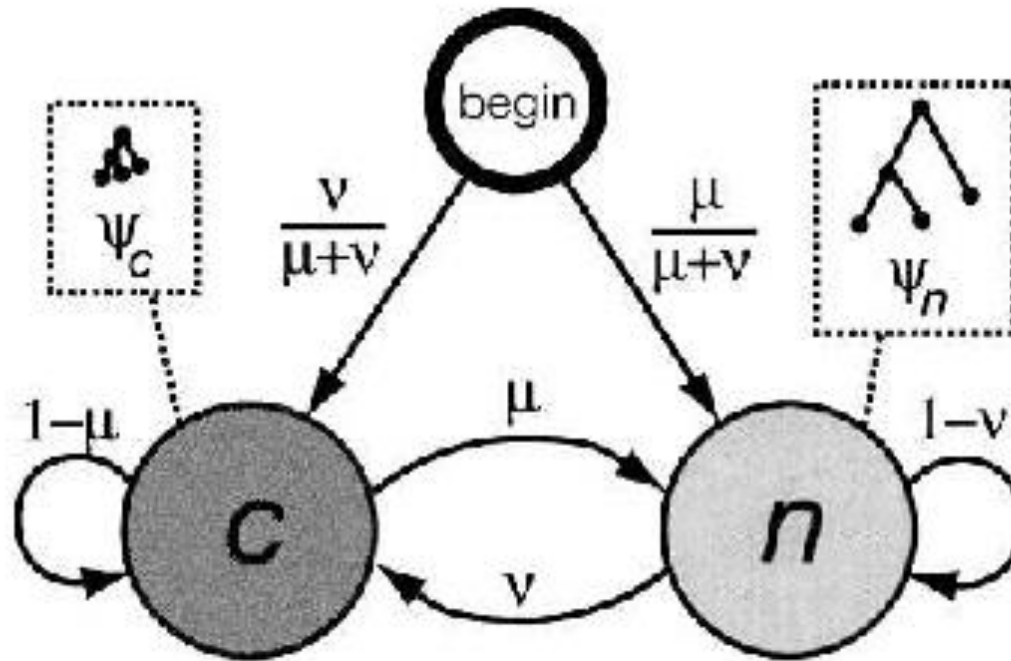


Today's Lecture

- PhastCons

PhastCons PhyloHMM



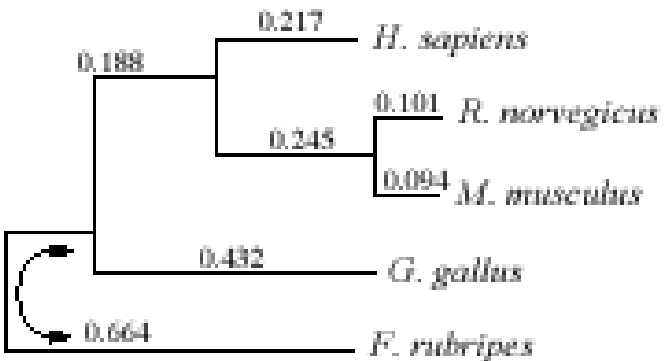
$$\mu = a_{cn}$$

$$v = a_{nc}$$

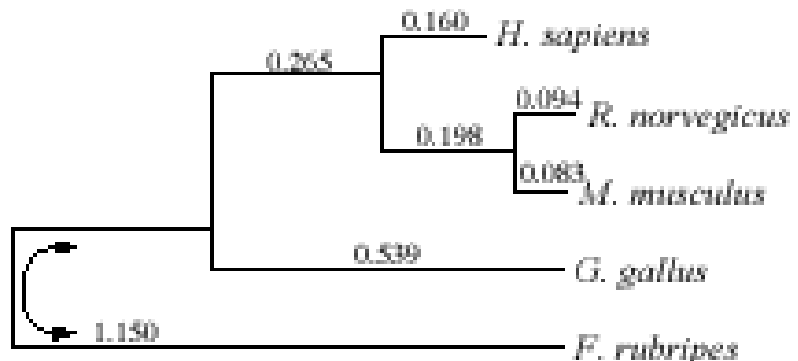
x = TCGCGACATATACGA . . .
TTGGGGGCATGTGGGT . . .
AGCAGACGTCCGCAA . . .

