Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#3:(Feb-02 & 09-2010)
Genome Wide Association Studies

Outline

1. Genetic Association Studies

2. Lac Operon

The law of causality ... is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ...

–Bertrand Russell, *On the Notion of Cause*. Proceedings of the Aristotelian Society 13: 1-26, 1913.

Outline

1 Genetic Association Studies

2 Lac Operon

Terminology

- Data used in population-based genetic studies have three components:
  1. the *genotype* of the organism under investigation;
  2. a single trait or multiple *traits* (also, called *phenotypes*) that are associated with disease progression or disease status; and
  3. patient specific *covariates*, including treatment history and additional clinical and demographic information.
- **Primary Aim**: Characterize the relationship between the first two of these components: the genotype and a trait
- *Pharmacogenomics*: analyzes how genotypes modify the effects of drug exposure (a *covariate*) on a trait.
- *Focus on certain epidemiological principles*: Confounding and effect mediation, effect modification and conditional association.
- *Haplotypes and Phase Ambiguities*

Genetic Information

- **Genotype**: Observed genetic sequence information — a categorical variable.
- **Genes**: Regions of DNA that are eventually translated into proteins, or involved in the regulation of transcription.
- **Candidate Gene Studies**: A set of genes under investigation is chosen based on known biological function.
- In whole or partial *genome-wide association studies (GWAS)*, segments of DNA across large regions of the genome are considered and may not be accompanied by an a priori hypothesis abut the specific pathways to disease.

Polymorphisms

- Polymorphism refers to multiple alleles of a gene (more generally, genomic region) within a population, usually (but not always) expressing different phenotypes.
- **SNP: Single Nucleotide Polymorphisms**: describe a single base-pair change that is variable across the general population at a frequency of at least 1%. Common SNPs (frequency $>$ 5%); Rare SNPs (5% $>$ frequency $>$ 1%).
- Regions of DNA are said to have *genetic variability* if the alleles within the region vary across a population. *Conserved regions* exhibit no variability in a population.

Multilocus Genotype

- **Multilocus Genotype** describes the observed genotype across multiple SNPs or genes... *Genotype and multilocus genotype are often used interchangeably*.
- A *locus* or *site* refer to the portion of the genome, which encodes a single gene or the location of a single nucleotide on the genome. Multilocus genotype data consist of a set/sequence of categorical variables with elements corresponding to the genotype at each of multiple sites on the genome.

Haplotype

- **Haplotype** refers to the specific combination of alleles that are in *alignment* on a single *homolog* (one of the two homologous chromosomes in humans).
- The corresponding pair of haplotypes is referred to as the individual's *diplotype*.
- **Missingness**: Unobserved haplotypes and haplotype ambiguities. (More to be said on this topic.)

Zygousity

- **Zygousity** is the comparative genetic makeup of two homologous chromosomes.
- An individual is said to be *homozygous* at a given locus (e.g., SNP) if two locus pairs are the same.
- An individual is said to be *heterozygous* at a given locus (e.g., SNP) if it has more than one allele at that site.
- The *loss of heterozygousity* (LOH) refers to the loss of function of an allele when the second allele is already inactive (through inheritance of the heterozygous genotype.)

Allele Frequency

- The *minor allele* frequency (also, referred to as the *variant allele* frequency) refers to the frequency of the less common allele at a variable site.
- Note: in genetics, **frequency** refers to a proportion in the population.
- The terms *homozygous rare* and *homozygous variant* refer to homozygous with two copies of the minor allele.

Traits

- **Population Based Genetic Association Studies**: aim to relate genetic information to a *clinical outcome* or *phenotype* — also, called a *trait*.
- **Quantitative Trait**: A trait described by a continuous variable (taking values in a range), e.g., cholesterol level.
- **Binary Trait**: A trait described by a binary variable (more generally discrete values), e.g., diseased or non-diseased; HIV-positive or HIV-negative.
- **Phenotype**: is defined as a physical attribute or the manifestation of a trait (e.g., measure of disease progression). For instance, IC_{50}, an *in vitro* measure called "50% inhibitory concentration" — Amount of drug required to reduce the replication rate of a virus by 50%.
- **Outcome**: Presence of a disease, but more generally, values taken by a random variable.

