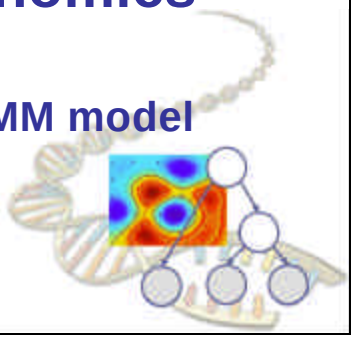
Molecular Evolution and Comparative Genomics

--- the phylogenetic HMM model



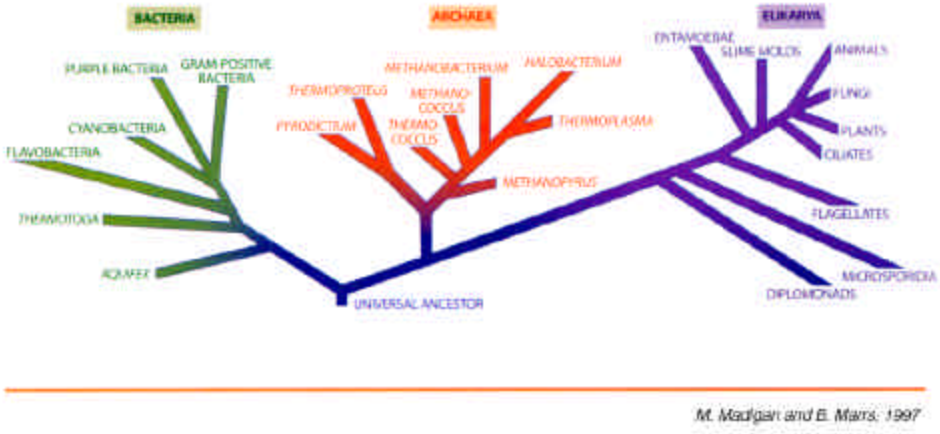
10-810, CMB lecture 5---Eric Xing

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



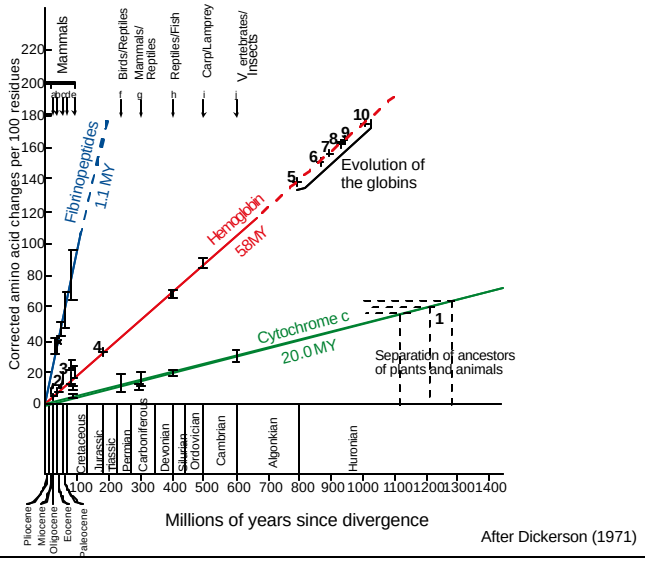
Two important early observations

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.)

An striking early version of these observations is next.

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)

1. Site independence.
2. Site homogeneity.
3. Markovian: given current base, future substitutions independent of past.
4. 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.

