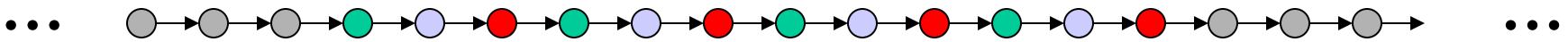
Today's Lecture

- More HMM examples
- Limitations of HMMs
- PhyloHMMs
- PhastCons

HMM Examples (cont'd)

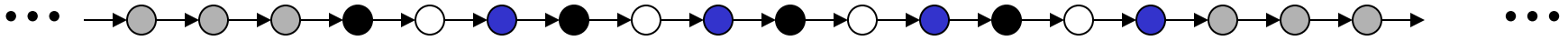
- Simple 7-state prokaryote genome model:
  - 1 state for intergenic regions
  - 3 states for codon positions in top-strand genes
  - 3 for codon positions in bottom-strand genes
- more complex models including sites (with states for each position in site) –
  - promoter elements
  - Shine-Dalgarno (translation 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 $b_{k+1} = c_k$

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
- 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)

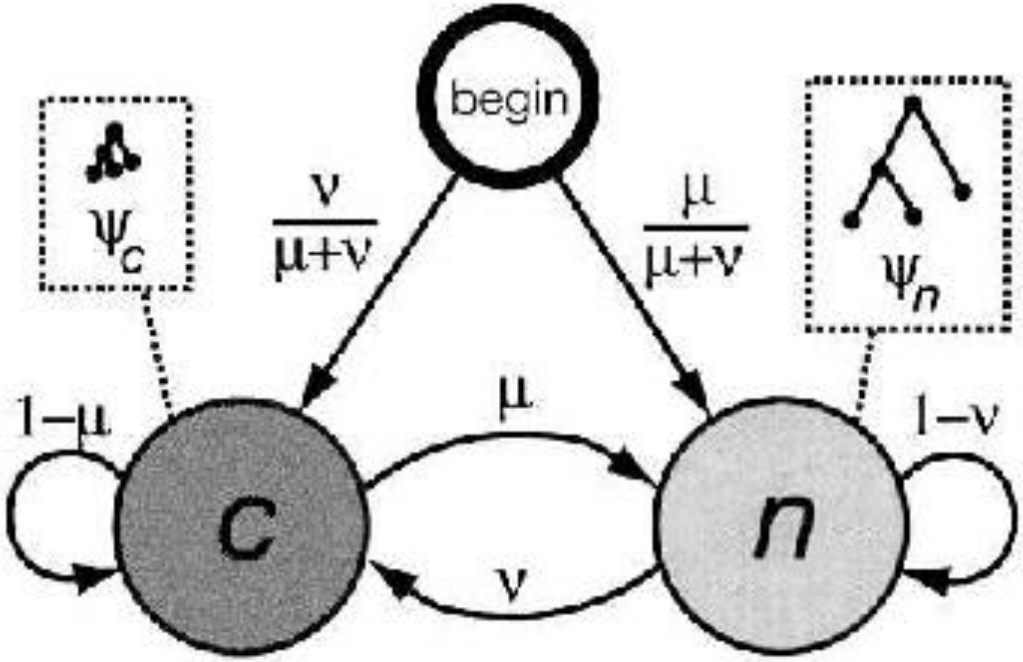
- Goal: starting from multiple genome sequence alignment, identify
  - conserved regions (regions under purifying selection),

