Lecture 15

- Detecting sequence conservation with PhyloHMMs
  - PhastCons

- 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)

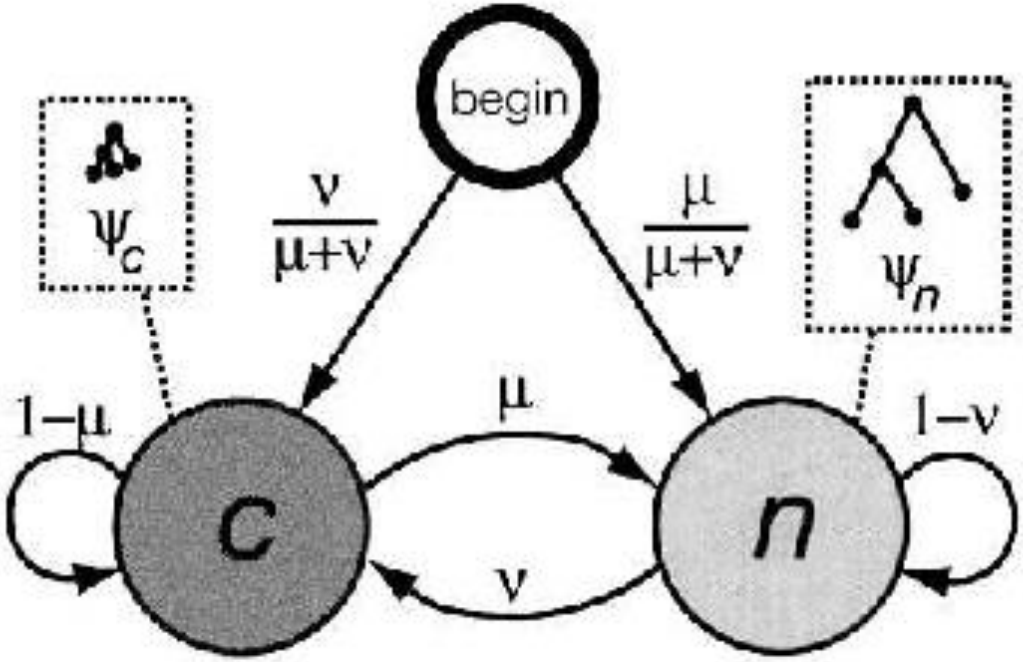
- Goal: starting from multiple genome sequence alignment, identify
  - conserved regions (regions under purifying selection),

  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
  - gaps in alignment treated as *missing data*