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: estimated number of substitutions 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)

```
                          10        20        30
alpha-human     - V L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T
beta-human      V H . T . E E . S A . T . L . . . . . N V D . V . G . . . G . L L V V Y . W .
delta-human     V H . T . E E . . A . N . L . . . . . N V D A V . G . . . G . L L V V Y . W .
epsilon-human   V H F T A E E . A A . T S L . S . M - - N V E . A . G . . . G . L L V V Y . W .
gamma-human     G H F T E E . . A T I T S L . . . . . N V E D A . G . T . G . L L V V Y . W .
myo-human       - G . . D G E W Q L . L N V . . . . E . D I P G H . Q . V . I . L . K G H . E .

                40        50                  60        70
alpha-human     K T Y F P H F - D L S H G S A - - - - - Q V K G H G K K V A D A L T N A V A H V
beta-human      Q R F . E S . G . . . T P D . V M G N P K . . A . . . . . L G . F S D G L . . L
delta-human     Q R F . E S . G . . . S P D . V M G N P K . . A . . . . . L G . F S D G L . . L
epsilon-human   Q R F . D S . G N . . S P . . I L G N P K . . A . . . . . L T S F G D . I K N M
gamma-human     Q R F . D S . G N . . S A . . I M G N P K . . A . . . . . L T S . G D . I K . L
myo-human       L E K . D K . K H . K S E D E M K A S E D L . K . . A T . L T . . G G I L K K K

                      80        90        100       110
alpha-human     D D M P N A L S A L S D L H A H K L R V D P V N F K L L S H C L L V T L A A H L
beta-human      . N L K G T F A T . . E . . C D . . H . . . E . . R . . G N V . V C V . . H . F
delta-human     . N L K G T F . Q . . E . . C D . . H . . . E . . R . . G N V . V C V . . R N F
epsilon-human   . N L K P . F A K . . E . . C D . . H . . . E . . . . . G N V M V I I . . T . F
gamma-human     . . L K G T F A Q . . E . . C D . . H . . . E . . . . . G N V . V T V . . I . F
myo-human       G H H E A E I K P . A Q S . . T . H K I P V K Y L E F I . E . I I Q V . Q S K H

                      120       130       140
alpha-human     P A E F T P A V H A S L D K F L A S V S T V L T S K Y R - - - - - -
beta-human      G K . . . . P . Q . A Y Q . V V . G . A N A . A H . . H . . . . . .
delta-human     G K . . . . Q M Q . A Y Q . V V . G . A N A . A H . . H . . . . . .
epsilon-human   G K . . . . E . Q . A W Q . L V S A . A I A . A H . . H . . . . . .
gamma-human     G K . . . . E . Q . . W Q . M V T A . A S A . S . R . H . . . . . .
myo-human       . G D . G A D A Q G A M N . A . E L F R K D M A . N . K E L G F Q G
```

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)

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

= ? $_j$ *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*]

= ? $_j$ $p_{i,j}$ *x* $p_{j,k}$ [*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

$$P'(t) = QP(t) = P(t)Q.$$

Interpretation of Q

Roughly, q_{ij} is the rate of transitions of i to j, while $q_{ii} = -\sum_{j \neq i} q_{ij}$, so each row sum is 0.

Now we have the short-time approximation:

$$p_{i \neq j}(t+h) = q_{ij}\,h + o(h), \quad \text{and} \quad p_{ii}(t+h) = 1 + q_{ii}\,h + o(h),$$

Now consider the Chapman-Kolmogorov relation:

(assuming we have a continuous-time Markov chain, and let $p_j(t) = pr(S_j \text{ at time } t)$)

$$\begin{aligned} p_j(t+h) &= \sum_i pr(S_i \text{ at } t,\ S_j \text{ at } t+h) \\ &= \sum_i pr(S_i \text{ at } t)\,pr(S_j \text{ at } t+h \mid S_i \text{ at } t) \\ &= p_j(t)x(1+hq_{jj}) + \sum_{i \neq j} p_i(t)x\,hq_{ij} \end{aligned}$$

$i.e.,\ h^{-1}[p_j(t+h) - p_j(t)] = p_j(t)q_{jj} + \sum_{i \neq j} p_i(t)q_{ij}$, which becomes $P' = QP$ as $h \to 0$.

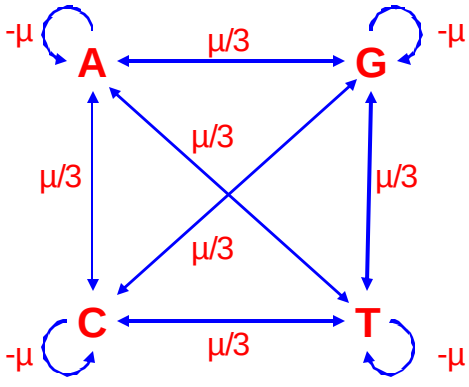
Important approximation: when t is small, $P(t) \approx I + Qt$.

Probabilistic models for DNA changes

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

The Jukes-Cantor model (1969)

Substitution rate:



the simplest sysmmetrical model for DNA evolution

Jukes-Cantor (cont.)

