Inferring Causal Phenotype Networks Using Structural Equation Models

G.J.M. Rosa[*], B.D. Valente[†], G. de los Campos[*], X.-L. Wu[*],
D. Gianola[*] and M.A. Silva[†]

Introduction

In biological systems, phenotypic traits may exert mutual effects which can be studied using statistical models that account for recursiveness and feedback. For example, high yield in dairy cows increases the liability to certain diseases and, conversely, the presence of disease may affect yield adversely. Likewise, the transcriptome may be a function of reproductive status in mammals and the latter may depend on other physiological variables. Knowledge of phenotype networks describing such interrelationships can be used to predict behavior of complex systems, e.g. biological pathways underlying complex traits such as diseases, growth and reproduction. Structural Equation Models (SEM) are used to study recursive and simultaneous relationships among phenotypes in multivariate systems such as genetical genomics, system biology, and multiple trait models in quantitative genetics. SEM can produce an interpretation of relationships among traits which differs from that obtained with standard multiple trait models (MTM), where all relationships are represented by symmetric linear associations among random variables, i.e., as measured by covariances. Unlike MTM, in SEM one trait can be treated as a predictor of another trait, providing a causal link between them.

In the last few years, genetics has been used as a device to infer phenotype networks, including causal relationships among them (Schadt et al. (2005)), and SEM or related methodologies have been employed for such task (e.g. Li et al. (2006); Liu et al. (2008); Chaibub Neto et al. (2008); Chaibub Neto et al. (2010)). These applications of SEM for phenotype network reconstruction considered genetical genomics studies with model species, making use of QTL, molecular marker, and or DNA sequence information to abet causal inference. In livestock, however, genetical genomics studies are not common due to its cost, and reliable information regarding QTL or even sequence information may not be available. More recently, Valente et al (2010) proposed a methodology that allows searching for recursive causal structures in the context of mixed models for genetic analysis of multiple traits, showing that under certain conditions it may be possible to infer phenotype networks and causal effects even without QTL or marker information. In this paper we briefly review SEM's and present some of their applications for phenotype network reconstruction in genetical genomics studies, in which both phenotypic and molecular information is available, as well as in the context of classical genetic analysis of multiple phenotypic traits.

[*] University of Wisconsin, Madison, Wisconsin USA 53706
[†] Federal University of Minas Gerais, Belo Horizonte, MG, Brazil 30123-970

Structural equation models

Structural Equation Models (Wright (1921); Haavelmo (1943)) provide a general statistical modeling technique for estimating and testing functional relationships among traits, which are often not revealed by standard linear models. In SEM, the causal structure can be represented as a directed graph in which variables (measured or unmeasured) constitute nodes and causal relationships are represented as directed edges between nodes. For example, consider the graph depicted in Figure 1, in which explanatory variables x and some additional (residual) variables e directly affect variables y, which have also some causal relationships among them.

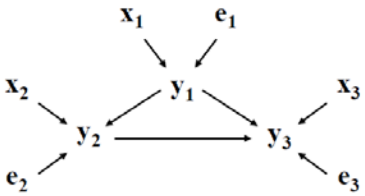

Figure 1: Example of a causal structure, in which y's represent measurements on three phenotypic traits, x's and e's represent known explanatory variables and residual factors affecting y's, respectively.

The graph in Figure 1 can be represented by a set of structural equations, given by:

$$\begin{cases} y_1 = \beta_1 x_1 + e_1 \\ y_2 = \lambda_{21} y_1 + \beta_2 x_2 + e_2 \\ y_3 = \lambda_{31} y_1 + \lambda_{32} y_2 + \beta_3 x_3 + e_3 \end{cases}$$

where β's are model parameters representing the "fixed effects" of the covariates x's on y's, and λ's are structural coefficients representing the magnitude of the casual effects among y's. Hence, in matrix notation, a SEM can be represented as $\mathbf{y} = \boldsymbol{\Lambda}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\Lambda}$ is a matrix with the coefficients λ, and **y**, **X**, **β** and **e** are appropriate vectors or matrices with the observations y's, explanatory variables x's, model parameters β's and residuals e's, respectively. Competing networks, presenting different causal structures among y's, may be compared using some model selection criteria, such as likelihood ratio tests (especially for nested models), AIC, BIC, or Bayesian model selection approaches.

