Causal Network Models for Correlated Quantitative Traits

Brian S. Yandell
UW-Madison
September 2013
www.stat.wisc.edu/~yandell/statgen
outline

• how are correlation and causation connected?
• hotspots: do many traits really map to the same locus?
• causal pairs: how to find causal drivers for hotspots?
• causal networks: how to infer signal cascades?
• how to scale up to larger problems
correlation & causation
“The old view of cause and effect ... could only fail; things are not in our experience either independent or causative. All classes of phenomena are linked together, and the problem in each case is how close is the degree of association.”

Karl Pearson (1911)

The Grammar of Science
“The ideal ... is the study of the direct influence of one condition on another ...[when] all other possible causes of variation are eliminated.... The degree of correlation between two variables ... [includes] all connecting paths of influence.... [Path coefficients combine] knowledge of ... correlation among the variables in a system with ... causal relations.

Sewall Wright (1921)

Correlation and causation. J Agric Res
"Causality is not mystical or metaphysical. It can be understood in terms of simple processes, and it can be expressed in a friendly mathematical language, ready for computer analysis."

Judea Pearl (2000)

Causality: Models, Reasoning and Inference
problems and controversies

• Correlation does not imply causation.
  – Common knowledge in field of statistics.

• Steady state (static) measures may not reflect dynamic processes.

• Population-based estimates may not reflect within-individual processes.
randomization and causation

• RA Fisher (1926) *Design of Experiments*
• control other known factors
• randomize assignment of treatment
  – no causal effect of individuals on treatment
  – no common cause of treatment and outcome
  – reduce chance correlation with unknown factors
• conclude (subsequent) outcome differences are caused by (due to) treatment
correlation and causation

• temporal aspect: cause before reaction
  – genotype (usually) drives phenotype
  – phenotypes in time series
  – *but* time order is not enough

• axioms of causality
  – transitive: if $A \rightarrow B$, $B \rightarrow C$, then $A \rightarrow C$
  – local (Markov): events have proximate causes
  – asymmetric: if $A \rightarrow B$, then not $B \rightarrow A$

• Shipley (2000) *Cause and Correlation in Biology*
causation casts probability shadows

• causal relationship
  – $Y_1 \rightarrow Y_2 \rightarrow Y_3$

• conditional probability
  – $\Pr(Y_1) \ast \Pr(Y_2 \mid Y_1) \ast \Pr(Y_3 \mid Y_2)$

• linear model
  – $Y_1 = \mu_1 + e$
  – $Y_2 = \mu_2 + \beta_1 \cdot Y_1 + e$

• adding in QTLs: $Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2$
  – $Y_1 = \mu_1 + \theta_1 \cdot Q_1 + e$
  – $Y_2 = \mu_2 + \beta_1 \cdot Y_1 + \theta_2 \cdot Q_2 + e$
organizing correlated traits

• functional grouping from prior studies
  – GO, KEGG; KO panels; TF and PPI databases
• co-expression modules (Horvath’s WGCNA)
• eQTL hotspots (here briefly)
• traits used as covariates for other traits
  – does one trait essentially explain QTL of another?
• causal networks (here and Horvath talk)
  – modules of highly correlated traits
strategy from hotspot to causality

• detect “real” hotspots
  – hotspot = locus where many traits map
  – use permutation test to assess

• find causal architecture for each hotspot
  – causal model selection tests for pairs of traits
  – do local traits (at hotspot) drive other traits?

• build causal network for small set of traits
  – cis (local) trait ideally is top of signal cascade
hotspots
hotspots of correlated traits

• multiple correlated traits map to same locus
  – is this a real hotspot, or an artifact of correlation?
  – use QTL permutation across traits

• references
genetic architecture of gene expression in 6 tissues
eQTL vs SNP architecture

No. eQTL per 5 cM

No. SNPs per 5 cM

Total SNPs
No. eQTLs in:
Islet
Adipose
Liver
Gastroc
Kidney
Hypothal

eQTL to SNP corr = 0.83

Chr 6 (cM)

Chr 8 (cM)

Jax SysGen: Yandell © 2013
correlated traits in a hotspot
why are traits correlated?