Prob transition matrix:

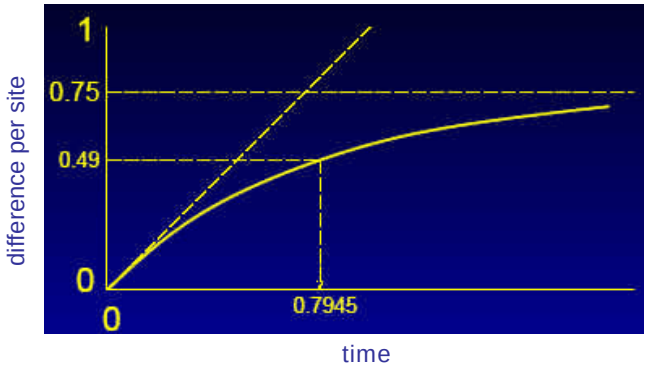
$$S(t) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix} \end{array}$$

Where we can derive:

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

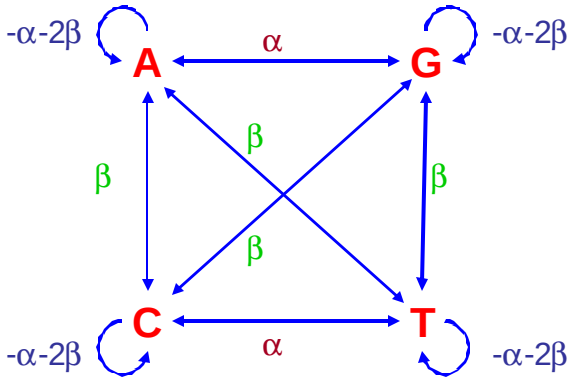
Jukes-Cantor (cont.)

Fraction of sites differences



Kimura's K2P model (1980)

Substitution rate:



which allows for different rates of transition and transversions.
Transitions (rate a) are much more likely than transversions (rate b).

Kimura (cont.)

Prob transition matrix:

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

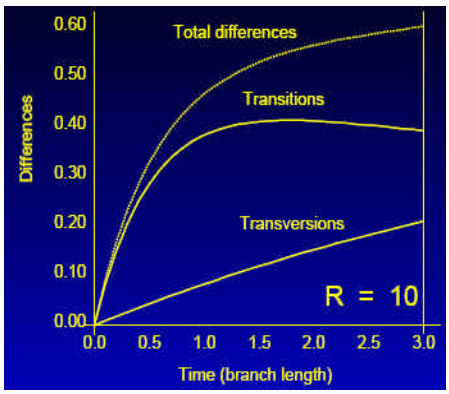
Where

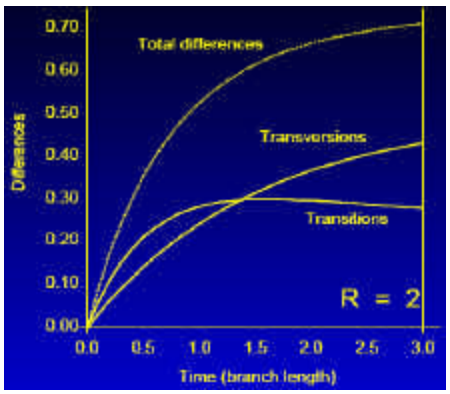
$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\beta t})$$
$$u(t) = \tfrac{1}{4}\,(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t})$$
$$r(t) = 1 - 2s(t) - u(t)$$

By proper choice of and 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:





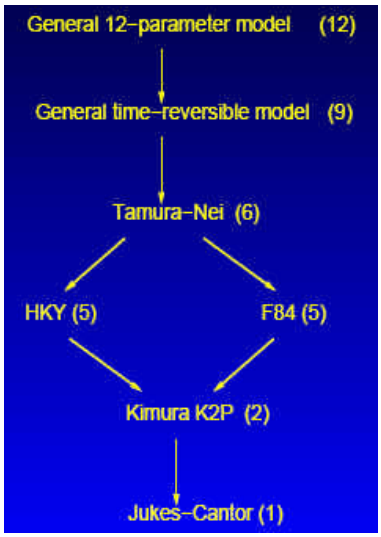
The general time-reversible model

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	?	ap_C	βp_G	$?p_T$
C	ap_A	?	dp_G	εp_T
G	βp_A	dp_C	?	$?p_T$
T	$?p_A$	εp_C	$?p_G$	?

