Predicting molecular phenotypes using statistical learning

James Baurley

BioRealm LLC
BINUS University AI R&D Center

NVIDIA AI Conference
23-24 October 2017

Motivation: Nicotine metabolizing enzymes and regulators

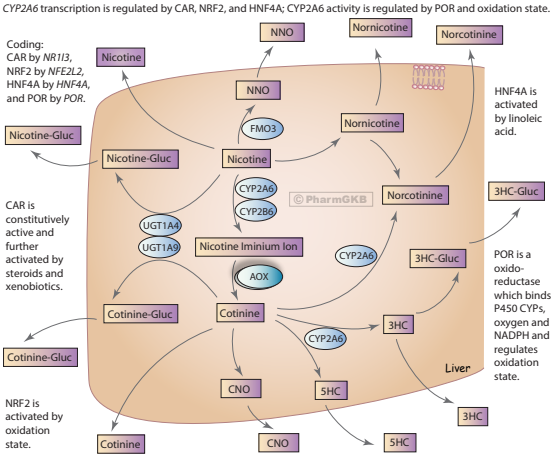


Figure: PharmGKB nicotine metabolism pathway with added annotations

Biosignature learning

Z		C_1	C_2	C_{P1}	G_1	G_2	G_3	G_{P2}
		1	0		0	2	0	
	~	0	1	...	2	1	0	...
		1	0		0	2	2	
		1	0		1	0	0	
...								

Biosignature applications

Y	Z_{pred}		C_1	C_2	C_{P1}	G_1	G_2	G_3	G_{P2}
0			1	0		0	2	0	
0									
1		~							
0						1	0	0	
0			1	1		0	2	0	
1			1	1		0	2	2	
...									

Predict Z using learned biosignatures

- Define biosignatures
- Genomic data availability
- Measuring molecular phenotypes may not be practical
- Assess many predicted molecular phenotypes (e.g. TWAS; see Gusev et al. 2016)
- Path for biomarker development (id subgroups at risk, select optimal treatments)

The conditional mean of Y_i, the NMR of individual i, depends on P explanatory variables through the link function $g(\cdot)$:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{P_1} \beta_{1j} C_{ij} + \sum_{j=1}^{P_2} \beta_{2j} G_{ij} + \sum_{j=1}^{P_3} \beta_{3j} Z_{ij},$$

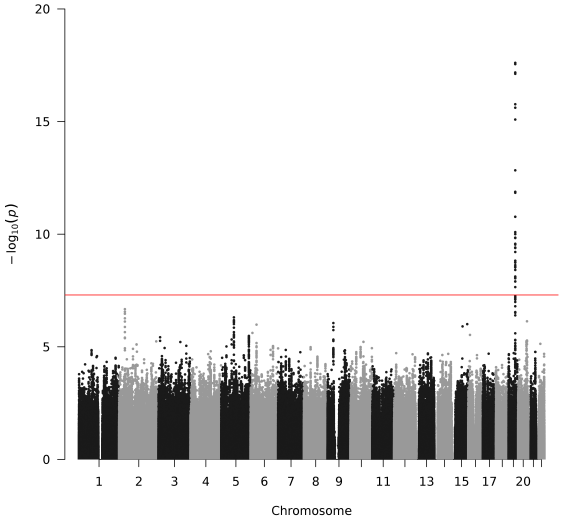
Notation:

- C_{ij} clinical factors
- G_{ij} genetic variants
- Z_{ij} derived variables
- β_{kj} regression coefficients

Application: Genotypes → nicotine metabolism → smoking cessation

- Nicotine metabolism influences:
  - development of dependence (Cannon, 2016; Chenoweth, 2016)
  - efficacy of treatment (Chen, 2014; Lerman, 2015)
- Nicotine metabolism is influenced by:
  - genetics ($h^2 = 0.74$ (Swan, 2009; Loukola, 2015))
  - ancestry (Wang, 2015)
  - age, sex, BMI, alcohol and cigarette consumption (Chenoweth, 2014)
- Data: Laboratory studies of nicotine metabolism (Baurley, 2016)
  - fixed dose nicotine administered, metabolites measured over time
  - $n = 49$ African Americans, $n = 51$ Asian Americans, $n = 212$ European Americans
  - genotyped DNA samples on the Smokescreen Genotyping Array

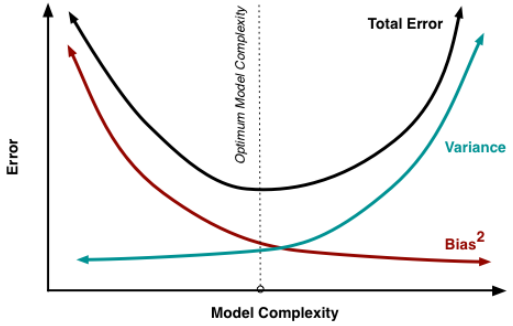
Nicotine Metabolism GWAS (Baurley, et al. 2016)



- NMR Application: $N = 312$, $P = 5.9$ million
- Given complex patterns of associations and $P >> N$, how do we get a prediction model?
- Reduce search space
  - used literature and ontologies to select 11 genomic regions (3,752 SNPs) coding for nicotine metabolic enzymes and transcription factors
- Reduce model complexity
  1. Machine learning (Penalized regression)
  2. Bayesian learning (ALPS)

A note on prediction error (Hastie, et al. 2009.)

