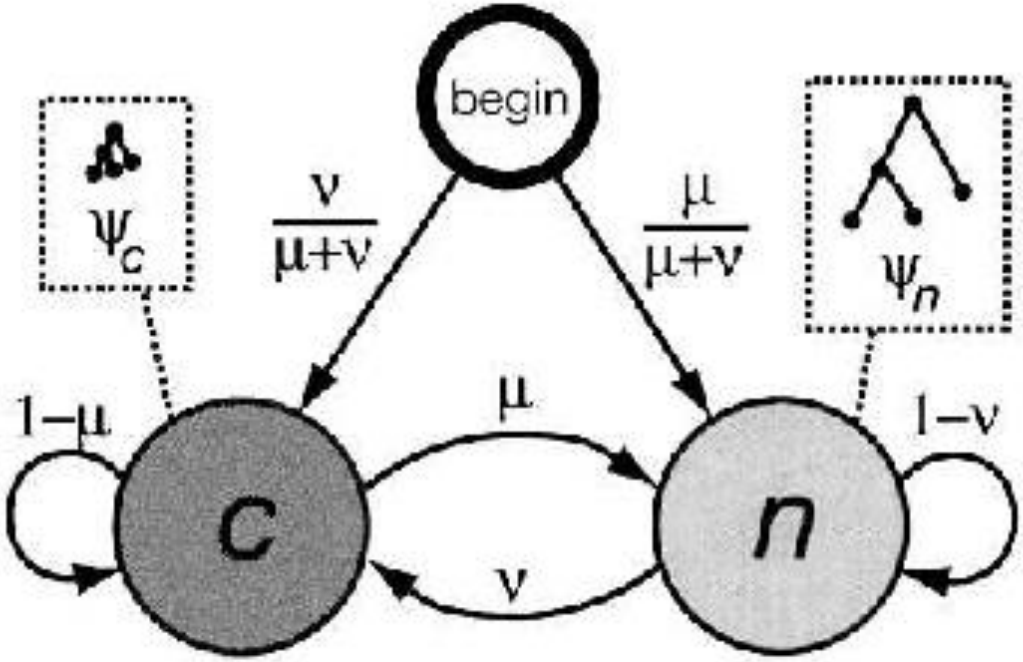
Today's Lecture

- PhastCons

PhastCons PhyloHMM

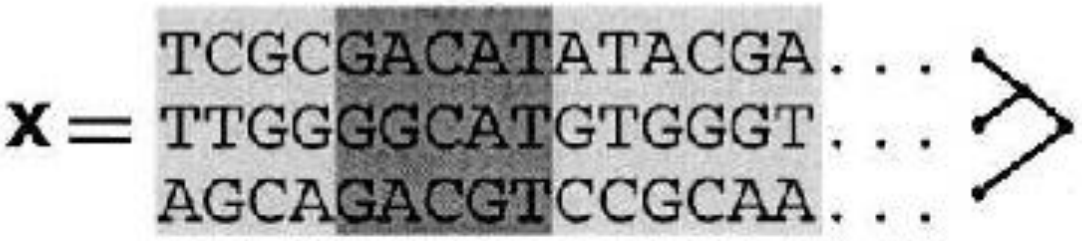
- model:
  - 2-state HMM
    - c: conserved state
    - n: neutral (or nonconserved) state
  - emitted symbols are *alignment columns*
  - emission probabilities based on *phylogenetic tree* relating sequences
    - discussed in Genome 541, or molecular phylogeny course
  - gaps in alignment treated as *missing data*