- for distantly related species, neutrally evolving regions no longer alignable
 - analyze 4D sites in coding sequences to estimate neutral rates
 - CDS alignments much more reliable, but
 - synonymous sites somewhat atypical (some selection; composition & mutation patterns)

PhastCons Nonconserved



Fourfold Degenerate



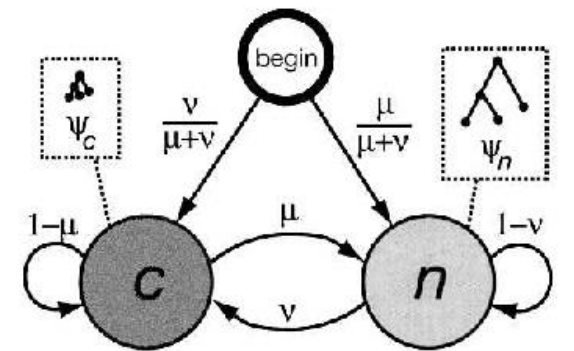
Notation

- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected 'coverage' γ (frac of genome that is conserved):

$$= \text{Elen}(\text{cons seg}) / (\text{Elen}(\text{cons seg}) + (\text{Elen}(\text{neut seg})))$$

$$= (1/\mu) / (1/\mu + 1/\nu)$$

$$= \nu / (\mu + \nu)$$



$\mathbf{x} =$
TCGCGACATATACGA...
TTGGGGCATGTGGGT...
AGCAGACGTCCGCAA... \gg

- transition probs imply *a priori* length dist'ns for conserved & non-conserved segments

- prob(cons seg has length n) is

$$(a_{cc})^{n-1}a_{cn} = (a_{cc})^{n-1}(1 - a_{cc})$$

- geometric distribution

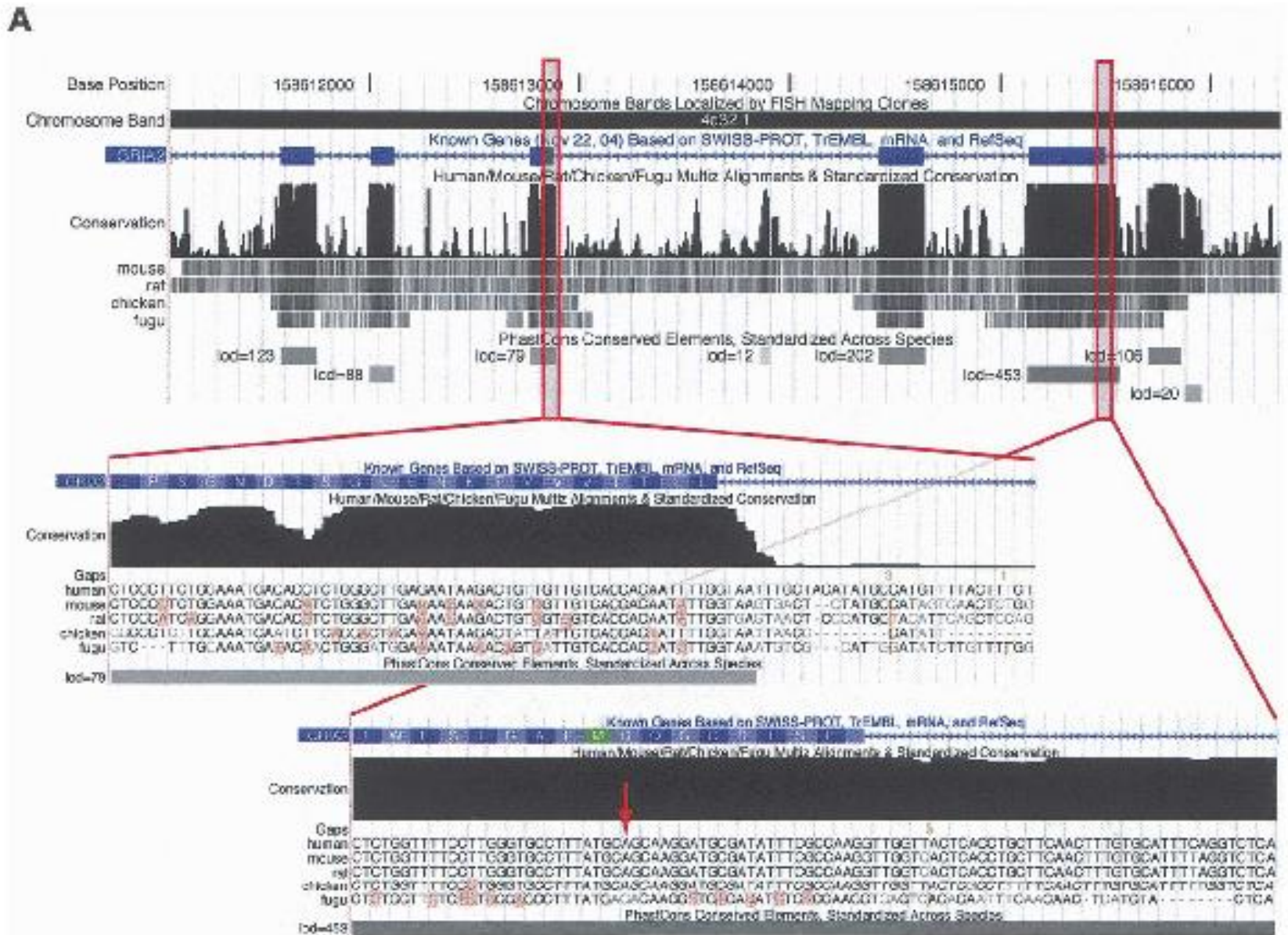
- expected length (Elen) ω of conserved segment is

