TOPICS COVERED
(order is approximate)

1. Evolution of statistical methods in quantitative genetics
2. Challenges from complexity and use of phenomic data
3. Brief review of Bayesian inference, Bayesian regression
4. Genome-enabled prediction: “Genomic Blup”; the alphabet: Bayes A, Bayes B, Bayes C, Bayes L
5. Principles of cross-validation
6. The problem of dealing with interactions
7. Introduction to non-parametric regression: LOESS, kernel regression, RKHS, radial basis functions, neural networks (NN)
8. Results from animals and plants
SOME BIBLIOGRAPHY: Bayesian

Statistical learning: general

Free software: WEKA
Gentle introductions to non-parametric regression…

ONLY IF YOU REALLY WANT TO GO DEEPLY…
1. EVOLUTION OF STATISTICAL METHODS IN QUANTITATIVE GENETICS

<table>
<thead>
<tr>
<th>Era</th>
<th>Methods</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaen</td>
<td>Visual appraisal (still widely used)</td>
<td>[Biblical times…]</td>
</tr>
<tr>
<td>Pathozoic</td>
<td>Fisher’s 1918, Path analysis, “Animal Breeding Plans”</td>
<td>[1918-1945]</td>
</tr>
<tr>
<td>Anovian</td>
<td>ANOVA (Method 1), least-squares, selection index</td>
<td>[1936-1943]</td>
</tr>
<tr>
<td>Post-anovian</td>
<td>Methods 2+3, MINQUE, MIVQUE</td>
<td>[1953-1973]</td>
</tr>
<tr>
<td>Blupassic</td>
<td>Mixed models, BLUP, animal model, multi-traits</td>
<td>[1948-1990]</td>
</tr>
<tr>
<td>Remlian</td>
<td>VCE, ASREML, DMU</td>
<td>[1971- 2009]</td>
</tr>
<tr>
<td>Posteriozoic</td>
<td>Threshold models, Survival, MCMC, QTLs</td>
<td>[1982-2008]</td>
</tr>
</tbody>
</table>


Chapter 20
D. Gianola
“Inferences from Mixed Models in Quantitative Genetics”
2. Challenges from complexity and use of phenomic data

Gene structure

Some genes do not have introns
Some genes are located within introns of other genes

Khatib (2011)
<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size</th>
<th># of genes</th>
<th>DNA/gene</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1.8 Mb</td>
<td>~1,700</td>
<td>~ 1 Kb</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4.6 Mb</td>
<td>~4,300</td>
<td>~ 1 Kb</td>
</tr>
<tr>
<td>Baker’s Yeast (Saccharomyces cerevisiae)</td>
<td>12.1 Mb</td>
<td>~6,000</td>
<td>~ 2 Kb</td>
</tr>
<tr>
<td>A worm (Caenorhabditis elegans)</td>
<td>97 Mb</td>
<td>~18,000</td>
<td>~5.4 Kb</td>
</tr>
<tr>
<td>Fruit fly (Drosophila melanogaster)</td>
<td>185 Mb</td>
<td>~14,000</td>
<td>~13 Kb</td>
</tr>
<tr>
<td>Human (Homo sapiens)</td>
<td>3,000 Mb</td>
<td>~25,000</td>
<td>~ 86 Kb</td>
</tr>
<tr>
<td>A flowering plant (Arabidopsis thaliana)</td>
<td>100 Mb</td>
<td>~25,000</td>
<td>~ 4 Kb</td>
</tr>
</tbody>
</table>

Khatib (2011)

1Mb = 1,000,000 bp

---

**The Phenomic data (phenotypes+genomic)**

1) Massive phenotypic data exist
2) Massive genomic data increasingly available

Example: SNPs (also gene expression)

- $10^7$ SNPs dbSNP 124 (Nat. Center Biotechnology)
- Perlegen: 1.58 million SNPs

Animals:

- Wong et al. (2004) -- chicken genetic variation map with 2.8 million SNPs
- Hayes et al. (2004) -- 2500 SNPs in salmon genome
- Poultry breeding companies -- Thousands of SNPs on sires/dams
- USA (2008) -- >50,000 SNPs in over 3000 Holstein sires
- All over developed world -- chips with 800,000 SNPs
SNP= DNA sequence variation occurring when a single nucleotide - A, T, C, or G in the genome differs between members of a species (or between paired chromosomes)