- Assume $Y = f(X) + \epsilon, \epsilon \sim N(0, \sigma_\epsilon)$
- We estimate the model $\hat{f}(X)$ of $f(X)$.
- The prediction error at x: $Err(x) = E[(Y - \hat{f}(x))^2]$
- Expand: $Err(x) = (E[\hat{f} - f)^2 + E[(\hat{f} - E[\hat{f}])^2] + \sigma_\epsilon^2$



Fortmann-Roe 2012

Minimize a penalized residual sum of squares:

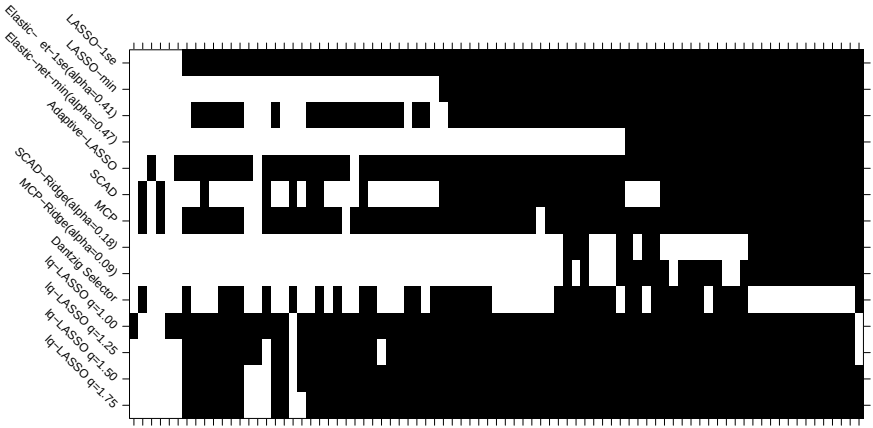
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{1}$$

- λ controls model complexity
- $q = 0$ is variable subset selection
- $q = 1$ is the lasso (variable selection)
- $q = 2$ is ridge regression (shrinkage)

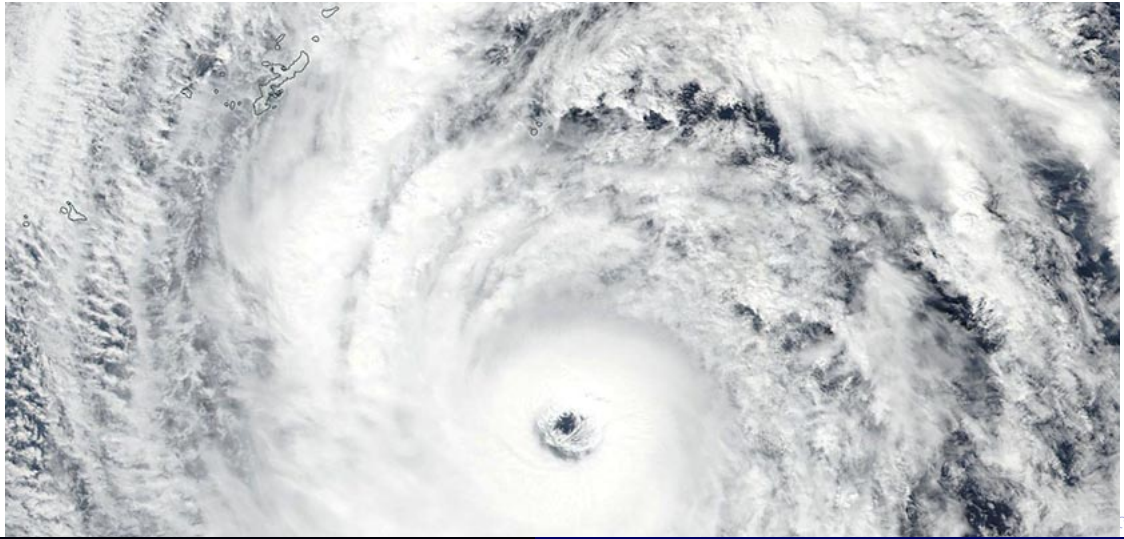
Elastic net replaces the penalty term with