– Environmental
  • hotspot is spurious

– Genetics and causal networks
  • One causal driver at locus
    – Traits organized in causal cascade
  • Multiple causal drivers at locus
    – Several closely linked driving genes
    – Correlation due to close linkage
    – Separate networks are not causally related
one causal driver

gene

gene product

chromosome

signal cascade of downstream traits
two linked causal drivers
pathways independent given drivers
hotspot permutation test
(Breitling et al. Jansen 2008 *PLoS Genetics*)

• for original dataset and each permuted set:
  – set single trait LOD threshold $T$
    • use Churchill-Doerge (1994) permutations
  – count number of traits ($N$) with LOD above $T$
    • count for every locus (marker or pseudomarker)
    • smooth counts if markers are dense

• find count with at most 5% of permuted sets above (critical value) as count threshold

• conclude original counts above threshold are real
Single trait permutation schema

1. shuffle phenotypes to break QTL
2. repeat 1000 times and summarize
Hotspot permutation schema

1. shuffle phenotypes by row to break QTL, keep correlation
2. repeat 1000 times and summarize

LOD at each locus for each phenotype over genome

count LODs at locus over threshold $T$

max count $N$ over genome
permutation across traits
(Breitling et al. Jansen 2008 PLoS Genetics)

right way

wrong way

break correlation between markers and traits

but

preserve correlation among traits
BxH ApoE-/- chr 2: hotspot

Data:
Ghazalpour et al. (2008)
*PLoS Genetics*

5% threshold on number of traits
motif matching
independent assay
---
bio validation in progress

green = hotspot size
red = causal
blue = reactive
black = independent

Attie et al. (unpublished)
quality vs. quantity in hotspots
(Chaibub Neto et al. 2012 Genetics)

• detect single trait with very large LOD
  – control FWER across genome and all traits
• find small hotspots with very significant traits
  – all traits have large LODs at same locus
  – maybe one strongly disrupted signal pathway?
• use sliding LOD threshold across hotspot sizes
  – small LOD threshold (~4) for large hotspots
  – large LOD threshold (~8) for small hotspots
••• causal pairs •••
causal architecture

• focus on one hotspot
• identify all traits physically near hotspot
  – local traits (called cis if it also maps to hotspot)
• what traits are up/downstream of local trait?
  – focal trait causal to downstream target traits
  – record count at Mb position of focal gene
  – red = downstream, blue = upstream
causal model selection choices in context of larger, unknown network

- **Causal**
  - Focal trait → Target trait

- **Reactive**
  - Focal trait ← Target trait

- **Correlated**
  - Focal trait ↔ Target trait

- **Uncorrelated**
  - Focal trait ↔ Target trait (dotted line)
causal architecture references

- BIC: Schadt et al. (2005) *Nature Genet*
- CIT: Millstein et al. (2009) *BMC Genet*
- CMST: Chaibub Neto et al. (2012) *Genetics*

Extends Vuong’s model selection tests to the comparison of 3, possibly *misspecified*, models.

\[(M_1)\]
\[Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2|_1\]

\[(M_2)\]
\[Q_1|_2 \rightarrow Y_1 \leftarrow Y_2 \leftarrow Q_2\]

\[(M_3)\]
\[Q_1 \rightarrow Y_1 \quad Y_2 \leftarrow Q_2\]
Liver expression data in a mice intercross.

3,421 transcripts and 1,065 markers.

261 transcripts physically located on chr 2.
Analysis restricted to 78 traits composing a hotspot around 54.2Mb.

This collection of traits enriches for “immune system process”.

*Pscdbp*, the local trait at 58.4Mb, is a transcription factor.
causal networks
QTL-driven directed graphs

• given genetic architecture (QTLs), what causal network structure is supported by data?
• R/qdg available at www.github.org/byandell
• references
partial correlation (PC) skeleton

true graph

1\textsuperscript{st} order partial correlations

drop edge

correlations

$y_2$ d-separates $y_1$ from $y_3$

$1 \perp 3 \mid 2$
partial correlation (PC) skeleton

genuine graph

1st order partial correlations

2nd order partial correlations

drop edge

(y2, y5) d-separate y1 from y4

1 \perp 4 \mid 2, 5

Jax SysGen: Yandell © 2013 36
edge direction: which is causal?