ABOVE: two sequenced DNA fragments AAGCCTA to AAGCTTA, contain a difference in a single nucleotide.

we say that there are two alleles: C and T

Copy number (CNV) of copy number polymorphisms (CNP): other source of information about genetic variation

- Individuals vary in number of copies of genomic regions
- Disease genes located in CNV regions

Higher number:
- Cancer cells
- Liability to HIV
Fluorescent map, genes in rows

SEQUENCES FOR THOUSANDS OF ANIMALS (WITHIN SPECIES) COMING SOON

A  
<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
<tr>
<td>dog</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
<tr>
<td>mouse</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
<tr>
<td>rat</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
<tr>
<td>cow</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>TACACAAGTGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
<tr>
<td>opossum</td>
<td>CGCAATCTGATGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>CGCAATCTGATGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
<td>CGCAATCTGATGATCTGGCCCTAAACTAGA-CCCCCGGGCTGATTCTGTG</td>
</tr>
</tbody>
</table>

X. laevis (RaxG4)  
X. laevis (AY250711)  
X. tropica (XIRaxG)  

Gene expression: Clustered gene sequences for thousands of animals (within species) coming soon.
Meuwissen, Hayes and Goddard (2001)  
“Genomic selection”  
Better terms:  
“Genome-enabled selection”  
“Genome-assisted selection”  

\[ y = \mu 1_n + \sum_i X_i g_i + e, \]

Effect of chromosomal segment, allelic, haplotype  

ANIMAL BREEDING:  
USE ALL SNP MARKERS IN MODELS  
FOR GENOMIC-ASSISTED EVALUATION  

QUESTION: BYE-BYE QTLS, PEDIGREES, GENES?.

POTATO GENOME  
(Nature 2011)  
- Final assembly 727 Mb  
- Genome size 844 Mb  
- 1 SNP every 40 bp  
- 1 indel every 394 bp (average 12.8 bp)  
- 24,051 genes cluster with at least one of 11 genomes
Essentials of genome-enabled prediction and selection

• Fit (train) some regression model (typically Bayesian) to a data set with markers and phenotypes
• Estimate marker effects
• Predict marked genetic value or phenotype in a new sample (testing or validation sample) for which only DNA information is available
• Once phenotype (or something related to phenotype) is observed, assess quality of prediction. For example, calculate predictive correlation or mean squared error of prediction (choice of metric?)
• Objective: gain reliability and if new sample is of juveniles, reduce generation interval. Dispense with progeny testing? Reduce frequency of phenotyping?

CROSS-VALIDATION

• Data available (genomic, phenotypic)
• Data generated according to unknown process
• Split into training (fitting)- testing (predictand) sets
• Fitting process essentially describes current data (model is typically wrong)
• Use training process to make statement about yet-to-be observed data (testing set)
• Prediction error (conditional and unconditional): point estimate
• Distribution of prediction errors (conditional or unconditional): interval estimate
Schaeffer (2006):

A potential drawback of genome-wide selection may be the existence of interactions or epistatic effects between QTL. If epistatic effects are large, then the accuracy of GEBV may never reach 0.75. A statistical model could be written to account for interactions, but this would likely be very difficult to compute.

YES, IT WOULD BE DIFFICULT!
SEE NEXT...
Dealing with epistatic interactions and non-linearities

- gene x gene
- gene x gene x gene
- gene x gene x gene x gene

(Alice in Wonderland)

**Fixed** effects models
(unravelling “physiological epistasis” a la Cheverud?)

- Lots of “main effects”
- Splendid non-orthogonality
- Lots of 2-factor interactions
- Lots of 3-factor interactions
- Lots of non-estimability
- Lots of uninterpretable high-order interactions
- Run out of “degrees of freedom”

Epistatic networks will probably involve a few genes of large effect
Example of epistatic network

Old fashioned, Ford-T car

Say one knows genes A, B, C, D. Do ANOVA:

A
B
C
D
AB $\rightarrow$ Significant at 0.05
AC $\rightarrow$ Significant at 0.01
AD
BC
BD $\rightarrow$ Significant at 0.01
CD $\rightarrow$ Significant at 0.001

Yawn. nobody will publish…

Modern Swedish car

Publish in Nature and claim new paradigm for epistasis

RANDOM EFFECTS MODELS FOR ASSESSING EPISTASIS REST ON:
Cockerham (1954) and Kempthorne (1954)