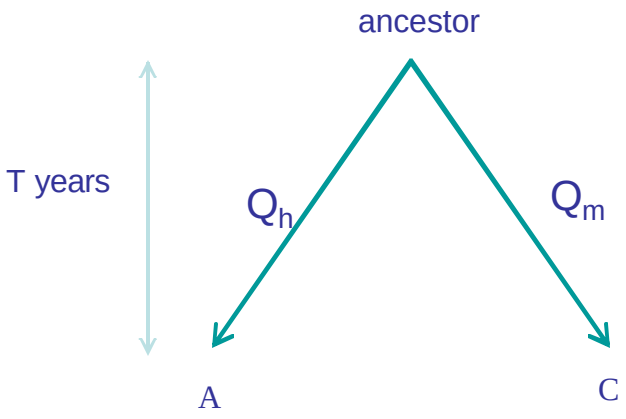
And there is of course the **general 12-parameter model** which has arbitrary rates for each of the 12 possible changes (from each of the 4 nucleotides to each of the 3 others). (Neither of these has formulas for the transition probabilities, but those can be done numerically.)

Relation between models



Phylogenetic trees

A pair of homologous bases



Typically, the ancestor is unknown.

More assumptions

- $Q_h = s_h Q$ and $Q_m = s_m Q$, for some positive s_h, s_m, and a rate matrix Q.
- The ancestor is sampled from the stationary distribution p of Q.
- Q is **reversible**: for $a, b, t \geq 0$

  $p(a)P(t,a,b) = P(t,b,a)p(b)$,

  (detailed balance).

The stationary distribution

A probability distribution p on $\{A,C,G,T\}$ is a **stationary distribution** of the Markov chain with transition probability matrix $P = P(i,j)$, if for all j,

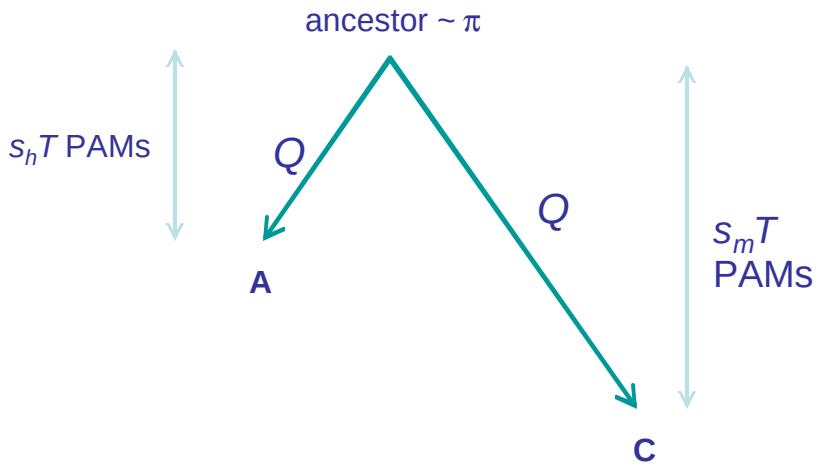
$$\Sigma_i \, p(i) \, P(i,j) = p(j).$$

Exercise. Given any initial distribution, the distribution at time t of a chain with transition matrix P converges to p as $t \to \infty$. Thus, p is also called an **equilibrium** distribution.

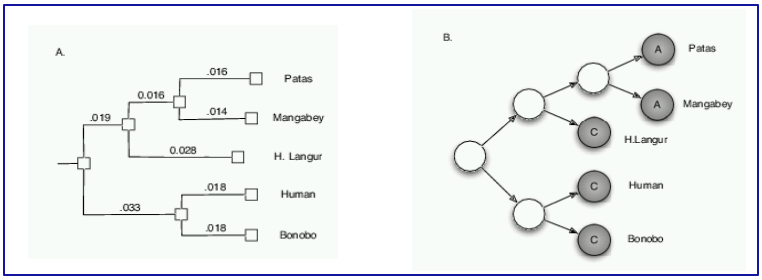
Exercise. For the Jukes-Cantor and Kimura models, the uniform distribution is stationary. (Hint: diagonalize their infinitesimal rate matrices.)

We often assume that the ancestor sequence is i.i.d π.

New picture



Phylogeny



- The shaded nodes represent the observed nucleotides at a given site for a set of organisms
- The unshaded nodes represent putative ancestral nucleotides
- Transitions between nodes capture the dynamic of evolution

Phylogeny methods

Basic principles:

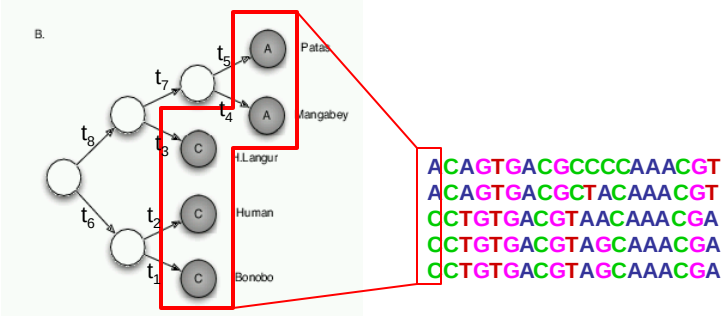
- Degree of sequence difference is proportional to length of independent sequence evolution
- Only use positions where alignment is pretty certain – avoid areas with (too many) gaps

Major methods:

- Parsimony phylogeny methods
- Likelihood methods

Likelihood methods

A tree, with branch lengths, and the data at a single site.