Measuring Traits

- Traits can be measured cross-sectionally or over multiple time points (spanning several weeks to years). Data measured over time are referred to as *longitudinal* or *multivariate* data.
- Data and sample size are important; they relate to the questions of *multiple-hypotheses testing* and *overfitting*.

Covariates

- Collection of other information on patient specific characteristics.
- Example: Trait = Cholesterol level among patients at risk for cardiovascular diseases; Genotype = SNP polymorphisms in coding regions of certain genes; Covariates: Gender, age, smoking status and BMI (body mass index).
- A covariate may be *confounding* (associated with the outcome, either because it is a secondary trait or because a genotype is a common cause of the trait and the covariate).
- A covariate may be *causal* (associated with the outcome only in conjunction with a genotype).

Necessary and Sufficient Conditions

- **Brian Skyrms**:
- Practical application of knowledge of causes consists either in (i) producing the cause in order produce the effect or in (ii) removing the cause in order to prevent the effect.
- The word “cause” is used to mean several different things... It is more useful to talk about *necessary conditions* and *sufficient conditions*

Necessary and Sufficient Conditions

Definition

A property F is a *sufficient condition* for a property G if and only if *whenever F is present, G is present.*

Definition

A property H is a *necessary condition* for a property I if and only if *whenever I is present, H is present.*

- Whenever we say that A causes B, we sometimes mean that A is a sufficient condition for B, sometimes that A is a necessary condition for B, sometimes that A is both necessary and sufficient for B, and sometimes none of these things.

Necessary and Sufficient Conditions

Theorem

1. If property A is a sufficient condition *for a property B, then B is a* necessary condition *for a property A.*

Theorem

2. If property C is a necessary condition *for a property D, then D is a* sufficient condition *for a property C.*

Necessary and Sufficient Conditions

Theorem

1'. If property A is a sufficient condition *for a property B, then ¬B is a* sufficient condition *for a property ¬A.*

Theorem

2'. If property C is a necessary condition *for a property D, then ¬D is a* necessary condition *for a property ¬C.*

Necessary and Sufficient Conditions

Theorem

1". If property A is a sufficient condition *for a property B, then* $\neg$*A is a* necessary condition *for a property* $\neg$*B.*

Theorem

2". If property C is a necessary condition *for a property D, then* $\neg$*C is a* sufficient condition *for a property* $\neg$*D.*

The Method of Agreement

- A, B, C ad D are *possible conditioning properties*.
- E is a *conditioned property*.

	A	B	C	D	E
occurrence (1)	P	P	P	A	P
occurrence (2)	P	A	P	P	P
occurrence (3)	A	P	P	A	P

- C is a necessary condition for E.
- Occurrence (1) shows that D cannot be a *necessary condition* for E. It can be eliminated.
- In the same manner, occurrences (2) and (3) eliminate B and A, respectively, as possible *necessary conditions*

The Inverse Method of Agreement

- A, B, C ad D are *possible conditioning properties*.
- E is a *conditioned property*.

	A	B	C	D	E
occurrence (1)	P	A	A	A	A
occurrence (2)	A	P	A	A	A
occurrence (3)	P	A	P	A	A

- D is a sufficient condition for E.
- Occurrence (1) shows that A cannot be a *sufficient condition* for E. It can be eliminated.
- In the same manner, occurrences (2) and (3) eliminate B and C, respectively, as possible *sufficient conditions*.

The Method of Difference

- A, B, C ad D are *possible conditioning properties*.
- E is a *conditioned property*.

	A	B	C	D	E
occurrence (*)	P	A	P	P	P
occurrence (1)	P	A	A	A	A
occurrence (2)	A	A	A	P	A

- We are interested in a sufficient condition for E, among the ones present in occurrence (*).
- Occurrence (*) thus eliminates B (it is absent).
- Occurrences (1) and (2) then eliminate A and D respectively as they are present when E is absent, and hence cannot be sufficient conditions.
- C is a sufficient condition for E.

The Combined Method

- A, B, C ad D are *possible conditioning properties*.
- E is a *conditioned property*.

	A	B	C	D	$\neg A$	$\neg B$	$\neg C$	$\neg D$	E
occurrence (1)	P	A	P	A	A	P	A	P	P
occurrence (1)	A	P	P	P	P	A	A	A	P
occurrence (1)	A	P	A	P	P	A	P	A	A
occurrence (1)	P	A	A	A	A	P	P	P	A

- C i a necessary and sufficient condition for E.

Outline

1 Genetic Association Studies

2 Lac Operon

LAC OPERON

- The modern study of gene regulation was initiated in the 1950's by Francois Jacob, Jacques Monod, Andre Lwoff, and many workers at the Institute Pasteur in Paris.
- They recognized that bacteria, like higher organisms, regulate expressions of their genes.
- For instance, the colon bacteria, *E. coli* does not express all its genes all of the time... even though it is a single-cell and has no developmental process to undergo.
- Jacob and his colleagues studied the ability of *E. coli* to grow on a wide array of different sugars — glucose and lactose.

β-Galactosidase

- The gene encoding the enzyme β-galactosidase is silent until its substrate – lactose – is added to the medium.
- The gene is then turned on and the enzyme is synthesized.
- Initially, there was some confusion whether this was an example of *adaptation* (a population property) or *mutation* (an individual property).
- The process seems too fast to be either a mutation or adaptation...
- Were it a mutation, then the evolution was too *Lamarckian* to be true.

A Simple Model

- One could imagine a model in which *E. coli* alternated between two states $\langle \{S_1, S_2\}, S_1, \mathcal{T} \rangle$, the transition from S_1 to S_2 being controlled by the "guard condition:"

  $$[\text{Glucose}] > [\text{Lactose}]$$

  and the reverse transition by the "guard condition:"

  $$[\text{Lactose}] > [\text{Glucose}] \; .$$

- However, the system acts somewhat differently (a process called, "diauxie") for this state-machine model to be true.

MIll-Skyrms Approach applied to Lac operon

- Diauxie (Meaning double growth)[1] is the phenomenon in which the *E. coli* consumes the glucose as its main carbon source, and switches to lactose (if present) when there is no glucose left.

	G	$\neg G \wedge L$	$Cons(G)$	$Cons(L)$
wildtype (1)	P	A	P	A
wildtype (2)	A	P	A	P

- $\neg G \wedge L \Rightarrow_N Cons(L)$
- Lac^-: This is a mutant which is unable to grow on Lactose.

[1] AUXESIA is the Greek goddess who grants growth and prosperity to the fields...

- $Lac^{-} \mapsto_{mut} Lac^{+}$
- Lac^{+} consumes lactose.
- Lac^{+} controls an enzyme β-galactosidase. It can be assayed with ONPG (o-nitrophenyl-β-D-galactosidase)

$$\begin{array}{llllllll} ONPG & + & \beta - \text{galactosidase} & \Rightarrow & galactose & + & o - nitrophenol \\ (colorless) & & & & & & (yellow) \end{array}$$

(*LacZ* gene makes the enzyme β-galactosidase.)

	$LacZ$	$Cons(L)$
Lac^{+}	< 0.1	A
Lac^{+}	100	P
Lac^{-}	100	A

- (i) $LacZ \Rightarrow_N Cons(L)$ and $LacZ \not\Rightarrow_S Cons(L)$...
- Not yet shown $\neg G \wedge L \Rightarrow_N LacZ$

- Other enzymes involved: *LacY* (β-galactoside permease) and *LacA* (β-galactoside transacetylase)
- Lac^- is a permease-mutant.

	$LacY$	$Cons(L)$
Lac^+	95	P
Lac^-	< 0.1	A

- (ii) $LacY \Rightarrow_N Cons(L)$ and $LacY \not\Rightarrow_S Cons(L)$...
- Similarly, (iii) $LacA \Rightarrow_N Cons(L)$ and $LacA \not\Rightarrow_S Cons(L)$...