SEM's have been intensively used in many fields, such as economics, psychometrics, social statistics, and biological sciences. More recently, it has been employed also in quantitative genetics in the context of mixed model analysis (e.g. Gianola and Sorensen (2004); Wu et al. (2010)) and on gene-phenotype network reconstruction, as discussed below.

QTL information and the randomization of alleles

Li et al. (2006) pointed out that genetically randomized experimental populations that segregate naturally occurring allelic variants can provide a basis for the inference of networks of causal associations among genetic loci, physiological phenotypes, and disease states. In particular, the randomization of alleles that occurs during meiosis provides a setting that is analogous to a randomized experimental design, such that causality can be inferred within the classical Fisherian statistical context.

In this context, Schadt et al. (2005) proposed a multi-step procedure to infer causal relationships between two phenotypic traits and a common QTL. More specifically, they tried to disentangle the causal path involving the expression of a particular gene (t), a cis-acting eQTL (g), and a complex trait c (e.g. a disease trait), to determine if they are related to each other following a causal, reactive or independent model (Figure 2). The model M_C depicted in Figure 2 refers to the simplest causal relationship with respect to t, in which allelic variations in g change c by changing the transcriptional activity t. Model M_R in the same figure represents the simplest reactive model with respect to t, in which the expression t is modulated by the trait c. Lastly, the independent model M_I represents a situation in which the QTL g controls t and c independently.

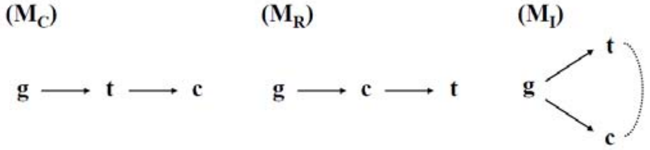

Figure 2: Some possible relationships between a QTL (g), transcriptional activity of a specific gene (t), and a complex trait (c), if both t and c are shown to be under control of g (adapted from Schadt et al. (2005)).

Schadt et al. (2005) proposed a likelihood-based causality model selection (LCMS) test that uses conditional correlation measures to determine which relationship among a trio of traits (a transcriptional trait, a complex phenotype, and a common QTL affecting both) is best supported by the data. Likelihoods associated with each of the models (causal, reactive and independent models) were constructed and maximized with respect to the model parameters, and the AIC criterion was used to select the model best supported by the data. More specifically, the joint probability distributions of the three models depicted in Figure 1 were described as:

$$\begin{cases} M_C : p(g,t,c) = p(g)p(t \mid g)p(c \mid t) \\ M_R : p(g,t,c) = p(g)p(c \mid g)p(t \mid c) \\ M_I : p(g,t,c) = p(g)p(t \mid g)p(c \mid g) \end{cases}$$

where t and c were assumed normally distributed about each genotypic mean at the common locus g. With those settings, model-specific likelihoods were obtained and standard maximum likelihood estimation methods were employed.

The authors applied their methodology to a mouse genetical genomics study comprised of large-scale genotypic, gene-expression and complex-trait data to identify genes related to obesity, and were able to identify known and new susceptibility genes for fat mass, and to successfully predict transcriptional response to perturbation in such genes. Their procedure, however, is restricted to simple gene-phenotypes networks, focusing on the identification of genes in the causal-reactive interval considering a trio of nodes comprising a common QTL affecting the expression of a specific gene and a complex trait. Evidently, gene and phenotype networks can be much more complex, as the causal-reactive genes may be also interacting in a broader network through an intricate cascade of genes and phenotypic traits.

More specifically with SEM, Li et al. (2006) presented a methodology for the analysis of multilocus, multitrait genetic data. Their method extends that of Schadt et al. (2005), not only by the number of loci and phenotypic traits studied, but also by different possible causal relationships among them, such that it provides a better characterization of the genetic architecture underlying complex traits. For instance, even if only a single locus and two correlated traits are considered, it allows for alternative recursive effects between phenotypes (Figure 3), outside the causal-reactive interval explored by Schadt et al. (2005).

