

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. $b_1b_2 \dots b_k$)
 - (unique) symbol emitted by $b_1b_2 \dots b_k$ is b_k
 - transition prob from $b_1b_2 \dots b_k$ to $c_1c_2 \dots c_k$ is non-zero only if
 - $c_1c_2 \dots c_{k-1} = b_2b_3 \dots b_k$, in which case it is $P(b_{k+1}|b_1b_2 \dots b_k)$ 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

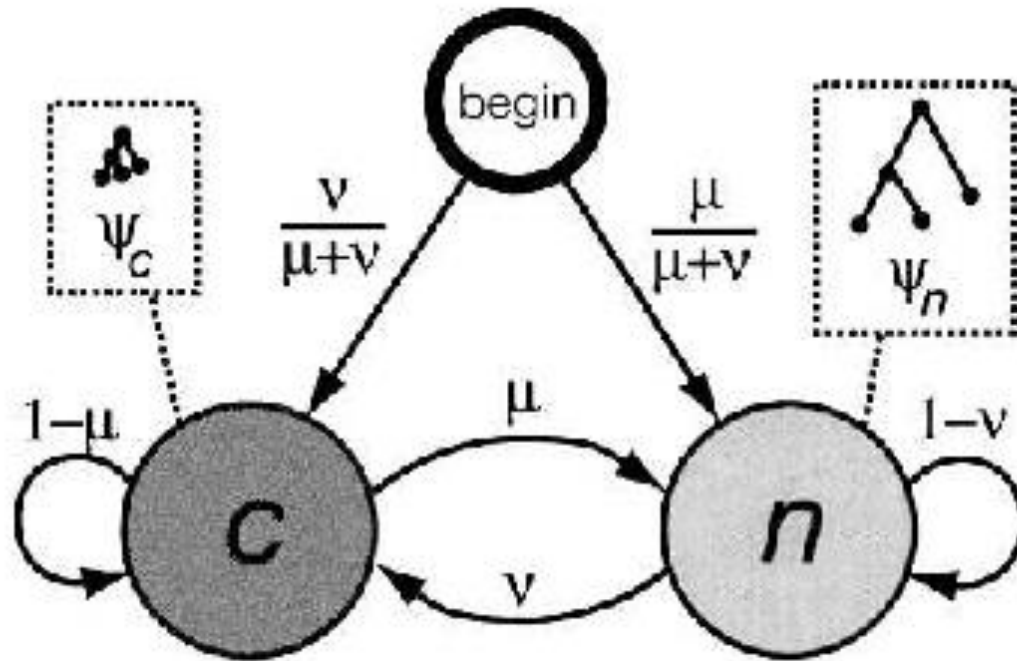
- PhyloHMMs: Yang 1995; Felsenstein & Churchill 1996
- Siepel A. *et al.* (2005): Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50
 - basis of PhastCons conservation scores (UCSC genome browser)

- 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*

PhastCons PhyloHMM



$$\mu = a_{cn}$$

$$v = a_{nc}$$

x = TCGCGACATATACGA . . .
TTGGGGGCATGTGGGT . . .
AGCAGACGTCCGCAA . . .