PhastCons PhyloHMM

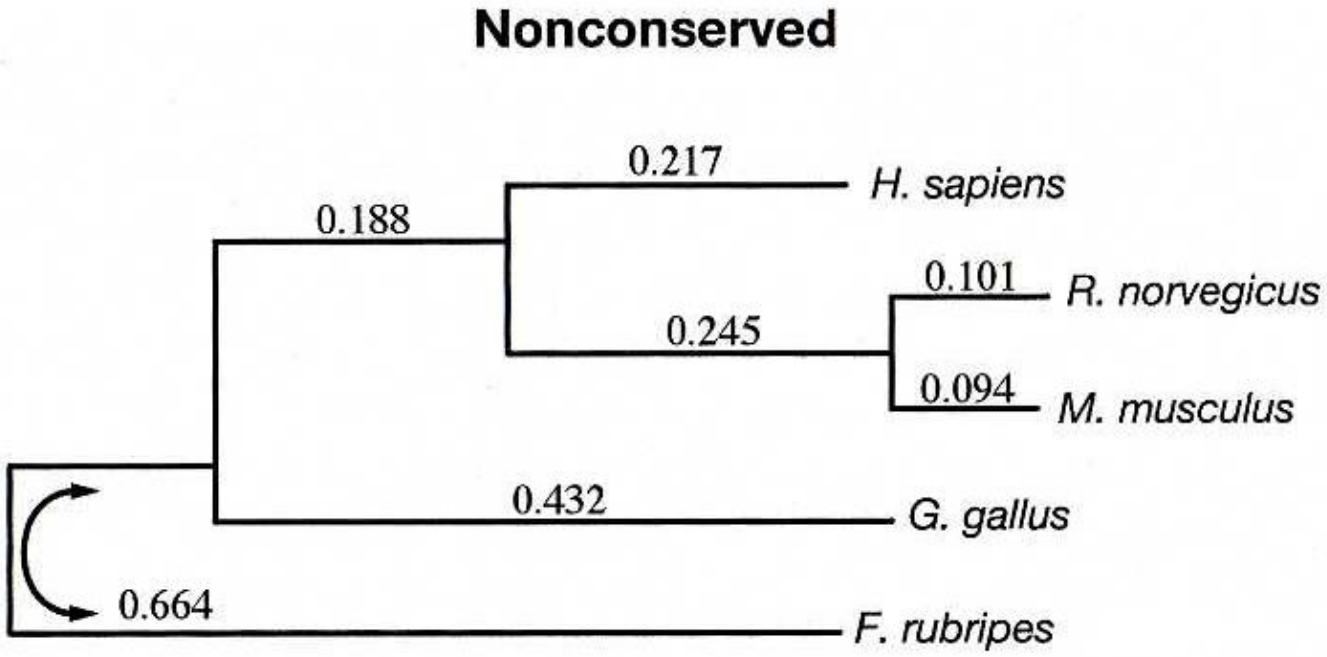


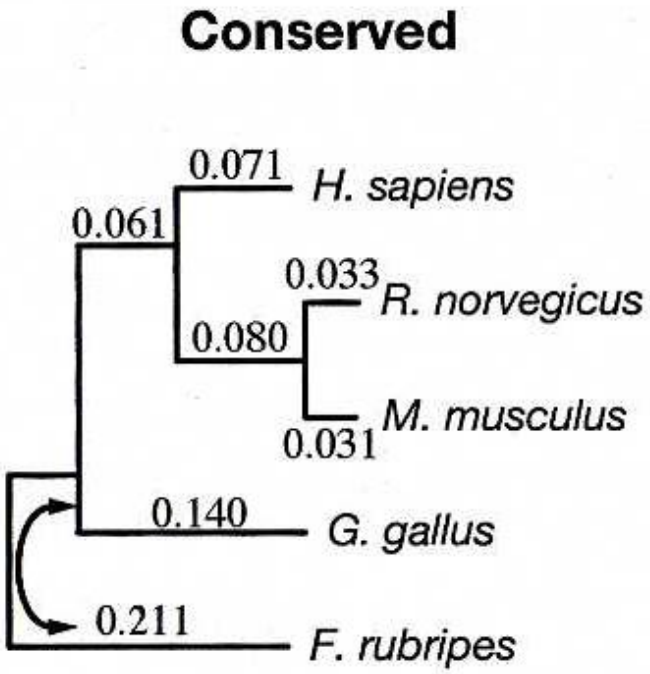
$\mu = a_{cn}$

$\nu = a_{nc}$