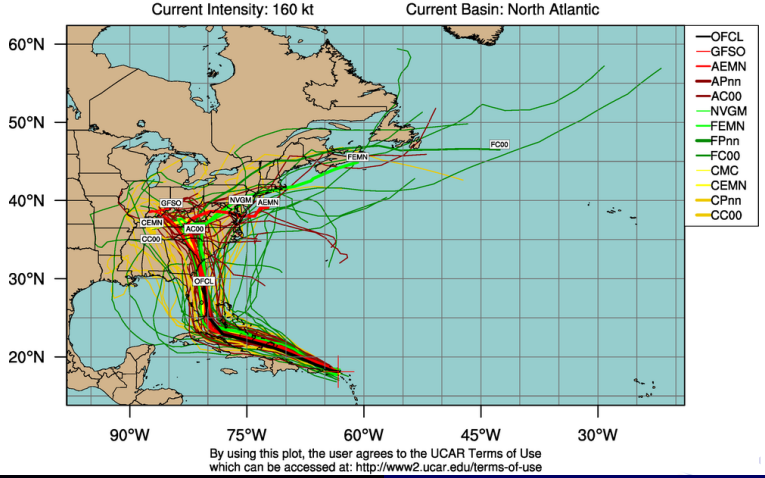
$$\lambda \sum_{j=1}^{p} \alpha\beta_j^2 + (1 - \alpha)|\beta_j|) \tag{2}$$

Nicotine Metabolism Biosignature

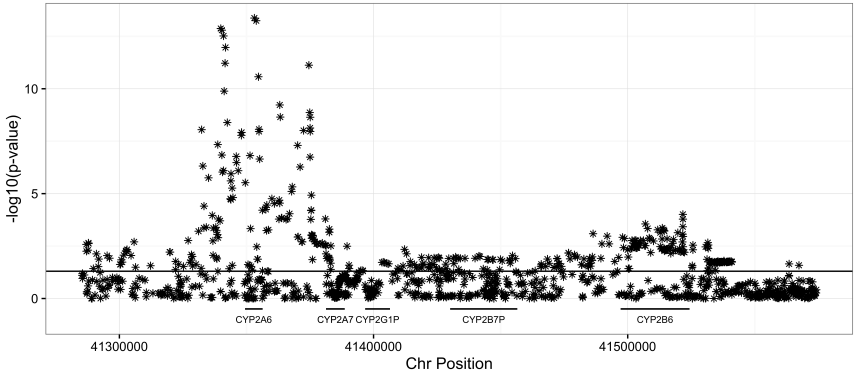


Question: Which model should we use for prediction?

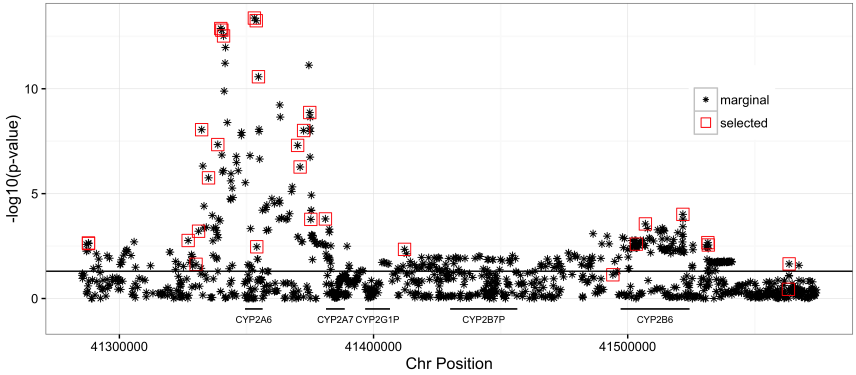



MAJOR HURRICANE IRMA (AL11)
EPS track guidance initialized at 1200 UTC, 06 September 2017
Current Intensity: 160 kt
Current Basin: North Atlantic
60°N
50°N
40°N
30°N
20°N
90°W
75°W
60°W
45°W
30°W
OFCL
GFSO
AEMN
APnn
AC00
NVGM
FEMN
FPnn
FC00
CMC
CEMN
CPnn
CC00
FC00
FEMN
GFSO
NVGM
AEMN
CEMN
AC00
CC00
OFCL
By using this plot, the user agrees to the UCAR Terms of Use
which can be accessed at: http://www2.ucar.edu/terms-of-use

chr19q13.2 Gene Region Marginal Results



Ensemble Selected SNPs, chr19q13.2



- Model diversity can improve prediction performance
- Bayesian approaches
  - account for uncertainty in model form and parameters
  - allows inclusion of existing evidence into the model
- The posterior probability (weight) of a model given data is given by

$$p(M|\mathbf{D}) = \frac{p(\mathbf{D}|M)p(M)}{\sum_{m \in \mathbf{M}} p(\mathbf{D}|m)p(m)}$$