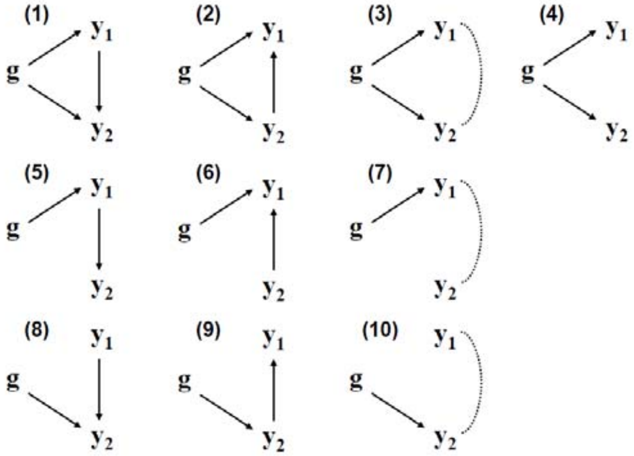

Figure 3: Causal relationships among a QTL (g) and two correlated phenotypes (y_1 and y_2). Arrows indicate the direction of causal effects and dotted lines represent unresolved associations between the two phenotypes (adapted from Li et al. (2006)).

The method of Li et al. (2006) comprises a series of 5 steps. First single locus genome scans are run for each individual phenotype using a LOD-based test. Next, conditional genome scans are performed using one trait as a covariate in the analysis of another trait. As the authors mention, the choice of which trait(s) to use as covariates can be performed extensively or, alternatively, it may be guided by known biological relationships among the traits. In this setting, traits that are known to be upstream in the causal pathways should be employed as conditioning variables. The comparison between results from unconditioned and conditioned scans can give a first insight into the causal relationships among the phenotypes. For example, in model (8) of Figure 3, g and y_1 are unconditionally independent; however, conditioning on y_2 will result in a nonzero partial correlation between them. By contrast, in model (9), g and y_1 are unconditionally correlated, and by conditioning on y_2 their dependence vanishes. When the QTL g and both traits y_1 and y_2 are causally connected, as in model (1)-(3), the raw and partial correlations between them will all be nonzero, but they will change in magnitude depending on the signs of the path coefficients (Li et al.

(2006)). A third step on Li et al. (2006)'s procedure refers to the construction of an initial path model and its respective SEM representation. In the graphical SEM, each measured trait is represented as a node, including the QTLs identified in steps 1 and 2. Edges should be directed from the QTLs to the corresponding traits, and edges should be added also from conditioning traits to the responses whenever a significant ΔLOD is observed. After the path models are constructed, they are assessed in terms of goodness-of-fit by comparing the predicted and observed covariance matrices and by significance tests for individual path coefficients. Finally, an additional step is performed to refine the model, by proposing and assessing alternative models, which are generated by adding or removing edges in the initial model, or by reversing the causal direction of an edge. The authors use a LRT approach for comparing such models, but they suggest also that alternative model criteria could be used, such as the AIC or variations thereof, or predictive ability assessed through some cross-validation approach. Steps 4 and 5 of model refinement and assessment may be also carried iteratively.

Li et al. (2006) carried out the genome scans with tests on every 2 cM using a permutation approach, followed by the SEM component of the analysis. They applied the methodology proposed to the analysis of body weight and weights of the inguinal, gonadal, peritoneal, and mesenteric fat pads of a SM × NZB intercross population with 260 females and 253 male mice raised on an atherogenic diet, and concluded that SEMs provide a powerful descriptive approach to the genetic analysis of multiple traits, allowing the characterization of pleiotropic and heterogeneous genetic effects of multiple loci on multiple traits, as well as the physiological interactions among traits.

Another application of SEM for phenotype causal network inference was presented by Liu et al. (2008), who proposed a methodology to search for a set of sparser structures within a putative directed network of causal regulatory relationships among gene expression levels and eQTL in genetical genomics studies. Their method encompasses three steps. First, eQTL mapping techniques are used to identify chromosomic regions modulating the expression of genes. Secondly, regulator-target pairs are identified, such that a directed network can be obtained. Finally, sparser optimal networks are sought within the initial directed network using a SEM approach. Liu et al (2008) apply their methodology to a genetical genomics data on yeast containing information on expression levels of 4589 genes and genotypes for 2956 markers on 112 haploid offspring originated from a cross between a laboratory and a wild strain. They detected a number of *cis*- and *trans*-acting eQTLs and regulator-target pairs, from which a directed network comprising 28K+ regulator-target pairs was constructed. Based on a partition of this initial network, which comprised 168 genes involved in a cycle genes and all genes connected to the cycle genes by up to 3 edges and all the eQTL associated with these genes, a SEM analysis was performed for its sparsification. The preliminary sub-network had 265 genes, 241 QTLs, 832 edges connecting genes, and 640 edges connecting eQTL to genes. The resulting SEM network contained 475 edges connecting genes, and 468 edges connecting eQTL to genes. Some additional analyses were performed to check for biological function lists of genes that were enriched on this network, revealing for example that 41.6% of the genes were involved in catalytic activity, and other 18% were involved in hydrolase activity.