PhastCons PhyloHMM

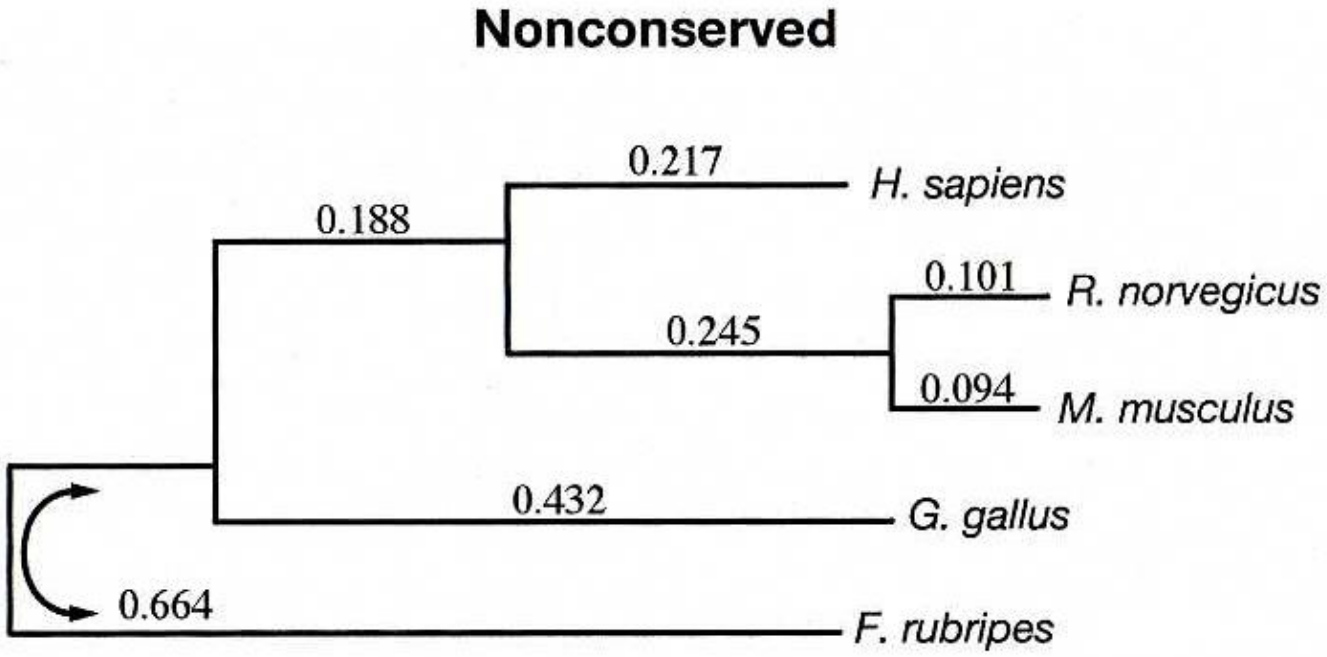


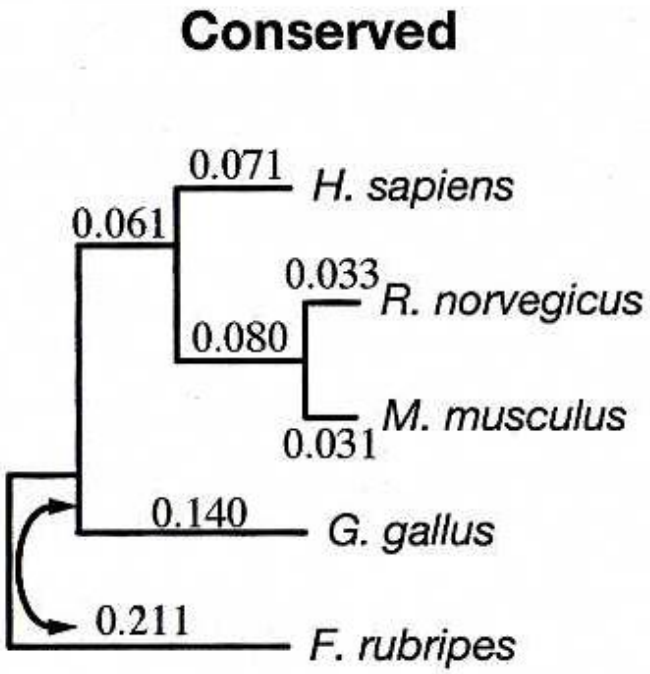
$\mu = a_{cn}$

$\nu = a_{nc}$

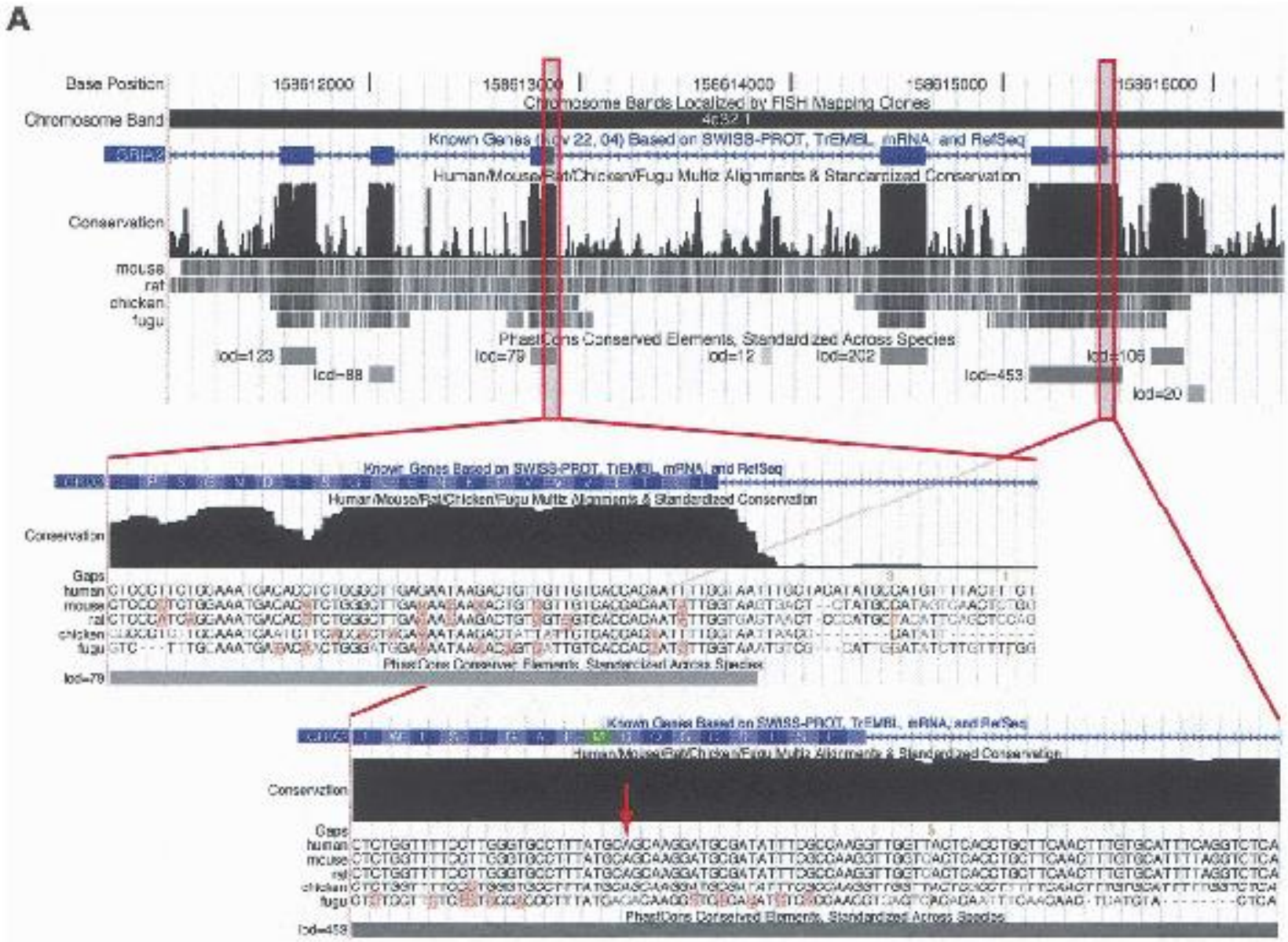
$$\mathbf{x} = \begin{matrix} \text{TCGCGACATATACGA}\ldots \\ \text{TTGGGGCATGTGGGT}\ldots \\ \text{AGCAGACGTCCGCAA}\ldots \end{matrix}$$

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

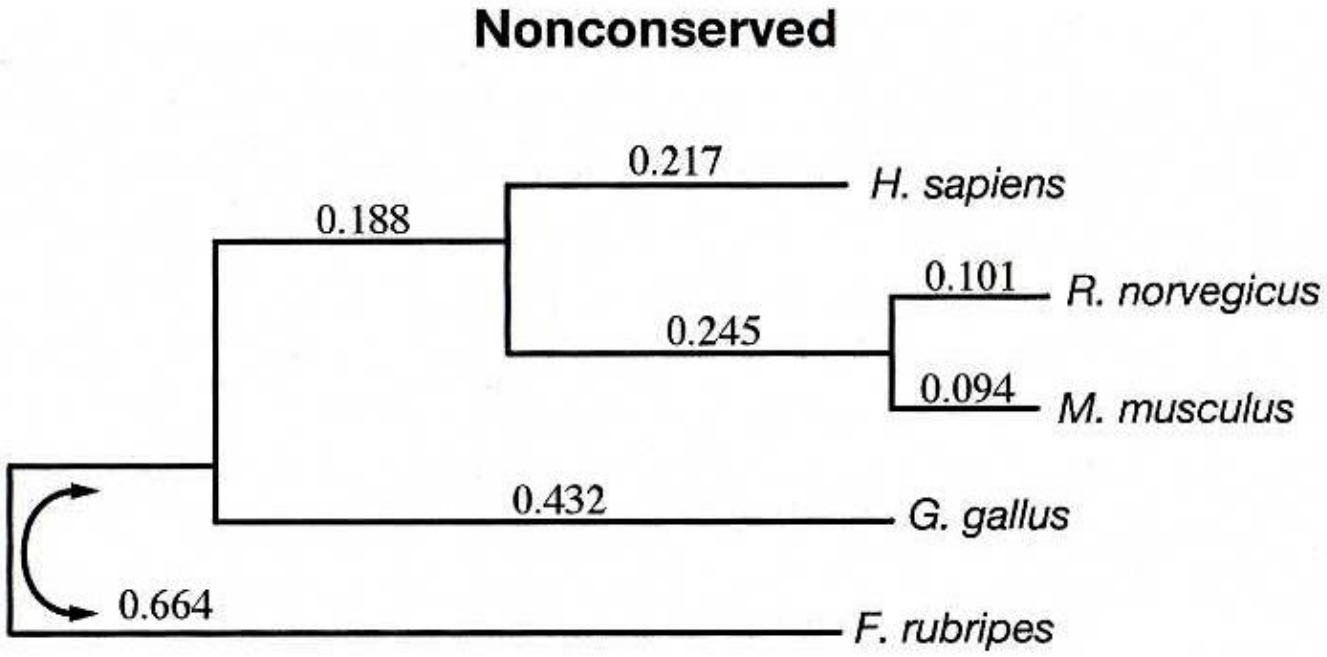

Nonconserved
0.217
H. sapiens
0.188
0.101
R. norvegicus
0.245
0.094
M. musculus
0.432
G. gallus
0.664
F. rubripes