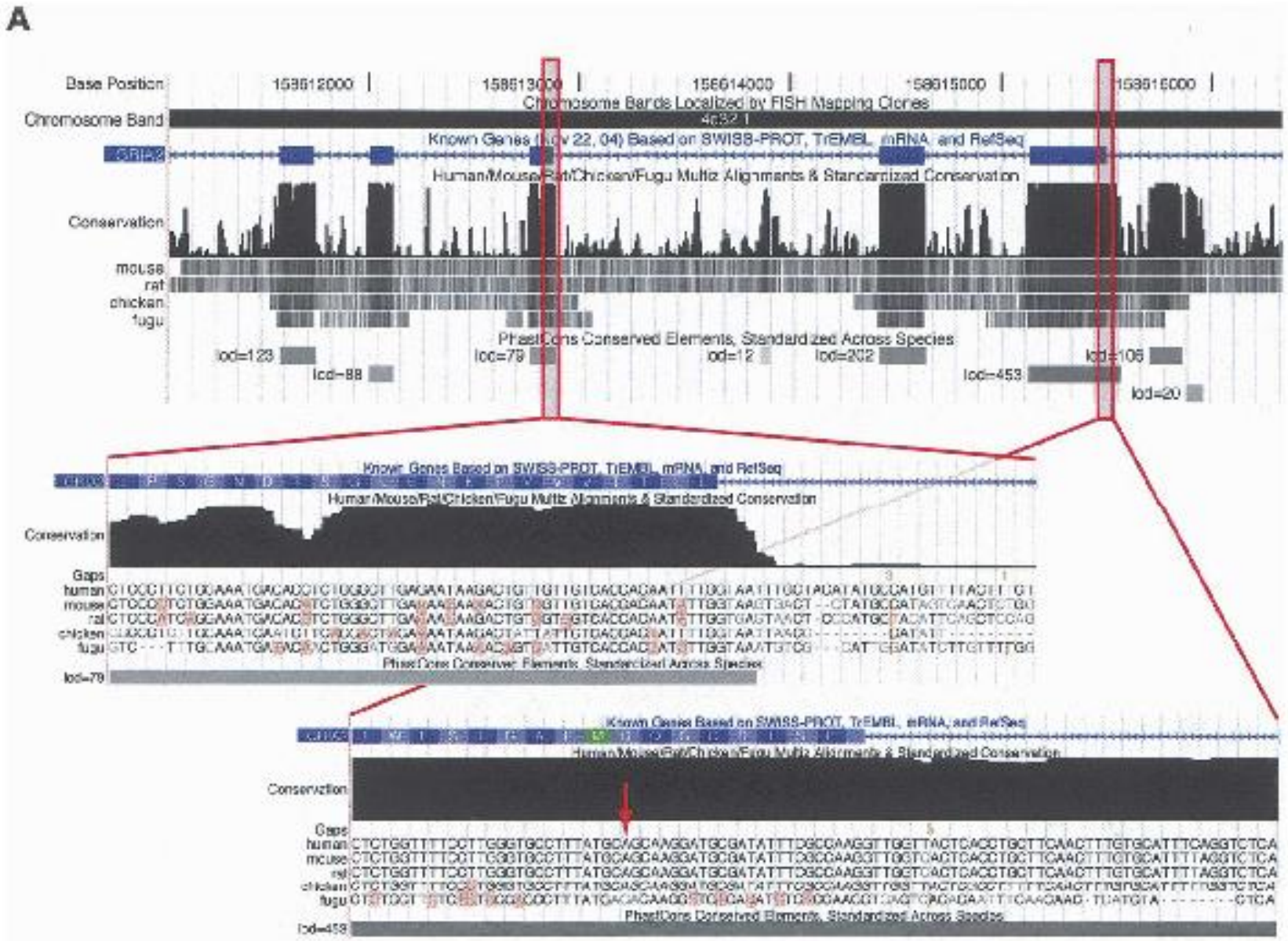


from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.


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.

