Today’s Lecture

• PhastCons
PhastCons PhyloHMM

\[ \mu = a_{cn} \]

\[ \nu = a_{nc} \]

• branch lengths:
  – Expected # substitutions/site over corresponding evolutionary time period
  – for neutral state, should reflect underlying mutation rate
  – for conserved state: mutation rate × scaling factor \( \rho \)
    • \( \rho = \text{frac of mutations that escape purifying selection} \)
    • \( \rho \approx 0.33 \) (for vertebrates)
Some general issues in applying probability models, in the PhyloHMM context

• Is the model computable?

• Is the model ‘reasonable’?
  – 2 states enough?
  – Markov condition on transition probabilities

• How good is the input data?
  – Alignability of neutral sequence
  – Accuracy of genome sequence alignments

• Are results reliable?
  – No true ‘test set’ – instead, putative false positive rate, and ‘biological plausibility’ of findings
Alignment issues

- **Multiz**: progressive pairwise alignments
- **accurate multiple genome alignment** *not* a solved problem!
  - ENCODE region alignment analyses: Margulies EH *et al.* 2007
  - major issues:
    - accurate gap placement (even for close species!!)
    - discrimination among paralogous sequences (e.g. repeats, duplications)
- **inaccurate alignments cause**
  - neutral rate to be *overestimated*
  - conserved segments to be *overidentified*
    - because more slowly mutating (or better aligned) neutral segments may be called conserved
• for distantly related species, neutrally evolving regions no longer alignable
  – analyze 4D sites in coding sequences to estimate neutral rates
    • CDS alignments much more reliable, but
    • synonymous sites somewhat atypical (some selection; composition & mutation patterns)
Notation

- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected ‘coverage’ $\gamma$ (frac of genome that is conserved):
  
  $\gamma = \frac{\text{Elen\ (cons seg)}}{\text{Elen\ (cons seg)} + (\text{Elen\ (neut seg)})}$
  
  $= \frac{1/\mu}{1/\mu + 1/\nu}$
  
  $= \nu / (\mu + \nu)$


- transition probs imply *a priori* length dist’ns for conserved & non-conserved segments
  
  - prob(cons seg has length \( n \)) is
    \[
    (a_{cc})^{n-1}a_{cn} = (a_{cc})^{n-1}(1 - a_{cc})
    \]
  
  - geometric distribution
  
  - expected length (Elen) \( \omega \) of conserved segment is
    \[
    1.0 / (1 - a_{cc}) = 1.0 / a_{cn}
    \]

  special case: \( a_{cc} = .5 = a_{nn} \Rightarrow \) positions are independent
PhastCons Parameter Estimation

- parameters estimated separately in 1 Mb windows using EM algorithm
  - full maximum likelihood analysis, or
  - constraining some parameters
  & averaged over genome

- full MLE results don’t match biologists’ intuition -- too much ‘smoothing’:
  - fewer, & larger, conserved elements
  - long, apparently non-conserved regions within conserved elements
  - attributed to fact that (prior) geometric length dist’n inappropriate
<table>
<thead>
<tr>
<th>Group</th>
<th>Method</th>
<th>Total no.</th>
<th>Ave. len.</th>
<th>Cov.</th>
<th>CDS cov.</th>
<th>$\mu$</th>
<th>$\nu$</th>
<th>$\omega$</th>
<th>$\gamma$</th>
<th>$L_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>vert.</td>
<td>MLE</td>
<td>561,103</td>
<td>216.1</td>
<td>4.2%</td>
<td>68.8%</td>
<td>0.018</td>
<td>0.004</td>
<td>55.4</td>
<td>0.191</td>
<td>30.4</td>
</tr>
<tr>
<td></td>
<td>55%</td>
<td>1,058,855</td>
<td>75.3</td>
<td>2.8%</td>
<td>56.8%</td>
<td>0.125</td>
<td>0.029</td>
<td>8.0</td>
<td>0.187</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>65%c</td>
<td>1,157,180</td>
<td>103.5</td>
<td>4.2%</td>
<td>66.1%</td>
<td>0.083</td>
<td>0.030</td>
<td>12.0</td>
<td>0.265</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>1,381,978</td>
<td>167.5</td>
<td>8.1%</td>
<td>76.6%</td>
<td>0.043</td>
<td>0.031</td>
<td>23.0</td>
<td>0.415</td>
<td>22.6</td>
</tr>
<tr>
<td>vert.</td>
<td>65%</td>
<td>1,157,180</td>
<td>103.5</td>
<td>4.2%</td>
<td>66.1%</td>
<td>18.0%</td>
<td>0.611</td>
<td>16.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4d</td>
<td>797,777</td>
<td>109.3</td>
<td>3.0%</td>
<td>64.2%</td>
<td>24.0%</td>
<td>0.854</td>
<td>11.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Instead: -- impose constraints

- coverage constraint:
  - 65% of coding bases covered by conserved elts
  - (target value based on earlier mouse/human analysis)

- smoothness constraint:
  - PIT (≡ expected min. amt of phylogenetic info required to predict a conserved element)
    \[ = 9.8 \text{ bits} \]
    - (forced to be same for all species groups)
• constraints met by ‘tuning’ $\gamma$ and $\omega$ (or equivalently transit probs)
  – choose $\gamma$ and $\omega$,
  – get ML estimates of other parameters by EM algorithm
  – see whether get desired coverage & PIT;
  – if not, adjust $\gamma$ and $\omega$ & redo
Notation

• $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)

