Today’s Lecture

- More HMM examples
- Limitations of HMMs
- PhyloHMMs & PhastCons
HMM Examples (cont’d)

• Simple 7-state prokaryote genome model:
  – 1 state for intergenic regions
  – 3 states for codon positions in top-strand genes
  – 3 for codon positions in bottom-strand genes

• more complex models including sites (with states for each position in site) –
  – promoter elements
  – Shine-Dalgarno (translation start site)
  – (in eukaryotes) splice sites, polyadenylation sites etc.
7-state model for prokaryote genomes

- intergenic
- first codon position – top strand coding sequence
- second codon position – top strand coding sequence
- third codon position – top strand coding sequence
- first codon position – bottom strand coding sequence
- second codon position – bottom strand coding sequence
- third codon position – bottom strand coding sequence

a (very short!) ‘bottom-strand’ gene, in a different region of the genome:
• N.B. the emitted symbols are always top strand nucleotides!
Other HMM examples (see Durbin *et al.*)

- protein families (like site models – but important to allow insertions & deletions)
- Pair HMMs
- protein structure (symbols emitted are structural elements)
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. $b_1 b_2 \ldots b_k$)
  – (unique) symbol emitted by $b_1 b_2 \ldots b_k$ is $b_k$
  – transition prob from $b_1 b_2 \ldots b_k$ to $c_1 c_2 \ldots c_k$ is non-zero only if
    • $c_1 c_2 \ldots c_{k-1} = b_2 b_3 \ldots b_k$, in which case it is
      $P(b_{k+1}|b_1 b_2 \ldots b_k)$ where $b_{k+1} = c_k$
D-segments & 2-state HMMs

- Consider 2-state HMM
  - states 1 & 2, transition probs $a_{11}, a_{12}, a_{21}, a_{22}$
  - observed symbols $\{r\}$, emission probs $\{e_1(r)\}, \{e_2(r)\}$

- Define
  - scores $s(r) = \log(e_2(r) a_{22}/(e_1(r) a_{11}))$
  - $S = -D = \log(a_{11}a_{22}/(a_{21}a_{12}))$

- Then if $S > 0$, the maximal D-segments in a sequence $(r_i)_{i=1,n}$ are the state-2 segments in the Viterbi parse.

- So via D-segment algorithm can get Viterbi parse in just one pass through the sequence!

- can allow for non-.5 initiation probs by starting cumul at non-zero value
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 (arising with 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
    - \( c \): conserved state
    - \( n \): neutral (or nonconserved) state
  - emitted symbols are *alignment columns*
  - emission *probabilities* based on *phylogenetic tree* relating sequences
    - discussed in Genome 541, or molecular phylogeny course
  - gaps in alignment treated as *missing data*