Molecular Evolution

Gina Cannarozzi

Outline

- Goal: measure amount of change since 2 sequences diverged (distance)
- Know: observed number of mutations
- Want: real number of mutations (distance)
- Solution: use models to correct for the difference in observed and expected differences
    - parametric or nonparametric,
    - amino acid, codon (synonymous or nonsynonymous) or single nucleotide
- Uses: phylogeny, selection

Why we have to correct

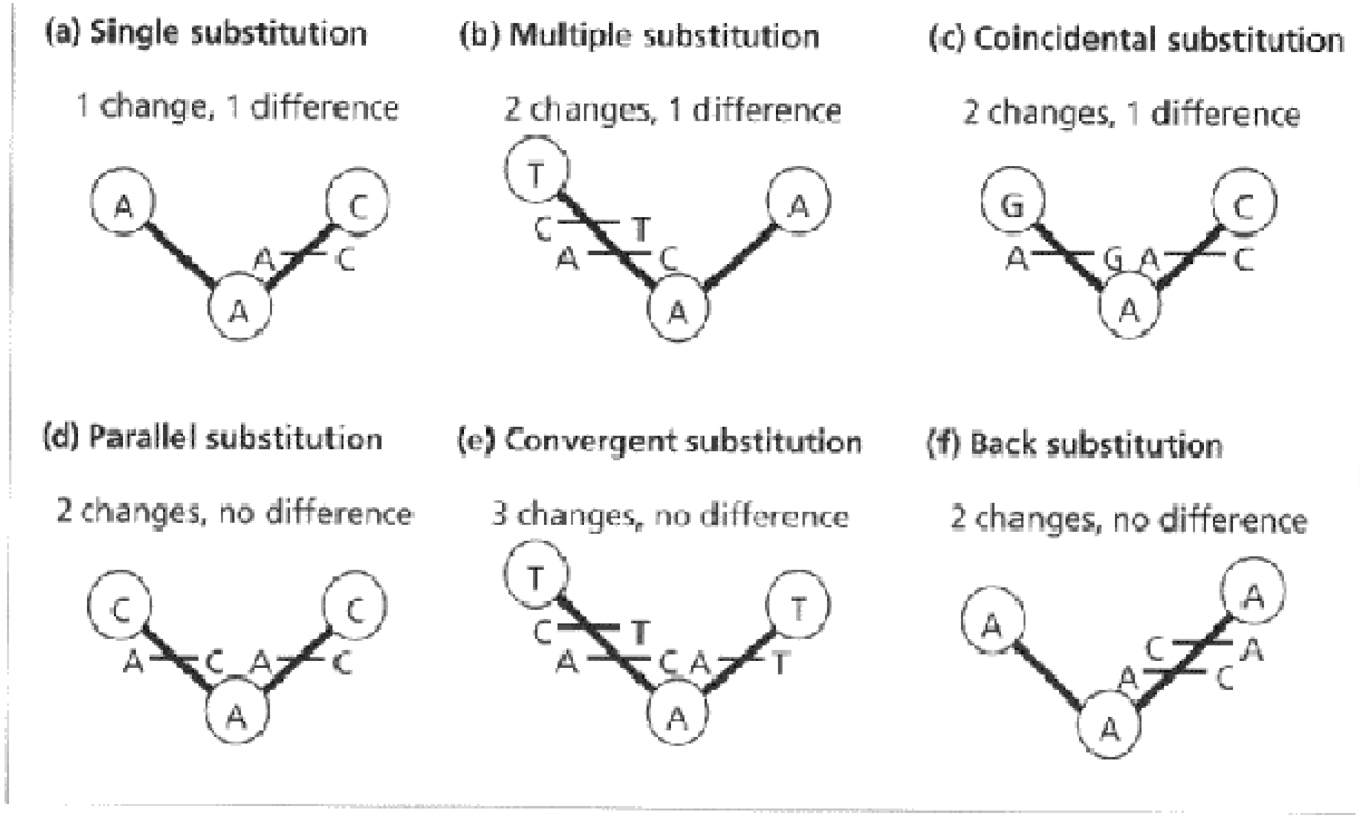
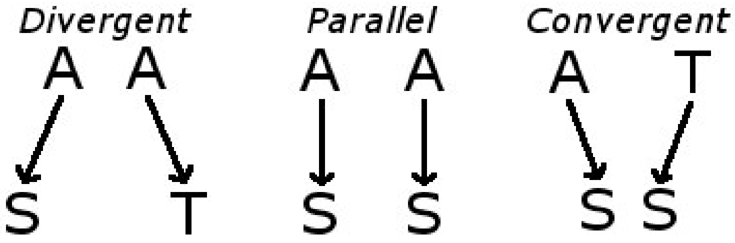


Fig. 5.9 Six kinds of nucleotide substitution. In each case the ancestral nucleotide was A. In all except the case of a single substitution, the number of substitutions that actually occurred is greater than would be counted if we just compared the two descendant sequences. In the lower three cases the nucleotides are identical in both descendant sequences, but this similarity has not been directly inherited from the ancestral sequence. Such similarity is termed 'homoplasious'.



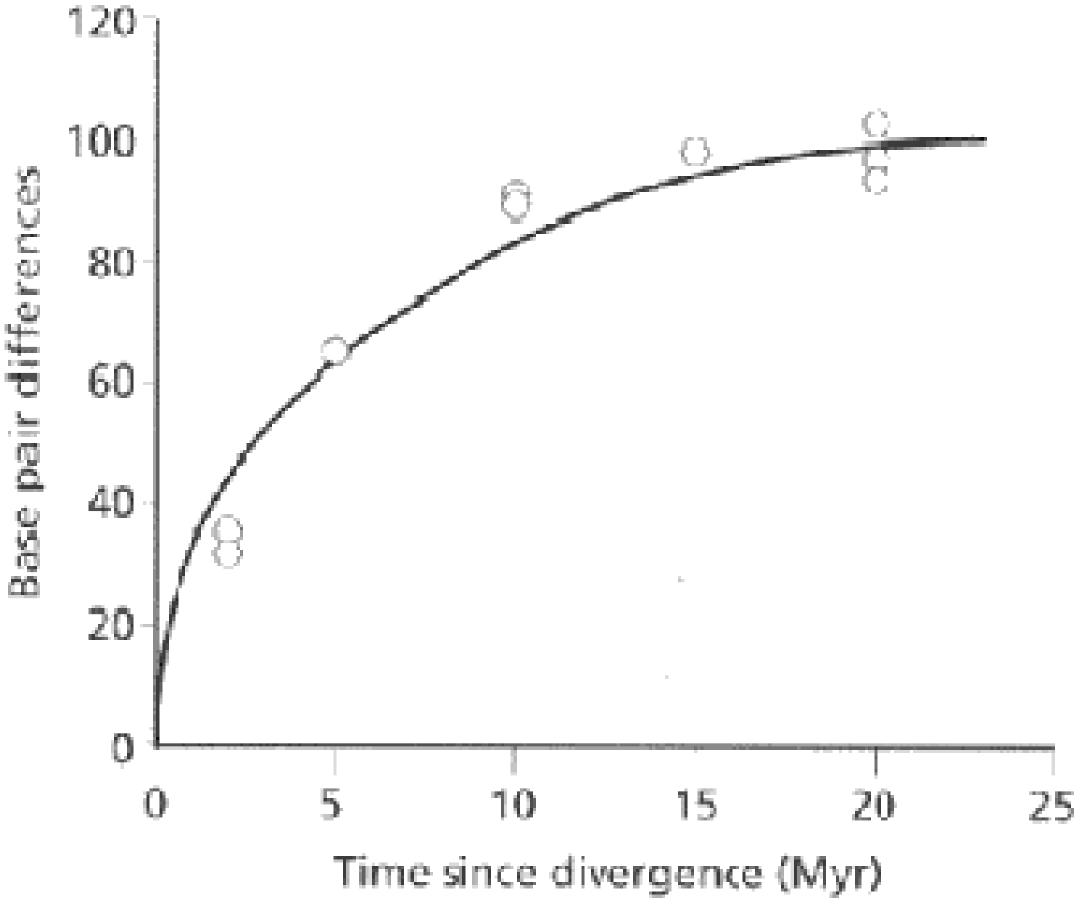


Fig. 5.11 Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al.* (1996).

p distance is the proportion of different sites ie 10 differences in two sequences of length 100 = 10% = .1

p distance is not an additive distance (on for which Dab + Dbc = Dac in expected value) so it is not a good distance measure and only usable for p < .05

All Models

- Both parametric (mechanistic) and nonparametric (empirical) models usually have two components 1) a tree and 2) a description of how sequences evolve by amino acid, nucleotide or codon replacement along the branches of the tree
- Markov model - independent sites, no memory, substitutions described by one rate matrix
- nonparametric (empirical) models - use properties calculated through comparisons of large numbers of observations
- parametric models - built on the basis of chemical or biological properties

Protein evolution

- usually empirical (nonparametric) such as our Dayhoff model
- substitution probabilities described by a 20 x 20 mutation matrix
- parameter values (substitution rates) are estimated once from empirical data and then fixed
- substitution rates are not influenced by the particular data set under consideration
- variants exist in which the frequency parameters of the data set under consideration are used in conjunction with the exchangeability parameters from a standard model usually denoted by +F suffix
- rate heterogeneity across sites is usually described by a gamma distribution and is denoted by +Γ

PAM distance

- Evolutionary distance (not time)
- definition: a 1 PAM transformation is an evolutionary step where 1% of the amino acids are expected to mutate
- M is a mutation matrix for which each element describes a probability of a mutation

$$M_{ij} = \Pr x_j \to x_i \ .$$

$$M = \begin{matrix} 0.98 & 0.01 & \dots & 0.01 \\ 0 & 0.99 & \dots & 0.002 \\ \vdots & \vdots & \ddots & \vdots \\ 0.001 & 0 & \dots & 0.97 \end{matrix}$$

$$\sum_{i=1}^{20} f_i(1 - M_{ii}) = 0.01$$

Similarity score

Our score compares two events- the probability of alignment by reasons of common ancestry divided by the probability of alignment by random chance

- -A- -

- -S- -

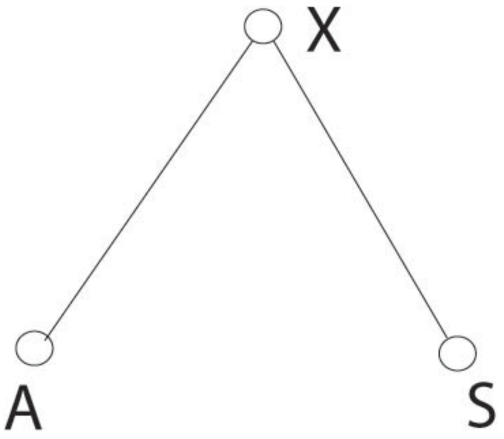
Match by Chance
$Pr\{A\}Pr\{S\}$
$= f_A f_S$

- -A- - sequence 1
- -X- - ancestor X.
- -S- - sequence 2

Pr{A and S from Ancestor X}
$\sum_X f_X Pr\{X \rightarrow A\} Pr\{X \rightarrow S\}$
$= \sum_X f_X M_{AX} M_{SX}$
$= \sum_X f_S M_{AX} M_{XS}$
$= f_S M^2_{AS}$
$= f_A M^2_{SA}$

where f_A is the frequency of A in nature
Compare Two Events

$$\frac{CommonAncestry}{Chance} = 10log_{10}\frac{f_A M^2_{AS}}{f_A f_S} = D_{AS}$$



Estimating distance via likelihood

```
ACRTES
AWKSDT
```

For this alignment, the score is $D^d_{AA} + D^d_{CW} + D^d_{RK}$ ...

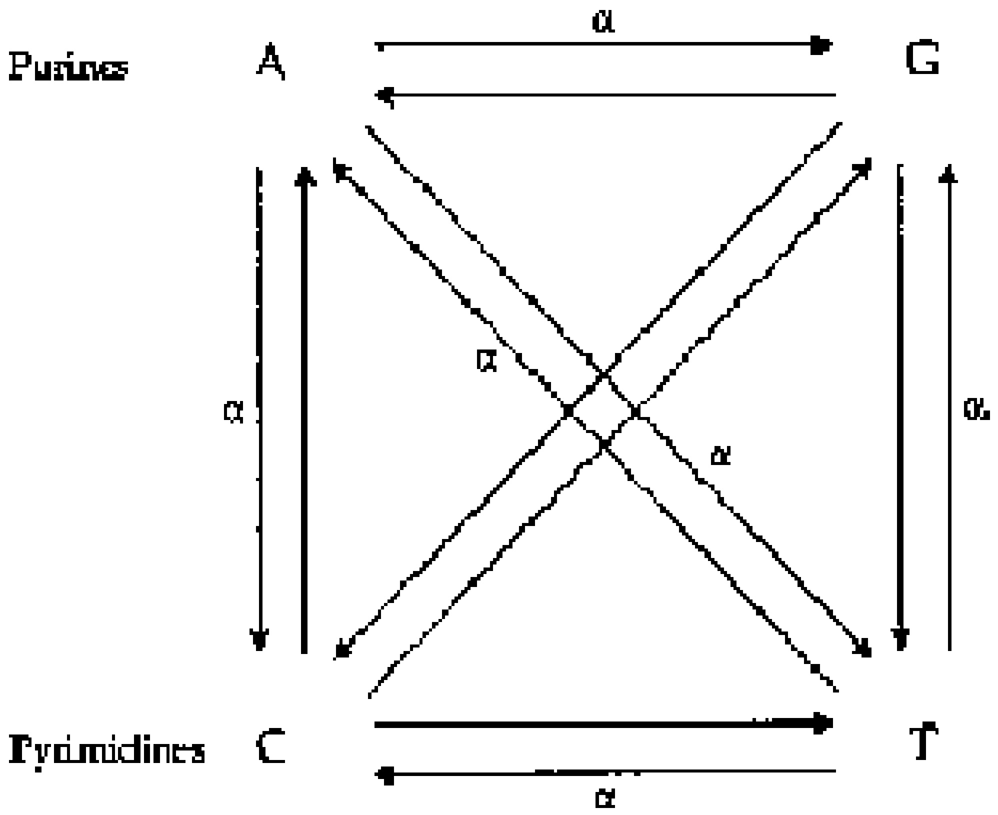
We score the alignment with a range of Dayhoff matrices and search the Dayhoff matrix, D (characterized by a distance, d) which maximizes the score. The distance of this matrix is taken as the distance of the alignment because it maximizes the likelihood of the alignment coming by reasons of ancestry as opposed to random chance.