\[ M_1 : \quad \begin{array}{c} \circled{y_1} \rightarrow \circled{y_2} \end{array} \quad M_2 : \quad \begin{array}{c} \circled{y_1} \leftarrow \circled{y_2} \end{array} \]

the above models are likelihood equivalent,

\[ f(y_1)f(y_2 \mid y_1) = f(y_1, y_2) = f(y_2)f(y_1 \mid y_2) \]

not likelihood equivalent due to QTL

\[ f(q_1)f(y_1 \mid q_1)f(y_2 \mid y_1, q_2)f(q_2) \neq f(q_2)f(y_2 \mid q_2)f(y_1 \mid y_2, q_1)f(q_1) \]
test edge direction using LOD score

\[
LOD = \log_{10} \left\{ \frac{\prod_{i=1}^{n} f(y_{1i} | q_{1i})f(y_{2i} | y_{1i}, q_{2i})}{\prod_{i=1}^{n} f(y_{2i} | q_{2i})f(y_{1i} | y_{2i}, q_{1i})} \right\}
\]

\[f(q_1)f(y_1 | q_1)f(y_2 | y_1, q_2)f(q_2) \neq \frac{f(q_2)f(y_2 | q_2)f(y_1 | y_2, q_1)f(q_1)}{}
\]

not likelihood equivalent because
reverse edges using QTLs
- We constructed a network from metabolites and transcripts involved in liver metabolism.

- We validated this network with *in vitro* experiments (Ferrara et al 2008). Four out of six predictions were confirmed.
causal graphical models in systems genetics

• what if genetic architecture and causal network are unknown?
  – jointly infer both using iteration


• R/qtlnet available from www.github.org/byandell

• Related references
basic idea of QTLnet

• iterate between finding QTL and network
• genetic architecture given causal network
  – trait y depends on parents pa(y) in network
  – QTL for y found conditional on pa(y)
    • Parents pa(y) are interacting covariates for QTL scan
• causal network given genetic architecture
  – build (adjust) causal network given QTL
  – each direction change may alter neighbor edges
missing data method: MCMC

• known phenotypes $Y$, genotypes $Q$
• unknown graph $G$
• want to study $\Pr(Y \mid G, Q)$
• break down in terms of individual edges
  – $\Pr(Y \mid G,Q) = \text{sum of } \Pr(Y_i \mid \text{pa}(Y_i), Q)$
• sample new values for individual edges
  – given current value of all other edges
• repeat many times and average results
MCMC steps for QTLnet

- propose new causal network $G$
  - with simple changes to current network:
    - change edge direction
    - add or drop edge

- find any new genetic architectures $Q$
  - update phenotypes when parents $pa(y)$ change in new $G$

- compute likelihood for new network and QTL
  - $\Pr(Y \mid G, Q)$

- accept or reject new network and QTL
  - usual Metropolis-Hastings idea
BxH ApoE-/- chr 2: causal architecture

- hotspot
- 12 causal calls
BxH ApoE-/- causal network for transcription factor Pscdbp

unpublished work of Elias Chaibub Neto
••• scaling up •••
scaling up to larger networks

• reduce complexity of graphs
  – use prior knowledge to constrain valid edges
  – restrict number of causal edges into each node

• make task parallel: run on many machines
  – pre-compute conditional probabilities
  – run multiple parallel Markov chains

• rethink approach
  – LASSO, sparse PLS, other optimization methods
graph complexity with node parents
how many node parents?

- how many edges per node? (fan-in)
  - few parents directly affect one node
  - many offspring affected by one node

BIC computations by maximum number of parents

<table>
<thead>
<tr>
<th>#</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1,300</td>
<td>2,560</td>
<td>3,820</td>
<td>4,660</td>
<td>5,120</td>
</tr>
<tr>
<td>20</td>
<td>23,200</td>
<td>100,720</td>
<td>333,280</td>
<td>875,920</td>
<td>10.5M</td>
</tr>
<tr>
<td>30</td>
<td>122,700</td>
<td>835,230</td>
<td>4.40M</td>
<td>18.6M</td>
<td>16.1B</td>
</tr>
<tr>
<td>40</td>
<td>396,800</td>
<td>3.69M</td>
<td>26.7M</td>
<td>157M</td>
<td>22.0T</td>
</tr>
<tr>
<td>50</td>
<td>982,500</td>
<td>11.6M</td>
<td>107M</td>
<td>806M</td>
<td>28.1Q</td>
</tr>
</tbody>
</table>
BIC computation