- The marginal likelihood is actually marginalizing over the parameters in the model.

$$p(\mathbf{D}|M) = \int_{\beta} p(\mathbf{D}|\beta, M)p(\beta)d\beta$$

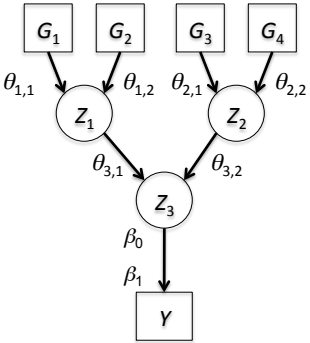
- Explore model space by Markov Chain Monte Carlo (MCMC) and approximate the marginal likelihood.

- ALPS considers sets of SNPs whose effects are combined based on tree structures Λ. See Baurley 2010, 2013.
- The output of each node of the tree is a derived variable
- θ's can represent logical ops. E.g., ADD, AND, OR's

$$Z_1 = (\theta_{1,1}G_1) + (\theta_{1,2}G_2) + (1-\theta_{1,1}-\theta_{1,2})G_1G_2$$
$$Z_2 = (\theta_{2,1}G_3) + (\theta_{2,2}G_4) + (1-\theta_{2,1}-\theta_{2,2})G_3G_4$$

$$Z_3 = (\theta_{3,1}Z_1) + (\theta_{3,2}Z_2) + (1-\theta_{3,1}-\theta_{3,2})Z_1Z_2$$

$$Y = \beta_0 + \beta_1 Z_3$$



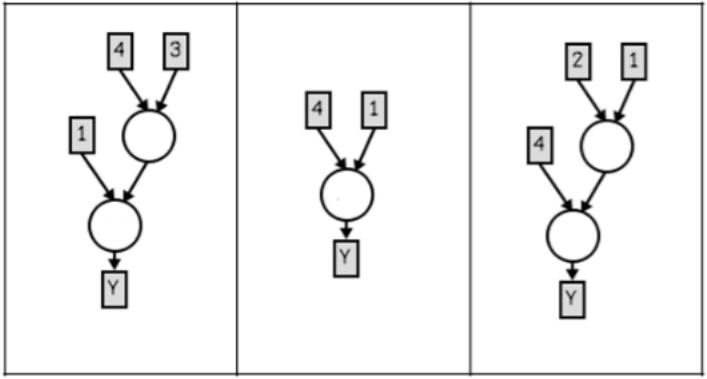
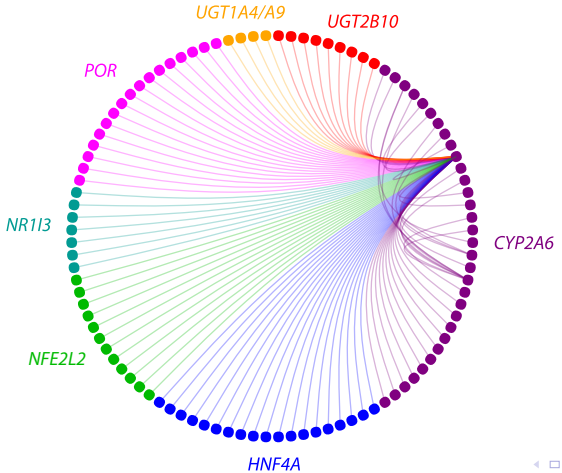


Figure 6.
Topology moves. From left to right, a node is removed deleting the edge to input 3. A new node is then added connecting input 1 and 2.

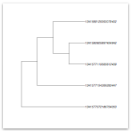
- Visited >6M Λ's from the 11 genomic regions of interest.
- Computed Bayes Factors, ratio of posterior to prior odds



Nicotine Metabolism: Top ALPS Pathway Trees


1


2


3


4

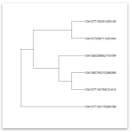

5


6


7

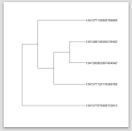

8


9

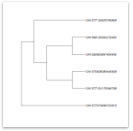

10


11


12


13


14


15


16


17


18

Genotypes $\rightarrow^1$ Molecular Phenotype $\rightarrow^2$ Outcome

- Approach not limited to genomics (e.g., phenotype panels, IoT)
- Model diversity can boost prediction performances: Ensemble methods, posterior predictive distribution
- Deep learning algorithms can discover new derived variables (e.g. control elements for gene expression)
- Refactoring is needed to GPU accelerate many statistical learning algorithms
- Invitation: Learn what's under the hood!
  - Offering 1-Week Short Course
  - May 2018 at BINUS AI R&D Center (Jakarta, Indonesia)
  - Contact Dr. Bens Pardamean: bpardamean@binus.edu