Today’s Lecture

• More HMM examples

• Limitations of HMMs

• PhyloHMMs

• PhastCons
HMM Examples (cont’d)

• Ordinary Markov chain model:
  – states = observed symbols
  – emission probs = 1 or 0
  – transition probs = prob of observing a symbol, given the preceding one.

• Order $k$ Markov model
  – states = length $k$ words (e.g. $b_1b_2 ... b_k$)
  – (unique) symbol emitted by $b_1b_2 ... b_k$ is $b_k$
  – transition prob from $b_1b_2 ... b_k$ to $c_1c_2 ... c_k$ is non-zero only if
    • $c_1c_2 ... c_{k-1} = b_2b_3 ... b_k$, in which case it is $P(b_{k+1}|b_1b_2 ... b_k)$ where $b_{k+1} = c_k$
Limitations of HMMs

• Markov chain cond’n on states is unrealistic
  – biological features have complex dependencies

• In particular, duration modelling frequently unrealistic –
  – can deal with this
    • Increase number of states
    • ‘generalized HMMs’
  – but at cost of speed & elegance

• Other issues (applicable to any complex models!)
  – Parameter estimation can be difficult and give suboptimal results
    • many local maxima in complex surface
  – Need to avoid overfitting
Detecting sequence conservation with PhyloHMMs

- 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
    - \textit{c}: conserved state
    - \textit{n}: neutral (or nonconserved) state
  - emitted \textit{symbols} are \textit{alignment columns}
  - emission \textit{probabilities} based on \textit{phylogenetic tree} relating sequences
    - discussed in Genome 541, or molecular phylogeny course
  - gaps in alignment treated as \textit{missing data}