This was an uebung exercise.

Single nucleotide evolution

- 4 state Markov model is used to estimate the real number of substitutions from the observed number
- distance is expressed as the expected number of substitutions per site
- nucleotide models are usually parametric models
- allow for parameters to be estimated for each data set under consideration
- parameters
  - base exchangeability parameters (up to 6)
  - base frequency parameters (GC content)

Jukes-Cantor Model



- one rate parameter alpha -rate of change between any two bases

Instantaneous rate matrix Q

$$Q = \{q_{ij}\} = \begin{array}{|c|c|c|c|}
\hline
T & C & A & G \\
\hline
-3\alpha & \alpha & \alpha & \alpha \\
\hline
\alpha & -3\alpha & \alpha & \alpha \\
\hline
\alpha & \alpha & -3\alpha & \alpha \\
\hline
\alpha & \alpha & \alpha & -3\alpha \\
\hline
\end{array}$$

$q_{ij}\Delta t$ = prob that i changes to j in an infinitely small time Δt

for $t > 0$

$P(t) = e^{Qt}$ $\{p_{ij}(t)\}$ prob that i will become j in time t

Transition probability matrix

$$P(t) = e^{Qt} = \begin{array}{c|cccc} & T & C & A & G \\ \hline T & p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ C & p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ A & p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ G & p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{array}$$

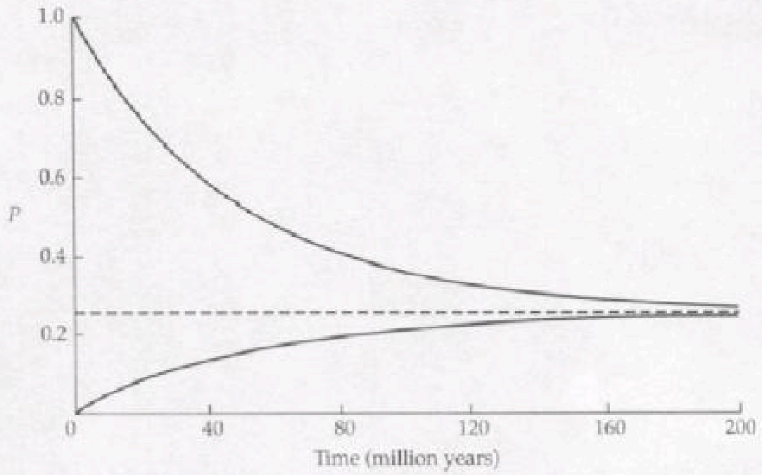


FIGURE 3.3 Temporal changes in the probability, P, of having a certain nucleotide at a position starting with either the same nucleotide (upper line) or with a different nucleotide (lower line). The dashed line denotes the equilibrium frequency ($P = 0.25$). $\alpha = 5 \times 10^{-9}$ substitutions per site per year.

$$p_0(t) = 1/4 + 3/4\, e^{-4\alpha t}$$

$$p_1(t) = 1/4 - 1/4\, e^{-4\alpha t}$$

- each column sums to 1 because the chain has to be in one of the 4 states
- $P(0) = \mathrm{I}$, the identity matrix, reflecting the case of no evolution
- rate, α, and time t occur only as a product αt. With no external information about rate or time, we can estimate the distance but not rate or time.
- at $t = \infty$, $p_{ij}(t) = 1/4$ for all i and j

How to estimate the distance

total substitution rate for any nucleotide is 3α

if two sequences are separated by time t (diverged t/2 ago) , the distance will be $d = 3\alpha t$ (the real number of substitutions)

if x of n sites are different $p=x/n$ (the observed substitutions)

the probability that the nucleotide in the two sequences are different is:

$p = 3p_1(t) = 3/4 - 3/4\ e^{-4\alpha t\ =\ 3/4} - 3/4\ e^{-4d/3}$

$d = -3/4 \log (1-(4/3)p)$

one way to derive Jukes Cantor probabilities

Assume a nucleotide at one site is A at time 0. What is the probability that the site is occupied by A at time *t*? This is denoted $P_{A(t)}$.

At $t = 0$, $P_{A(0)} = 1$

At $t = 1$, $P_{A(1)} = 1-3\alpha$

3α is the probability of change and $1 - 3\alpha$ is the probability of A remaining unchanged.

At $t = 2$, $P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha[1 - P_{A(1)}]$

At $t = 2$, there are two possibilities

1) the nucleotide has remained unchanged from time 0 to 2

2) the nucleotide changed to T, G or C at t=1 but changed back at $t = 2$

	1st possibility	2nd possibility
$t = 0$	A	A
$t = 1$	A	not A
$t = 2$	A	A

These give rise to the two terms at t=2.

$P_{A(t+1)} = (1-3\alpha)P_{A(t)} + \alpha[1-P_{A(t)}]$

for discrete time:

$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3\alpha P_{A(t)} + \alpha[1-P_{A(t)}] = -4\alpha\, P_{A(t)} + \alpha$

for continuous time:

$dP_{A(t)}/dt = -4\alpha P_{A(t)} + \alpha$

which is a first order linear differential equation with solution given by:

$P_{A(t)} = 1/4 + (P_{A(0)} - 1/4)\, e^{-4\alpha t}$

if $A(0) = 1$, then $P_{A(t)} = 1/4 + (3/4)\, e^{-4\alpha t}$

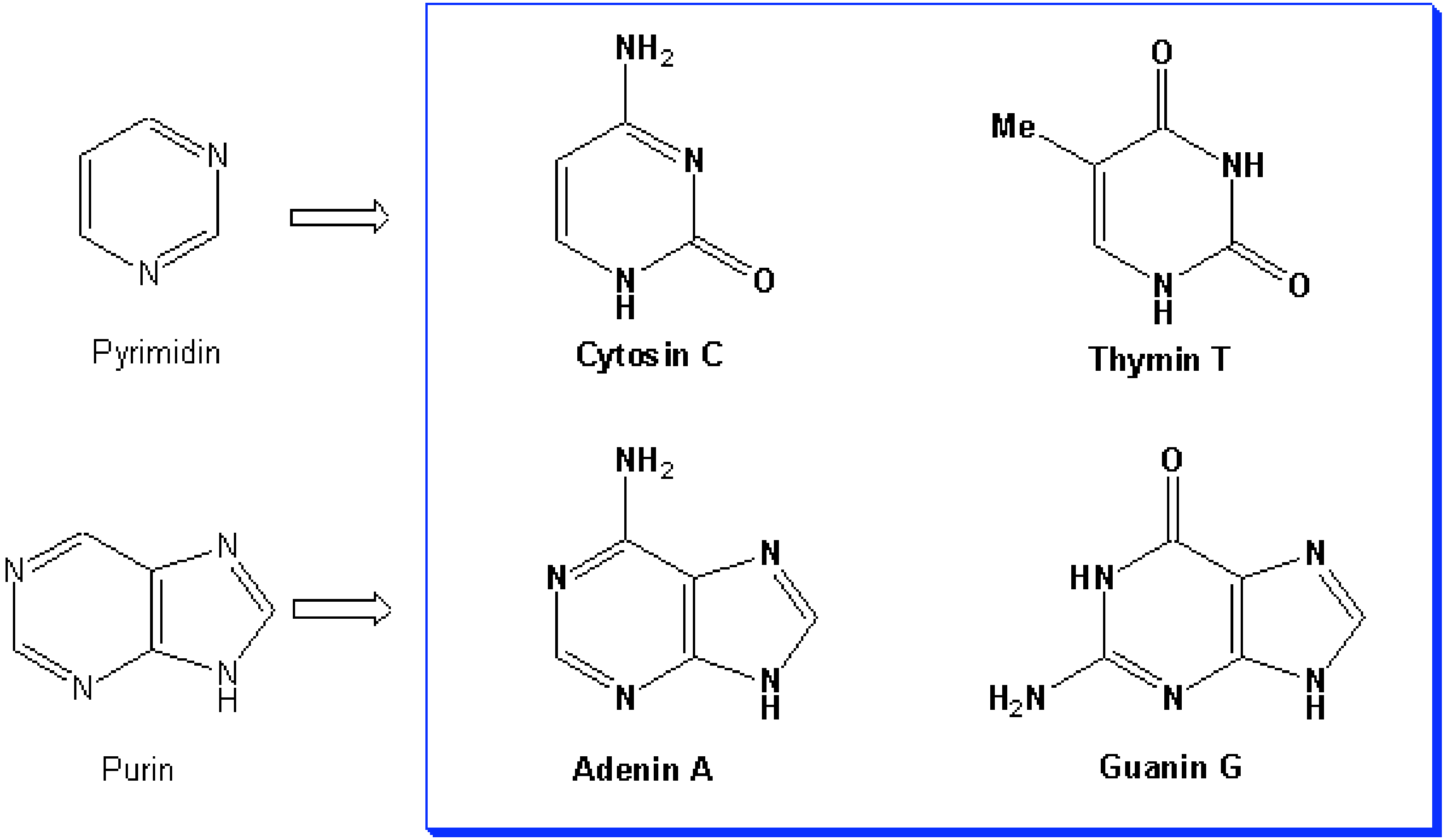
if $A(0) = 0$, then $P_{A(t)} = 1/4 - (1/4)\, e^{-4\alpha t}$

Idealized Mutations

- jump to idealized mutations biorecipe
- another way to come to the same conclusion