LacZ + LacY + LacA

$$[LacZ \wedge LacY \wedge LacA \Rightarrow_N Cons(L)] \text{ and } [LacZ \perp LacY \perp LacA]$$

- One can build z^+y^-, z^-y^+ and z^+y^+ mutants to show that $LacZ \perp LacY$

	$LacZ$	$LacY$	$Cons(L)$
z^+y^+	100	95	P
z^+y^-	100	< 0.1	A
z^-y^+	< 0.1	95	A

$$LacZ \not\Rightarrow_N LacY \text{ and } LacZ \not\Rightarrow_S LacY;$$

$$LacY \not\Rightarrow_N LacZ \text{ and } LacY \not\Rightarrow_S LacZ.$$

More Mutants

- There are two more systems to study: I^+ (wildtype) and I^- – In I^-, *LacZ* is expressed constitutively.

	G	$\neg G \wedge L$	$LacZ$
$I+$	P	A	A
$I+$	A	P	P
$I-$	P	A	P
$I-$	A	P	P

$$(\neg G \wedge L) \perp LacZ | \{LacI, O\}$$

	G	$\neg G \wedge L$	$LacZ$
$I+O^{-}$	P	A	P
$I+O^{-}$	A	P	P
$I-O^{+}$	P	A	P
$I-O^{+}$	A	P	P

- In $I^{+}O^{-}$ and $I^{-}O^{+}$, $LacZ$ is expressed constitutively.

Diploid Mutant

- To analyze regulatory mutants of the lac operon, Jacob developed a system by which a second copy of the lac genes (*LacI* with its promoter, and *LacZYA* with promoter and operator) could be introduced into a single cell.
- A culture of such bacteria, which are diploid for the lac genes but otherwise normal, is then tested for the regulatory phenotype.
- In particular, it is determined whether LacZ and LacY are made even in the absence of any interference (e.g. treatment with IPTG).
- This experiment, in which genes or gene clusters are tested pairwise, is called a **complementation test**.

Diploid Mutant

	G	$\neg G \wedge L$	$LacI$	O_B	$LacZ$
$I^-O^+Z^+$	P	A	A	P	P
$I^-O^+Z^+$	A	P	A	P	P
$I^+O^-Z^+$	P	A	P	A	P
$I^+O^-Z^+$	A	P	P	A	P
$I^+O^+Z^-$	P	A	P	P	A
$I^+O^+Z^-$	A	P	P	P	A
$I{+}O^+Z^+; I^-O^+Z^+$	P	A	{P,A}	P	A
$I{+}O^+Z^+; I^-O^+Z^+$	A	P	{P,A}	P	P
$I{+}O^+Z^+; I^+O^-Z^+$	P	A	P	{P,A}	P
$I{+}O^+Z^+; I^+O^-Z^+$	A	P	P	{P,A}	P
$I{+}O^+Z^-; I^+O^-Z^+$	P	A	P	{P,A}	P
$I{+}O^+Z^-; I^+O^-Z^+$	A	P	P	{P,A}	P

- How do we interpret this?

Interpretation

$$
\begin{aligned}
\neg G \wedge L \quad &\Rightarrow_N \quad \neg LacI(trans) \\
LacI \quad &\Rightarrow_S \quad O_B(cis) \\
\neg O_B \quad &\Rightarrow_N \quad \left\{ \begin{array}{l} LacZ \\ LacY \\ LacA \end{array} \right.
\end{aligned}
$$

Interpretation

$$
\begin{aligned}
G \vee \neg L \quad &\Rightarrow_S \quad LacI \\
&\Rightarrow_S \quad O_B \\
&\Rightarrow_S \quad \begin{cases} \neg LacZ \\ \neg LacY \\ \neg LacA \end{cases}
\end{aligned}
$$

$$LacZ \wedge LacY \wedge LacA \Rightarrow_N Cons(L)$$

How the Story is Told

- The*lac operon* is an operon required for the transport and metabolism of lactose in *Escherichia coli* and some other enteric bacteria.
- It consists of three adjacent structural genes, a promoter, a terminator, and an operator. The lac operon is regulated by several factors including the availability of glucose and of lactose.
- Gene regulation of the *lac operon* was the first complex genetic regulatory mechanism to be elucidated and is one of the foremost examples of prokaryotic gene regulation.