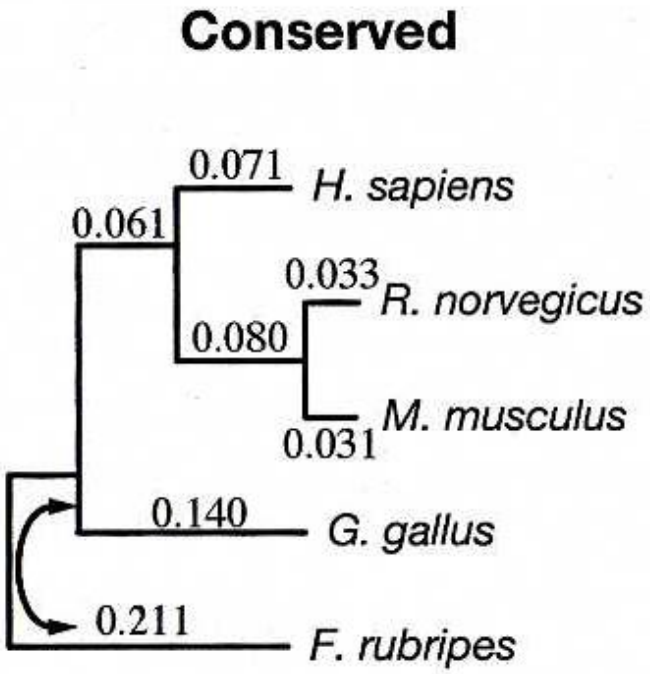

Conserved
0.071
H. sapiens
0.061
0.033
R. norvegicus
0.080
M. musculus
0.031
0.140
G. gallus
0.211
F. rubripes

- branch lengths:
  - Expected # substitutions/site over corresponding evolutionary time period
  - for neutral state, should reflect underlying mutation rate
  - for conserved state: mutation rate $\times$ scaling factor ρ
    - ρ = frac of mutations that escape purifying selection
    - $\rho \approx .33$ (for vertebrates)



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

Probability calculations on evolutionary tree (lecture 11)