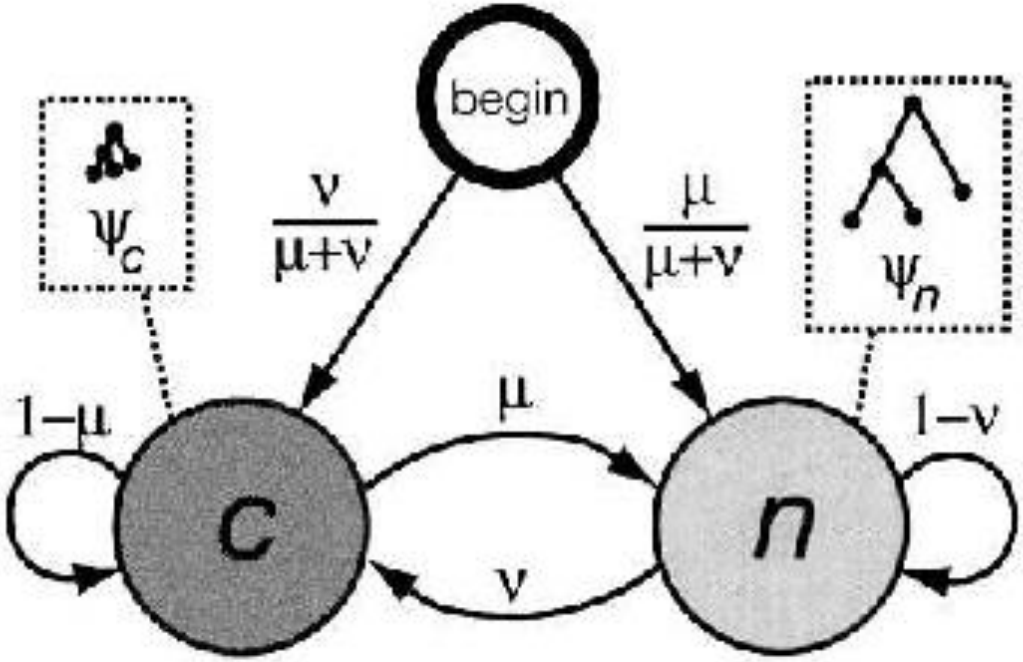
Some general issues in applying probability models, in the PhyloHMM context

- Is the model computable?
- Is the model ‘reasonable’?
  - 2 states enough?
  - Markov condition on transition probabilities
- How good is the input data?
  - Alignability of neutral sequence
  - Accuracy of genome sequence alignments
- Are results reliable?
  - No true ‘test set’ – instead, putative false positive rate, and ‘biological plausibility’ of findings

Alignment issues

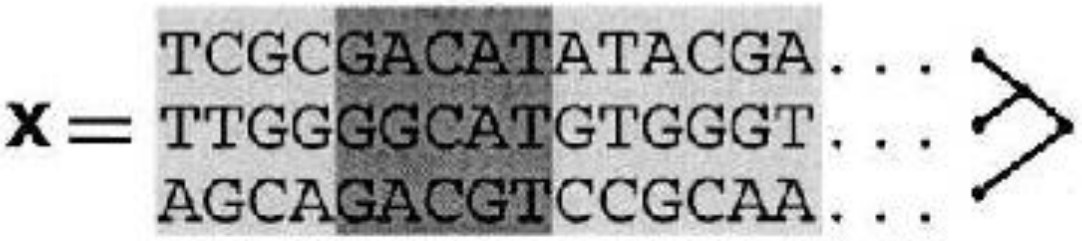
- Multiz: progressive pairwise alignments
- accurate multiple genome alignment *not* a solved problem!
  - statistical assessment: Prakash & Tompa (2005, 2007, 2009)
  - ENCODE region alignment analyses: Margulies EH *et al.* 2007
  - major issues:
    - accurate gap placement (even for close species!!)
    - discrimination among paralogous sequences (e.g. repeats, duplications)
- inaccurate alignments cause
  - neutral rate to be *overestimated*
  - conserved segments to be *overidentified*
    - because more slowly mutating (or better aligned) neutral segments may be called conserved