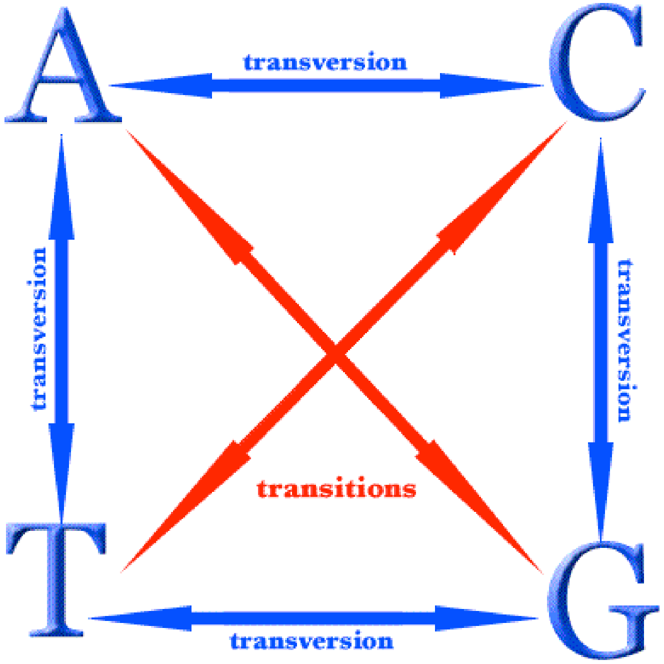
Chemistry of bases

- Two kinds of bases- purines and pyrimidines



Base substitutions

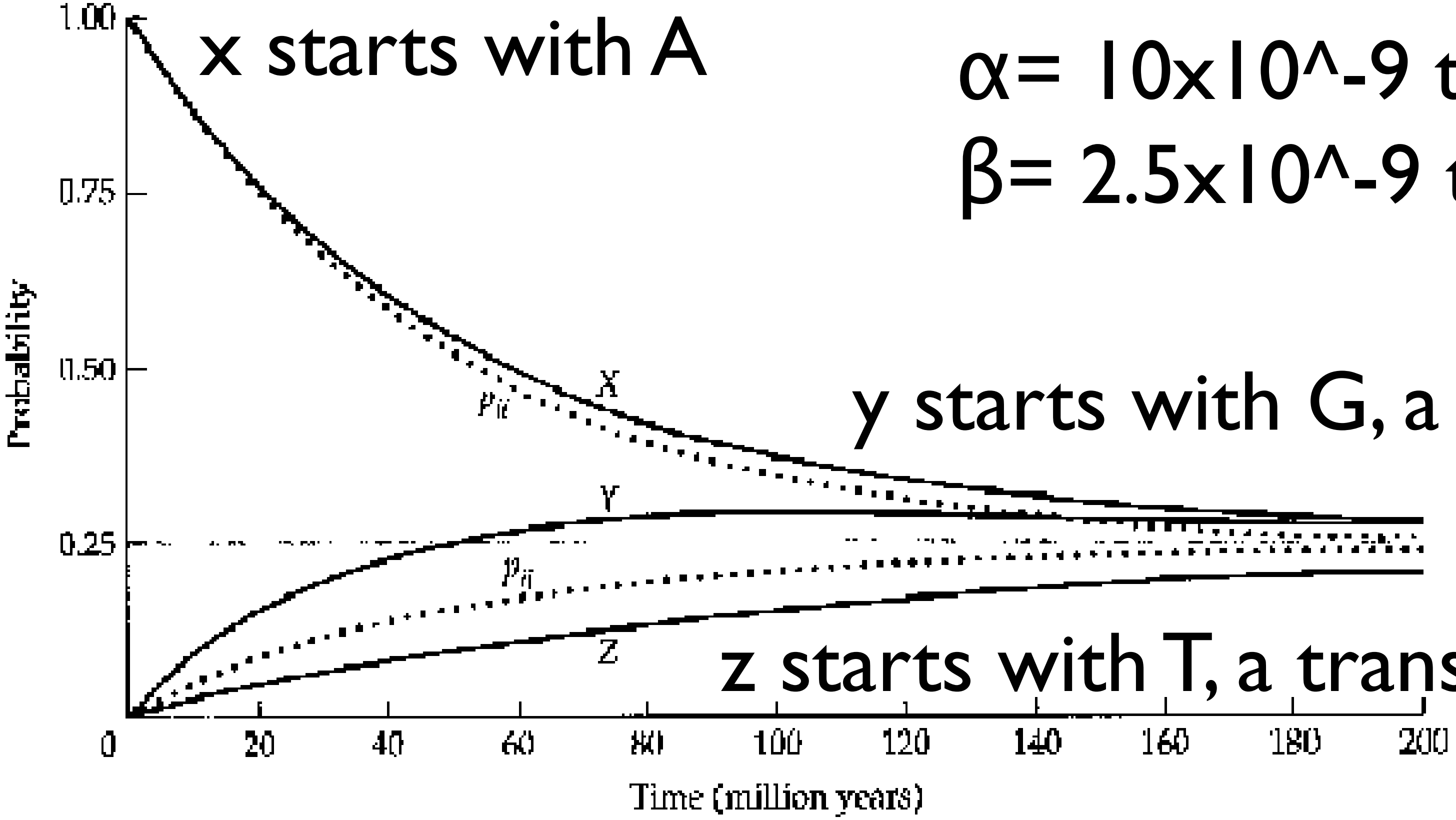
- transition - purine-purine or pyrimidine-pyrimidine substitution
- trasversion- substitution between a purine and a pyrimidine
- purines - R - nucleotides A and G
- pyrimidines - Y - nucleotides C and T



Kimura’s two parameter model

- two parameters α describe transition rate and β the transversion rate
- is identical to the JC model when $\alpha = \beta$
- parameters are estimated from the data

Probability with K2P



prob of having an A

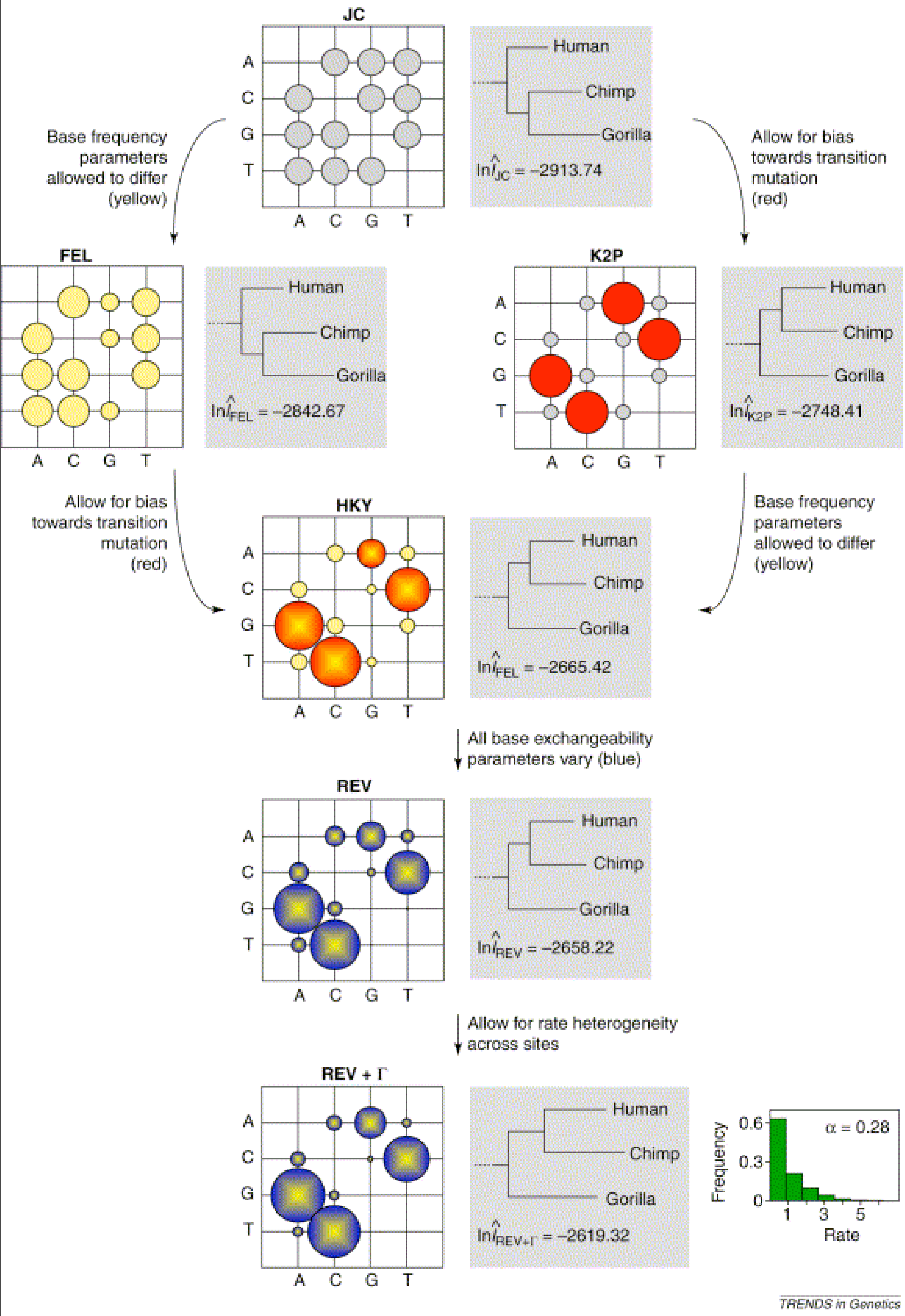
x starts with A

α= 10x10^-9 transitions
β= 2.5x10^-9 transversions

y starts with G, a transition

z starts with T, a transversion

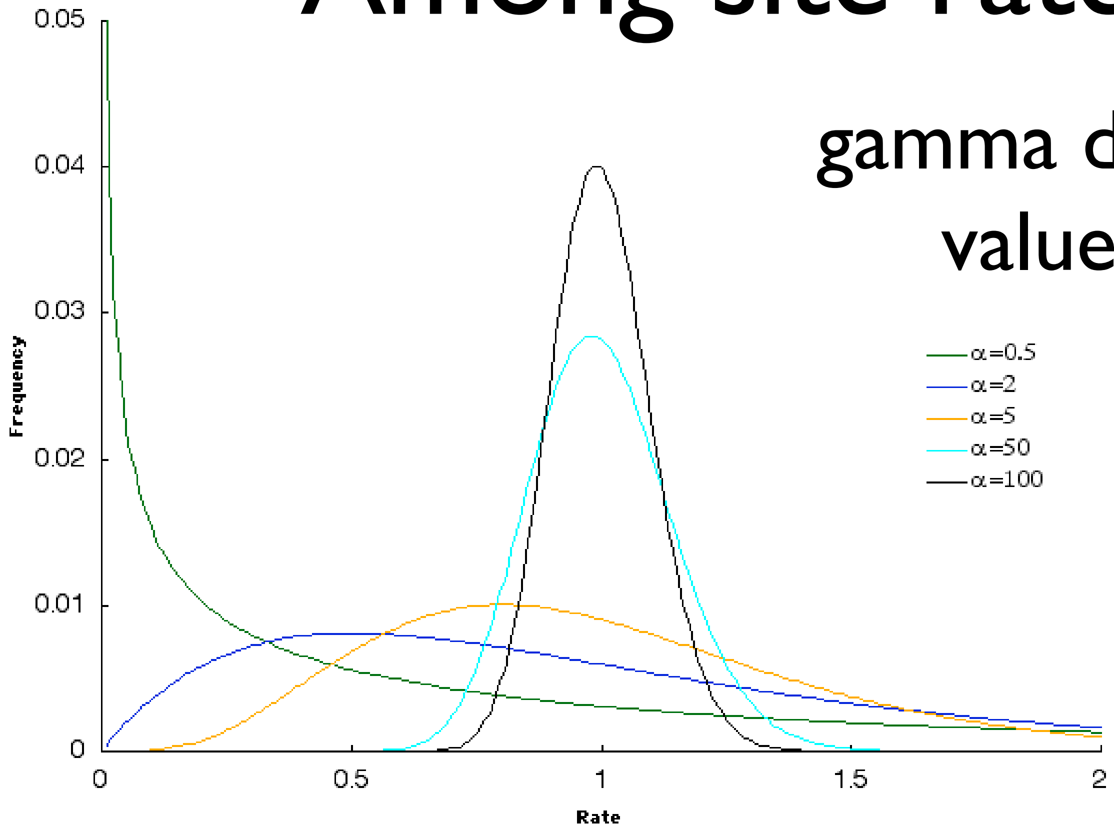
Figure 3.3 Temporal changes in the probability of having a certain nucleotide, say A, at a given nucleotide site. The two dotted lines are computed under the one-parameter model with $\alpha = 5 \times 10^{-9}$ substitutions/site/year. The line denoted by p_{ii} starts with the same nucleotide (i.e., A) while the line denoted by p_{ij} starts with a different nucleotide (i.e., T, C, or G). The three solid lines are computed under Kimura's two-parameter model with $\alpha = 10 \times 10^{-9}$ and $\beta = 2.5 \times 10^{-9}$ substitutions/site/year. The line denoted by X starts with A, the line denoted by Y starts with G (a transition), and the line denoted by Z starts with T (or C; a transversion). The dashed line denotes the equilibrium frequency (0.25).


JC
A
C
G
T
A C G T
Human
Chimp
Gorilla
ln$\hat{l}_{JC}$ = −2913.74
Base frequency parameters allowed to differ (yellow)
Allow for bias towards transition mutation (red)
FEL
Human
Chimp
Gorilla
ln$\hat{l}_{FEL}$ = −2842.67
K2P
Human
Chimp
Gorilla
ln$\hat{l}_{K2P}$ = −2748.41
Allow for bias towards transition mutation (red)
Base frequency parameters allowed to differ (yellow)
HKY
Human
Chimp
Gorilla
ln$\hat{l}_{FEL}$ = −2665.42
All base exchangeability parameters vary (blue)
REV
Human
Chimp
Gorilla
ln$\hat{l}_{REV}$ = −2658.22
Allow for rate heterogeneity across sites
REV + Γ
Human
Chimp
Gorilla
ln$\hat{l}_{REV+Γ}$ = −2619.32
Frequency
0.6
0.3
0
α = 0.28
1
3
5
Rate

TRENDS in Genetics

- each model described by a likelihood
- likelihood gets bigger as models describe the evolution better
- notice how the phylogeny and branch lengths change depending on the model
- nested models are indicated by arrows
- must use correct method to compare likelihoods of models with different numbers of parameters

Among site rate variation



gamma distribution with different values of the parameter α

- want to allow rates to vary from one site to another
- among site rate variation is often modeled by choosing rate at each site from a gamma distribution
- a single parameter α describes the shape of the distribution and is estimated from the data set
- α= infinity is the same as no variation of rates between sites

Codon Based models

- synonymous substitutions - the codon changes but the amino acid does not change
- nonsynonymous substitutions - amino acid changes as well as the codon

```
 P  M
CCCATG
CCAATA
 P  I
```

We want to estimate the real distance from the observed number of mutations but now we need to estimate two distances, one for synonymous and one for nonsynonymous.