How the Story is Told

- In its natural environment, *lac operon* is a complex mechanism to digest lactose efficiently.
- The cell can use lactose as an energy source, but it must produce the enzyme β-galactosidase to digest it into glucose. It would be inefficient to produce enzymes when there is no lactose available, or if there is a more readily-available energy source available (e.g. glucose).
- The lac operon uses a two-part control mechanism to ensure that the cell expends energy producing β-galactosidase, galactoside permease and thiogalactoside transacetylase only when necessary. It achieves this with the **lac repressor**, which halts production in the absence of lactose, and the Catabolite activator protein (CAP), which assists in production in the absence of glucose.

- This dual control mechanism causes the sequential utilization of glucose and lactose in two distinct growth phases, known as **diauxie**.
- Similar diauxic growth patterns have been observed in bacterial growth on mixtures of other sugars as well, such as glucose and xylose or glucose and arabinose, etc. The genetic control mechanisms underlying such diauxic growth patterns are known as *xyl operon* and *ara operon*, etc.

Why was it so Hard to Interpret?

- For Hinshelwood the repressor (LacI) did not exist. It was a misinterpretation:
- “*If we accept the view that all other effects on the cell of a nonmetabolizable inducer are entirely trivial and secondary, then a purely negative action of this kind would follow. This would mean, however, that the regulatory gene had no real function other than to cut off a normally quite unnecessary process, namely the formation of inducible enzyme. If it had no other function, then natural selection would seem to have done its work very badly, leaving two genes with no function but to frustrate one another.*”

Causation and Correlation

- A fallacy, known as *cum hoc ergo propter hoc* (Latin for "with this, therefore because of this"): Correlations do not imply causation.
- Statements associated with *necessity* and *sufficiency*
- **The INUS condition**: An Insufficient but Non-redundant part of an Unnecessary but Sufficient condition.
- **The Probability Raising condition**
- **Temporal Priority**

Regularity Theories (David Hume)

- **Causes are invariably followed by their effects**: “We may define a cause to be an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.”
- Attempts to analyze causation in terms of invariable patterns of succession are referred to as “regularity theories” of causation.
- There are a number of well-known difficulties with regularity theories, and these may be used to motivate probabilistic approaches to causation.

Imperfect Regularities

- The first difficulty is that most causes are not invariably followed by their effects.
- **Penetrance**: The presence of a disease allele does not always lead to a disease phenotype.
- **Probabilistic theories of causation**: simply requires that *causes raise the probability of their effects*; an effect may still occur in the absence of a cause or fail to occur in its presence.
- Thus smoking is a cause of lung cancer, not because all smokers develop lung cancer, but because smokers are more likely to develop lung cancer than non-smokers.

Imperfect Regularities: INUS condition

- **John Stuart Mill and John Mackie** offer more refined accounts of the regularities that underwrite causal relations.
- **An INUS condition**: for some effect is an *insufficient but non-redundant part of an unnecessary but sufficient condition.*
- **Complexity**: raises problems for the epistemology of causation.

INUS condition

- Suppose, for example, that a lit match causes a forest fire. The lighting of the match, by itself, is not sufficient; many matches are lit without ensuing forest fires. The lit match is, however, a part of some constellation of conditions that are jointly sufficient for the fire. Moreover, given that this set of conditions occurred, rather than some other set sufficient for fire, the lighting of the match was necessary: fires do not occur in such circumstances when lit matches are not present.
- Epistasis, and gene-environment interaction.

Asymmetry

- If A causes B, then, typically, B will not also cause A.
- Causation is usually asymmetric.
- This poses a problem for regularity theories, for it seems quite plausible that if smoking is an INUS condition for lung cancer, then lung cancer will be an INUS condition for smoking.
- One way of enforcing the asymmetry of causation is to stipulate that causes precede their effects in time.