PhastCons PhyloHMM



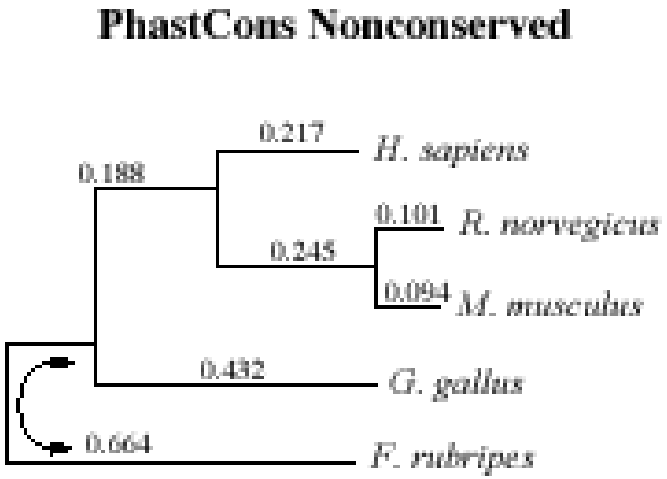
$\mu = a_{cn}$

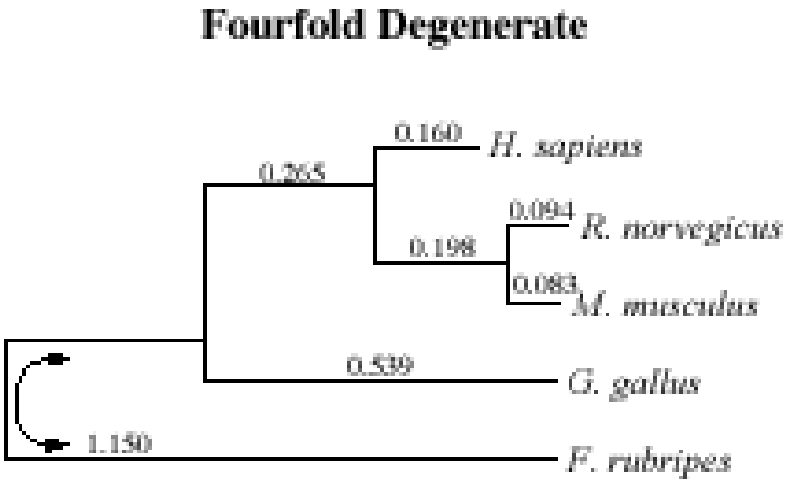
$\nu = a_{nc}$



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

- 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)





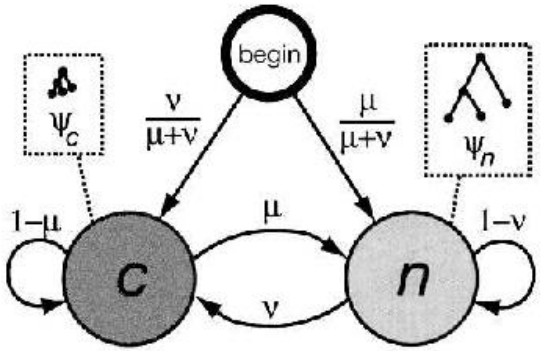
Notation

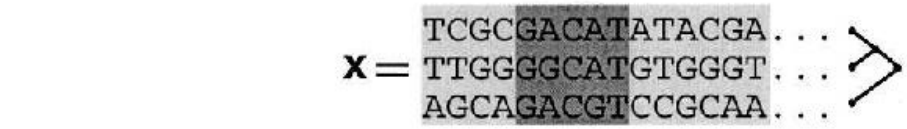
- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected 'coverage' γ (frac of genome that is conserved):

  = Elen (cons seg) / (Elen(cons seg) + (Elen(neut seg))