  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

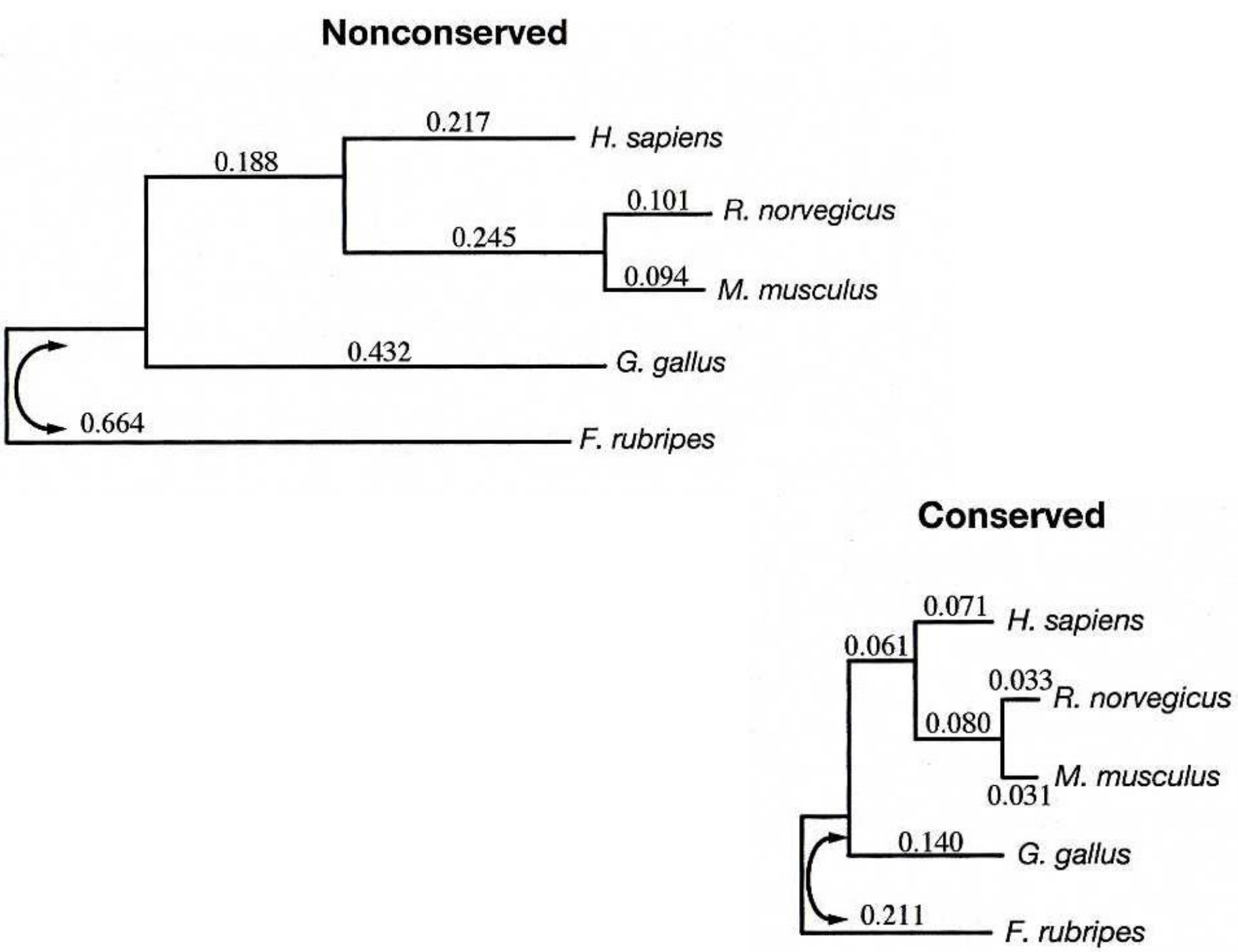


$\mu = a_{cn}$

$\nu = a_{nc}$

$$\mathbf{x} = \begin{matrix} \text{TCGCGACATATACGA}\ldots \\ \text{TTGGGGCATGTGGGT}\ldots \\ \text{AGCAGACGTCCGCAA}\ldots \end{matrix}$$

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.


Nonconserved
0.217
H. sapiens
0.188
0.101
R. norvegicus
0.245
0.094
M. musculus
0.432
G. gallus
0.664
F. rubripes
Conserved
0.071
H. sapiens
0.061
0.033
R. norvegicus
0.080
M. musculus
0.031
0.140
G. gallus
0.211
F. rubripes