--Orthogonal partition of genetic variance into additive, dominance, additive x additive, etc. ONLY if

- No selection
- No inbreeding
- No assortative mating
- No mutation
- No migration
- Linkage equilibrium

JUST CONSIDER Linkage disequilibrium

ALL ASSUMPTIONS VIOLATED!
CLOSE ENCOUNTERS OF THE PREHISTORIC KIND

A prevailing view, and for good reasons (Hill et al., 2008; Crow, 2010; Hill, 2010)

- Fisher’s theorem of natural selection
- Interactions are second-order effects; likely tiny and hard to detect
- Epistasis probably arises with genes of large effects, unlikely to be observed in outbred populations
- Epistatic systems generate additive variance and “release” it, so why worry?
A much less popular view  
(Gianola and a few others)

- If everything behaves as additive, can additive models allow us to learn about “genetic architecture”?
- In areas where phenotypic prediction is crucial (medicine, precision mating) can the exploitation of interaction have added value?
- Is so, should we consider enriching our battery of statistical tricks?

A VIEW OF LINEAR MODELS  
(as employed in q. genetics)

Mathematically, can be viewed as a “local” approximation of a complex process

\[
f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots
\]

- Linear approximation
- Quadratic approximation
- \(n^{th}\) order approximation

FELDMAN and LEWONTIN (1975)
CHEVALET (1994)
How good are linear and quadratic approximations? A Taylor series provides a local approximation only…

\[ y = g(x) + e \]

\[ g(x) = \sin(x) + \cos(x) \]

Structuralism? Systems analysis?

Levi-Strauss (1908-2009)


Foucault (1926-1984)

Althusser (1918-1990)
Will “systems biology” help?

- von Bertalanffy (1968) wrote:

  "There exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relation or 'forces' between them.

  It seems legitimate to ask for a theory, not of systems of a more or less special kind, but of universal principles applying to systems in general.

  In this way we postulate a new discipline called General System Theory. Its subject matter is the formulation and derivation of those principles which are valid for 'systems' in general.

  Concepts like those of organization, wholeness, directiveness, teleology, and differentiation are alien to conventional physics. However, they pop up everywhere in the biological, behavioural and social sciences, and are, in fact, indispensable for dealing with living organisms or social groups. Thus, a basic problem posed to modern science is a general theory of organization.

  General system theory is, in principle, capable of giving exact definitions for such concepts and, in suitable cases, of putting them to quantitative analysis…"

Systems analysis is not new in the animal sciences…

**MODELING BEEF PRODUCTION SYSTEMS**

G. E. Jouvent and T. C. Cartwright

*Texas A&M University, College Station*

**THE USE OF SYSTEMS ANALYSIS IN ANIMAL SCIENCE WITH EMPHASIS ON ANIMAL BREEDING**

T. C. Cartwright

*Texas A&M University, College Station*

**Where is the beef?**
GENERAL McCHRISTAL: “By the time we understand this slide, the war will be over” (and he was sacked by Obama after The Rolling Stones)

WHAT CAN WE EXPECT FROM SYSTEMS ANALYSIS?
SYSTEMS ANALYSIS IN ACTION: PENTAGON “SYSTEMS” VIEW OF THE WAR IN AFGHANISTAN

What is machine learning?
Automatically produce models, such as rules and patterns, from data. Closely related to data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.

Pattern recognition
Neural networks
Data mining
Universal approximations
Machine learning

Cross-validation designs
Kernel methods
Random forest algorithms
Sampling methods
Support vector machines
Ensemble Methods: bagging
Ensemble Methods: boosting
Non-parametric prediction
Bayesian networks

Do not fight over methods (Gonzalez-Recio)
Distinctive aspects of non-parametric fitting

- Investigate patterns free of strictures imposed by parametric models
- Regression coefficients appear but (typically) do not have an obvious interpretation
- Often: very good predictive performance in cross-validation
- Tuning methods and algorithms (maximization, MCMC) similar to those of parametric methods
- Often produce surprising results

PENALIZED and BAYESIAN METHODS for functional inference play a role

- The idea of “penalty is ad-hoc
- It does not arise “naturally” in classical inference
- It appears very naturally in Bayesian inference
  - $L_2$ penalty: equivalent to Gaussian prior
  - $L_1$ penalty: equivalent to double exponential prior
  - Penalties on covariance matrices equivalent to priors (e.g., inverse Wishart)

Bayesian methods arise naturally in predictive inference