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

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





- 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

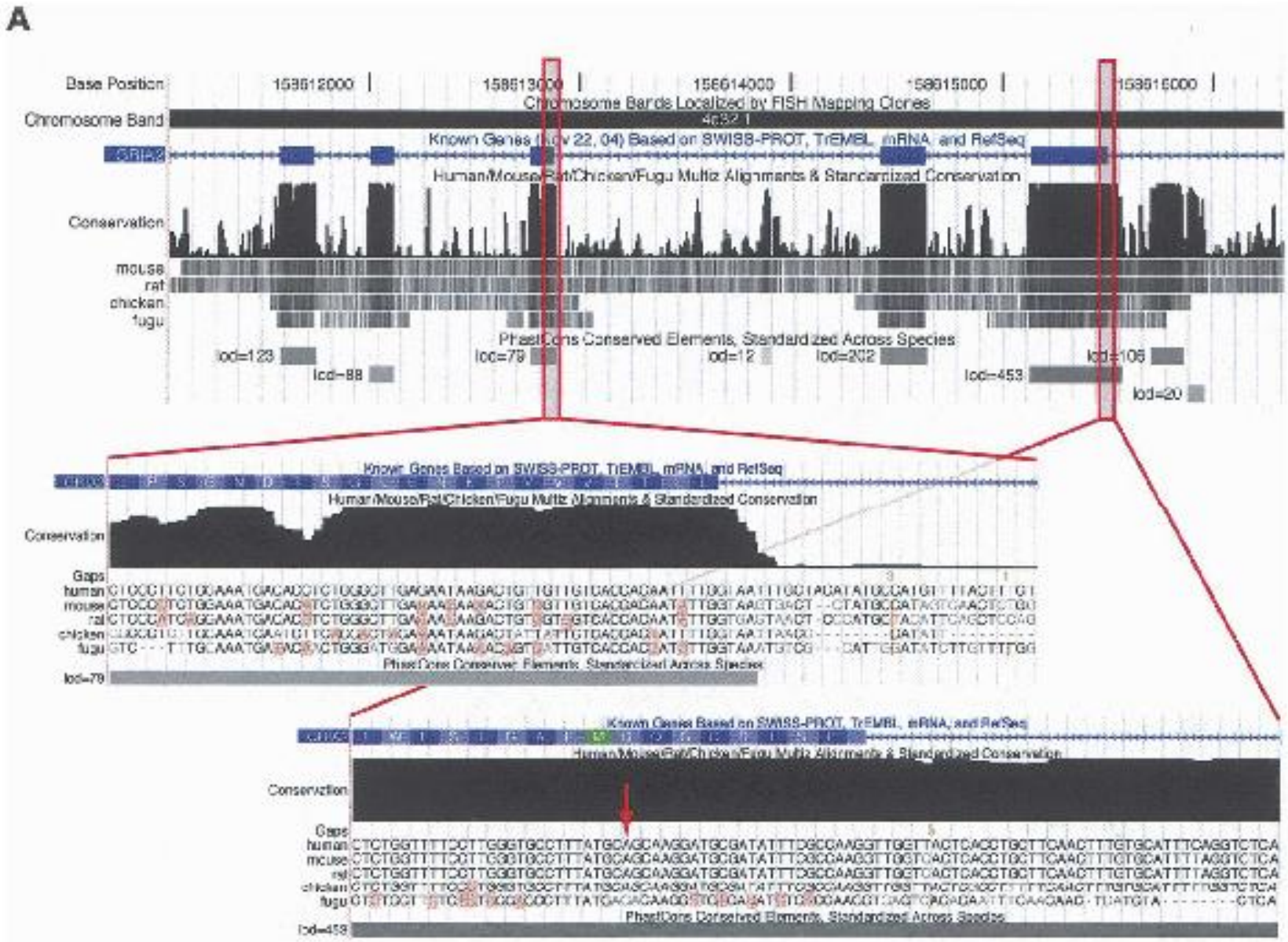
$$1.0 / (1 - a_{cc}) = 1.0 / a_{cn}$$

  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

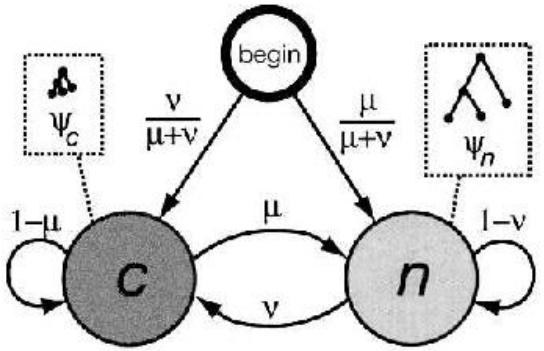
Notation

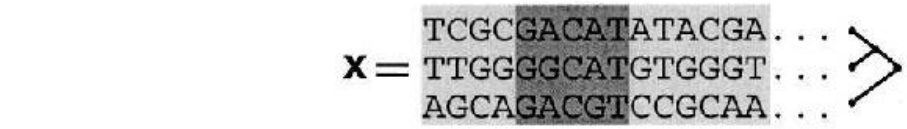
- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected 'coverage' γ (frac of genome that is conserved):

  = Elen (cons seg) / (Elen(cons seg) + (Elen(neut seg))

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

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





Instead: -- impose constraints

- coverage constraint:
  - 65% of coding bases covered by conserved elts
  - (target value based on earlier mouse/human analysis)
- smoothness constraint:
  - PIT (≡ 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

- $L_{\min}$: expected min length of a conserved segment that could appear in a Viterbi path
- at $L_{\min}$,

  expected loglike of staying in state n

  = expected loglike of switching to c & back again, so

$$(L_{\min}+1)\log(1-\nu)+L_{\min}\sum_x P(x|\psi_c)\log P(x|\psi_n)$$

$$=\log\nu+\log\mu+(L_{\min}-1)\log(1-\mu)+L_{\min}\sum_x P(x|\psi_c)\log P(x|\psi_c)$$