Synonymous substitutions:

```
 P  M
CCCATG
CCAATA
 P  I
```

- do not change the encoded amino acid
- as selection acts on the protein, they are evolving under fewer functional constraints
- + have almost clock-like behavior for short distances, -saturate
- ratio of synonymous to nonsynonymous change used to identify selection

Rates of Change

	nonsynonymous rate	synonymous rate
Ribosomal Proteins		
S14	.02 $+-$ .02	2.16$+-$0.42
S17	.06 $+-$ .04	2.69$+-$0.53
Contractile System Proteins		
actin α	.01 $+-$ .01	2.92 $+-$ 0.34
Mysoin β heavy chain	.10 $+-$ .01	2.15 $+-$ 0.13
Misc		
Relaxin	2.59 $+-$ 0.51	6.39 $+-$ 3.75
γ interferon	3.06$+-$ 0.37	5.50 $+-$ 1.45

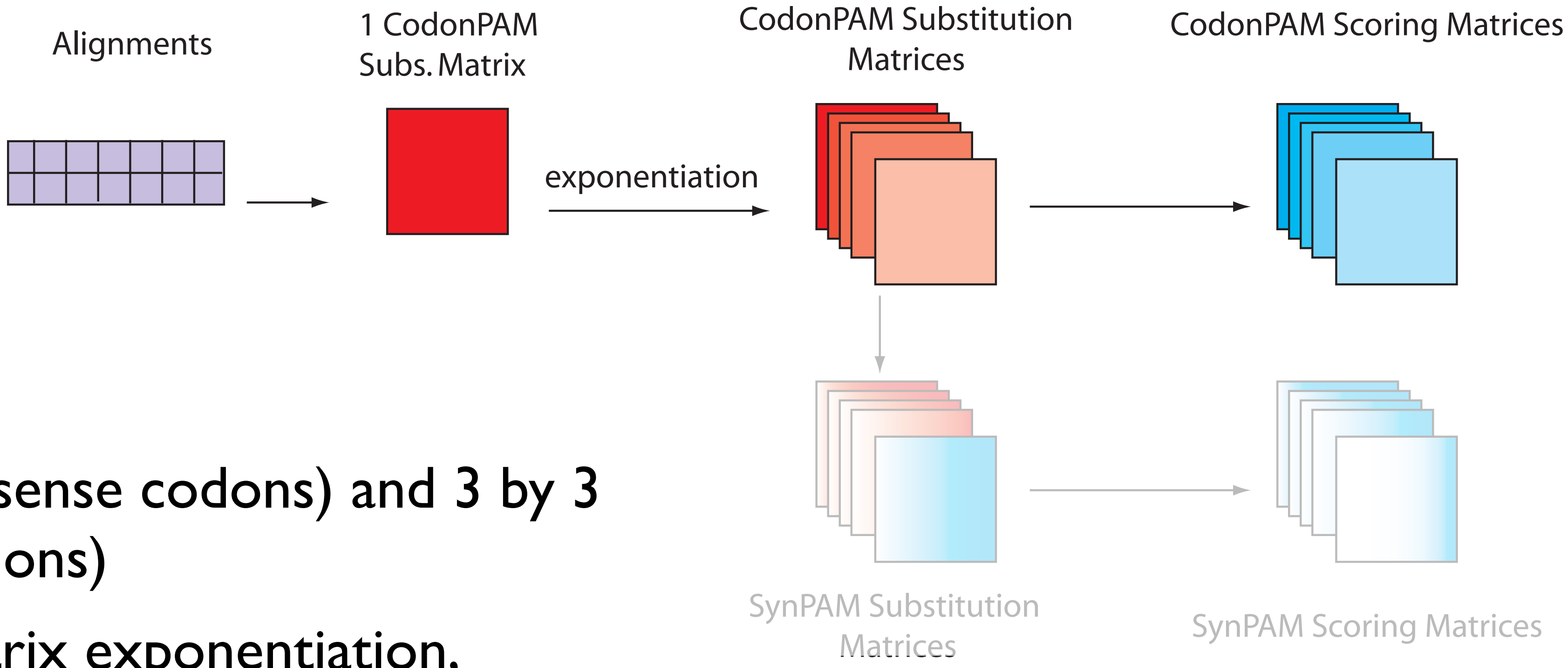
based on comparison of human and mouse/rat (divergence time set at 80 MYA) rates are in units of substitutions per site per 10^9 years.

synonymous varies less than nonsynonymous

Models of codon evolution

- Nonparametric (empirical) -Schneider, Cannarozzi, Gonnet
- Parametric (Yang, Pupko)

Codon substitution matrices


Alignments
1 CodonPAM Subs. Matrix
exponentiation
CodonPAM Substitution Matrices
CodonPAM Scoring Matrices
SynPAM Substitution Matrices
SynPAM Scoring Matrices

- 61 x 61 (sense codons) and 3 by 3 (stop codons)
- using matrix exponentiation, CodonPAM matrices representing different evolutionary distances are constructed
- Scoring matrices are then derived via the method of Dayhoff

Result: Maximum Likelihood estimation of the distance based on codon substitutions

part of the semester work of Adrian Schneider

Results

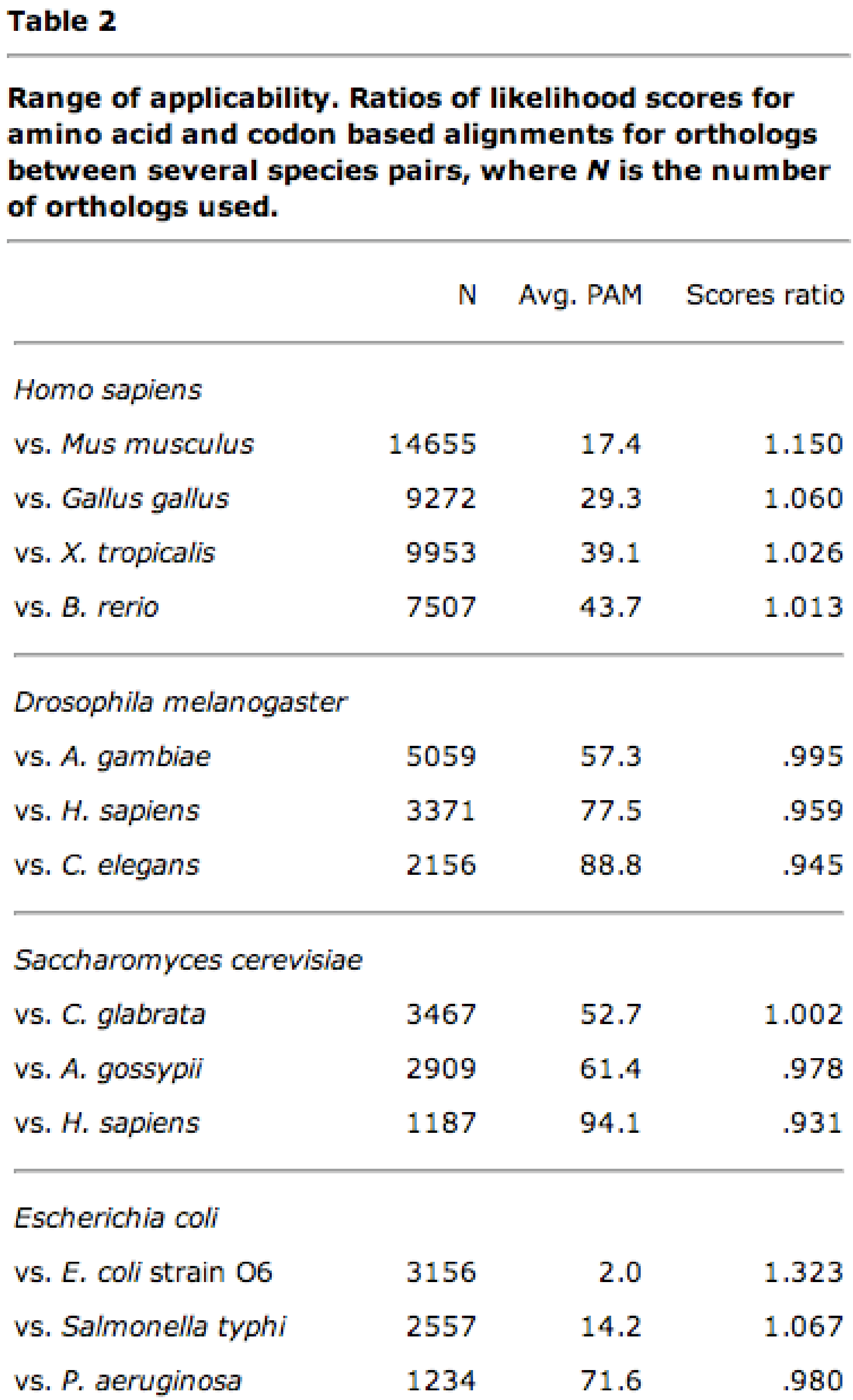


Table 2

Range of applicability. Ratios of likelihood scores for amino acid and codon based alignments for orthologs between several species pairs, where *N* is the number of orthologs used.

	N	Avg. PAM	Scores ratio
Homo sapiens			
vs. *Mus musculus*	14655	17.4	1.150
vs. *Gallus gallus*	9272	29.3	1.060
vs. *X. tropicalis*	9953	39.1	1.026
vs. *B. rerio*	7507	43.7	1.013
Drosophila melanogaster			
vs. *A. gambiae*	5059	57.3	.995
vs. *H. sapiens*	3371	77.5	.959
vs. *C. elegans*	2156	88.8	.945
Saccharomyces cerevisiae			
vs. *C. glabrata*	3467	52.7	1.002
vs. *A. gossypii*	2909	61.4	.978
vs. *H. sapiens*	1187	94.1	.931
Escherichia coli			
vs. *E. coli* strain O6	3156	2.0	1.323
vs. *Salmonella typhi*	2557	14.2	1.067
vs. *P. aeruginosa*	1234	71.6	.980

Synonymous vs Nonsynonmous

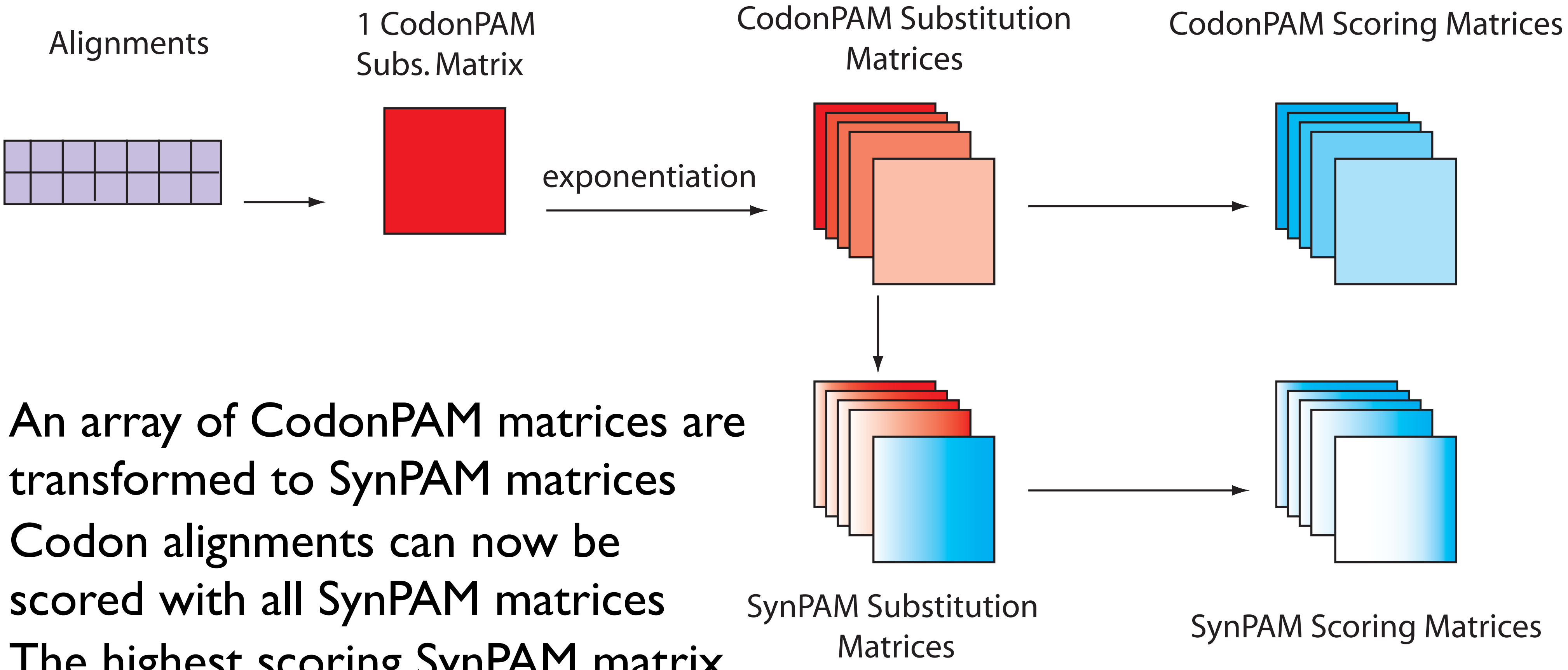
		Lysine		Asparagine	
		AAA	AAG	AAC	AAT
Lys	AAA	0.802	0.112	0.003	0.006
	AAG	0.127	0.826	0.003	0.004
Asn	AAC	0.003	0.002	0.811	0.119
	AAT	0.004	0.002	0.105	0.782