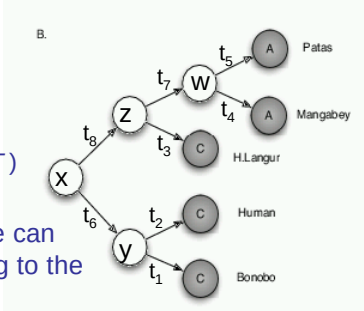
Since the sites evolve independently on the same tree,

$$L = P(D \mid T) = \prod_{i=1}^{m} P(D^{(i)} \mid T)$$

Likelihood at one site on a tree

We can compute this by summing over all assignments of states x, y, z and w to the interior nodes:

$$P(D^{(i)} \mid T) = \sum_{x}\sum_{y}\sum_{z}\sum_{w} P(A,A,C,C,C,x,y,z,w \mid T)$$



Due to the Markov property of the tree, we can factorize the complete likelihood according to the tree topology:

$$P(A,A,C,C,C,x,y,z,w \mid T) =$$
$$P(x)\ P(y \mid x,t_6)\ P(A \mid y,t_1)\,P(C \mid y,t_2)$$
$$P(z \mid x,t_8)\ P(C \mid y,t_3)$$
$$P(w \mid z,t_7)\,P(C \mid y,t_4)\,P(C \mid y,t_5)$$

Summing this up, there are 256 terms in this case!

Getting a recursive algorithm

when we move the summation signs
as far right as possible:

$$P(D^{(i)} \mid T) = \sum_x \sum_y \sum_z \sum_w P(A,A,C,C,C,x,y,z,w \mid T) =$$

$$\sum_x P(x)$$
$$\left(\sum_y P(y \mid x, t_6) \quad P(A \mid y, t_1) P(C \mid y, t_2) \right)$$
$$\left(\sum_z P(z \mid x, t_8) \quad P(C \mid z, t_3) \right.$$
$$\left. \left(\sum_w P(w \mid z, t_7) P(C \mid w, t_4) P(C \mid w, t_5) \right) \right)$$

Felsenstein's Pruning Algorithm

To calculate $P(x_1, x_2, \ldots, x_N \mid T, t)$

Initialization:

Set $k = 2N - 1$

Recursion: Compute $P(L_k \mid a)$ for all $a \in \Sigma$

If k is a leaf node:

Set $P(L_k \mid a) = 1(a = x_k)$

If k is not a leaf node:

1. Compute $P(L_i \mid b)$, $P(L_j \mid b)$ for all b, for daughter nodes i, j

2. Set $P(L_k \mid a) = \sum_{b,\, c} P(b \mid a, t_i) P(L_i \mid b)\ P(c \mid a, t_j)\ P(L_j \mid c)$

Termination:

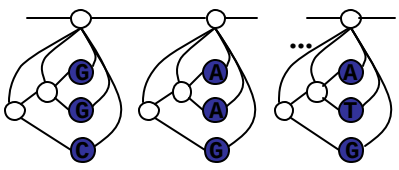
Likelihood at this column = $P(x_1, x_2, \ldots, x_N \mid T, t) = \sum_a P(L_{2N-1} \mid a) P(a)$

Finding the ML tree

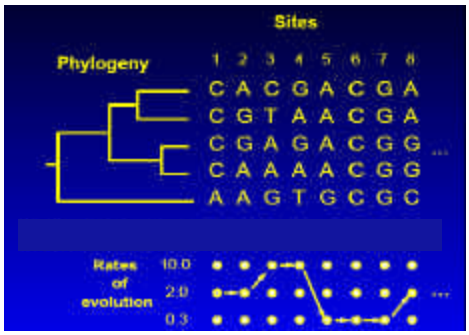
So far I have just talked about the computation of the likelihood for one tree with branch lengths known.