Also using QTL information to orient edges connecting phenotypes, Chaibub Neto et al. (2008) proposed a methodology comprised of two main steps. First, an association network is constructed using either an undirected dependency graph (UDG; Shipley (2002)) or a skeleton derived from the PC algorithm of Spirtes et al. (2000). Second, LOD score tests are used to determine causal direction for every edge that connects a pair of phenotypes, conditional on QTLs affecting the phenotypes. They assessed the performance of their methodology in simulations studies, showing that it is able to recover network edges and infer their causal direction correctly at a high rate. However, although their method can be applied to human studies and outbred populations, it heavily depends on the availability of reliable information regarding QTLs affecting the phenotypic traits of interest. Nonetheless, as discussed by Chaibub Neto et al. (2010), traditional QTL mapping approaches are based on single trait analyses, in which the network structure among phenotypes is not taken into account. Such single-trait analyses may detect QTLs that directly affect each phenotype, as well as QTLs with indirect effects, which affect phenotypes upstream to the specific phenotype being analyzed. For example, consider the causal graph depicted in Figure 4, consisting of 5 phenotypes (y_1-y_5) and 3 QTLs (q_1-q_3). The outputs of single-trait analyses under this scenario are given in Figure 5. Now, when a multi-trait QTL analysis is performed according to the actual phenotype causal network, detecting indirect-effect QTLs is avoided by simply performing mapping analysis of each phenotype conditional on their parents. For example, in Figure 4, if a QTL analysis for phenotype y_3 is performed conditionally on trait y_2, only QTL q_3 will be detected because y_3 is conditionally orthogonal to q_1 and q_2, the two QTLs with indirect effects (through y_1 and y_2) on y_3.

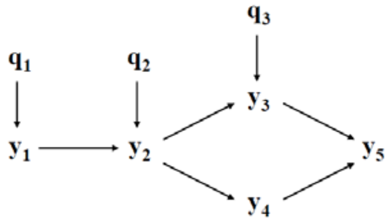

Figure 4: Example of network with five phenotypes and three QTLs.

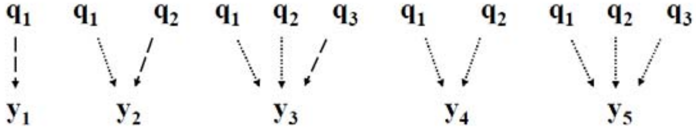

Figure 5: Expected output of single-trait QTL analyses for phenotypes in Figure 4. Dashed and pointed arrows represent direct and indirect effects of QTLs on phenotypes, respectively.

Hence, traditional QTL mapping approaches that ignore the phenotype network result in poorly estimated genetic architecture of phenotypes, which may hamper correct inferences regarding causal relationships among phenotypes. In view of this drawback of traditional QTL analyses and phenotype network reconstruction methods, Chaibub et al. (2010) suggested a methodology that simultaneously infers a causal phenotype network and its associated genetic architecture. Their approach is based on jointly modeling phenotypes and QTLs using homogeneous conditional Gaussian regression models and a graphical criterion for model equivalence. The concept of randomization of alleles during meiosis and the unidirectional relationship from genotype to phenotype are used to infer causal effects of QTLs on phenotypes. Subsequently, causal relationships among phenotypes are inferred using the QTL nodes, which might enable distinguishing among phenotype networks that would otherwise be distribution equivalent.

Inferring causal phenotype networks with no QTL information

The phenotype network reconstruction approaches discussed so far are all reliant on information regarding QTLs affecting the phenotypes, or on the availability of genetic marker information for the joint inference regarding phenotype network and genetic architecture. Such QTLs are used as parent nodes on putative networks, facilitating inferences on the remaining of the network, either on the construction of preliminary undirected graphs or on the establishment of causal relationships. However, it may be argued that even without information on QTL it may be possible to infer (at least partially) the causal relationships among phenotypic traits.

