Today’s Lecture

• PhastCons

• Karlin-Altschul theory
Notation

• $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
• $\nu = a_{nc}$
• expected ‘coverage’ $\gamma$ (frac of genome that is conserved):
  
  $= \frac{\text{Elen (cons seg)}}{\text{Elen (cons seg)} + \text{Elen (neut seg)}}$
  
  $= \frac{(1/\mu)}{(1/\mu + 1/\nu)}$
  
  $= \frac{\nu}{(\mu + \nu)}$
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 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
• $L_{\text{min}}$: 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
  – $\text{PIT} = L_{\text{min}} H(\psi_c\|\psi_n) = 9.8$ 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>$\mu$</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%</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$^a$</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)
Context for Karlin-Altschul Theory for Maximal Segment Analysis

- Linked list, with labels attached to edges, e.g.
  - a sequence graph: labels = sequence residues
  - (ungapped) aligned pair of seqs: labels = possible alignment columns (pairs of residues)
- edge weights depend only on labels:
  - each label is assigned a weight $W(s) = w_s$

```
-2  1  1  1  1  -2  1  1  1  -2  -2  1
A  C  C  G  C  T  G  C  G  A  A  G
```
• in backgd model, each label $s$ occurs with probability $P(s) = p_s$ where
  – $P = \text{prob dist’n on sample space } S = \{\text{labels}\}$
Methods for Computing Statistical Significance of Maximal Segment Scores

1. exact prob dist’n
2. approximate formula (Karlin-Altschul)
3. from simulated sequences
4. from real biological ‘background’ sequences
   – i.e. not having feature in question

1, 2, 3 require prob model approximating biological reality; 4 requires an appropriate dataset
2 is faster than 1 or 3, but involves add’l approximations (ignores ‘edge effects’)
1 requires more complex algorithm
Exact Score Dist’n for Segments in WLLs

- Exact score dist’n (following proof allows position-specific scores and probabilities):
  - Let $P_{k,m}^{(i)}$ = prob that:
    - highest-scoring path ending at position $i$ has score $k$, and also
    - highest scoring path ending at any pos’n $\leq i$ has score $m$
  - special cases:
    - $P_{k,m}^{(i)} = 0$ if $k < 0$ or $m < k$;
    - $P_{0,0}^{(0)} = 1$,
    - $P_{k,m}^{(0)} = 0$ if $k$ or $m \neq 0$
  - dist’n of maximum score is $P_m = \sum_{k \leq m} P_{k,m}^{(N)}$.
    ($N =$ seq length)
• Algorithm to compute \( \{ P_{k,m}^{(i)} \} \) from \( \{ P_{k,m}^{(i-1)} \} \):
  
  – If \( 0 < k < m \)
    
    • \( \implies \) best path ending at position \( i \) cannot start at \( i \), and best path ending at position \( \leq i - 1 \) must have score = \( m \)

    \[
    P_{k,m}^{(i)} = \sum_j p_j^{(i)} P_{k-j,m}^{(i-1)}
    \]

  – if \( 0 < k = m \)
    
    • \( \implies \) best path ending at position \( \leq i - 1 \) may have score \( \leq m \)

    \[
    P_{k,m}^{(i)} = \sum_j p_j^{(i)} \sum_{n\leq m} P_{k-j,n}^{(i-1)}
    \]

  – \( P_{0,m}^{(i)} = \sum_j p_j^{(i)} \sum_{n\leq-j} P_{n,m}^{(i-1)} \)

  – stop when \( i \) reaches \( N \)
• Can incorporate Markov chain dependencies in sequence probs:
  – just keep track of preceding residue $r$ as well as $k,m : P_{r,k,m}^{(i)}$.

• Reduce required memory by truncating for large $m$, with appropriate modifications.

• Would like to have generalization to arbitrary DAG (e.g. edit graphs for sequence alignment)!
  – Difficult, because $P_{k,m}^{(v)}$ not independent for different parent vertices $v$
Why Is *Approximation* to Exact Score Distribution of Interest?

- faster to compute: useful for database searches
- gives better intuition for score behavior
- *Form* of approximation extends to other situations
  - e.g. gapped alignments
where exact dist’n currently unavailable
Approximate Score Distribution for High-Scoring Segments in WLLs: Karlin-Altschul theory

- Main reason why BLAST is most widely used computational biology tool!
- Ideas closely related to
  - classical random walk and gambler’s ruin problems in probability theory
  - sequential sampling in statistics