• each trait (node) has a linear model
  – $Y \sim \text{QTL} + \text{pa}(Y) + \text{other covariates}$

• BIC = LOD – penalty
  – BIC balances data fit to model complexity
  – penalty increases with number of parents

• limit complexity by allowing only 3-4 parents
parallel phases for larger projects

Phase 1: identify parents

Phase 2: compute BICs

Phase 3: store BICs

Phase 4: run Markov chains

Phase 5: combine results
parallel implementation

• R/qtlnet available at www.github.org/byandell
• Condor cluster: chtc.cs.wisc.edu
  – System Of Automated Runs (SOAR)
    • ~2000 cores in pool shared by many scientists
    • automated run of new jobs placed in project

![SOAR Job Progress](image-url)
BIC samples for 100 MCMC runs

neighborhood for reversals only

burnin
limits of causal inference

unfaithful: false positive edges

\[ \lambda = \min |\text{cor}(Y_i, Y_j)| \]
\[ \lambda = c \cdot \sqrt{dp/n} \]
\[ d = \text{max degree} \]
\[ p = \# \text{nodes} \]
\[ n = \text{sample size} \]

Uhler, Raskutti, Buhlmann, Yu (2012 arxiv)
how to use functional information?

• functional grouping from prior studies
  – may or may not indicate direction
  – gene ontology (GO), KEGG
  – knockout (KO) panels
  – protein-protein interaction (PPI) database
  – transcription factor (TF) database

• methods using only this information

• priors for QTL-driven causal networks
  – more weight to local (cis) QTLs?
modeling biological knowledge

• infer graph $G$ from biological knowledge $B$
  
  \[
  \text{Pr}(G \mid B, W) = \exp(-W \cdot |B-G|) / \text{constant}
  \]
  
  \[B = \text{prob of edge given TF, PPI, KO database}\]
  
  • derived using previous experiments, papers, etc.
  
  \[G = 0-1\text{ matrix for graph with directed edges}\]

• $W = \text{inferred weight of biological knowledge}$
  
  \[W=0: \text{no influence}; W \text{ large: assumed correct}\]
  
  \[P(W \mid B) = \phi \exp(-\phi W) \text{ exponential}\]

• Werhli and Husmeier (2007) *J Bioinfo Comput Biol*
combining eQTL and bio knowledge

• probability for graph $G$ and bio-weights $W$
  – given phenotypes $Y$, genotypes $X$, bio info $B$
  
  $$\Pr(G, W | Y, Q, B) = \Pr(Y|G,Q)\Pr(G|B,W)\Pr(W|B)$$
  
  – $\Pr(Y|G,Q)$ is genetic architecture (QTLs)
    • using parent nodes of each trait as covariates
  
  – $\Pr(G|B,W)$ is relation of graph to biological info
    • see previous slides
    • put priors on QTL based on proximity, biological info

• related ref: Kim et al. Przytycka (2010) *RECOMB*
ROC curve simulation

open = QTLnet

closed = phenotypes only
integrated ROC curve

2x2:
genetics pathways

probability classifier ranks true > false edges

\( \delta = \text{accuracy of } B \)
QTL software on CRAN

- **R/qtlhot**: hotspots & causal architecture
  - map hotspots, permutation tests
  - causal model selection tests
- **R/qtlnet**: QTL-driven phenotype networks
  - infer QTLs and directed graphs
  - coming: prior biological information
- **R/qtlbim**: Bayesian Interval Mapping for QTL
  - multiple QTL inference, graphical diagnostics
  - see earlier Jax talks for details
many thanks!

• NIH/NIGMS with Karl Broman, Nengjun Yi
• NIH/NIDDK with Alan Attie, Mark Keller
• Other collaborators:
  – Mark Keller (Attie Lab Scientist)
  – Chris Plaisier (Institute for Systems Biology, Seattle)
  – Elias Chaibub Neto (Sage Bionetworks, Seattle)
  – Jee Young Moon (grad student)
  – Xinwei Deng (VA Tech Asst Prof)
  – and many more!