• Detecting sequence conservation with PhyloHMMs
  – PhastCons
• PhyloHMMs: Yang 1995; Felsenstein & Churchill 1996

  – basis of PhastCons conservation scores (UCSC genome browser)
Goal: starting from multiple genome sequence alignment, identify
– conserved regions (regions under purifying selection), against background of
– neutrally evolving regions
PhastCons PhyloHMM

• model:
  – 2-state HMM
    c: conserved state
    n: neutral (or nonconserved) state
  – emitted symbols are alignment columns
  – emission probabilities based on phylogenetic tree relating sequences
  – gaps in alignment treated as missing data
• 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 $\rho$
    • $\rho = \text{frac of mutations that escape purifying selection}$
    • $\rho \approx .33$ (for vertebrates)
Probability calculations on evolutionary tree (lecture 11)

- Given:
  1. a set of observed residues at the leaves
     (a gap-free alignment column of the sequences)
  2. \( \{P_e(s \mid r)\} \) and \( \{P_{root}(r)\} \)

compute prob of observed residues

- Still exponentially many (in \( n_{anc} \)) possibilities for ancestral residues!
- But can use dynamic programming on a WDAG…
urnode

\[ P_{\text{root}}(r) \]

\[ P_{g}(s \mid r) \]

\[ P_{e}(A \mid r) \]

\[ P_{f}(G \mid r) \]

\[ P_{h}(G \mid r) \]
cf. WDAG for 3-state HMM length n sequence (lecture 13)

weights are emission probabilities $e_k(b_i)$ for $i^{th}$ residue $b_i$

weights are transition probabilities $a_{kl}$

$b_{i-1}$  
position $i-1$

$b_i$  
position $i$

$b_{i+1}$  
position $i+1$
Prob calcs in HMMs (lecture 14):

For each vertex $v$, let $f(v) = \sum_{\text{paths } p \text{ ending at } v} \text{weight}(p)$, where weight($p$) = product of edge weights in $p$. Only consider paths starting at ‘begin’ node.

Compute $f(v)$ by dynam. prog: $f(v) = \sum_{i} w_i f(v_i)$, where $v_i$ ranges over the parents of $v$, and $w_i = \text{weight of the edge from } v_i \text{ to } v$.

Similarly for $b(v) = \sum_{p \text{ beginning at } v} \text{weight}(p)$

The paths beginning at $v$ are the ones ending at $v$ in the reverse (or inverted) graph.
\[ f(v)b(v) = \text{sum of the path weights of all paths through } v. \]

\[ f(v')wb(v) = \text{sum of the path weights of all paths through the edge } (v',v) \]
urnode

\[ P_{\text{root}}(r) \]

\[ P_g(s \mid r) \]

\[ P_e(A \mid r) \]

\[ P_f(G \mid r) \]

\[ P_h(G \mid r) \]
• Compute overall probability of leaf residues (nucleotides) by dynamic programming on WDAG:

• Let, for each node \( v \), \( f(v) = \text{prob of leaf nucs below } v \) (i.e. tree-descendants, or WDAG-ancestors, of \( v \)), given \( v \)’s nuc

\[ f_{\text{left}}(v) = \text{prob of leaf nucs below and to left} \]

\[ f_{\text{right}}(v) = \text{prob of leaf nucs below and to right} \]

then \( f(v) = f_{\text{left}}(v) f_{\text{right}}(v) \)
• Compute these values node-by-node, visiting (WDAG-)parents before children:
  – starting at leaf nodes (setting $f(v) = 1$), ending at urnode

$$f_{\text{left}}(v) = \sum_{\text{left} - u} w(u, v) f(u) \quad \text{where}$$

  – $u$ ranges over parent nodes to the left
  – $w(u, v) = \text{weight on edge from } u \text{ to } v$

  ($= \text{mutation prob from } v \text{ to } u$)

Similarly for $f_{\text{right}}(v)$

$$f(v) = f_{\text{left}}(v) f_{\text{right}}(v)$$

  – For $v = \text{urnode}$, view all parents as being to ‘left’ and $f(v) = f_{\text{left}}(v)$

• $f(\text{urnode}) = \text{probability of the observed leaf nucs}$
• a ‘forward-backward’ calc gives posterior prob of having
  – a particular nuc at an ancestral node, or
  – a particular mutational change along an edge
• can use these as *fractional counts* to estimate $P$’s (EM algorithm)
Siepel *et al* evolutionary model

- single, reversible, infinitesimal mutation process across tree
- branches differ only in their lengths
- selection strength same across tree and sites