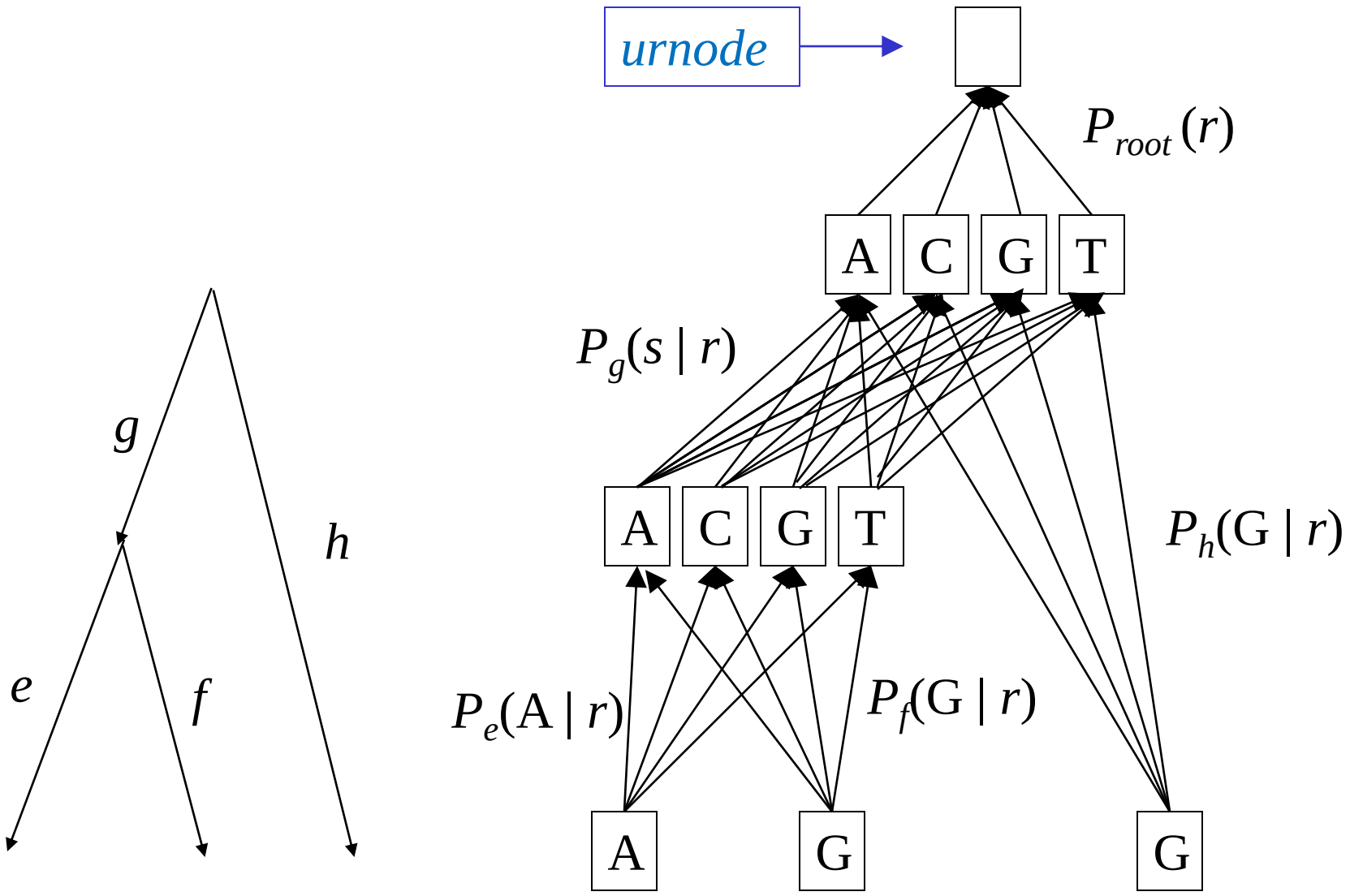
- Given:
  1. a set of observed residues at the leaves

     (a gap-free alignment column of the sequences)
  2. $\{P_e(s \mid r)\}$ and $\{P_{root}(r)\}$

