Modeling

Announcements

- you've received a "bio glossary" for this course to ensure no one is impeded by not knowing a term.
- This material was previously covered in the assigned reading, and in Discussion section, but the glossary puts it all in one place in concise form.
- This glossary will be included during your exams, along with a summary sheet of standard equations.
- This class is not about memorization, but instead understanding how to use the concepts!
- Reminder: participation in the discussions is required, and counts as 5% of your grade.

Modeling Principles

(covered in the reading)

- models go from hidden to observable variables
- choose assumptions that are “as simple as possible but no simpler”
- distinguish random variables from constants
- conditional independence enables factoring of probability summation

Phenotype Sequencing Causality?

You are modeling a phenotype sequencing experiment in which a parental strain is subjected to a chemical that causes mutations at random throughout the genome, and many independent mutants are screened for a specific phenotype. Assume there are one or more target genes in which mutations can cause the selected phenotype, and that mutants with this phenotype are equally likely to be mutated in any of the target genes.

Propose causal descriptions (i.e. what causes what) of the pattern of mutation in a *target gene* vs. a *non-target gene*, that explain how they can be distinguished by sequencing multiple independent mutants. Define your variables and write a list of "A causes B" statements.

TARGET

T : NUMBER OF TARGET GENES

k : " $\left(\lambda = \frac{1}{\tau}\right)$ $\lambda_i = \frac{L_i}{S_i}$

τ, λ CAUSES k

S

Phenotype Sequencing Information Graph

Write the information graph for a target gene vs. a non-target gene.

Phenotype Sequencing Information Graph Answer

non-target: mu -> k

target: tau -> k

Forensic Test Causality

You are modeling a forensic test in which two samples, one from a crime scene (X), and another from a suspect (S), are compared using a microarray that scores a large panel of SNPs by measuring the fraction of the sample that had each SNP. Assume that you are given the frequency θ of each SNP in the general population.

.

Propose causal descriptions (i.e. what causes what) of two models for a single SNP: a *match* model that assumes both samples came from the same person; vs. a *mismatch* model that assumes they came from unrelated people. Define your variables and write a list of "A causes B" statements.

Forensic Test Causality Answer

Clearly the microarray measurements approximate the copy number in the person(s) sampled. So we need to include that factor and the likelihoods of getting the same copy number in two unrelated people by random chance. This evidently depends on how frequent the SNP is in the general population. So we choose the following variables:

- θ: frequency of the SNP
- κ, λ: copy number of the SNP in a given person

And the causal flow is

- match: θ causes κ; κ causes X; κ causes S
- mismatch: θ causes κ; κ causes X; θ causes λ; λ causes S

Note that the model follows how we would actually calculate the probability: given θ we can calculate a likelihood for κ; given κ we can calculate a likelihood for X etc.

S

K

X

S

MATCH

K

X

S

MISMATCH

Forensic Test Factoring

Based on the information graphs for a. the *match* model; b. the *mismatch* model, indicate how you could factor the joint probability $p(X, S|\theta)$, by writing the equation for this with appropriate parentheses to indicate which terms if any can be factored.

Forensic Test Factoring Answer

For the match model

$$p(X, S|\theta) = \sum_{\kappa} p(\kappa|\theta)p(X|\kappa)p(S|\kappa)$$

For the mismatch model

$$p(X, S|\theta) = \left(\sum_{\kappa} p(\kappa|\theta)p(X|\kappa)\right)\left(\sum_{\lambda} p(\lambda|\theta)p(S|\lambda)\right)$$