- $$L_{\min}=\frac{\log\nu+\log\mu-\log(1-\nu)-\log(1-\mu)}{\log(1-\nu)-\log(1-\mu)-H(\psi_c\|\psi_n)}$$

- where

$$H(\psi_c||\psi_n) = \sum_x P(x|\psi_c) \log \frac{P(x|\psi_c)}{P(x|\psi_n)}$$

= rel entropy of *c*-state emission prob dist'n w.r.t.

n-state dist'n

- PIT (phylogenetic information threshold)

= $L_{\min} H(\psi_c||\psi_n)$

= 'expected min amt of phylogenetic info required to predict conserved element'

- Final param estimates (for vertebrates):
  - $\gamma = 0.265$
  - $\omega = 12.0$ bp
  - $H(\psi_c \| \psi_n) = .608$ bits / site
  - $L_{\min} = 16.1$ bp
  - PIT = $L_{\min} H(\psi_c \| \psi_n) = 9.8$ bits

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

Estimating false positive rates

- simulate 1 Mb alignment
  - by sampling 4D sites (with replacement) from aligned CDSs
  - caveat: these not typical of all neutral sites!
- predict cons elts (using prev param estimates)
- frac of bases in cons elts:

Group	65%	75%	MLE
vertebrate	0.00279[a]	0.00362	0.00005
insect	0.00286	0.01026	0.00152
worm	0.00000	0.00000	0.00000
yeast	0.00006	0.00042	0.00023

- does not address (important) issue of rate of false positive bases within, or flanking, true conserved elements
- also: genes more G+C rich than genome average, & have somewhat higher mutation rate (due in part to more frequent CpGs)