$$1.0 / (1 - a_{cc}) = 1.0 / a_{cn}$$

special case: $a_{cc} = .5 = a_{nn} \Rightarrow$ positions are independent

PhastCons Parameter Estimation

- parameters estimated separately in 1 Mb windows using EM algorithm
 - full maximum likelihood analysis, or
 - constraining some parameters& averaged over genome
- full MLE results don't match biologists' intuition -- too much 'smoothing':
 - fewer, & larger, conserved elements
 - long, apparently non-conserved regions within conserved elements
 - attributed to fact that (prior) geometric length dist'n inappropriate



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

Group	Method	Total no. ^a	Ave. len. ^b	Cov. ^c	CDS cov. ^d	μ	ν	ω	γ	L_{\min}
vert.	MLE	561,103	216.1	4.2%	68.8%	0.018	0.004	55.4	0.191	30.4
	55%	1,058,855	75.3	2.8%	56.8%	0.125	0.029	8.0	0.187	12.9
	65% ^e	1,157,180	103.5	4.2%	66.1%	0.083	0.030	12.0	0.265	16.0
	75%	1,381,978	167.5	8.1%	76.6%	0.043	0.031	23.0	0.415	22.6
Group	Method	Total no. ^a	Ave. len. ^b	Cov. ^c	CDS cov. ^d	CDS frac. ^e	$H(\psi_c \psi_n)$	L_{\min}		
vert.	65%	1,157,180	103.5	4.2%	66.1%	18.0%	0.611	16.0		
	4d	797,777	109.3	3.0%	64.2%	24.0%	0.854	11.0		

Instead: -- impose constraints

- coverage constraint:
 - 65% of coding bases covered by conserved elts
 - (target value based on earlier mouse/human analysis)
- smoothness constraint:
 - PIT (\equiv expected min. amt of phylogenetic info required to predict a conserved element)
= 9.8 bits
 - (forced to be same for all species groups)

- constraints met by ‘tuning’ γ and ω (or equivalently transit probs)
 - choose γ and ω ,
 - get ML estimates of other parameters by EM algorithm
 - see whether get desired coverage & PIT;
 - if not, adjust γ and ω & redo