To find a ML tree, we must search the space of tree topologies, and for each one examined, we need to optimize the branch lengths to maximize the likelihood.

Phylogenetic HMM

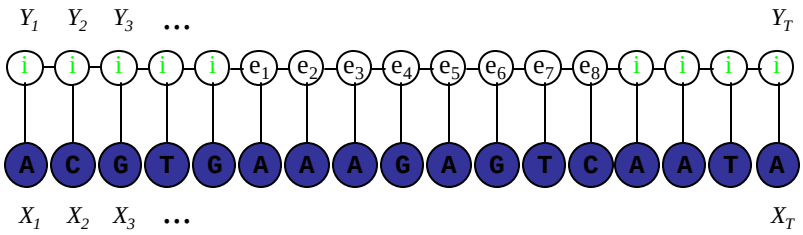


Modeling rate variation among sites



- There are a finite number of rates (denote rate i as r_i).
- There are probabilities p_i of a site having rate i.
- A process not visible to us ("hidden") assigns rates to sites.
- The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting approproately by their probabilities of occurrence.

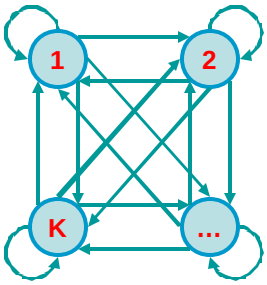
Rocall the HMM



- The shaded nodes represent the observed nucleotides at particular sites of an organism's genome
- For discrete Y_i, widely used in computational biology to represent segments of sequences
  - gene finders and motif finders
  - profile models of protein domains
  - models of secondary structure

Definition (of HMM)

Definition: A hidden Markov model (HMM)



- Observation alphabet $\Sigma = \{ b_1, b_2, \ldots, b_M \}$
- Set of hidden states $Q = \{ 1, \ldots, K \}$
- Transition probabilities between any two states

  $a_{ij} = P(y=j|y=i)$
  $a_{i1} + \ldots + a_{iK} = 1$, for all states $i = 1 \ldots K$

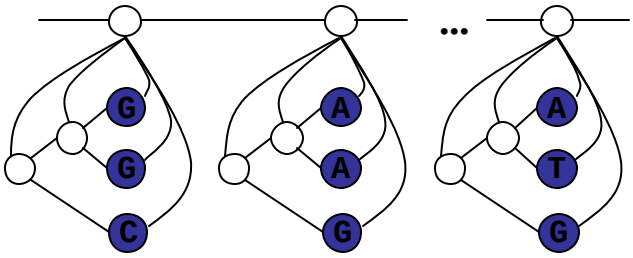
- Start probabilities a_{0i}

  $a_{01} + \ldots + a_{0K} = 1$

- Emission probabilities associated with each state

  $e_{ib} = P(x_i = b \mid y_i = k)$
  $e_{i,b1} + \ldots + e_{i,bM} = 1$, for all states $i = 1 \ldots K$

Hidden Markov Phylogeny



- This yields a gene finder that exploits evolutionary constraints
- Based on sequence data from 12-15 primate species, McAuliffe et al (2003) obtained sensitivity of 100%, with a specificity of 89%.
  - Genscan (state-of-the-art gene finder) yield a sensitivity of 45%, with a specificity of 34%.

The Forward Algorithm

We can compute $f_k(t)$ for all k, t, using dynamic programming!

Initialization:

$f_0(0) = 1$
$f_k(0) = 0$, for all $k > 0$

Iteration:

$f_l(t) = e_l(\mathbf{X}_t) \Sigma_k f_k(t-1) a_{kl}$ (a_{0k} is a vector of initial probability)

Termination:

$P(x) = \Sigma_k f_k(T)$

The Backward Algorithm

We can compute $b_k(t)$ for all k, t, using dynamic programming

Initialization:

$b_k(T) = 1$, for all k

Iteration:

$b_k(t) = \Sigma_l e_l(\mathbf{X}_{t+1}) a_{kl} b_l(t+1)$

Termination:

$P(x) = \Sigma_l a_{0l} e_l(x_1) b_l(1)$

Open questions (philosophical)

Observation:

- Finding a good phylogeny will help in finding the genes.
- Finding the genes will help to find biologically meaningful phylogenetic trees

Which came first, the chicken or the egg?

Open questions (technical)

- How to learn a phylogeny (topology and transition prob.)?
- Should different site use the same phylogeny? Function-specific phylogeny?
- Other evolutionary events: duplication, rearrangement, lateral transfer, etc.