Applications of Graphical Models in Genetics and Genomics

Guilherme J. M. Rosa
University of Wisconsin-Madison

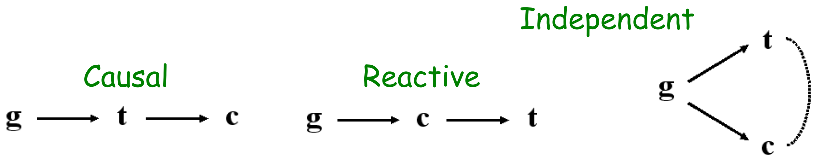
Francisco Peñagaricano
University of Florida

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal meaning of genomic predictors
  - Visualization and model selection tool
- Concluding Remarks

Applications: Flow of Information from DNA to Phenotype

- Example with 3 nodes (Schadt et al. 2005): polymorphism (g), expression (t) and disease outcome (c)
- Causal, reactive and independent models:



- Likelihood-based causality model selection (LCMS):

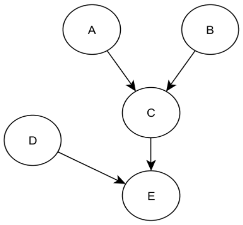
$$\begin{cases} C: p(g,t,c) = p(g)p(t \mid g)p(c \mid t) \\ R: p(g,t,c) = p(g)p(c \mid g)p(t \mid c) \\ I: p(g,t,c) = p(g)p(t \mid g)p(c \mid g,t) \end{cases}$$

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal meaning of genomic predictors
  - Visualization and model selection tool
- Concluding Remarks

Applications: Parsimonious Models for Multi-Trait Analysis

- k traits (nodes) → k(k - 1)/2 covariances
- Matrix Λ of SEM potentially with fewer parameters
- Model comparison using traditional techniques such as AIC, BIC, DIC etc.
- Example:



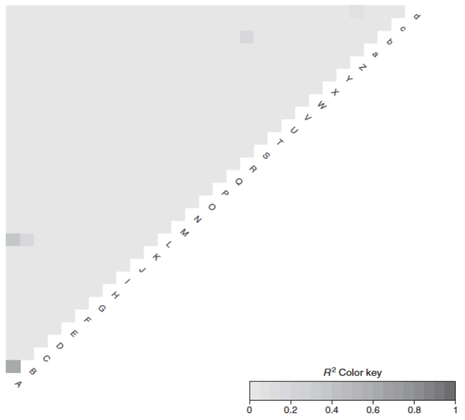
- Structure matrix Σ with 10 covariance parameters
- Matrix Λ with 4 unconstrained parameters

Example: Multilocus Linkage Disequilibrium

- 4898 Holstein bulls genotyped with the Illumina BovineSNP50 Bead Chip, provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD, USA)
- 36 778 SNP markers after editing and imputation using fastPHASE 1.4.0
- BN using SNPs with highest effects (BLasso)
- Tabu search algorithm with BDe scoring metric, and IAMB algorithm with X^2 test

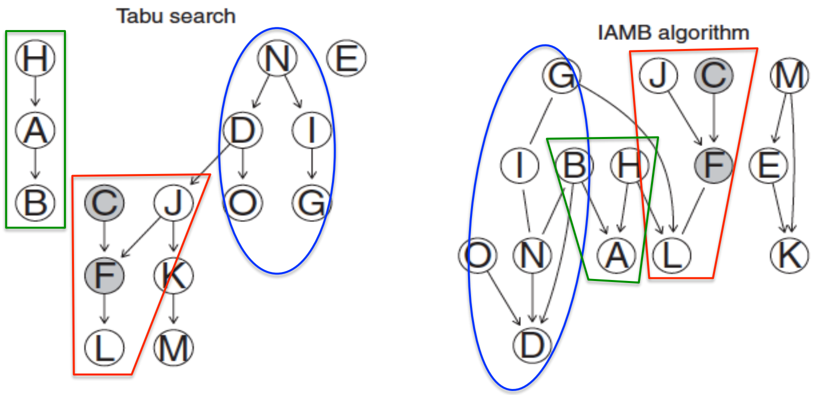
Morota G, Valente BD, Rosa GJM, Weigel KA and Gianola D. An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. *Journal of Animal Breeding and Genetics* 129: 474-487, 2012.