20 CodonPAM
matrix

- diagonal around .8
- synonymous - higher probability than nonsynonymous
- nonsynonymous have low probability

- **New Goal**: Maximum likelihood estimation of the distance based on only *synonymous* substitutions
- **Problem**: We could use the same formalism but the amount of synonymous substitutions decreases with increasing distance
- **Solution**: Transform the CodonPAM matrices to describe only the relative synonymous substitution probabilities

Estimating Synonymous Change


Alignments
1 CodonPAM Subs. Matrix
exponentiation
CodonPAM Substitution Matrices
CodonPAM Scoring Matrices
SynPAM Substitution Matrices
SynPAM Scoring Matrices

- An array of CodonPAM matrices are transformed to SynPAM matrices
- Codon alignments can now be scored with all SynPAM matrices
- The highest scoring SynPAM matrix corresponds to the ML estimation of the SynPAM distance
- 1 SynPAM is the distance where 1% of the synonymous positions undergo substitution

diplomarbeit of
Adrian Schneider

Example:

		Lysine		Asparagine	
		AAA	AAG	AAC	AAT
Lys	AAA	0.802	0.112	0.003	0.006
	AAG	0.127	0.826	0.003	0.004
Asn	AAC	0.003	0.002	0.811	0.119
	AAT	0.004	0.002	0.105	0.782

20 CodonPAM matrix

	AAA	AAG	AAC	AAT
AAA	0.864	0.120	0	0
AAG	0.136	0.880	0	0
AAC	0	0	0.885	0.132
AAT	0	0	0.115	0.868

transformed probabilites

	AAA	AAG	AAC	AAT
AAA	2.7	-5.9	0	0
AAG	-5.9	2.2	0	0
AAC	0	0	2.2	-6.1
AAT	0	0	-6.1	2.7

Synonymous scoring matrix

Other measures of synonymous distance

- dS - synonymous substitutions per synonymous site
- NED or TREx- Benner- look only at conserved amino acids with exactly 2 codons; model decay of percent identity as exponential; (R-R transitions- Glu, Gln, Lys Y-Y transitions- Cys, Asp, Asn, Tyr, Phe, His)
- We used the ML dS implementation of Yang's PAML and implemented NED in Darwin.

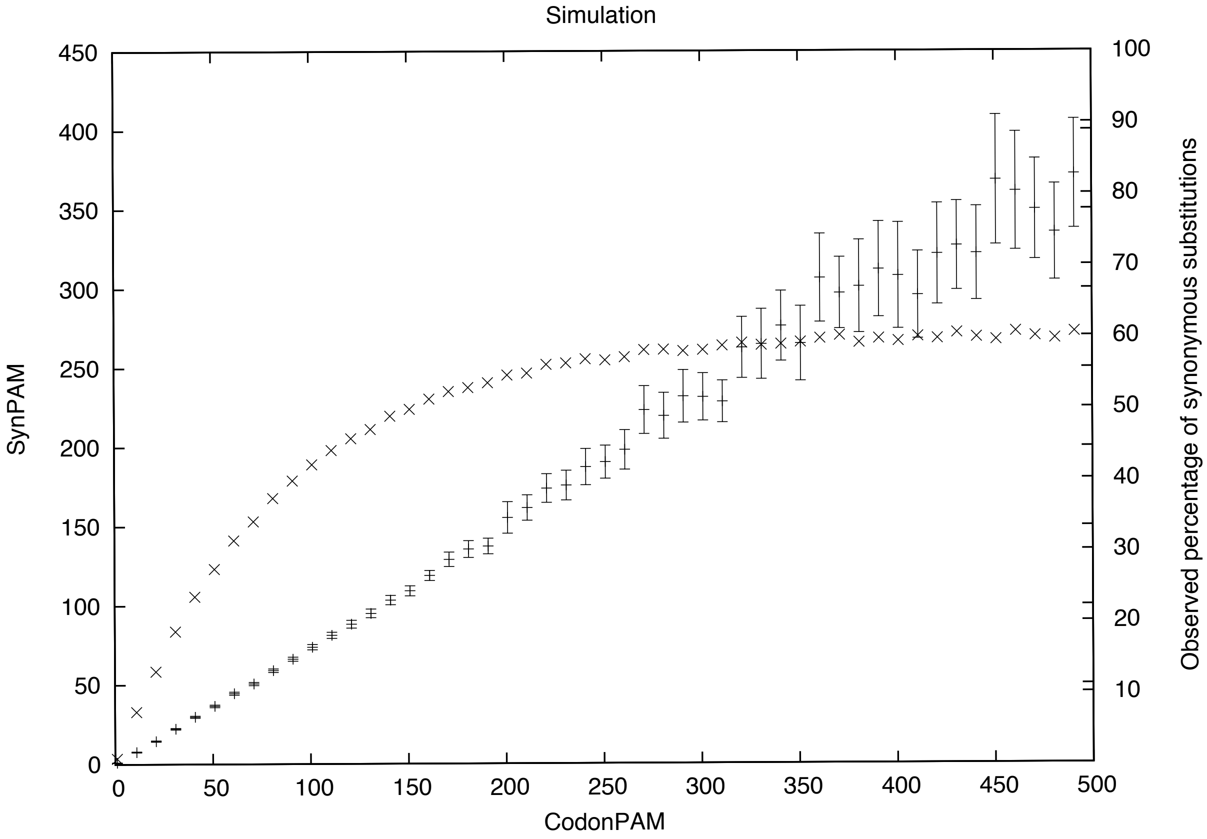
Results

Two ways to compare distance measures:

- Simulated data
  - + the true result is known
  - - what model of evolution to use?
- Real data
  - + most realistic
  - - how to compare the results?

Simulated Data

dog/human 40
chicken/human 110
fish/human 160
ciona/human 290
mosquito/human 300



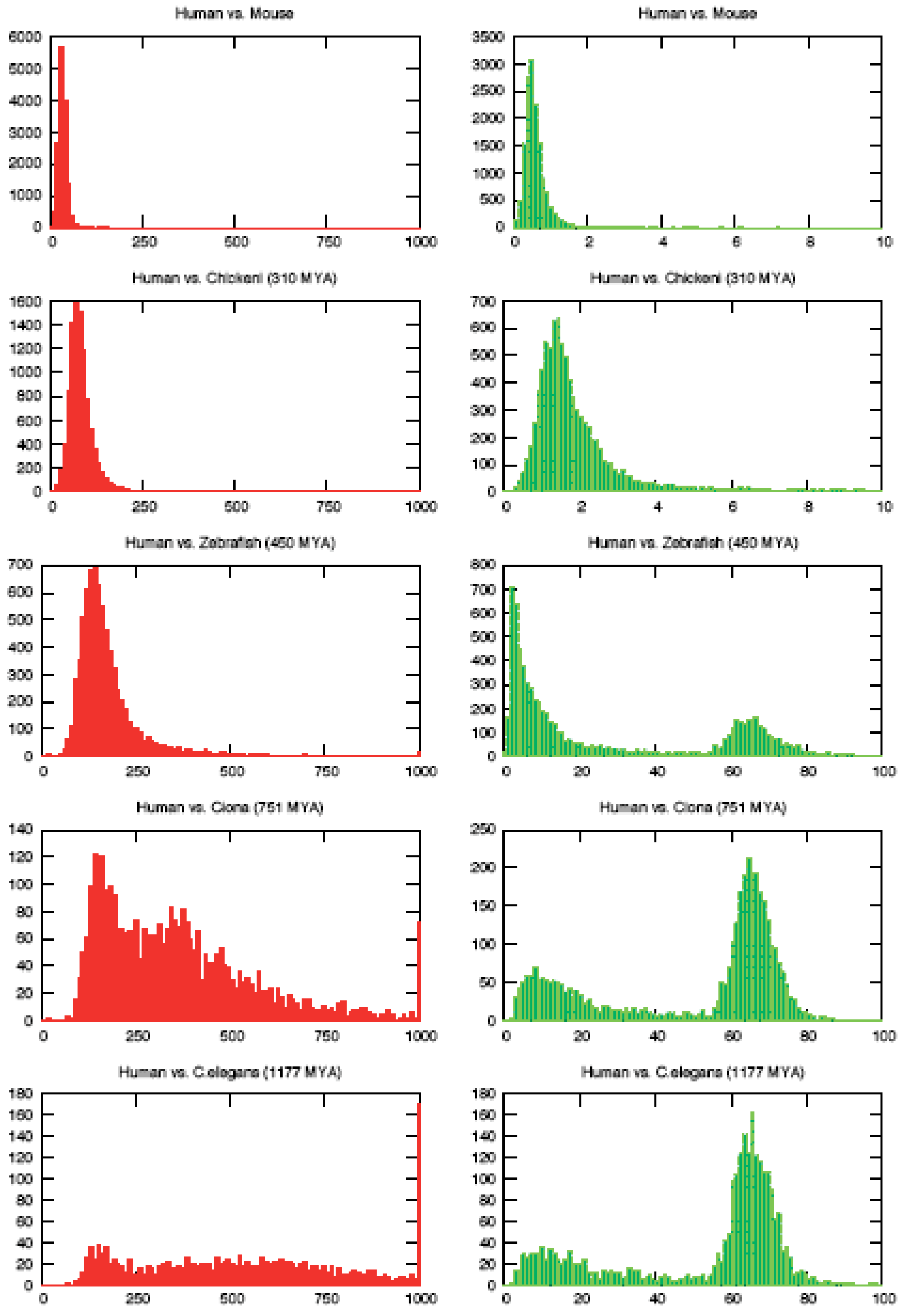
- Result of 100 repetitions of mutating 500 codons over different CodonPAM distances
- Verification of SynPAM is a semester project for Barbara Keller

Biological Data

- Apply the methods to orthologous sequence pairs as they should have the same divergence time (OMA project)*
- Compute for all methods the variances of the pairwise distance estimates of the orthologs
- Take the unbiased estimator with the lowest variance as the best

* C Dessimoz, GM Cannarozzi, M Gil, D Margadant, A Roth, A Schneider and GH Gonnet, 2005: **OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements,** LNCS 3678: Comparative Genomics: RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland.

SynPAM and dS for all orthologs



human-mouse

human-chicken
(310 MYA)

human-zebrafish
(450 MYA)

human-ciona
(751 MYA)

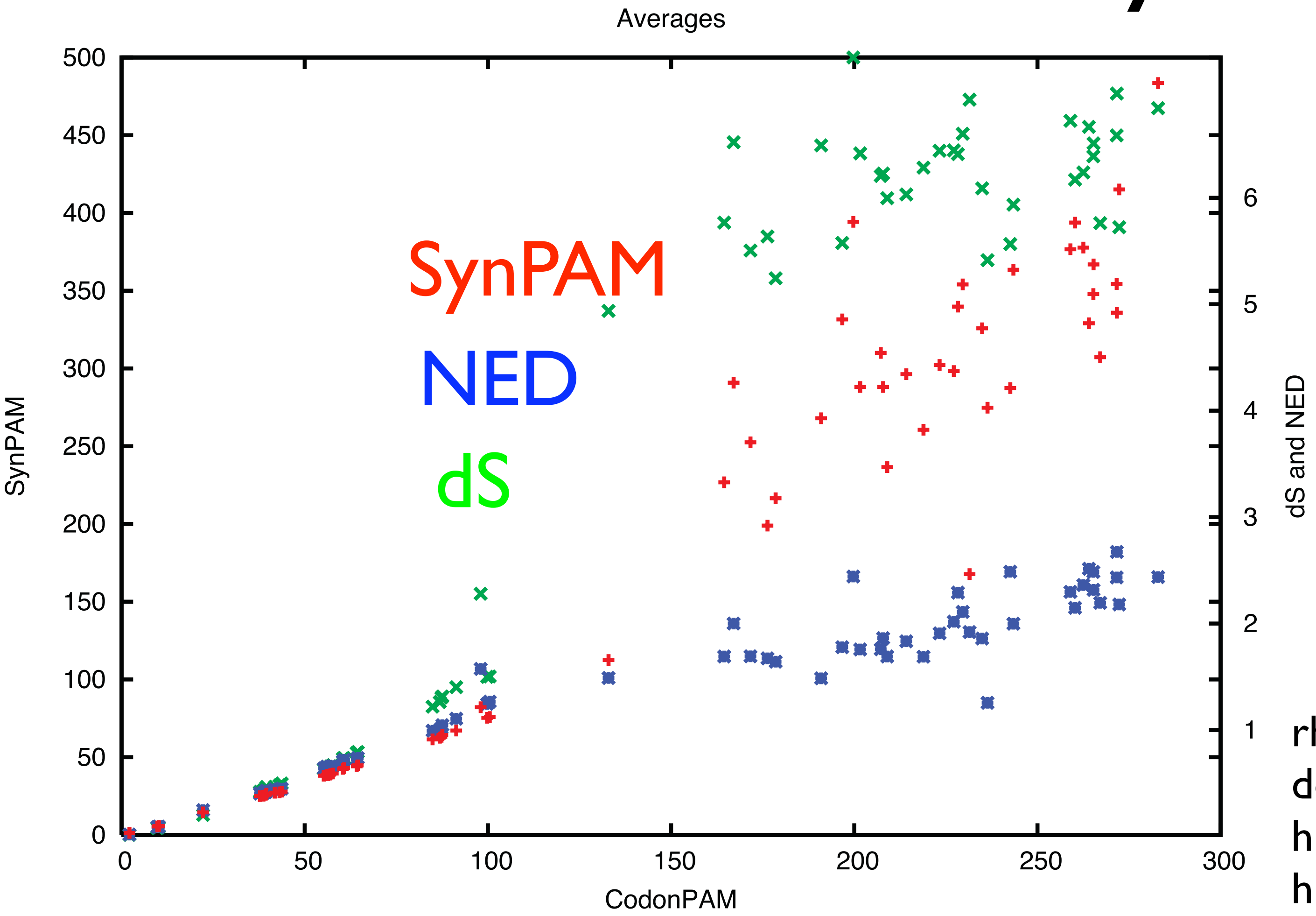
human-C. elegans
(1177 MYA)