• $\nu = a_{nc}$

• expected ‘coverage’ $\gamma$ (frac of genome that is conserved):

  $= \frac{\text{Elen (cons seg)}}{\text{Elen (cons seg)} + \text{Elen (neut seg)}}$

  $= \frac{1/\mu}{1/\mu + 1/\nu}$

  $= \frac{\nu}{\mu + \nu}$

\[ x = \text{TCGGCACTATACGA...} \]
• $L_{\text{min}}$: expected min length of a conserved segment that could appear in a Viterbi path

• at $L_{\text{min}}$, expected loglike of staying in state $n$
  
  = expected loglike of switching to c & back again, so

$$(L_{\text{min}} + 1) \log(1 - \nu) + L_{\text{min}} \sum_x P(x|\psi_c) \log P(x|\psi_n)$$

  
  $= \log \nu + \log \mu + (L_{\text{min}} - 1) \log(1 - \mu) + L_{\text{min}} \sum_x P(x|\psi_c) \log P(x|\psi_c)$$

• $L_{\text{min}} = \frac{\log \nu + \log \mu - \log(1 - \nu) - \log(1 - \mu)}{\log(1 - \nu) - \log(1 - \mu) - H(\psi_c||\psi_n)}$
where
\[ H(\psi_c \| \psi_n) = \sum_x P(x|\psi_c) \log \frac{P(x|\psi_c)}{P(x|\psi_n)} \]
= rel entropy of \(c\)-state emission prob dist’n w.r.t. \(n\)-state dist’n

PIT (phylogenetic information threshold)
\[ = L_{\min} H(\psi_c \| \psi_n) \]
= ‘expected min amt of phylogenetic info required to predict conserved element’
• Final param estimates (for vertebrates):
  – $\gamma = 0.265$
  – $\omega = 12.0$ bp
  – $H(\psi_c \ || \ \psi_n) = .608$ bits / site
  – $L_{\text{min}} = 16.1$ bp
  – PIT $= L_{\text{min}} \ H(\psi_c \ || \ \psi_n) = 9.8$ bits
<table>
<thead>
<tr>
<th>Group</th>
<th>Method</th>
<th>Total no.</th>
<th>Ave. len.</th>
<th>Cov.</th>
<th>CDS cov.</th>
<th>μ</th>
<th>ν</th>
<th>ω</th>
<th>γ</th>
<th>$L_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>vert.</td>
<td>MLE</td>
<td>561,103</td>
<td>216.1</td>
<td>4.2%</td>
<td>68.8%</td>
<td>0.018</td>
<td>0.004</td>
<td>55.4</td>
<td>0.191</td>
<td>30.4</td>
</tr>
<tr>
<td></td>
<td>55%</td>
<td>1,058,855</td>
<td>75.3</td>
<td>2.8%</td>
<td>56.8%</td>
<td>0.125</td>
<td>0.029</td>
<td>8.0</td>
<td>0.187</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>65%</td>
<td>1,157,180</td>
<td>103.5</td>
<td>4.2%</td>
<td>66.1%</td>
<td>0.083</td>
<td>0.030</td>
<td>12.0</td>
<td>0.265</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>1,381,978</td>
<td>167.5</td>
<td>8.1%</td>
<td>76.6%</td>
<td>0.043</td>
<td>0.031</td>
<td>23.0</td>
<td>0.415</td>
<td>22.6</td>
</tr>
<tr>
<td>vert.</td>
<td>65%</td>
<td>1,157,180</td>
<td>103.5</td>
<td>4.2%</td>
<td>66.1%</td>
<td>18.0%</td>
<td></td>
<td></td>
<td>0.611</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>4d</td>
<td>797,777</td>
<td>109.3</td>
<td>3.0%</td>
<td>64.2%</td>
<td>24.0%</td>
<td></td>
<td></td>
<td>0.854</td>
<td>11.0</td>
</tr>
</tbody>
</table>
Estimating false positive rates

- simulate 1 Mb alignment
  - by sampling 4D sites (with replacement) from aligned CDSs
  - caveat: these not typical of all neutral sites!
- predict cons elts (using prev param estimates)
- frac of bases in cons elts:

<table>
<thead>
<tr>
<th>Group</th>
<th>65%</th>
<th>75%</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>vertebrate</td>
<td>0.00279</td>
<td>0.00362</td>
<td>0.00005</td>
</tr>
<tr>
<td>insect</td>
<td>0.00286</td>
<td>0.01026</td>
<td>0.00152</td>
</tr>
<tr>
<td>worm</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>yeast</td>
<td>0.00006</td>
<td>0.00042</td>
<td>0.00023</td>
</tr>
</tbody>
</table>
• does not address (important) issue of rate of false positive bases within, or flanking, true conserved elements

• also: genes more G+C rich than genome average, & have somewhat higher mutation rate (due in part to more frequent CpGs)

  ⇒ underestimating false pos rate

• also: randomization procedure destroys underlying mutation rate variation

  ⇒ underestimating false pos rate
Characteristics of phastCons predicted conserved elements

• 1.18 million elements
• constitute 4.3% of human sequence
  – 66% of coding bases
    • 88% of coding exons overlap predicted elt
  – 23% of 5’UTR bases
    • 63% of exons
  – 18% of 3’UTR bases
    • 64% of exons
  – 42% of RNA gene bases
    • 56% of genes
  – 3.6% of intronic bases
  – 2.7% of intergenic bases
  – < 1% of mammalian ‘ancestral repeats’ (ARs)