Pairwise LD among SNPs (r^2)

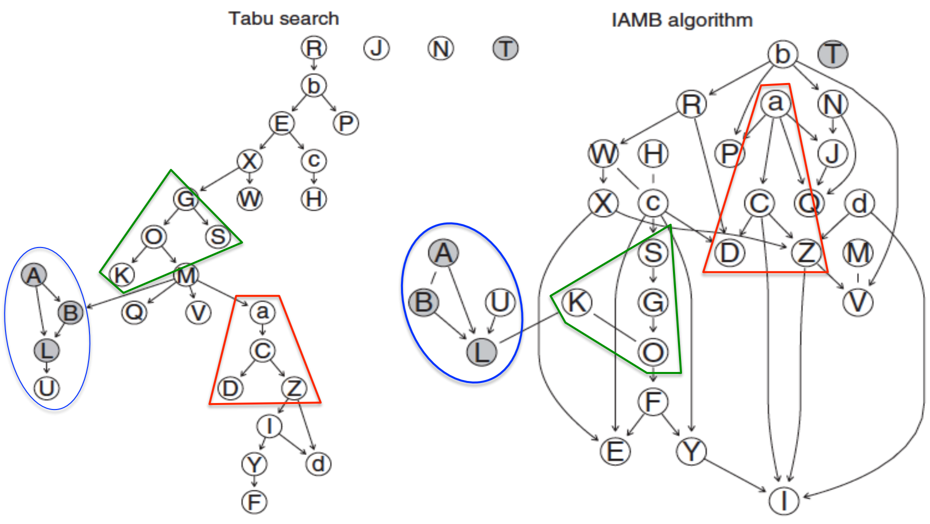


LD among the top 30 SNPs with the largest absolute posterior means using the r^2 metric.

Bayesian LD network



BNs learned by the Tabu and the IAMB searches for the 15 SNPs with the largest absolute posterior means.
Grey-filled nodes are SNPs located in chromosome 14.



Similarly as before, for the 30 SNPs with the largest absolute posterior means.

Outline

- Introduction about Networks
- Brief Overview of Graphical Models
- Usefulness and Applications
  - Flow of information from DNA to phenotype
  - Parsimonious models for multi-trait analysis
  - Prediction, Markov Blanket
  - Causal meaning of genomic predictors
  - Visualization and model selection tool
- Concluding Remarks

Example: Egg Production in Poultry

- Two strains (L1 and L2) of European Quail
- 31 traits (female quails):
  - Body weight
  - Weight gain
  - Age at first egg
  - Egg production
  - Egg quality traits

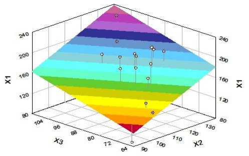
Felipe VPS, Silva MA, Valente BD and Rosa GJM. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry Science* 94(4): 772-780, 2015.

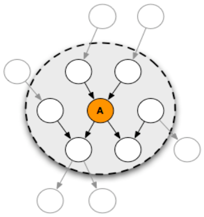
Material & Methods

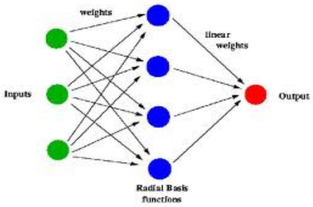
- Sample sizes (training and test sets):
  - Line 1 (90 + 90), Line 2 (102 + 103)
- Traits:
  - Weekly body weight (birth to 35 d, BW1 to BW6)
  - Weight gain (0-35 and 21-35 d, WG1 and WG2)
  - Age at first egg (AFE)
  - Egg quality traits, four time points: 125, 170, 215, 260 d
    Egg Weight – Ew, Yolk Weight – Y, Egg Shell Weight – ES
    Egg White Weight – EW, Egg Specific Gravity - DENS
  - Partial Egg Production (35-80d, EP1) and
    Total egg production (35-260d, TEP)