Coefficient of Variance

$$CV(X) = \sigma(x)/\mu(x) = \frac{\text{standard deviation}}{\text{mean}}$$

Human vs.	# pairs	SynPAM	NED	dS
Chimp	14565	1.87	2.31	5.63
Dog	15439	0.37	0.48	0.70
Mouse	15265	0.29	0.38	0.59
Opossum	12513	0.35	0.47	0.70
Chicken	8031	0.37	0.45	0.67
Frog	3131	0.39	0.43	0.56
Zebrafish	2627	0.50	0.46	0.49
Ciona	201	0.73	0.42	0.32
Drosophila	101	0.89	0.59	0.32

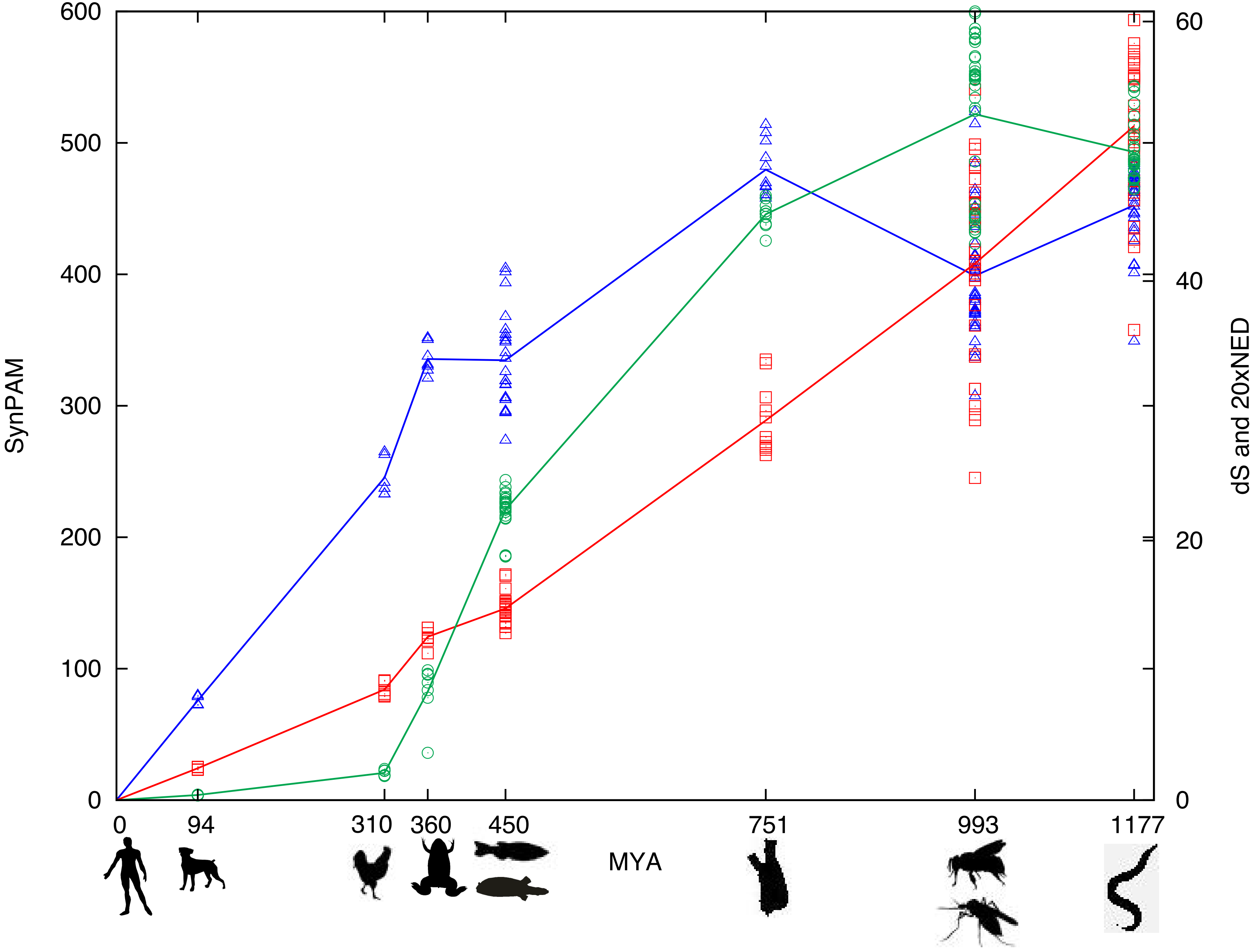
CodonPAM vs SynPAM



human-chimp	17948	2.3
rhesus monk.-human	18686	11.3
dog-human	16092	40.3
human-cow	15903	45.1
human-rat	15571	60.4
human-mouse	16336	60.0
opossum-human	13958	94.4
chicken-human	10372	121.9
honey bee-human	3469	260.5

all orthologs- average CodonPAM vs average SynPAM

Molecular dating



Selection

- the ratio of nonsynonymous to synonymous change can indicate selection
- high amounts of accepted nonsynonymous changes indicate pressure to change amino acid composition and thus function
- positive or directional selection increases the frequency of a beneficial mutation
- negative or purifying selection is the selective removal of rare alleles that are deleterious. This can result in the maintenance of conserved gene sequences between species over long periods of evolutionary time.

Remember that we have two sequences and we are trying to estimate the synonymous and nonsynonymous distances.

Detecting selection with dN/dS

- dN or Ka = number of nonsynonymous substitutions/nonsynonymous site
- dS or Ks = number of synonymous substitutions/synonymous site
- measure the amount of selection by looking at the ratio of dN/dS or Ka/Ks
- (dN/dS >1 = positive selection)

What has been done to measure dN/dS or Ka/Ks

- heuristic based counting methods
- ML method of Yang
- Alessandro Rigazzi is doing a semesterarbeit investigating PAM/SynPAM as a measure of evolution/ time. One strength of this approach is that it is based on a Markov model.

History (started around 1980 with first DNA sequencing)

1980 Miyata and Yasunaga - assumed a simple model with equal rates between nucleotides and weighted using amino acid similarity

1986 Nei and Gojobori used equal weighting

1993 Li et al. pointed out the importance of transition/transverion rate differences and treated it by putting codon positions into different degeneracy classes

improved on by Li (1993), Pamilo and Bianchi (1993), Comeron (1995) and Ina (1995)

Counting methods

involve 3 steps

- 1) Count synonymous and nonsynonymous sites
- 2) Count synonymous and nonsynonymous differences
- 3) Calculate the proportions of differences and correct for multiple hits

What is a synonymous site? What is a nonsynonymous site? How do you count synonymous and nonsynonymous differences when the codons differ by more than one nucleotide?

1) Counting Sites

Each codon has 3 nucleotide sites, divided into synonymous and nonsynonymous categories. For example TTT (Phe) has 9 neighbors:

`TTC (Phe)`
`TTA (Leu)`
`TTG (Leu)`
`TCT (Ser)`
`TAT (Tyr)`
`TGT (Cys)`
`CTT (Leu)`
`ATT (Ile)`
`GTT (Val)`

Only 1, TTC, codes for the same amino acid thus there are $3 * (1/9) = 1/3$ synonymous sites and $3 * (8/9) = 8/3$ nonsynonymous sites in codon TTT. (Transitions to stop codons are not allowed.)

Apply this procedure to all codons in sequence 1 to obtain numbers of syn and nonsyn sites. Do the same for sequence 2. Use the average of the two sequences (average synonymous and average nonsynonymous).

2) Counting differences

observed differences are partitioned into two categories- synonymous and nonsynonymous.   3 cases:

- 1) codons do not differ (TTT vs TTT) the number of synonymous and nonsynonymous differences are zero)
- 2) codons differ by one nucleotide (TTC vs TTA) - if the encoded amino acid changes, it is nonsynonymous. If not, it is synonymous.
- 3) codons differ by two or three nucleotides, there are 4 or 6 evolutionary pathways from one to the other. consider all pathways, either weighted or unweighted.

Consider the two pathways between codons CCT and CAG:
CCT (Pro) ↔ CAT (His) ↔ CAG (Gln)

CCT (Pro) ↔ CCG (Pro) ↔ CAG (Gln)   .5 syn diffs and 1.5 nonsyn diffs

If synonymous rate is higher than the nonsynonymous rate, the second pathway should be more likely.   Counting is done codon by codon across the 2 aligned sequences to produce the numbers of synonymous and nonsynonymous differences.

3) Correct for multiple hits

- S_d = # of synonymous differences; S = number of synonymous sites
- N_d = number of nonsynonymous differences; N = number of nonsynonymous sites
- Use a model to correct for multiple hits (e.g. Jukes Cantor)
- $p_S = S_d/S$; $d_S = -3/4 \log (1-4/3p_S)$
- $p_N = N_d/N$; $d_N = -3/4 \log (1-4/3p_N)$

Mechanistic Codon Models

- Yang Mol. Biol. Evol. 19(6):908-917 2002

$$Q = q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than 1 position} \\ \mu\pi_j & \text{for synonymous transversion} \\ \mu\kappa\pi_j & \text{for synonymous transition} \\ \mu\omega\pi_j & \text{for nonsynonymous transversion} \\ \mu\omega\kappa\pi_j & \text{for nonsynonymous transition} \end{cases}$$

q_{ij}= rate of subs. from i to j

κ is the transition/transversioin rate ratio

π_j is the equilibrium codon frequency

μ is a factor defined so that the average substitution rate is one

ω is the nonsynonymous/synonymous substituion rate ratio = dN/dS

$$P(t) = e^{Qt}$$

estimate parameters to maximize probability of observed data

How do we use this information?

Mutations can be silent (does not change the amino acid) or nonsilent (changes the amino acid). Silent changes are assumed to not be under selective pressure. Consider the following 3 possibilities:

- Consider a protein that is perfectly optimized for its function and its function remains constant over time. Expressed mutations will diminish its functionality and be removed from the gene pool. Silent changes will not be removed and will accumulate with clock-like behavior. Expected is a low ratio of expressed to silent changes.
- Consider a protein acquiring a new derived function. The amino acid sequence at the beginning of the episode will be optimized for the ancestral function. Amino acids will have to change for the protein to be optimized for the derived function therefore there will be positive selection for changes that improve the ability to perform the new function. A high ratio of expressed/silent changes is expected.
- Pseudogenes (genes that used to be functional but now are no longer) will accept synonymous and nonsynonymous changes

Look at the dN/dS ratio- does it indicate negative selection (case 1), positive selection (case 2) or neutral drift (case 3)?

An example- Selection in lysozymes

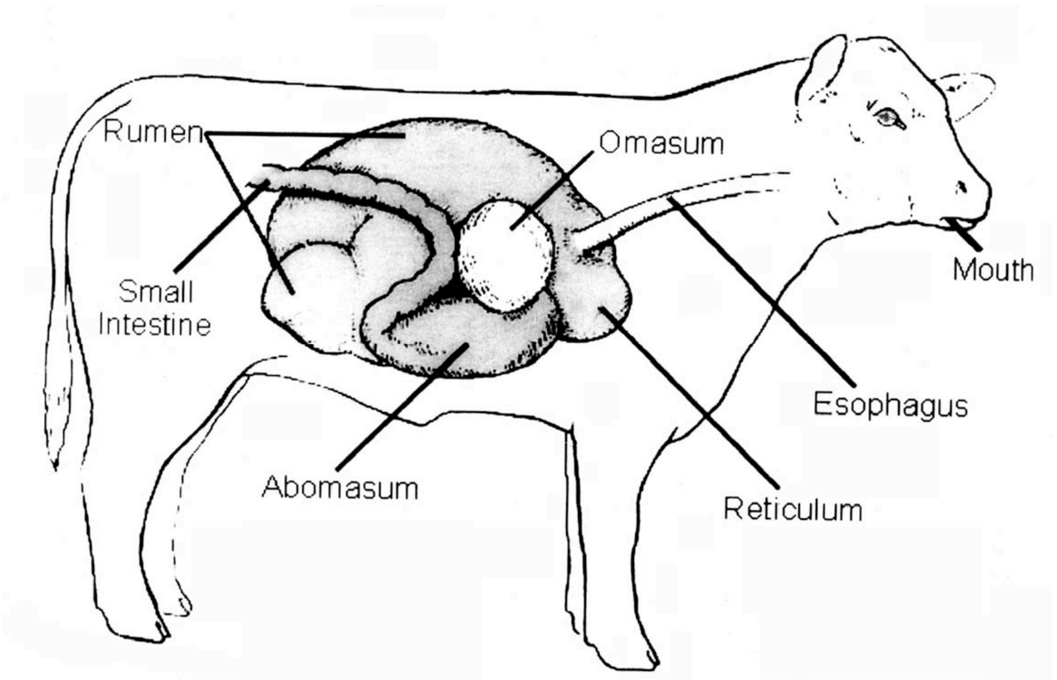
- difficult to demonstrate rigorously that amino-acid differences have adaptive significance
- two tests for positive selection: sequence convergence and neutral rate violation
- Lysozymes from the stomachs of cows and langur monkeys (forgut fermentors) show amino-acid convergence
- Messier and Stewart (Nature 1997 385(6612):151) used ancestral reconstruction and tests of neutral rate violation to document positive selection on the lineage leading to the common ancestor of the foregut fermenting monkeys

Lysozyme

- Lysozyme is a 130 amino acid long enzyme, whose catalytic function is to cleave the β(1-4) glycosidic bonds between N-acetyl glucoseamine and N-acetyl muramic acid in the cell walls of eubacteria, thereby depriving the bacteria of their protection against osmotic pressure and subsequent lysis.
- By virtue of its catalytic function and its expression in body fluids, such as saliva, serum, tears, avian egg white and mammalian milk, lysozyme usually serves as a first-line defense against bacterial invasion.