Spurious Regularities

- Suppose that a cause is regularly followed by two effects. For instance, a particular allele A is pleiotropic... It causes a disease trait, but also transcription of another gene B. B may be mistakenly thought to be causing the disease.
- B is also an INUS condition for disease state. But it's not a cause.
- Whenever the barometric pressure drops below a certain level, two things happen: First, the height of the column of mercury in a barometer drops . Shortly afterwards, a storm occurs. Then, it may well also be the case that whenever the column of mercury drops, there will be a storm.

Causes raise the probability of their effects.

- This can be expressed formally using the apparatus of conditional probability.
- Let A, B, C, $\ldots$ represent factors that potentially stand in causal relations.
- Let Pr be a probability function... such that $Pr(A)$ represents the empirical probability that factor A occurs or is instantiated.
- Let $Pr(B|A)$ represent the conditional probability of B, given A.

$$Pr(B|A) = \frac{Pr(A \wedge B)}{Pr(A)}.$$

- If $Pr(A)$ is 0, then the ratio in the definition of conditional probability is undefined. (There are other ways of handling this formally).
- "A raises the probability of B" is that

$$Pr(B|A) > Pr(B|\neg A).$$

PR Axiom

PR: A causes B if and only if$Pr(B|A) > Pr(B|\neg A)$.

Problems

- Probability-raising is **symmetric**: if $Pr(B|A) > P(B|\neg A)$, then $Pr(A|B) > P(A|\neg B)$. The causal relation, however, is typically asymmetric.
- Probability-raising has trouble with spurious correlations. If A and B are both caused by some third factor, C, then it may be that $Pr(B|A) > Pr(B|\neg A)$ even though A does not cause B.
- Those with yellow-stained fingers are more likely to suffer from lung cancer ... smoking tends to produce both effects. Because individuals with yellow-stained fingers are more likely to be smokers, they are also more likely to suffer from lung cancer.
- Intuitively, the way to address this problem is to require that causes raise the probabilities of their effects *ceteris paribus*.

Spurious Correlations

- **Screening off**: If $Pr(B|A \wedge C) = P(B|C)$, then C is said to screen A off from B.
- Equivalently $(A \perp B)|C$...
  $[Pr(A \wedge B|C) = Pr(A|C)Pr(B|C))]$ ... Intuitively, C renders A probabilistically irrelevant to B.
- To avoid the problem of spurious correlations, add a '*no screening off*' (NSO)

NSO

Factor A occurring at time t, is a cause of the later factor B if and only if:

$$Pr(B|A) > Pr(B|\neg A)$$

There is no factor C, occurring earlier than or simultaneously with A, that screens A off from B.

Yule-Simpson Effect

- *NSO does not suffice to resolve the problem of spurious correlations*
- Suppose, for example, that smoking is highly correlated with exercise: those who smoke are much more likely to exercise as well. Smoking is a cause of heart disease, but suppose that exercise is an even stronger preventative of heart disease. Then it may be that smokers are, over all, less likely to suffer from heart disease than non-smokers.
- $A \mapsto$ smoking, $C \mapsto$ exercise, and $B \mapsto$ heart disease, $Pr(B|A) < Pr(B|\neg A)$. Note, however, that if we conditionalize on whether one exercises or not, this inequality is reversed:

$$Pr(B|A \wedge C) > Pr(B|\neg A \wedge C)$$

$$Pr(B|A \wedge \neg C) > Pr(B|\neg A \wedge \neg C).$$

Test Situations

- Causes must raise the probability of their effects in test situations:

TS

TS: A causes B if $Pr(B|A \wedge T) > Pr(B|\neg A \wedge T)\ \ \forall$ test situation T.

- A test situation is a conjunction of factors, which are “held fixed.” This suggests that in evaluating the causal relevance of A for B, we need to hold fixed other causes of B, either positively or negatively.

[End of Lecture #3]