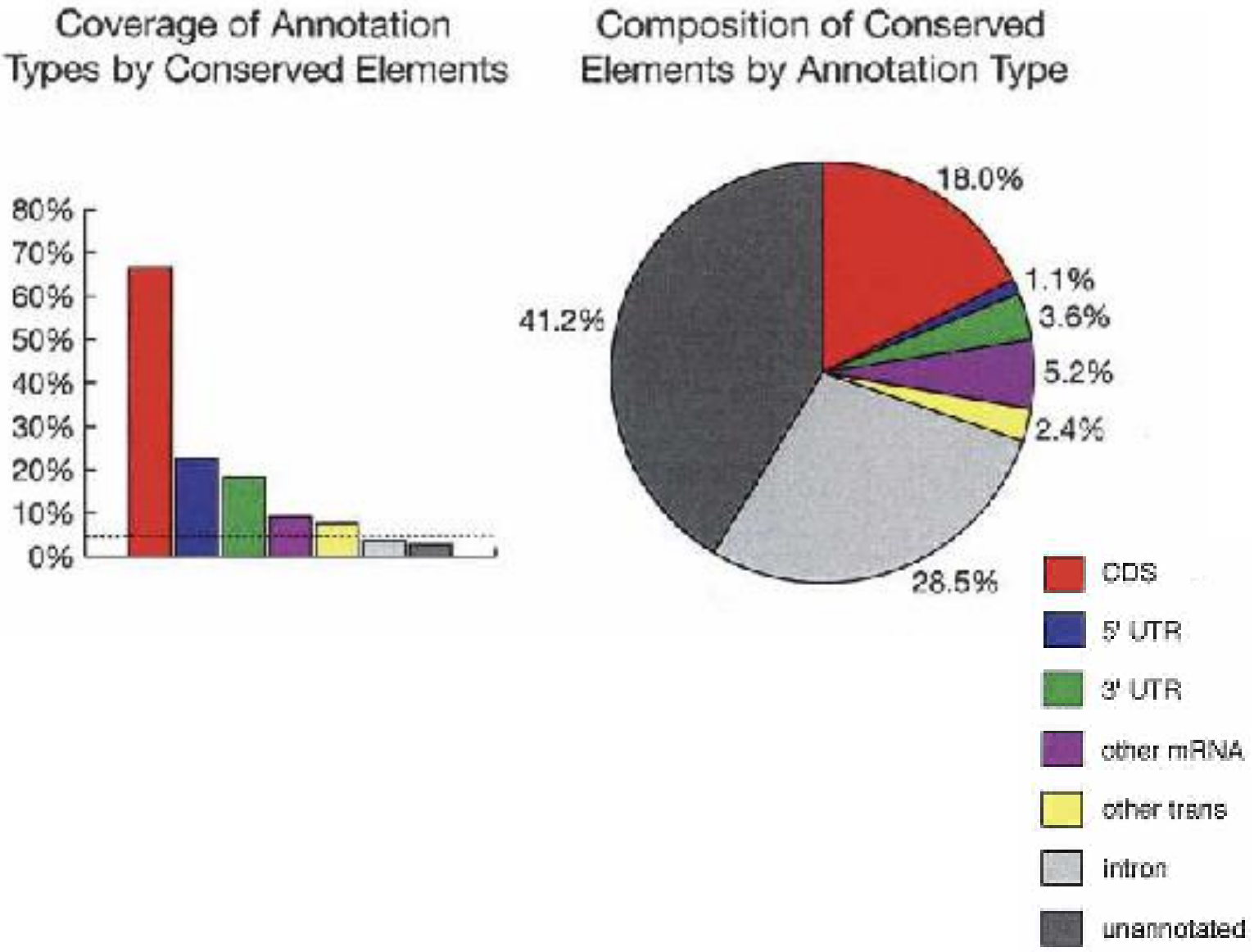
  $\Rightarrow$ *underestimating* false pos rate
- also: randomization procedure destroys underlying mutation rate variation

  $\Rightarrow$ *underestimating* false pos rate

Characteristics of phastCons predicted conserved elements

- 1.18 million elements
- constitute 4.3% of human sequence
  - 66% of coding bases
    - 88% of coding exons overlap predicted elt
  - 23% of 5'UTR bases
    - 63% of exons
  - 18% of 3'UTR bases
    - 64% of exons
  - 42% of RNA gene bases
    - 56% of genes
  - 3.6% of intronic bases
  - 2.7% of intergenic bases
  - < 1% of mammalian 'ancestral repeats' (ARs)

Vertebrate
(H. sapiens)



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