Foregut fermenters

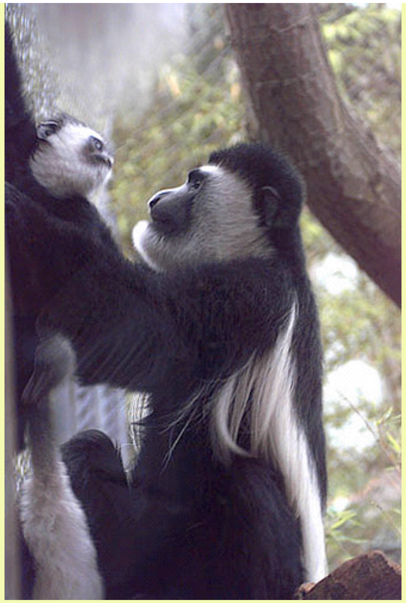
Animals in which the anterior part of the stomach functions as a chamber for bacterial fermentation of ingested plant matter.



Ruminants: cows, deer, sheep, giraffe

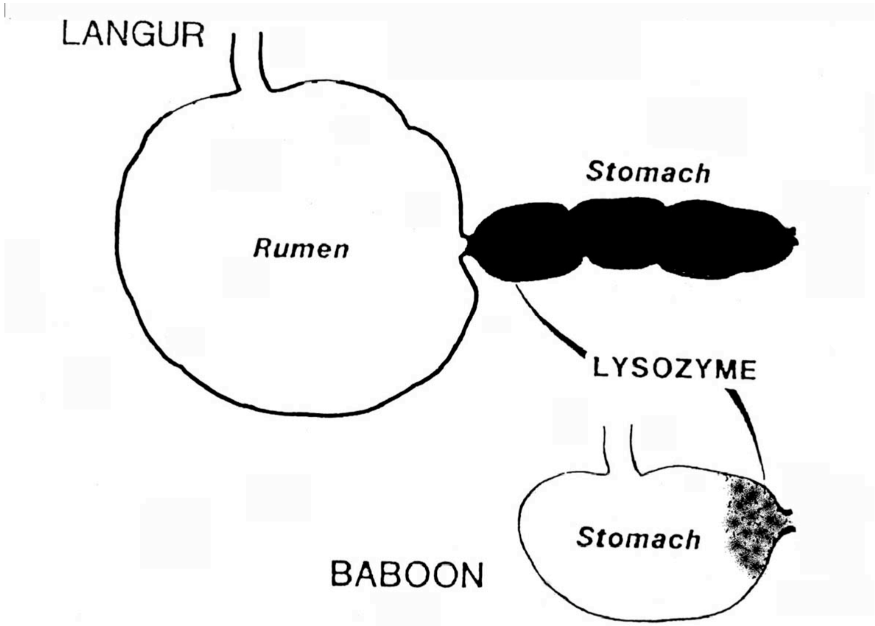
Why? because there is "room in it"

Foregut fermentation is not limited to artiodactyls

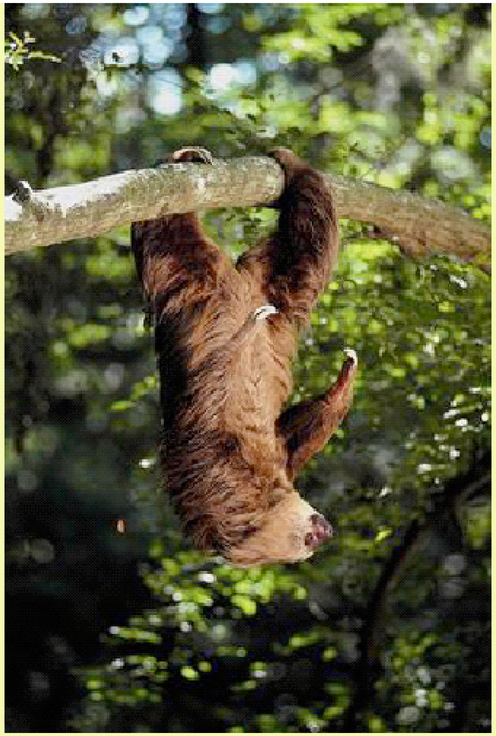


Colobus guereza

colobine monkeys
(e.g. langurs)