Material & Methods

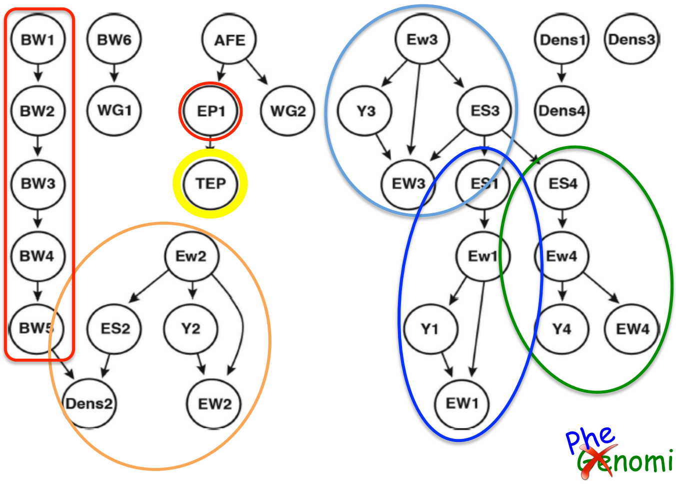
- Multiple regression analysis
  - Step-wise OLS
- Bayesian Networks
  - MB detection
- Artificial Neural Networks
  - Machine learning tool to map relationship between inputs and output



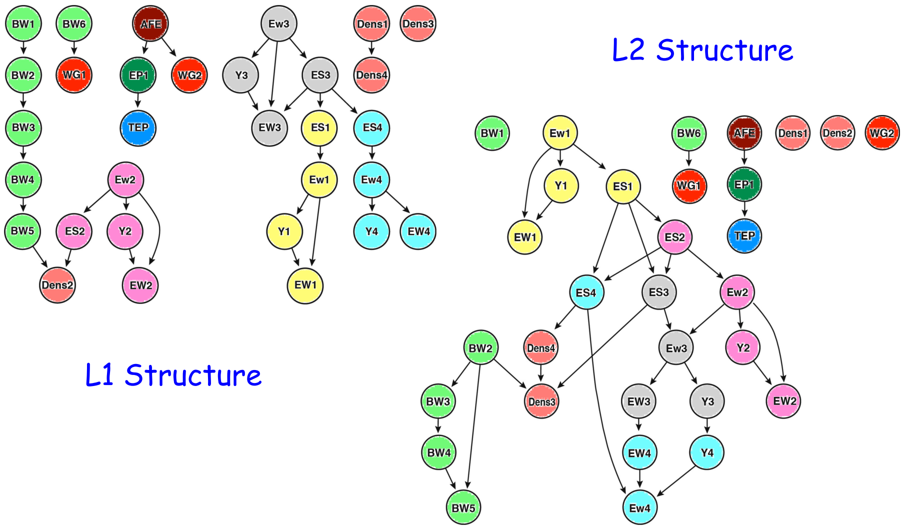




- Structure Learning (L1): Given EP1, TEP is independent from the other traits



Phe~~Ge~~nomic Prediction!



L2 Structure

L1 Structure

GENETICS | GENOMIC SELECTION

The Causal Meaning of Genomic Predictors and How It Affects Construction and Comparison of Genome-Enabled Selection Models

Bruno D. Valente,[*,†,1] Gota Morota,[†] Francisco Peñagaricano,[†] Daniel Gianola,[*,†,‡] Kent Weigel,[*] and Guilherme J. M. Rosa[†,‡]
[*]Departments of Dairy Science, [†]Animal Sciences, and [‡]Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706

Genetics, Vol. 200, 483–494 June 2015

Causal meaning of genomic predictors

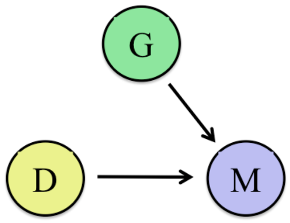
□ prediction versus causal inference

□ construction and comparison of selection models

Example 1: Simulation Settings

⇨ Consider the following causal network involving a

- Genetics component (G)
- Disease incidence (D), and
- Milk yield (M)