For example, there are algorithms that use the notion of d-separation (Pearl (2000)) to explore the space of causal hypotheses so as to arrive to a causal structure (or a class of observationally equivalent causal structures) that is capable of generating the observed pattern of conditional probabilistic independencies between variables. Within this context, Valente et al. (2010) proposed an approach to use one of such algorithms (IC algorithm: Verma and Pearl (1990), Pearl (2000)) to search for acyclic causal structures in a quantitative genetics mixed models framework. In this scenario, each trait is affected by additive genetic effects, which may be correlated (Figure 6). Correlated genetic effects act as an additional source of phenotypic covariance, which may confound the search for causal structures. As an example, given the causal structure among phenotypes in Figure 6, y_1 would be expected to be independent of y_3 given y_2, but this may not hold because of the correlation between u_1 and u_3.

To restore the connection between causal structures and joint density, Valente et al. (2010) proposed to perform the search based on the joint distribution of phenotypes conditionally on additive genetic effects (e.g., y_1 is expected to be independent of y_3 given y_2 and all the additive genetic effects). To obtain such distribution, the authors proposed to fit a multiple trait model where additive genetic effects could be predicted based on pedigree information. This allows one to access the covariance matrix of the phenotypic traits given the additive genetic effects, which is the residual covariance matrix for this model. Next, the IC algorithm can be applied to this matrix, returning a class of equivalent causal structures (i.e. causal structures that results in the same conditional independencies in the joint probability distribution). The authors validate their methodology using simulated data with different causal structures and sample sizes, showing that it can indeed recover the underlying causal structure among phenotypic traits. Details on their methods and results will be presented by Valente et al. (2010) in this congress.

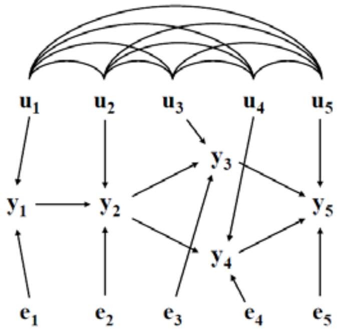

Figure 6: Example of network involving five phenotypic (observable) traits, and their corresponding additive genetic (u's) and residual (e's) effects; arcs connecting u's represents genetic correlations (adapted from Valente et al. (2010)).

Concluding remarks

SEMs provide a flexible and powerful approach for the genetic analysis of multiple traits, allowing the characterization of pleiotropic and heterogeneous genetic effects of multiple loci on multiple traits, as well as causal relationships among phenotypes, which can be used to predict behavior of complex systems, e.g. biological pathways underlying disease traits. More specifically with livestock, SEMs can be used to infer phenotype networks in the genetics analysis of quantitative traits, such that the effect of external interventions can be better predicted. This may foster the development of more efficient breeding programs and optimal decision-making strategies regarding farm management practices.

References

Chaibub Neto, E., Ferrara, T. C., Attie, A. D. *et al.* (2008). *Genetics* 179: 1089-1100.

Chaibub Neto, E., Keller, M. P, Attie, A. D. *et al.* (2010). *Ann. Appl. Stat.* (in press)

Gianola, D., and Sorensen D. (2004). *Genetics* 167: 1407-1424.

Haavelmo, T. (1943). *Econometrica* 11: 1–12.

Li, R., Tsaih, S. W., Shockley, K. *et al.* (2006). *PLoS Genetics* 2: e114.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008). *Genetics* 178: 1763-1776.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.*

Schadt, E. E., Lamb, J., Yang, X. *et al.* (2005). *Nature Genetics* 37: 710-717.

Shipley, B. (2002). *Cause and Correlation in Biology*.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*.

Valente, B. D., Rosa, G. J. M., de los Campos, G. *et al.* (2010) *Genetics* (submitted).

Verma, T. and Pearl, J. (1990). In *6th UAI*, pages 220-227.

Wright, S. (1921). *J. Agric. Res.* 201: 557–585.

Wu, X.-L., Heringstad, B., and Gianola, D. (2010). *J. Anim. Breed. Genet.* 127: 3-15.