Natural Selection

02-715 Advanced Topics in Computational Genomics
Composite Likelihood Test  
(Nielsen et al., 2005)

• Likelihood models for null and alternative hypotheses  
  – Examines the allele frequency spectrum

• Incorporates a scheme for correcting for the ascertainment bias
Composite Likelihood Test 1

- \( p = \{p_1, \ldots, p_{n-1}\} \), where \( p_j \) is probabilities of observing a derived allele frequency \( j \) for \( n \) samples

- Likelihood model under neutral evolution, for \( k \) SNPs in the whole chromosome

\[
CL_1(p) \equiv \prod_{i=1}^{k} p_{X_i} = \prod_{j=1}^{n-1} p_{j}^{k_j}.
\]

- Likelihood model under selective sweep, in a region \((v,b)\)

\[
CL_1(p; v \leftrightarrow b) \equiv \prod_{i=v}^{b} p_{X_i}.
\]

- Test statistic

\[
T_1 = 2\{\log CL_1(\hat{p}_{v \leftrightarrow b}; v \leftrightarrow b) - \log CL_1(\hat{p}; v \leftrightarrow b)\}
\]
Composite Likelihood Test 2

- Incorporate spatial distribution in allele frequencies along the chromosome due to recombinations.

- Assumption: each ancestral lineage in the genealogy has an i.i.d. probability of *escaping* a selective sweep through recombination onto the selected background.
Composite Likelihood Test 2

• The probability of escaping through recombination

\[ P_e = 1 - e^{-\alpha d}. \]

- \( d \): distance \( d \) between a given locus and the selected variant
- \( \alpha \): a parameter that is a function of recombination rate, effective population size, selection coefficient of the selected mutation (e.g., \( \alpha = r \ln(2N)/s \))
Composite Likelihood Test 2

- The probability that $k$ ($0<k<n$) out of $n$ gene copies escaped the sweep:

$$P_e(k) = \binom{n}{k} P_e^k (1 - P_e)^{n-k}$$

- The probability of observing $B$ mutant alleles after a sweep

$$p_B^* = P_e(n)p_B + \sum_{k=0}^{n-1} P_e(k) \left( p_{B+1-n+k,k+1} \frac{B + 1 - n + k}{k + 1} + p_{B,k+1} \frac{k + 1 - B}{k + 1} \right)$$
Simulation Study

- Distribution of test statistics under null hypothesis: test 2 is more robust with respect to recombination rates
Correcting for Ascertainment Bias

- Likelihood for allele frequencies after conditioning on ascertainment (i.e., unobserved true allele frequencies)

\[
L(\theta) \propto \Pr(X_i = \chi_i \mid \theta; \text{Asc}_i) = \frac{\Pr(\text{Asc}_i \mid X_i = \chi_i, \theta)\Pr(X_i = \chi_i \mid \theta)}{\Pr(\text{Asc}_i \mid \theta)}
\]

\[
\Pr(\text{Asc}_i \mid X_i = \chi_i, \theta) = 1 - \frac{\left(\frac{\chi_i}{d}\right) + \left(\frac{n - \chi_i}{d}\right)}{\binom{n}{d}}
\]

\[
\Pr(\text{Asc}_i \mid \theta) = \sum_{j=1}^{n-1} \Pr(X_i = j \mid \theta) \Pr(\text{Asc}_i \mid X_i = j)
\]
Correcting for Ascertainment Bias
(Nielson et al., 2004)

- Illustration through simulation study
HapMap Data Analysis

• HapMap chromosome 2

• Test 1: requires a choice of window size

• Test 2: no need to fix the window size
Ascertainment Bias from HapMap Analysis
Evolution

02-715 Advanced Topics in Computational Genomics
Cross-species Sequence Analysis

• Functional regions of genomes are conserved across species

• Cross-species sequence conservation is believed to occur because of negative (purifying) selection

• About 5% or more of bases in mammalian genomes are under purifying selection

• Protein coding genes account for 1.5% of the regions under purifying selection
Phylo-HMM

• Parse aligned sequences into two classes
  – Conserved vs. nonconserved

• Maximum likelihood estimation of parameters of Phylo-HMM
Phylo-HMM

- $\mu, \nu$: transition probabilities

- States
  - c: conserved region
  - n: non-conserved region

- $\psi_n, \psi_c$: emission probabilities as a tree
Phylo-HMM

- $\psi_n, \psi_c$: emission probabilities as a phylogenetic model
  - Identical phylogenetic model structure for two states
    - $\rho$: scaling factor for branch length $0 \leq \rho \leq 1$
        - Average substitution rate
Datasets

• Vertebrate species
  – Human, mouse, rat, chicken, fugu rubripes
  – Alignment with human as reference sequence

• Insect species
  – Three species of Drosophila and Anopheles gambiae
  – Alignment with D. melanogaster as reference sequence

• Two species of Caenorhabditis
  – Alignment with C. elegans as reference sequence

• Seven species of saccharomyces
  – Alignment with S. cerevisiae as reference sequence
Phylogenetic Models: Assumed Topologies and Estimated Branch Lengths
Estimated Conserved Elements

- More complex organisms have more conserved regions outside of coding regions.
Conservation Around GRIA2 in Human
Extreme Conservation

- Extreme conservation at the 3’ end of the *ELAVL4* gene
Key Observations

• Conserved regions
  – 3%-8% of the human genome conserved in vertebrates and other mammals
  – 37-53% in D. melanogaster
  – 18-37% in C. elegans
  – 47-68% in S. cerevisiae

• Highly conserved regions (HCE)
  – 42% of HCEs overlap with exons in vertebrate genomes
  – >93% for insects, worms, yeasts

• Extreme conservations in 3’ UTRs
  – Post-transcriptional regulation?

• HCEs in intron regions
  – Enriched for RNA secondary structure: encoding functional RNAs?