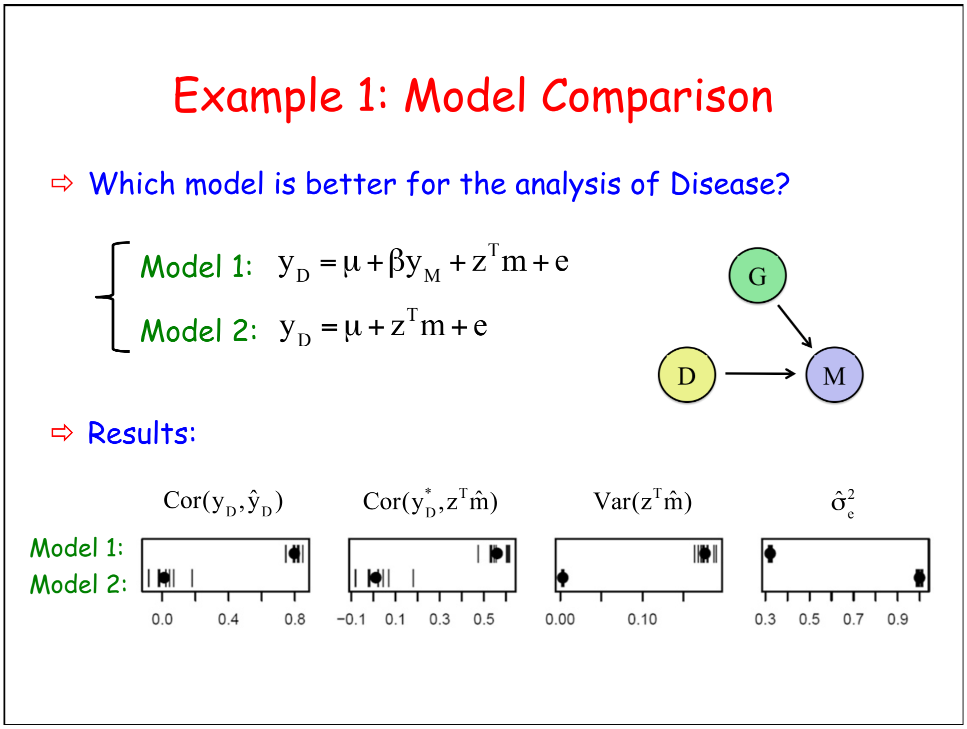
⇨ The following model was used to simulate data:

$$\begin{cases} y_D = \mu_D + e_D \\ y_M = \mu_M - 1.5y_D + u_M + e_M \end{cases}$$

with

$$\mathrm{Var}\left[u_M\right] = 1.0$$

$$\mathrm{Var}\begin{bmatrix} e_D \\ e_M \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

