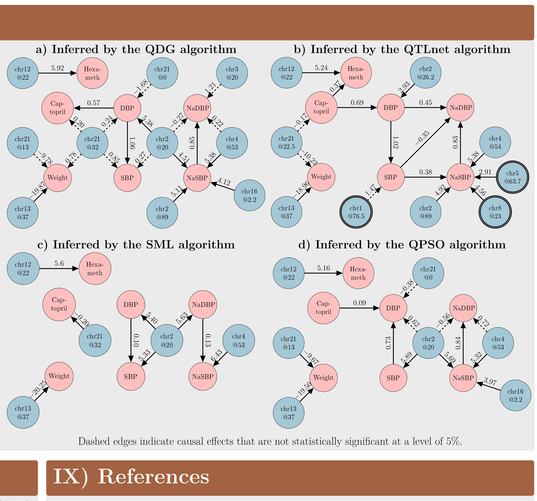

Correlation matrix of the seven phenotypic variables.

| | Weight | SBP | DBP | NaSBP | NaDBP | Captopril |
|---|---|---|---|---|---|---|
| SBP | 0.17 | | | | | |
| DBP | 0.12 | **0.90** | | | | |
| NaSBP | 0.08 | 0.47 | 0.50 | | | |
| NaDBP | 0.10 | 0.42 | 0.52 | **0.90** | | |
| Captopril | 0.14 | 0.64 | **0.66** | 0.43 | 0.44 | |
| Hexamethonium | 0.10 | 0.24 | 0.27 | 0.16 | 0.11 | 0.33 |

LOD score statistics of the QTLs identified by a Haley–Knott regression interval mapping, implemented in the R/QTL package, with a threshold of 2.5.

| chr | position (cM) | Marker 1 (position) | Marker 2 (position) | Weight | SBP | DBP | NaSBP | NaDBP | Captopril | Hexamethonium |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 20.20 | R5129 (18.45) | R155 (29.45) | - | 2.83 | 3.11 | 4.48 | 2.70 | - | - |
| 2 | 89.20 | FIBG2 (75.13) | R520 (96.53) | - | - | - | 3.05 | - | - | - |
| 3 | 19.94 | R6954 (12.40) | R796 (19.94) | - | - | - | - | 2.60 | - | - |
| 4 | 53.00 | R514 (49.83) | TGFAA (54.37) | - | - | - | 6.44 | 6.25 | - | - |
| 12 | 22.60 | R1053 (12.59) | PAH1AA (22.70) | - | - | - | - | - | - | 2.69 |
| 13 | 37.00 | R578 (30.60) | R267 (40.75) | 3.07 | - | - | - | - | - | - |
| 16 | 2.16 | R762 (1.16) | R220 (5.71) | - | - | - | 3.15 | - | - | - |
| 21 | 0.00 | R5108/AUR2A (0.00) | PFKFB (0.08) | - | - | 2.72 | - | - | - | - |
| 21 | 13.00 | AR (12.70) | R41/INHBBN (14.89) | 7.62 | - | - | - | - | - | - |
| 21 | 32.00 | PLANH (22.7) | R5660 (32.00) | 4.64 | 2.52 | 2.85 | - | - | 4.04 | - |

The QTLs identified only by the QTLnet algorithm, using a LOD score threshold of 2.5, are illustrated in double-lined nodes and are consistent with the literature.




Dashed edges indicate causal effects that are not statistically significant at a level of 5%.

The association network among phenotypes provided to the QDG and QPSO algorithm was estimated with a significance level of 1% by the PC skeleton algorithm. Except for the SML algorithm, the effects sizes were estimated using the R/lavaan package.

## VIII) Conclusions

Due to the joint inference of the QTL mapping and the phenotype network structure, the QTLnet algorithm had great QTL detection power and robustness to perturbations in the genetic architecture. However, it can only infer acyclic structure networks with up to 20 phenotypes.

The QDG algorithm could correctly infer acyclic and cyclic structures, but only of small networks. Since it provides LOD score statistics of each edge direction, it would be possible to eliminate the spurious edges that are associated with scores very close to zero.

The QPSO algorithm, unlike the QDG algorithm, can infer network structures even when some phenotypes have not QTLs associated. However, it is the most time consuming algorithm and had the worst accuracy. It was unexpected that simple structures, such as the reactive and fork, could not be correctly inferred, and that the solution depends on the order in which the variables are given as input to the algorithm.

The SML algorithm had good performance in inferring acyclic and cyclic structures, even in networks with many phenotypes. However, it tends to return an excessively sparse solution and is highly dependent on a threshold value used to eliminate weak causal effects representing spurious associations. In addition, the restriction that all phenotypes must be affected by only one QTL critically affects the accuracy of the inferred networks.

It would be very valuable for the field new methodologies that do not make very restrictive assumptions, such as normality, causal sufficiency, time-invariant dynamic system, linear relationships, acyclic network structure, and independent error components.

## IX) References


[1] Ribeiro, AH, Soler, JMP, Neto, EC, and Fujita, A. Causal Inference and Structure Learning of Genotype-Phenotype Networks using Genetic Variation. In *Big Data Analytics in Genomics*. Springer, New York (2016 - in press).

[2] Neto, EC, Ferrara, CT, et al. Inferring causal phenotype networks from segregating populations. *Genetics*, 179, 2 (2008), 1089-1100.

[3] Neto, EC, Ferrara, CT, et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Stat*, 4, 1 (2010), 320.

[4] Cai, X, Bazerque, JA, and Giannakis, GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol*, 9, 5 (2013), e1003068.

[5] Wang, H and van Eeuwijk, FA. A new method to infer causal phenotype networks using QTL and phenotypic information. *PLOS One*, 9, 8 (2014), e103907.

[6] Schork, NJ, Krieger, JE, Trollíet, MR, et al. A biometrical genome search in rats reveals the multigenic basis of blood pressure variation. *Genome Res*, 5 (1995), 164-172.



**Funding**: this work was supported by FAPESP (2013/01715-3, 2014/09576-5, and 2015/01587-0), CNPq (306319/2010-1 and 473063/2013-1), CAPES, and NAP eScience—PRP—USP.