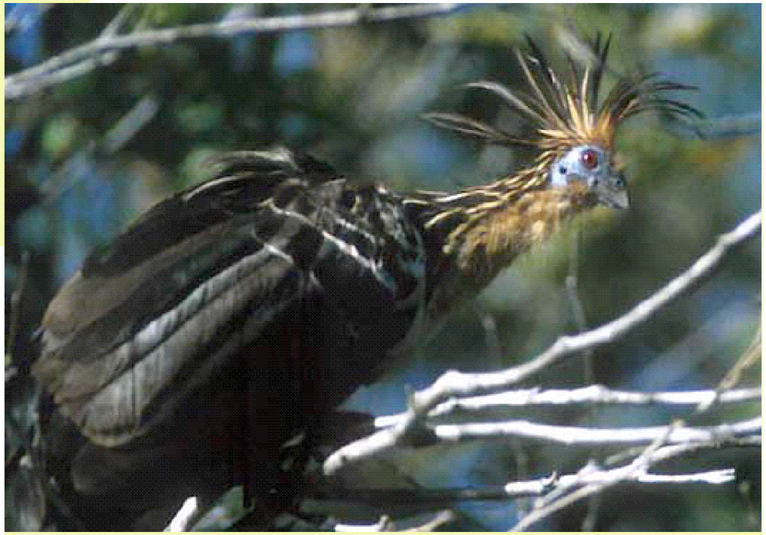


Others



Choloepus didactylus

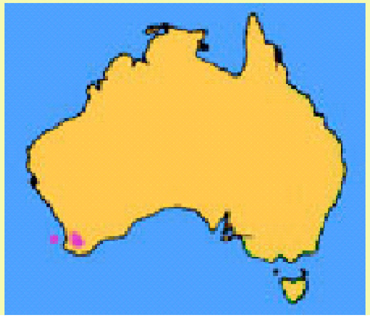
South American bird, Hoatzin



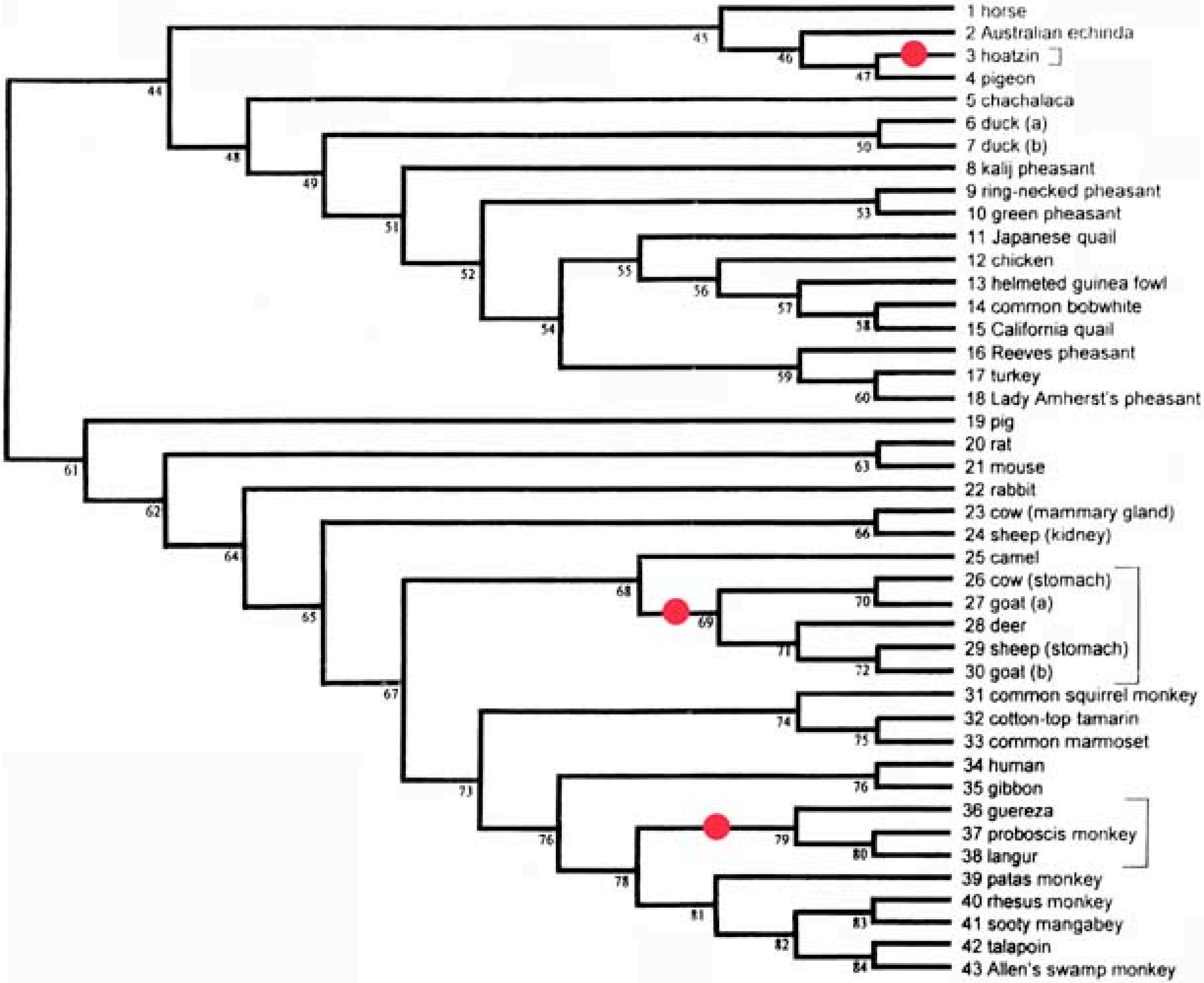
Opisthocomus hoazin

Quokka

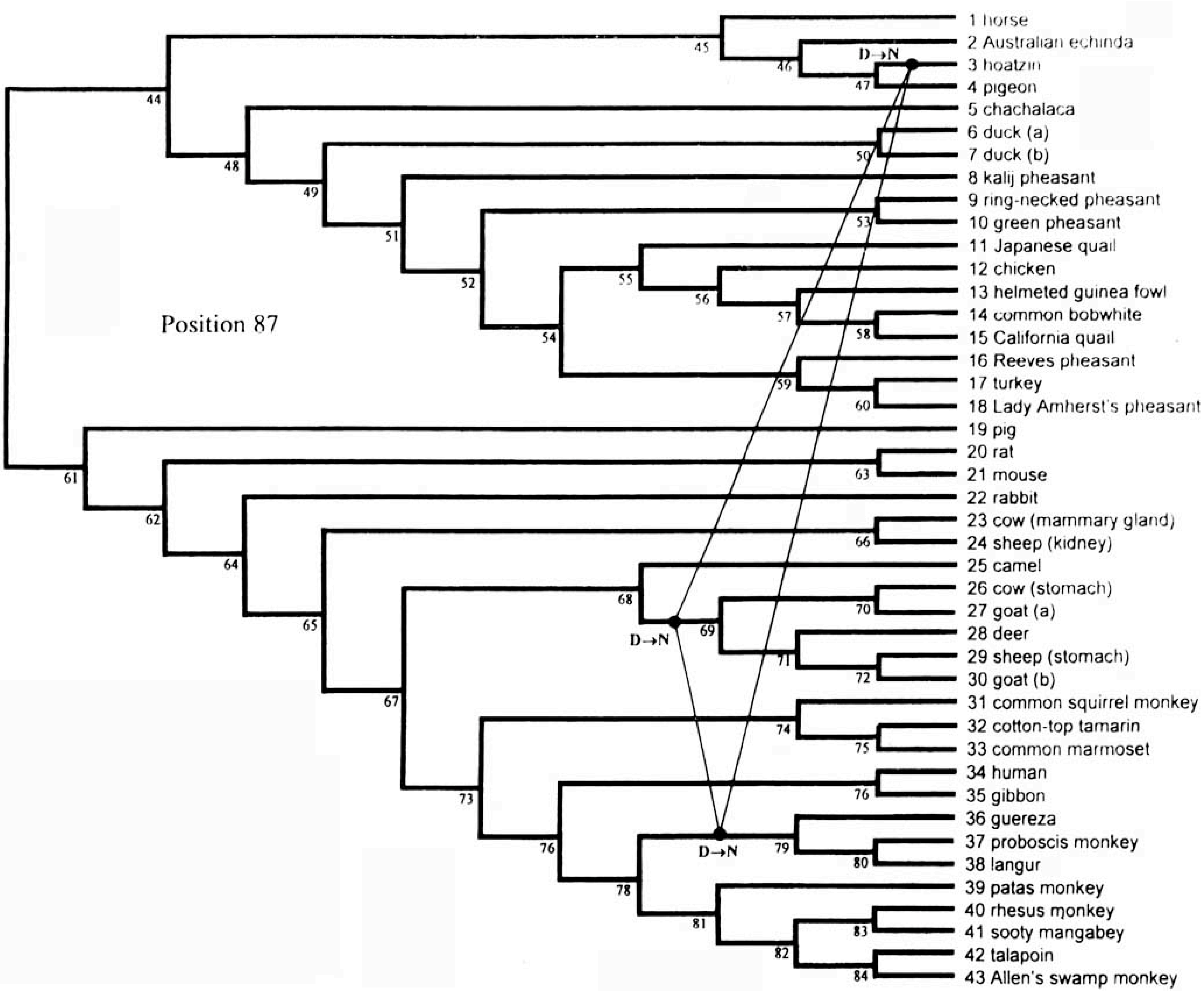
Setonix brachyurus



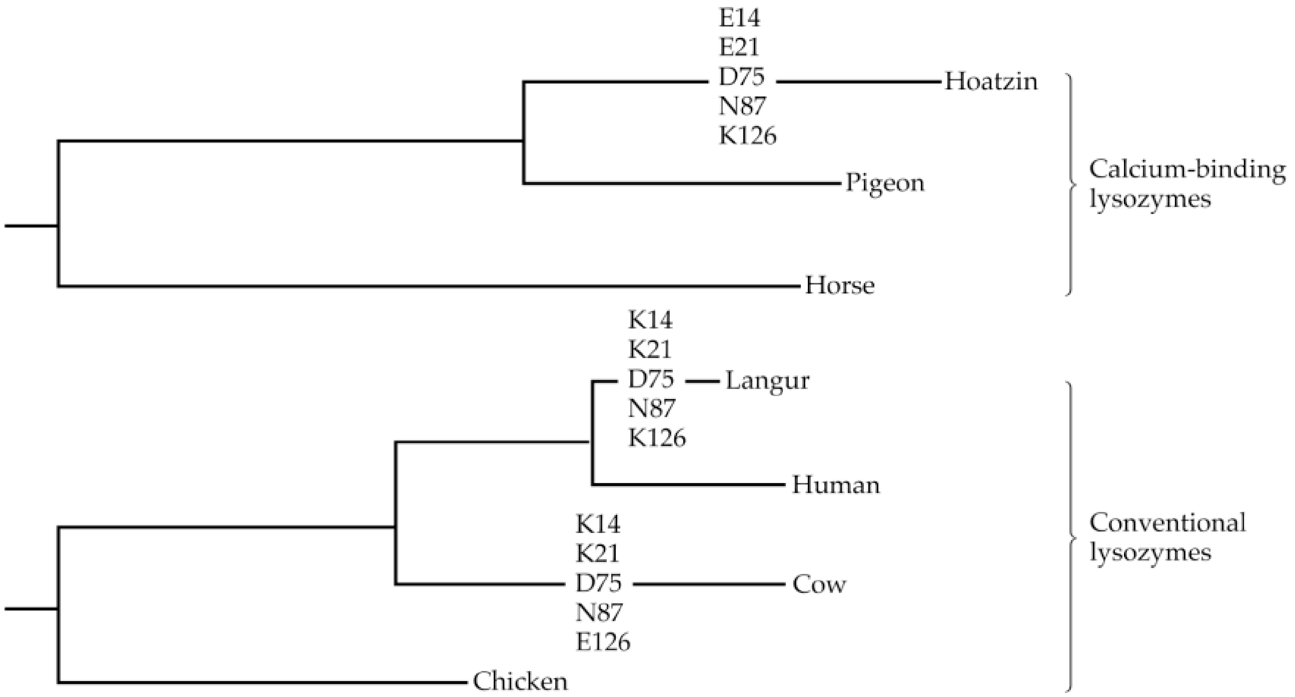
Branches leading to foregut fermenters



Adaptive D to N substitution



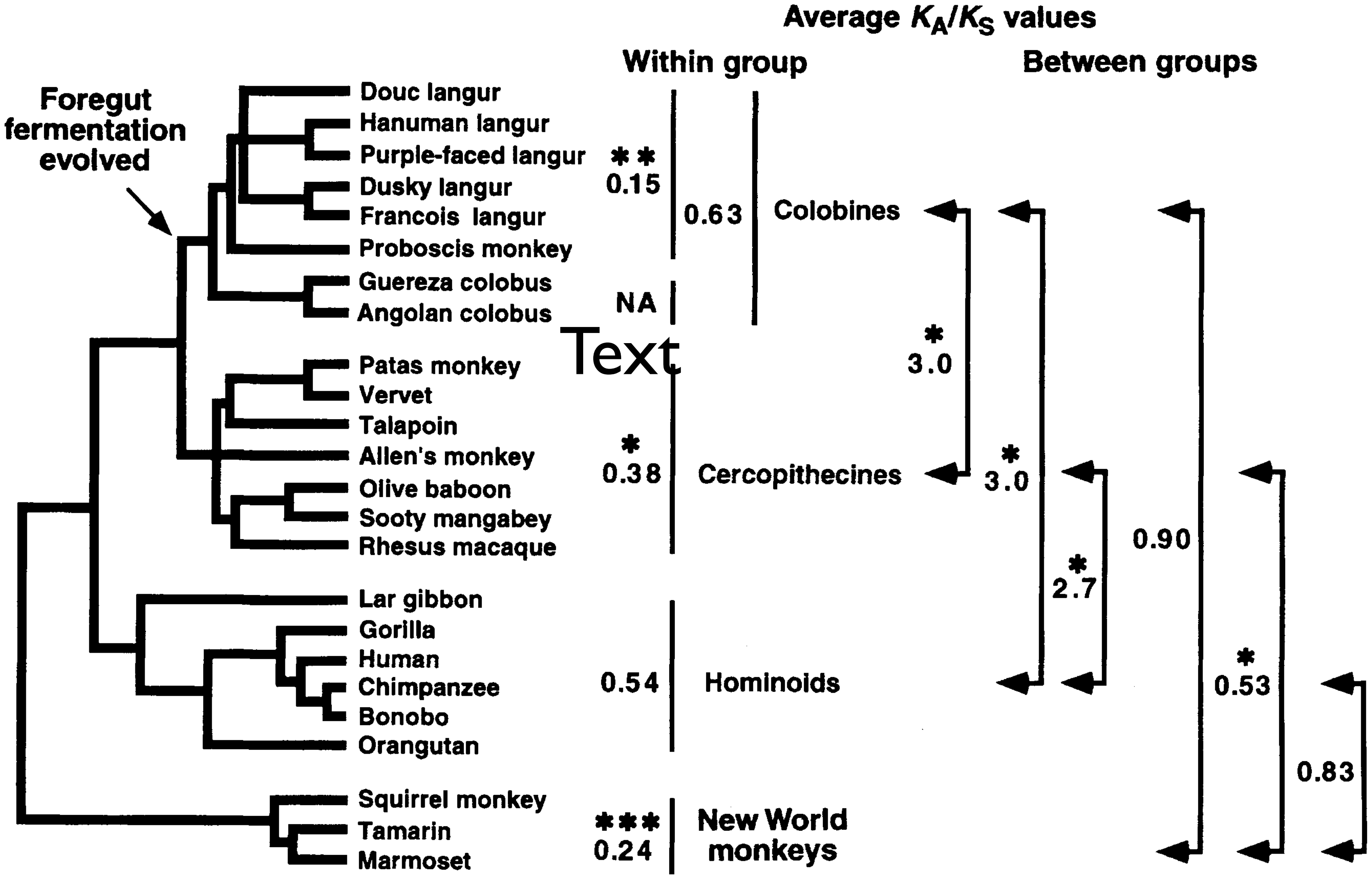
Convergent amino acid substitutions



Convergent amino-acid replacements in lysozymes from the foregut of cow, langur and hoatzin. The lengths of the branches are proportional to the total numbers of amino-acid replacements along them. Only convergent replacements are shown, denoted by a one-letter abbreviation of the resultant amino acid followed by the position number at which the replacement occurred.

- Adaptive replacements contribute to a better performance at low pH and confer protection against the proteolytic activities of the stomach.

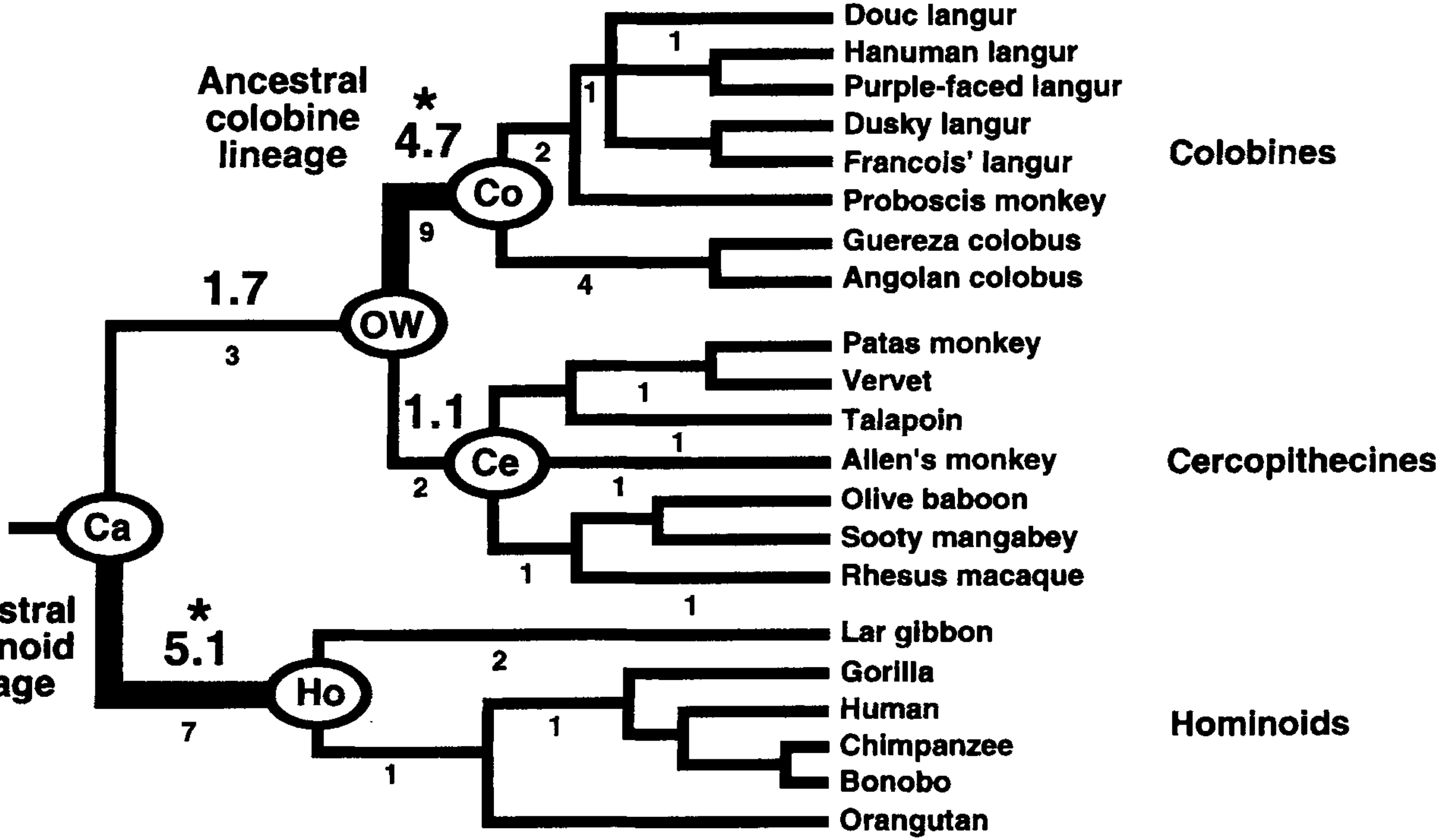
Messier Stewart



Text

Messier Stewart, Nature 9 January 1997 page 151

via ancestral reconstruction



Messier Stewart, Nature 9 January 1997 page 151