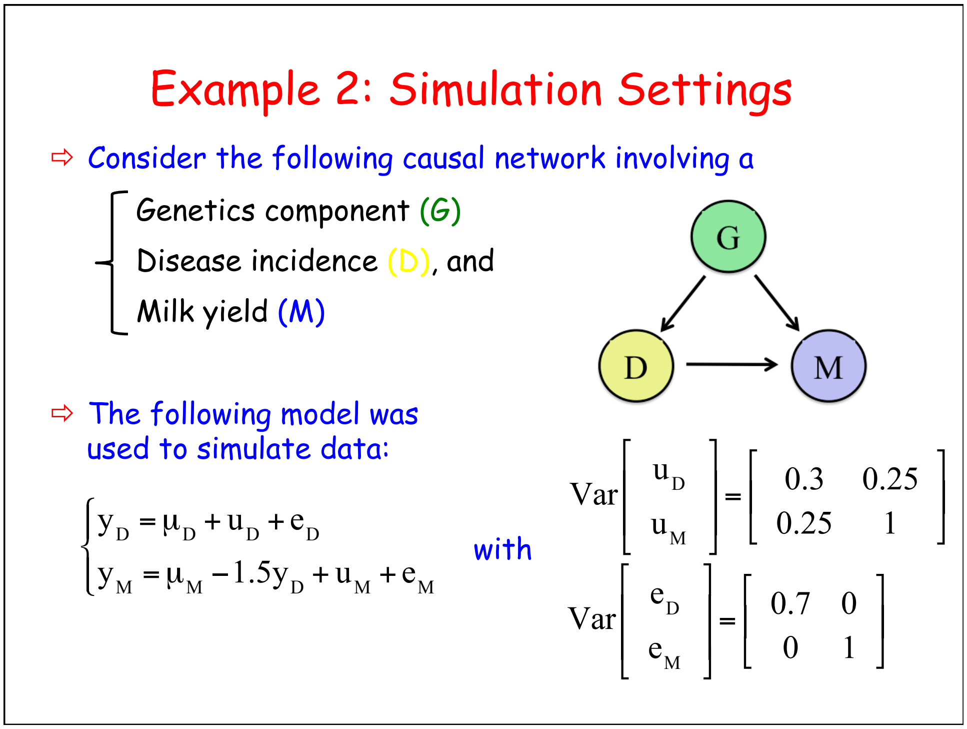


Example 1: Model Comparison

⇨ Which model is better for the analysis of Disease?

- Model 1: $y_D = \mu + \beta y_M + z^T m + e$
- Model 2: $y_D = \mu + z^T m + e$

⇨ Results:



Example 2: Simulation Settings

⇨ Consider the following causal network involving a

- Genetics component (G)
- Disease incidence (D), and
- Milk yield (M)

⇨ The following model was used to simulate data:

$$\begin{cases} y_D = \mu_D + u_D + e_D \\ y_M = \mu_M - 1.5 y_D + u_M + e_M \end{cases}$$

with

$$\mathrm{Var}\begin{bmatrix} u_D \\ u_M \end{bmatrix} = \begin{bmatrix} 0.3 & 0.25 \\ 0.25 & 1 \end{bmatrix}$$

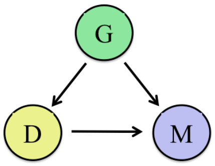
$$\mathrm{Var}\begin{bmatrix} e_D \\ e_M \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 2: Model Comparison

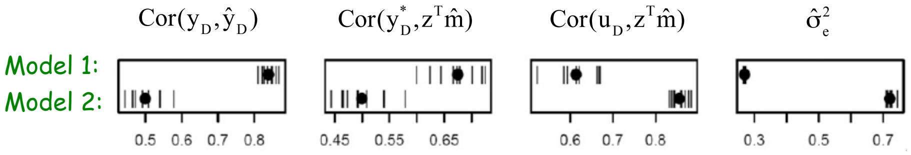
⇨ Which model is better for the analysis of Disease?

Model 1: $y_D = \mu + \beta y_M + z^T m + e$

Model 2: $y_D = \mu + z^T m + e$

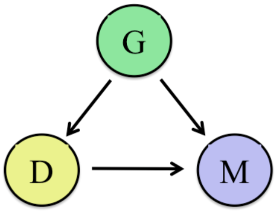


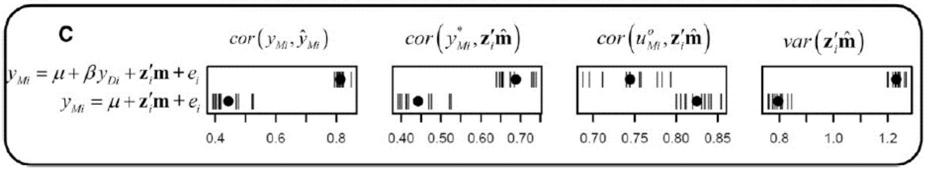
⇨ Results:



Direct effect or Total effect?

⇨ Which model is better for the analysis of **Milk?**

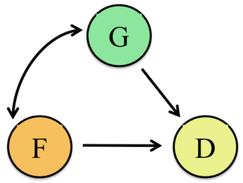




Example 4: Simulation Settings

⇨ Consider the following causal network involving a

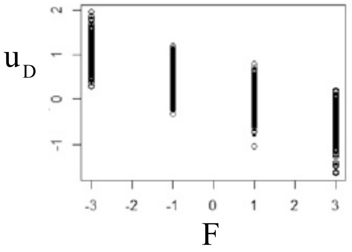
- Genetics component (G)
- Farm effect (F), and
- Disease incidence (D)



⇨ The following model was used to simulate data:

$$y_D = \mu_D + F + u_D + e_D$$

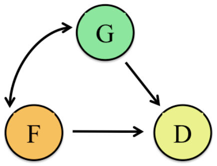
with
$$\text{Var}\left[u_D\right] = 0.30$$
$$\text{Var}\left[e_D\right] = 0.70$$



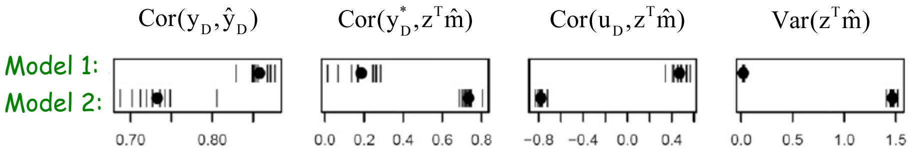
Example 4: Model Comparison

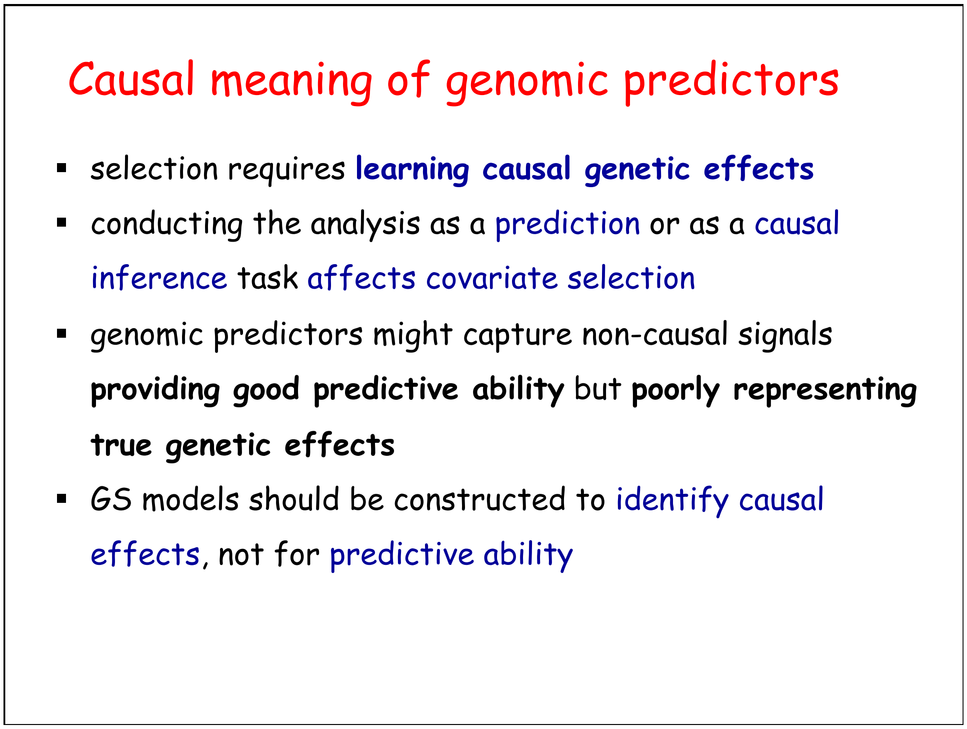
⇨ Which model is better for the analysis of Disease?

- Model 1: $y_D = \mu + F + z^T m + e$
- Model 2: $y_D = \mu + z^T m + e$



⇨ Results:





Causal meaning of genomic predictors

- selection requires **learning causal genetic effects**
- conducting the analysis as a prediction or as a causal inference task affects covariate selection
- genomic predictors might capture non-causal signals **providing good predictive ability** but **poorly representing true genetic effects**
- GS models should be constructed to identify causal effects, not for predictive ability