Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Bayesian Inference

- Bayesian inference derives the posterior probability P(H|E) as a consequence of three antecedents, a prior probability P(H), a likelihood P(E|H), and a normalizing Constant P(E) . Bayesian inference computes the posterior probability according to Bayes' rule (diachronic interpretation):

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- Posterior probability P(H|E) = Our update in the probability of the hypothesis H given some evidence E
- Prior probability P(H) = Probability of our hypothesis H before we saw the evidence
- Likelihood P(E|H) = Probability of seeing the evidence E if our hypothesis H is true
- Normalizing Constant P(E) = Probability of the E evidence under any circumstances

The Cookie Problem

- Suppose there are two bowls of cookies.
  - Bowl 1 contains 30 vanilla cookies and 10 chocolate cookies.
  - Bowl 2 contains 20 of each.

- Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla.

- What is the probability that it came from Bowl 1?

- This is a conditional probability; we want p(Bowl 1 | vanilla), but it is not obvious how to compute it.

- If I ask a different question—the probability of a vanilla cookie given Bowl 1—it would be easy: p(vanilla | Bowl 1) = 3/4

- We use Bayes Theorem!

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \qquad \mathbf{P(B_1|V)} = \frac{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)}{\frac{5}{8}}$$

- which reduces to 3/5.

Bayesian Chaining

- Bayes' Rule can written as follows:

Impact of E on P(H)

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)} = P(H) \cdot \left(\frac{P(E|H)}{P(E)}\right)$$

- Now if we have 2 types of independent evidence E_1, E_2 that affects P(H) then we can chain their impacts on H together by having the Posterior Probability of E_1 be the Prior Probability before E_2:

$$P(H|E_1, E_2) = P(H) \cdot \left(\frac{P(E_1|H)}{P(E_1)}\right) \cdot \left(\frac{P(E_2|H)}{P(E_2)}\right)$$

- In general, if we have N independent evidences $E_1,\ldots,E_N$ then we have:

$$P(H|E_1, \ldots, EN) = P(H) \cdot \left(\frac{\prod_{i=1}^{N} P(Ei|H)}{\prod_{i=1}^{N} P(Ei)}\right)$$

eBayes R Package

- Uses Bayesian methods to shrink the estimated sample variances towards a pooled estimate, resulting in far more stable inference when the number of arrays is small.
  - Compromise between unpooled and pooled t-Tests (i.e. - a method of partial-pooling)
  - Uses the evidence about the information from the total ensemble of genes to assist in the inference about each gene individually
- A number of summary statistics are computed by the eBayes() function for each gene and each contrast:
  - M-value (M) is the log2-fold expression or fold change for a gene.
  - A-value (A) is the average expression level for a gene across all the arrays and channels.
  - Moderated t-statistic (t) is the ratio of the M-value over its posterior residual standard deviation (instead of standard deviation). Has the same interpretation as an ordinary t-statistic except that the standard deviations have been moderated across genes, borrowing information from the ensemble of genes to aid with inference about each individual gene (i.e. intelligent partial-pooling).
  - The moderated t-statistic follows a t-distribution with augmented degrees of freedom.
  - p-value for the moderated t-statistic, usually after some multiple hypothesis correction.
  - Moderated F-statistic (F) also borrows information from the ensemble
  - B-statistic (lods or B) is the posterior log-odds that a gene is differentially expressed.

Contrasts

- Sometimes there may be comparisons between the levels of a treatment factor that you are particularly keen to assess. In this case you can test the significance of these individual comparisons using contrasts. Within the ANOVA table, the sums of squares and significance of these comparisons will be printed for each factor.

- Contrast Factor = the factor for which the contrasts are to be applied in the ANOVA.

- Contrast Matrix = a matrix containing the contrasts to be applied in the ANOVA when the selected contrast type is either Regression or Comparison.
  - Each row in the matrix represents a separate contrast, and the columns in the matrix correspond to the factor levels.
  - The row labels in the matrix will be used to label the contrasts in the ANOVA table.

- Contrasts have the property that the sum of the values making up the contrast should be zero.

- The simplest contrast is an individual difference between two levels of a factor and these would be given values -1 and 1.

- Typically a contrast table is used when you have more than 2 levels

Contrast Table

The following table gives common sets of contrasts in a 4 level factor. Any treatment level that is not involved in the contrast is give a value of 0.

Factor level				Contrast Type
A	B	C	D	
-1	1	0	0	Difference between A and B
0	0	-1	1	Difference between C and D
-1	-1	1	1	Difference between average of A and B and that of C and D
-2	1	1	0	Difference between A and the average of B & C
-3	1	1	1	Difference between A and the average of B,C and D
-3	-1	1	3	Linear trend across A, B, C and D

Gene Ontology (GO)

- “Ontologies" consist of a representation of things that are detectable or directly observable, and the relationships between those things.

- The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains:

  - cellular component = the parts of a cell or its extracellular environment
  - molecular function = the elemental activities of a gene product at the molecular level, such as binding or catalysis
  - biological process = operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

- The GO ontology is structured as a directed acyclic graph (DAG), and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.

- The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

Example GO term

- id: GO:0000016
- name: lactase activity
- namespace: molecular_function
- def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
- synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
- synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
- xref: EC:3.2.1.108
- xref: MetaCyc:LACTASE-RXN
- xref: Reactome:20536
- is_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds