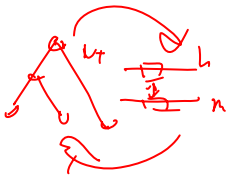
Computational Genomics

Molecular Evolution:
nucleotide substitution models



Eric Xing

Lecture 13, February 27, 2007

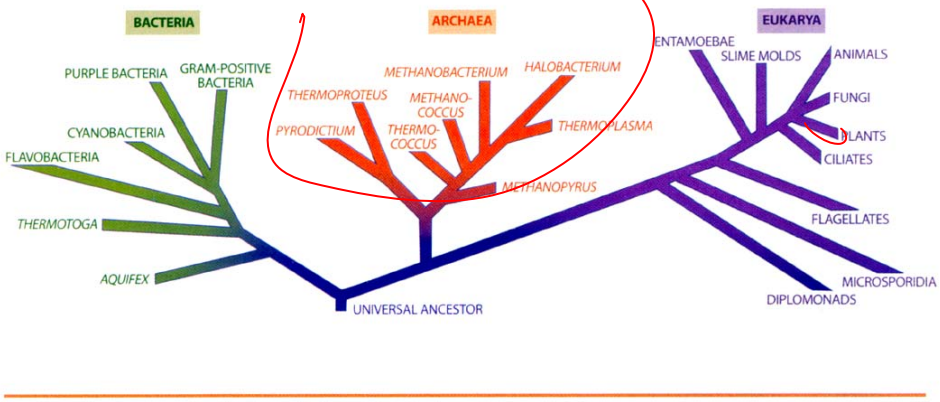
Reading: DTW book, Chap 12
DEKM book, Chap 8

Some important dates in history (billions of years ago)

• Origin of the universe	15 ±4
• Formation of the solar system	4.6
• First self-replicating system	3.5 ±0.5
• Prokaryotic-eukaryotic divergence	1.8 ±0.3
• Plant-animal divergence	1.0
• Invertebrate-vertebrate divergence	0.5
• Mammalian radiation beginning	0.1

(86 CSH Doolittle et al.)

The three kingdoms

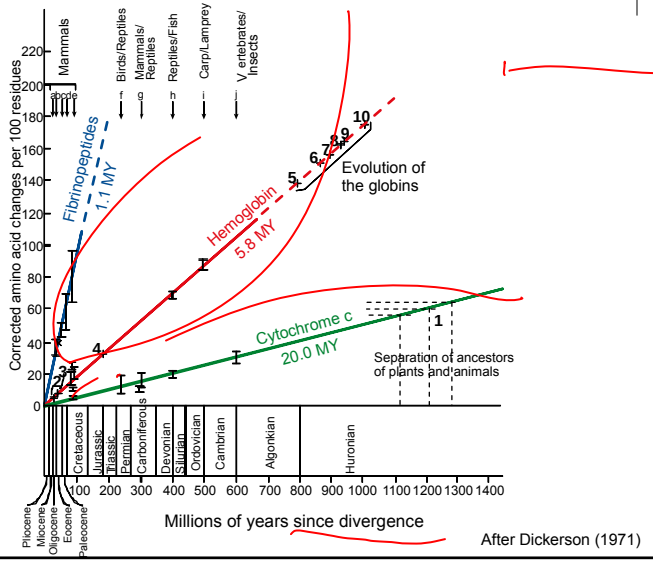


M. Madigan and B. Marrs, 1997

Two important early observations

- Different proteins evolve at different rates, and this seems more or less independent of the host organism, including its generation time.
- It is necessary to adjust the observed percent difference between two homologous proteins to get a distance more or less linearly related to the time since their common ancestor. (Later we offer a rational basis for doing this.)
- See nest slide …

Rates of macromolecular evolution



After Dickerson (1971)

How does sequence variation arise?

- **Mutation**:
  - (a) Inherent: DNA replication errors are not always corrected.
  - (b) External: exposure to chemicals and radiation.
- **Selection**: Deleterious mutations are removed quickly. Neutral and rarely, advantageous mutations, are tolerated and stick around.
- **Fixation**: It takes time for a new variant to be established (having a stable frequency) in a population.

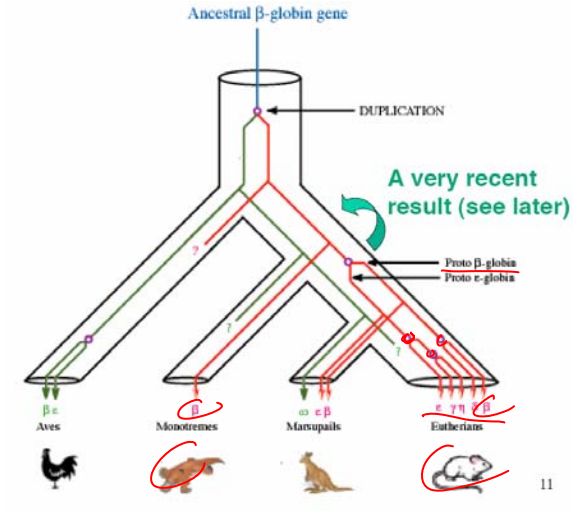
Modeling DNA base substitution

- Standard assumptions (sometimes weakened)
  - Site independence.
  - Site homogeneity.
  - Markovian: given current base, future substitutions independent of past.
  - Temporal homogeneity: stationary Markov chain.
- Strictly speaking, only applicable to regions undergoing little selection.

Some terminology

- In evolution, homology (here of proteins), means similarity due to common ancestry.
- A common mode of protein evolution is by duplication. Depending on the relations between duplication and speciation dates, we have two different types of homologous proteins. Loosely,
- **Orthologues**: the "same" gene in different organisms; common ancestry goes back to a speciation event.
- **Paralogues**: different genes in the same organism; common ancestry goes back to a gene duplication.
- Lateral gene transfer gives another form of homology.

Speciation vs. duplication



Beta-globins (orthologues)

```
                    10        20        30        40
BG-human    M V H L T P E E K S A V T A L W G K V N V D E V G G E A L G R L L V V Y P W T Q
BG-macaque  . . . . . N . . . T . . . . . . . . . . . . . . . . . . . . . . . . . . . .
BG-bovine   - - M . . A . . . A . . . . F . . . . K . . . . . . . . . . . . . . . . . .
BG-platypus . . . . S G G . . . . . N . . . . . . I N . L . . . . . . . . . . . . . . .
BG-chicken  . . . W . A . . . Q L I . G . . . . . . . A . C . A . . . A . . . I . . . . .
BG-shark    . . . W S E V . L H E I . T T . K S I D K H S L . A K . . A . M F I . . . . . T

                    50        60        70        80
BG-human    R F F E S F G D L S T P D A V M G N P K V K A H G K K V L G A F S D G L A H L D
BG-macaque  . . . . . S . . . . . . . . . . . . . . . . . . . . . . . . . . N . . .
BG-bovine   . . . . . A . . . . N . . . . . . . . . . . . D S . . N . M K . . .
BG-platypus . . . . A . . . . . S A G . . . . . . . . . . . A . . . T S . G . A . K N . .
BG-chicken  . . . A . . . N . . S . T . I L . . . M . R . . . . . . T S . G . A V K N . .
BG-shark    . Y . G N L K E F T A C S Y G - - - - - - - E . A . . . T . . L G V A V T . . G

                    90        100       110       120
BG-human    N L K G T F A T L S E L H C D K L H V D P E N F R L L G N V L V C V L A H H F G
BG-macaque  . . . . . Q . . . . . . . . . . . . . . . . . . K . . . . . . . . . . . . . .
BG-bovine   D . . . . . A . . . . . . . . . . . . . . . . . K . . . . . . . V . . . R N . .
BG-platypus D . . . . . K . . . . . . . . . . . . . . . . . N R . . . . . I V . . . R . . S
BG-chicken  . I . N . . S Q . . . . . . . . . . . . . . . . . . . . . D I . I I . . . A . . S
BG-shark    D V . S Q . T D . . K K . A E E . . . . V . S . K . . A K C F . V E . G I L L K

                    130       140
BG-human    K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H
BG-macaque  . . . . . Q . . . . . . . . . . . . . . . . . . . . . . .
BG-bovine   . . . . . V L . . D F . . . . . . . . . . . . . R . .
BG-platypus . D . S . E . . . . W . . L . S . . . H . . G . . . .
BG-chicken  . D . . . E C . . . W . . L . R V . . H . . . R . . .
BG-shark    D K . A . Q T . . I W E . Y F G V . V D . I S K E . .
```

. means same as reference sequence

\- means deletion

Beta-globins: uncorrected pairwise distances

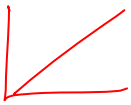
- DISTANCES between protein sequences (calculated over: 1 to 147)
  - Below diagonal: observed number of differences
  - Above diagonal: number of differences per 100 amino acids

	hum	mac	bov	pla	chi	sha
hum	----	**5**	**16**	**23**	**31**	**65**
mac	7	----	**17**	**23**	**30**	**62**
bov	23	24	----	**27**	**37**	**65**
pla	34	34	39	----	**29**	**64**
chi	45	44	52	42	----	**61**
sha	91	88	91	90	87	----

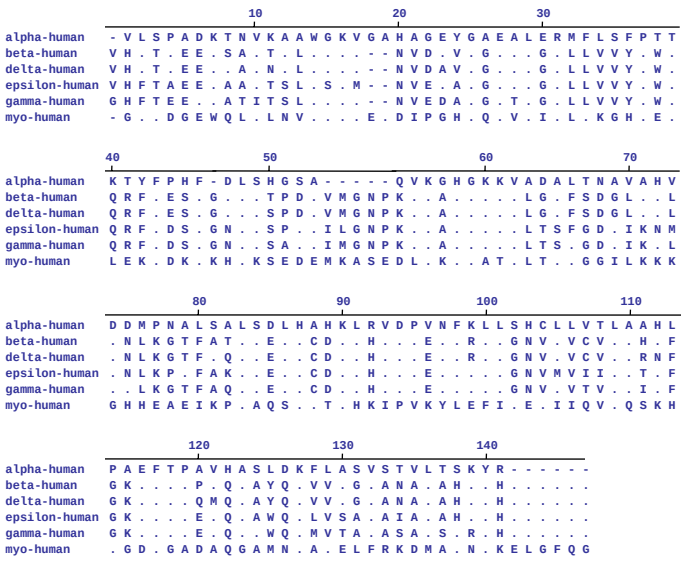
Beta-globins: corrected pairwise distances

- DISTANCES between protein sequences (calculated over: 1 to 147)
  - Below diagonal: observed number of differences
  - Above diagonal: number of differences per 100 amino acids
  - Correction method: Jukes-Cantor

	hum	mac	bov	pla	chi	sha
hum	----	**5**	**17**	**27**	**37**	**108**
mac	7	----	**18**	**27**	**36**	**102**
bov	23	24	----	**32**	**46**	**110**
pla	34	34	39	----	**34**	**106**
chi	45	44	52	42	----	**98**
sha	91	88	91	90	87	----



Human globins (paralogues)



Human globins: corrected pairwise distances

- DISTANCES between protein sequences (calculated over 1 to 141)
  - Below diagonal: observed number of differences
  - Above diagonal: estimated number of substitutions per 100 amino acids
  - Correction method: Jukes-Cantor

	alpha	beta	delta	epsil	gamma	myo
alpha	----	**281**	**281**	**281**	**313**	**208**
beta	82	----	**7**	**30**	**31**	**1000**
delta	82	10	----	**34**	**33**	**470**
epsil	89	35	39	----	**21**	**402**
gamma	85	39	42	29	----	**470**
myo	116	117	116	119	118	----

Correcting distances between DNA and protein sequences

- Why it is necessary to adjust observed percent differences to get a distance measure which scales linearly with time?
- This is because we can have multiple and back substitutions at a given position along a lineage.
- All of the correction methods (with names like Jukes-Cantor, 2-parameter Kimura, etc) are justified by simple probabilistic arguments involving Markov chains whose basis is worth mastering.
- The same molecular evolutionary models can be used in scoring sequence alignments.

Markov chain

- State space = {A,C,G,T}.

  p(i,j) = pr(next state S_j | current state S_i)

- Markov assumption:

  p(next state S_j | current state S_i & any configuration of states before this) = p(i,j)

  Only the *present* state, not previous states, affects the probs of moving to next states.

The multiplication rule

pr(state after next is S_k | current state is S_i)

= $\sum_j$ *pr*(state after next is S_k, next state is S_j | current state is S_i) [*addition rule*]

= $\sum_j$ *pr*(next state is S_j | current state is S_i) x *pr*(state after next is S_k | current state is S_i, next state is S_j) [*multiplication rule*]

= $\sum_j p_{i,j} \times p_{j,k}$ [*Markov assumption*]

= (i,k)-element of P^2, where $P=(p_{i,j})$.

More generally,

pr(state t steps from now is S_k | current state is S_i) = i,k element of P^t

Continuous-time version

- For any (s, t):
  - Let $p_{ij}(t) = pr(S_j \text{ at time } t+s \mid S_i \text{ at time } s)$ denote the stationary (time-homogeneous) transition probabilities.
- Let $P(t) = (p_{ij}(t))$ denote the matrix of $p_{ij}(t)$'s.
  - Then for any (t, u): $P(t+u) = P(t)\,P(u)$.
- It follows that $P(t) = \exp(Qt)$, where $Q = P'(0)$ (the derivative of $P(t)$ at $t = 0$).
- Q is called the infinitesimal matrix (transition rate matrix) of $P(t)$, and satisfies

  $P'(t) = QP(t) = P(t)Q.$
- Important approximation: when t is small,

  $P(t) \approx I + Qt.$

Interpretation of Q

- Roughly, q_{ij} is the **rate** of transitions of *i* to *j*, while $q_{ii} = -\Sigma_{j \neq i} q_{ij}$, so each row sum is *0* (Why?).
- Now we have the short-time approximation:

$$p_{i \neq j}(t+h) = q_{ij}h + o(h) \qquad p_{i=j}(t+h) = 1 + q_{ii}h + o(h)$$

where $p_{ij}(t+h)$ is the probability of transitioning from *i* at time *t* to *j* at time *t+h*

- Now consider the Chapman-Kolmogorov relation: (assuming we have a continuous-time Markov chain, and let $p_j(t) = pr(S_j \text{ at time } t)$):

$$p_j(t+h) = \sum_i pr(S_i \text{ at } t, S_j \text{ at } t+h)$$
$$= \sum_i pr(S_i \text{ at } t)\, pr(S_j \text{ at } t+h \mid S_i \text{ at } t)$$
$$= p_j(t) \times (1 + q_{jj}h) + \sum_{i \neq j} p_i(t) \times h q_{ij}$$

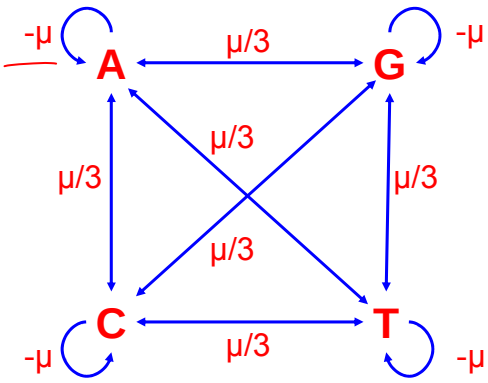
i.e., $h^{-1}\left(p_j(t+h) - p_j(t)\right) = p_j(t) q_{jj} + \sum_{i \neq j} p_i(t) q_{ij}$, which becomes: $p' = Qp$ as $h \to 0$.

Probabilistic models for DNA changes

Orc:	ACAGTGACGCCCCAAACGT
Elf:	ACAGTGACGCTACAAACGT
Dwarf:	CCTGTGACGTAACAAACGA
Hobbit:	CCTGTGACGTAGCAAACGA
Human:	CCTGTGACGTAGCAAACGA

The Jukes-Cantor model (1969)

- **Substitution rate:**



the simplest symmetrical model for DNA evolution

Transition probabilities under the Jukes-Cantor model

- IID assumption:
  - All sites change independently
  - All sites have the same stochastic process working at them
- Equiprobablity assumption:
  - Make up a fictional kind of event, such that when it happens the site changes to one of the 4 bases chosen at random equiprobably
- Equilibrium condition:
  - No matter how many of these fictional events occur, provided it is not zero, the chance of ending up at a particular base is 1/4 .
- Solving differentially equation system $P' = QP$

Transition probabilities under the Jukes-Cantor model (cont.)

- **Prob transition matrix:**

$$P(t) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{bmatrix} \end{array}$$

Where we can derive:

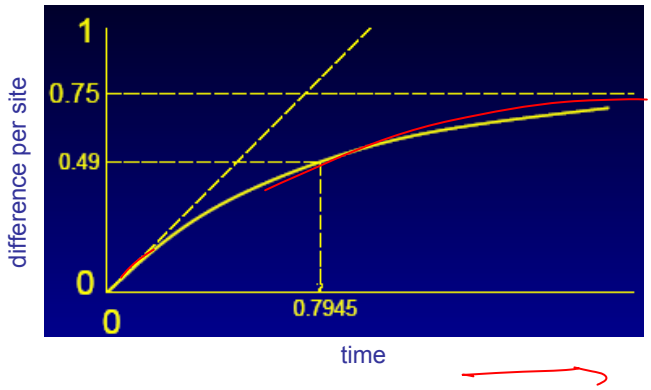
$$r(t) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}\mu t}\right)$$

$$s(t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right)$$

Homework!

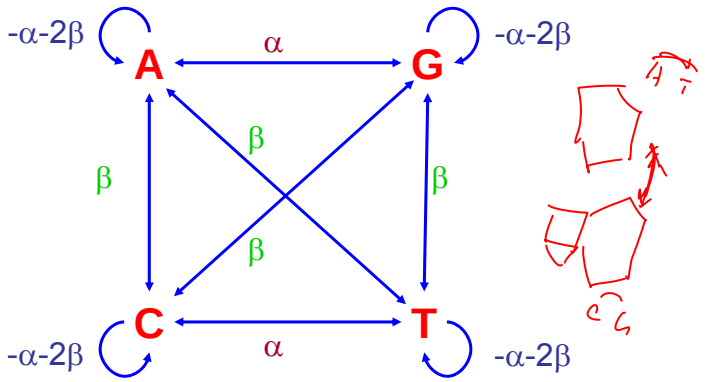
Jukes-Cantor (cont.)

- **Fraction of sites differences**



Kimura's K2P model (1980)

- **Substitution rate:**



- which allows for different rates of transition and transversions.
- Transitions (rate α) are much more likely than transversions (rate β).

Kimura (cont.)

- **Prob transition matrix:**

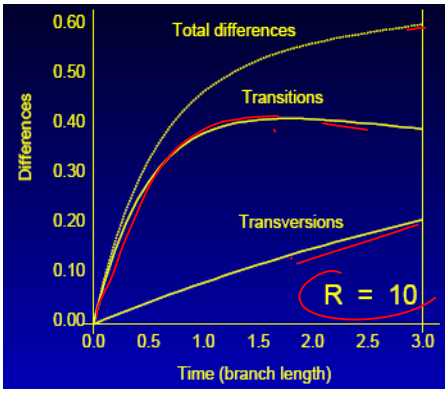
$$P(t) = \begin{pmatrix} r(t) & s(t) & u(t) & s(t) \\ s(t) & r(t) & s(t) & u(t) \\ u(t) & s(t) & r(t) & s(t) \\ s(t) & u(t) & s(t) & r(t) \end{pmatrix}$$

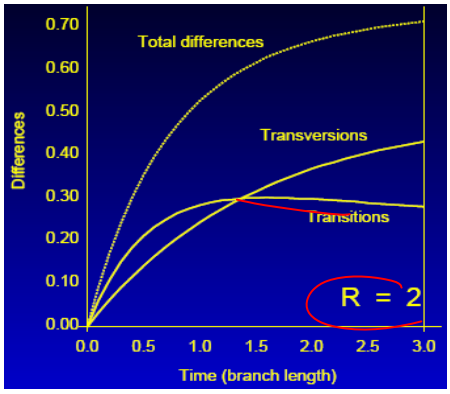
Where
$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\beta t})$$
$$u(t) = \tfrac{1}{4}\,(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t})$$
$$r(t) = 1 - 2s(t) - u(t)$$

- By proper choice of α and β one can achieve the overall rate of change and Ts=Tn ratio R you want (*warning: terminological tangle*).

Kimura (cont.)

- Transitions, transversions expected under different R:





Other commonly used models

- Two models that specify the equilibrium base frequencies (you provide the frequencies A; C; G; T and they are set up to have an equilibrium which achieves them), and also let you control the transition/transversion ratio:
- The **Hasegawa-Kishino-Yano (1985) model**:

to : from :	A	G	C	T
A	$-$	$\alpha\pi_G + \beta\pi_G$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\pi_A$	$-$	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	$-$	$\alpha\pi_T + \beta\pi_T$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\pi_C$	$-$

Other commonly used models

- The **F84 model** (Felsenstein)

to : from :	A	G	C	T
A	–	$\alpha\pi_G + \beta\frac{\pi_G}{\pi_R}$	$\alpha\pi_C$	$\alpha\pi_T$
G	$\alpha\pi_A + \beta\frac{\pi_A}{\pi_R}$	–	$\alpha\pi_C$	$\alpha\pi_T$
C	$\alpha\pi_A$	$\alpha\pi_G$	–	$\alpha\pi_T + \frac{\beta\pi_T}{\pi_Y}$
T	$\alpha\pi_A$	$\alpha\pi_G$	$\alpha\pi_C + \beta\frac{\pi_C}{\pi_Y}$	–

  - where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ (The equilibrium frequencies of purines and pyrimidines)

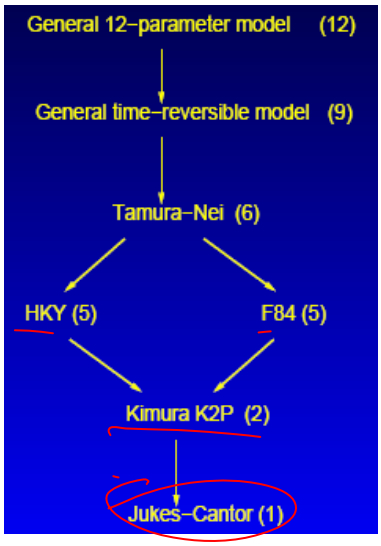
The general time-reversible model

- It maintains "detailed balance" so that the probability of starting at (say) A and ending at (say) T in evolution is the same as the probability of starting at T and ending at A:

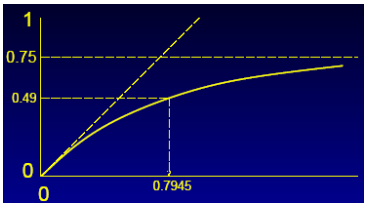
	A	C	G	T
A	–	$\alpha\pi_C$	$\beta\pi_G$	$\gamma\pi_T$
C	$\alpha\pi_A$	–	$\delta\pi_G$	$\varepsilon\pi_T$
G	$\beta\pi_A$	$\delta\pi_C$	–	$\nu\pi_T$
T	$\gamma\pi_A$	$\varepsilon\pi_C$	$\nu\pi_G$	–

- And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others).
- (Neither of these has formulas for the transition probabilities, but those can be done numerically.)

Relation between models



Adjusting evolutionary distance using base-substitution model



The Jukes-Cantor model

Common ancestor of human and orang

t time unit

Human (now)

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

$$P = \begin{bmatrix} r & s & s & s \\ s & r & s & s \\ s & s & r & s \\ s & s & s & r \end{bmatrix}$$

Consider e.g. the 2nd position in a-globin2 Alu1.

$r = (1+3e^{-4\alpha t})/4, \quad s = (1 - e^{-4\alpha t})/4.$

Definition of PAM

- Let $P(t) = \exp(Qt)$. Then the A,G element of $P(t)$ is

  $pr(G \text{ now} \mid A \text{ then}) = (1 - e^{-4\alpha t})/4.$

  - Same for all pairs of different nucleotides.
  - Overall rate of change $k = 3\alpha t$.

- ***PAM* = accepted point mutation**
  - When $k = .01$ described as *1 PAM*
  - Put $t = .01/3\alpha = 1/300\alpha$. Then the resulting $P = P(1/300\alpha)$ is called the *PAM(1)* matrix.

- Why use *PAMs?*

Evolutionary time, PAM

- Since sequences evolve at different rates, it is convenient to rescale time so that *1 PAM* of evolutionary time corresponds to *1%* expected substitutions.
- For Jukes-Cantor, $k = 3\alpha t$ is the expected number of substitutions in *[0,t],* so is a distance. (Show this.)
  - Set $3\alpha t = 1/100$, or $t = 1/300\alpha$, so $1\ PAM = 1/300\alpha$ years.

Distance adjustment

- For a pair of sequences, $k = 3\alpha t$ is the desired metric, but not observable. Instead, *pr(different)* is observed. So we use a model to convert *pr(different)* to *k*.
- This is completely analogous to the conversion of

  $\theta = pr$(recombination)

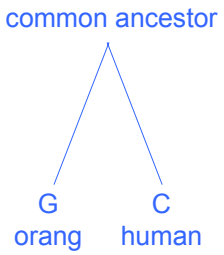
  to genetic (map) distance (= expected number of crossovers) using the Haldane map function

  $\theta = 1/2 \times (1 - e^{-2d})$,

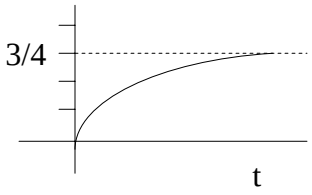
  assuming the no-interference (Poisson) model.

Towards Jukes-Cantor adjustment

- E.g., 2nd position in a-globin Alu 1
- Assume that the common ancestor has A, G, C or T with probability 1/4.



- Then the chance of the nt differing
  $p_{\neq} = 3/4 \times (1 - e^{-8\alpha t})$
  $= 3/4 \times (1 - e^{-4k/3})$, since $k = 2 \times 3\alpha t$



Jukes-Cantor adjustment

- If we suppose all nucleotide positions behave identically and independently, and $n_{\neq}$ differ out of n, we can invert this, obtaining

$$\hat{k} = -\frac{3}{4} \times \log\left(1 - \frac{4}{3} n_{\neq} / n\right)$$

- This is the corrected or adjusted fraction of differences (under this simple model). × 100 to get PAMs
- The analogous simple model for amino acid sequences has

$$\hat{k} = -\frac{19}{20} \times \log\left(1 - \frac{20}{19} n_{\neq} / n\right)$$

× 100 for PAM.

Illustration

1. Human and bovine beta-globins are aligned with no deletions at 145 out of 147 sites. They differ at 23 of these sites. Thus $n_{\neq}/n = 23/145$, and the corrected distance using the Jukes-Cantor formula is (natural logs)

   $-19/20 \times \log(1 - 20/19 \times 23/145) = 17.3 \times 10^{-2}$.

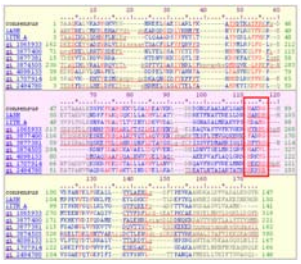
2. The human and gorilla sequences are aligned without gaps across all 300 bp, and differ at 14 sites. Thus $n_{\neq}/n = 14/300$, and the corrected distance using the Jukes-Cantor formula is

   $-3/4 \times \log(1 - 4/3 \times 14/300) = 4.8 \times 10^{-2}$.

Correspondence between observed a.a. differences and the evolutionary distance (Dayhoff et al., 1978)

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

Scoring matrices for alignment



How scoring matrices work

```
134 LQQGELDLVMTSDILPRSELHYSPMFDFEVRLVLAPDHPLASKTQITPEDLASETLLI
    |     |||          |       |         ||||||    |    || ||
137 LDSNSVDLVLMGVPPRNVEVEAEAFMDNPLVVIAPPDHPLAGERAISLARLAEETFVM
```

D:D = +6

D:R = -2

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

From Henikoff 1996

Statistical motivation for alignment scores

Alignment: AGCTGATCA . . .
AACCGGTTA . . .

Hypotheses: H = **homologous** (indep. sites, Jukes-Cantor)
R = **random** (indep. sites, equal freq.)

$$\Pr(data \mid H) = \Pr(AA \mid H)\Pr(GA \mid H)\Pr(CC \mid H)\ldots$$
$$= (1-p)^a p^d, \text{where } a = \#\text{agreements}, d = \#\text{disagreements}, p = \frac{3}{4}(1-e^{-8\alpha t}).$$

$$\Pr(data \mid R) = \Pr(AA \mid R)\Pr(GA \mid R)\Pr(CC \mid R)\ldots$$
$$= (\frac{1}{4})^a (\frac{3}{4})^d$$

$$\Rightarrow \quad \log\{\frac{\Pr(data \mid H)}{\Pr(data \mid R)}\} = a \log\frac{1-p}{1/4} + d \log\frac{p}{3/4} = a \times \sigma + d \times (-\mu).$$

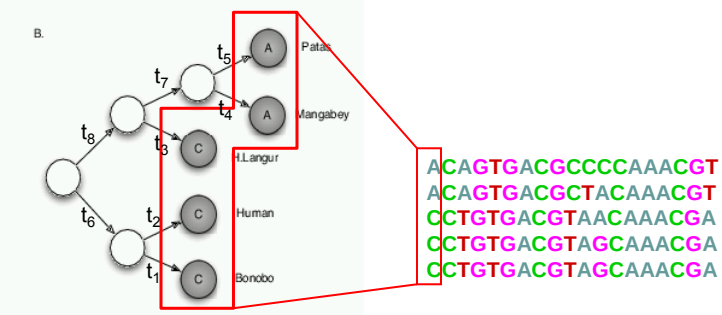
- Since $p<3/4$, $\sigma = \log((1-p)/(1/4))>0$, while $-\mu = \log(p/(3/4))<0$.
- Thus the alignment score = $a \times \sigma + d \times (-\mu)$, where the match score $\sigma > 0$, and the mismatch penalty is $-\mu < 0$.

Large and small evolutionary distances

- Recall that
  - $p = (3/4)(1-e^{-8\alpha t})$,
  - $\sigma = \log((1-p)/(1/4))$,
  - $-\mu = \log(p/(3/4))$.
- Now note that if $\alpha t \approx 0$,
  - then $p \approx 6\alpha t$, and $1-p \approx 1$, and so $\sigma \approx \log 4$, while $-\mu \approx \log 8\alpha t$ is large and negative.
  - That is, we see a big difference in the two values of σ and μ for small distances.
- Conversely, if αt is large,
  - $p = (3/4)(1-\varepsilon)$, hence $p/(3/4) = 1-\varepsilon$, giving $\mu = -\log(1-\varepsilon) \approx \varepsilon$, while $1-p = (1+3\varepsilon)/4$, $(1-p)/(1/4) = 1+3\varepsilon$, and so $\sigma = \log(1+3\varepsilon) \approx 3\varepsilon$.
  - Thus the scores are about 3 (for a match) to 1 (for a mismatch) for large distances. This makes sense, as mismatches will on average be about 3 times more frequent than matches.
- the matrix which performs best will be the matrix that reflects the evolutionary separation of the sequences being aligned.

What about multiple alignment

- Phylogenetic methods: a tree, with branch lengths, and the data at a single site.



- See next lecture for how to compute likelihood under this hypothesis

Acknowledgments

- **Terry Speed**: for some of the slides modified from his lectures at UC Berkeley
- **Phil Green** and **Joe Felsenstein**: for some of the slides modified from his lectures at Univ. of Washington