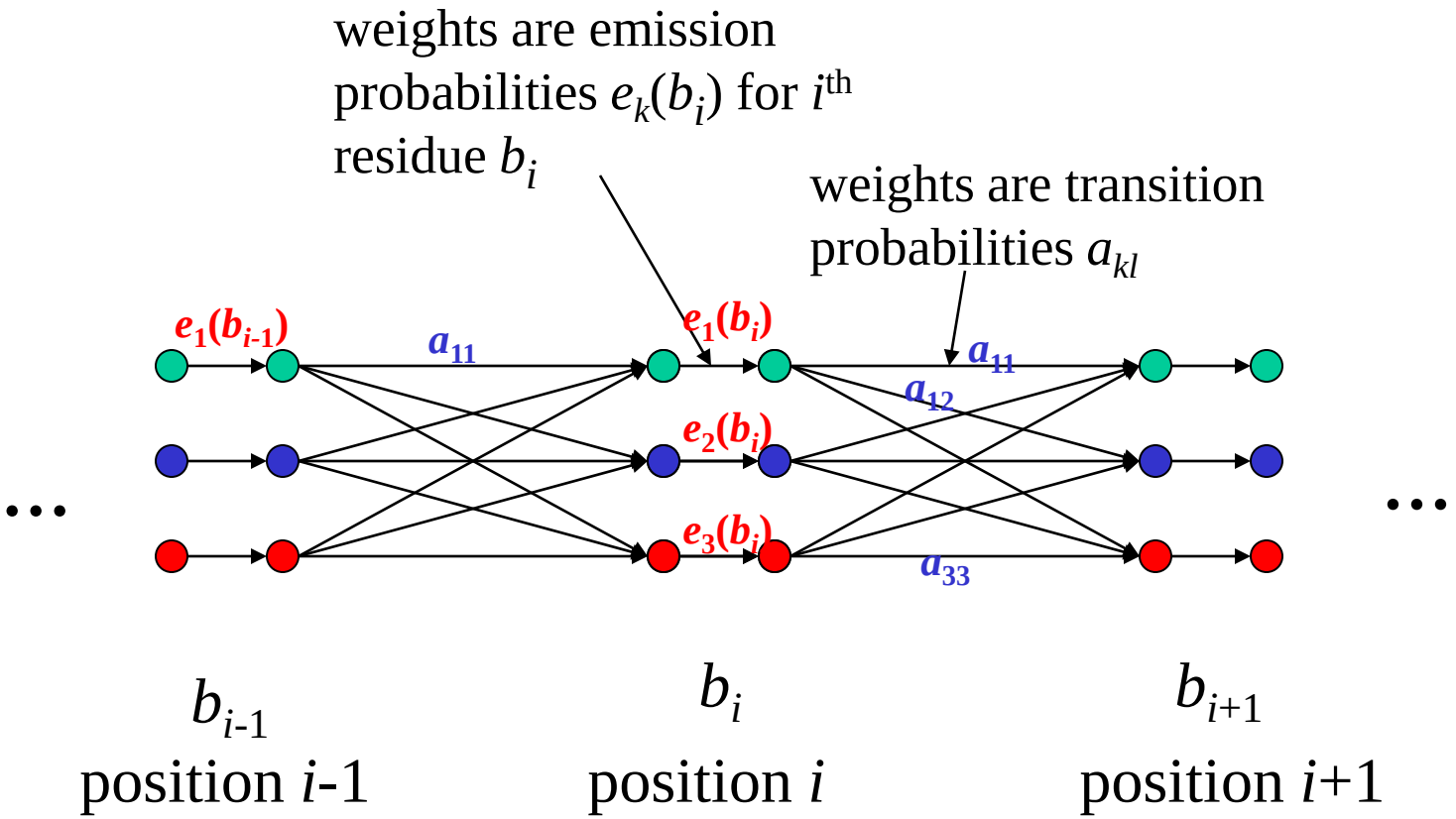
compute prob of observed residues

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

  …


urnode
$P_{root}(r)$
A C G T
$P_g(s \mid r)$
g
h
e
f
A C G T
$P_h(\mathrm{G} \mid r)$
$P_e(\mathrm{A} \mid r)$
$P_f(\mathrm{G} \mid r)$
A
G
G

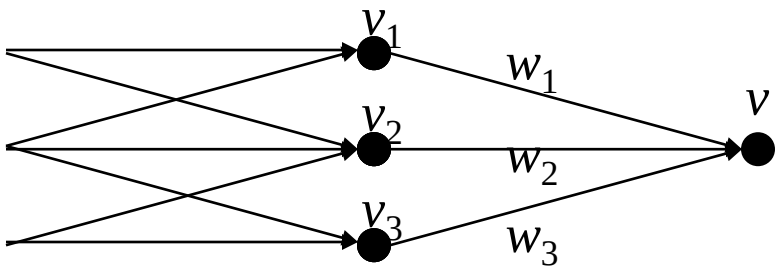
cf. WDAG for 3-state HMM length n sequence (lecture 13)



Prob calcs in HMMs (lecture 14):

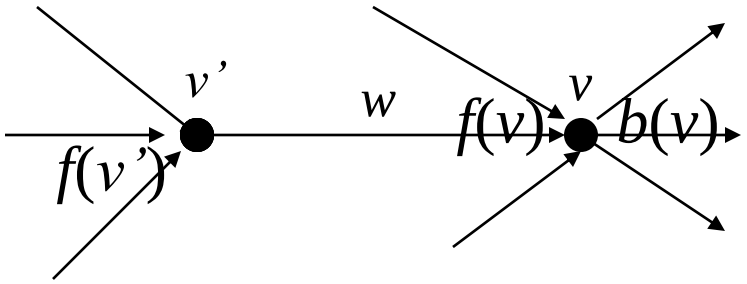
For each vertex v, let $f(v) = \sum_{\text{paths } p \text{ ending at } v} \text{weight}(p)$, where weight($p$) = *product* of edge weights in p. Only consider paths starting at 'begin' node.

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 = weight of the edge from v_i to v.



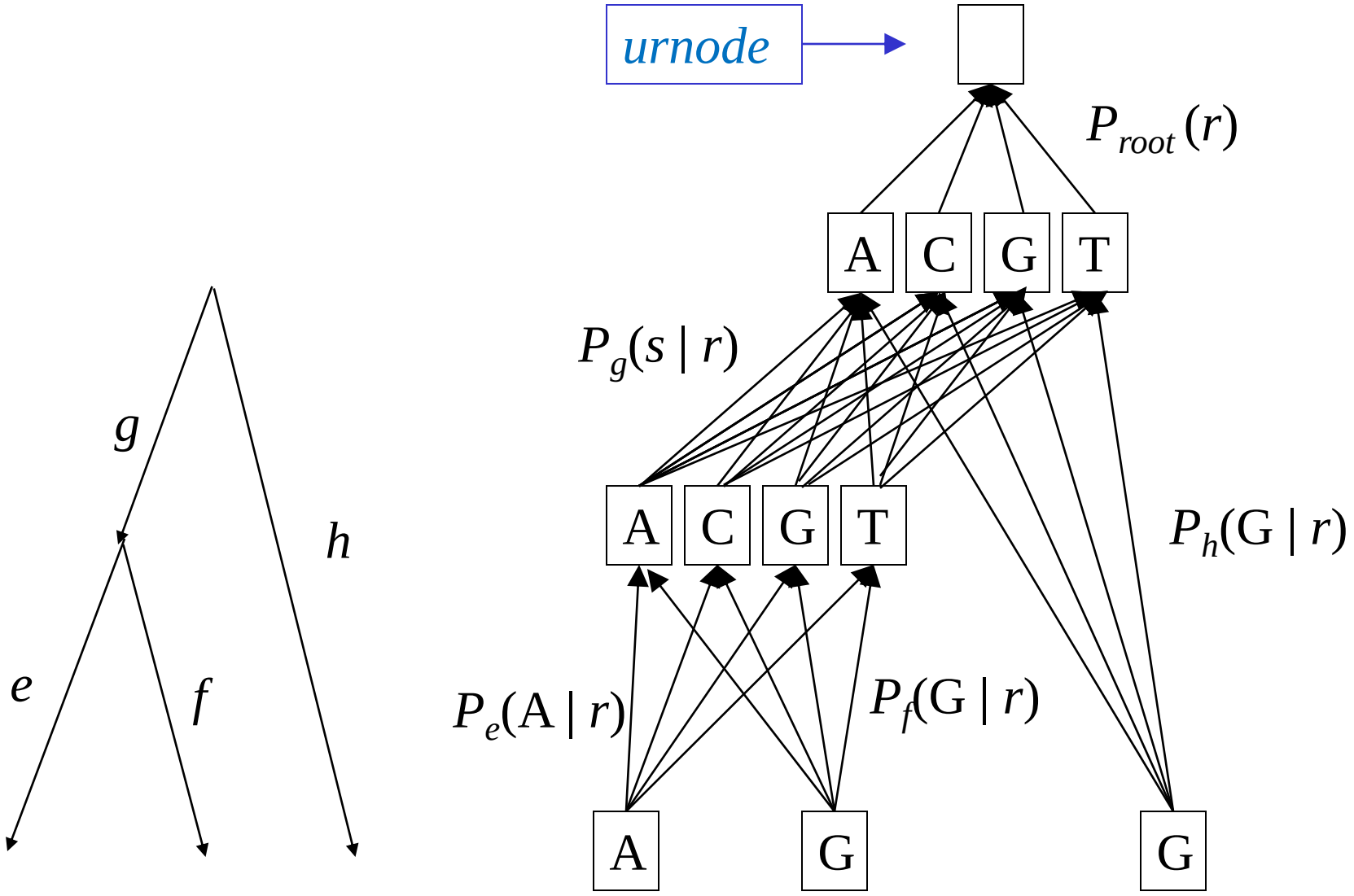
Similarly for $b(v) = \sum_{p \text{ beginning at } v} \text{weight}(p)$

The paths *beginning* at v are the ones *ending* at v in the *reverse (or inverted) graph*



$f(v)b(v)$ = sum of the path weights of all paths *through v*.

$f(v')wb(v)$ = sum of the path weights of all paths *through the edge (v',v)*


urnode
$P_{root}(r)$
A C G T
$P_g(s \mid r)$
g
h
e
f
A C G T
$P_h(G \mid r)$
$P_e(A \mid r)$
$P_f(G \mid r)$
A
G
G

- Compute overall *probability* of leaf residues (nucleotides) by *dynamic programming* on WDAG:

- Let, for each node v, $f(v)$ = prob of leaf nucs *below* v (i.e tree-descendants, or WDAG-ancestors, of v), given v's nuc

  $f_{left}(v)$ = prob of leaf nucs *below* and to *left*

  $f_{right}(v)$ = prob of leaf nucs *below* and to *right*

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

- Compute these values node-by-node, visiting (WDAG-)parents before children:
  - *starting* at leaf nodes (setting $f(v) = 1$), *ending* at urnode

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

  - u ranges over parent nodes to the left
  - $w(u,v)$ = weight on edge from u to v
    (= mutation prob from v to u)

Similarly for $f_{right}(v)$

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

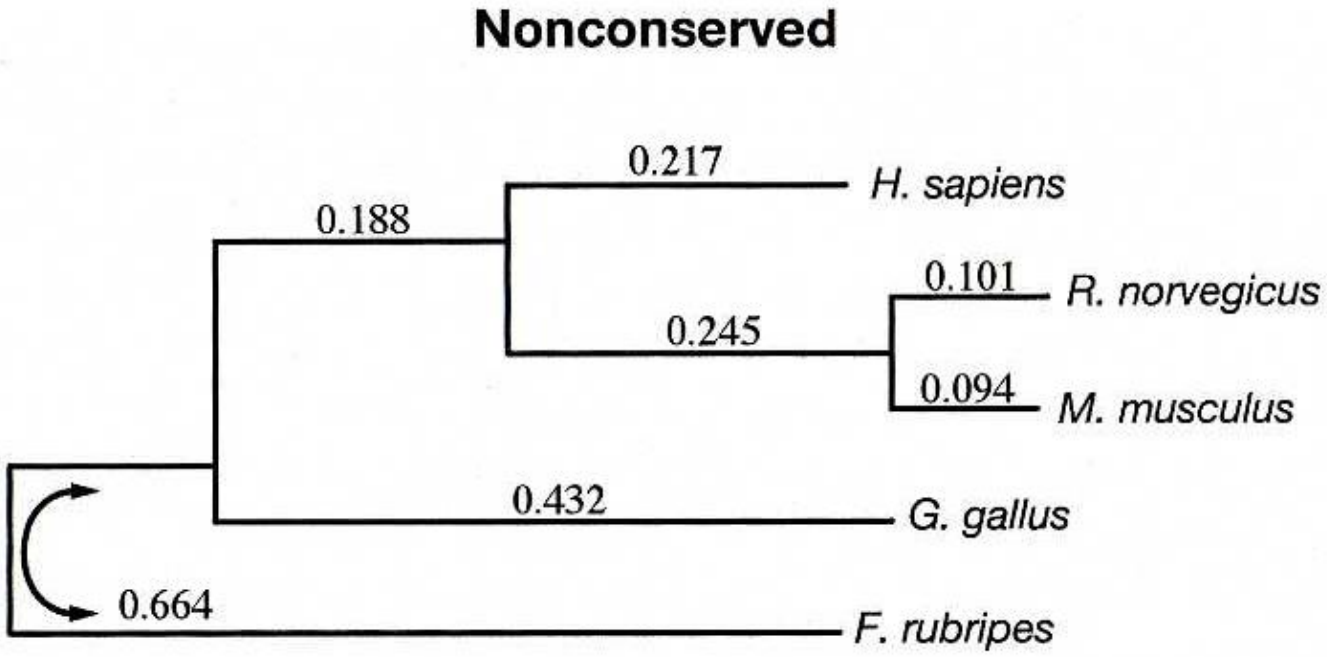
  - For v = urnode, view *all* parents as being to 'left' and $f(v) = f_{left}(v)$

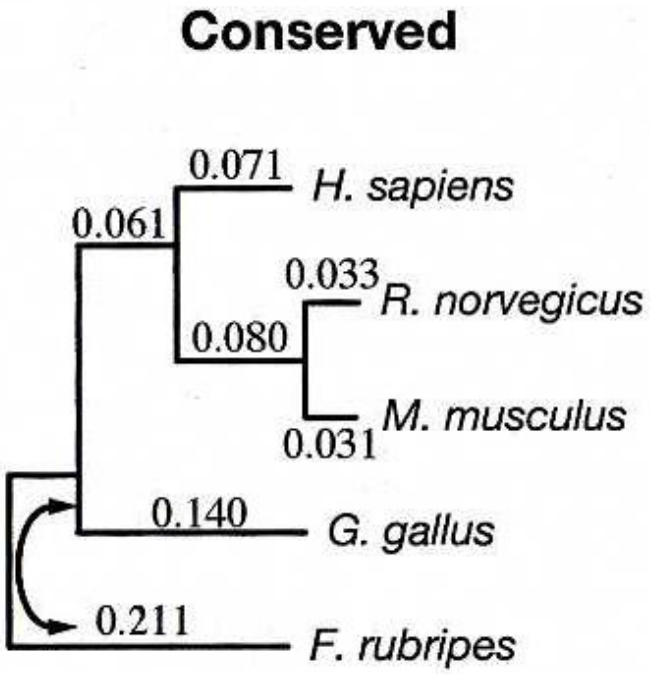
- f(urnode) = probability of the observed leaf nucs

- 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


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