Computational Genomics

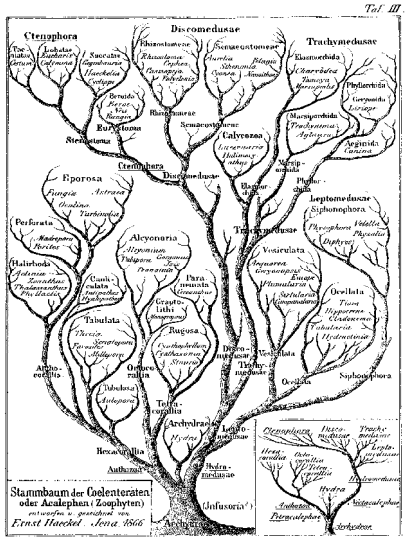
Molecular Evolution: Phylogenetic trees

Eric Xing

Lecture 14, March 6, 2007

Reading: DTW book, Chap 12
DEKM book, Chap 7, 8

Phylogeny



In 1866, Ernst Haeckel coined the word "phylogeny" and presented phylogenetic trees for most known groups of living organisms.

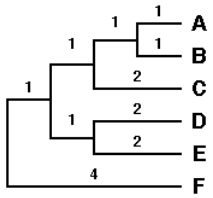
Ernst Haeckel (1834-1919)

Phylogenetic Inference

- Given a multiple alignment, how do we construct the tree?

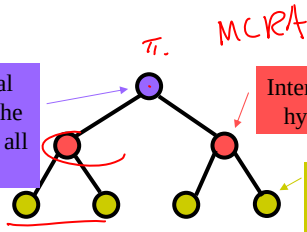
```
A  -  GCTTGTCCGTTACGAT
B  -  ACTTGTCTGTTACGAT
C  -  ACTTGTCCGAAACGAT
D  -  ACTTGACCGTTTCCTT
E  -  AGATGACCGTTTCGAT
F  -  ACTACACCCTTATGAG
```

?



What Is a Tree?

- A tree is a mathematical structure which represents a model of an actual evolutionary history of a group of sequences or organisms.
  - In other words, it is an evolutionary hypothesis.
- A tree consists of nodes connected by branches.



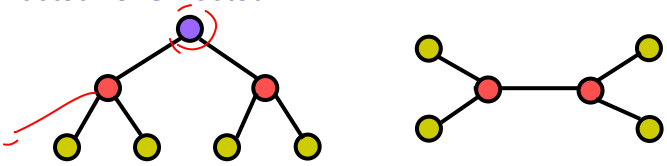
One unique internal node is the root of the tree: the ancestor of all the sequences.

Internal nodes represent hypothetical ancestors

Terminal nodes represent sequences or organisms for which we have data. **Each is typically called a "Operational Taxonomical Unit" or OTU.**

Types of Trees

- Rooted vs. Unrooted



		Branches	Nodes
Rooted	Interior	M – 2	M – 1
	Total	2M – 2	2M – 1
Unrooted	Interior	M – 3	M – 2
	Total	2M – 3	2M – 2

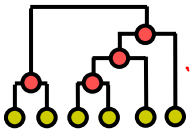
- M is the number of OTU's

The number of rooted and unrooted trees:

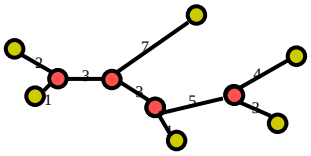
Number of OTU's	Possible Number of	
	Rooted trees	**Unrooted trees**
2	**1**	**1**
3	**3**	**1**
4	**15**	**3**
5	**105**	**15**
6	**945**	**105**
7	**10395**	**945**
8	**135135**	**10395**
9	**2027025**	**135135**
10	**34459425**	**2027025**

More different kinds of trees

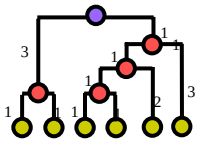
- Different kinds of trees can be used to depict different aspects of evolutionary history



1. Cladogram:
   simply shows relative order of common ancestry



2. Additive trees:
   a cladogram with branch lengths,
   also called phylograms and metric trees



3. Ultrametric trees:
   (dendograms) special kind of additive tree in which the tips of the trees are all equidistant from the root

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:

Clustering methods	UPGMA	Neighbor-joining
	WPGMA	
	Single linkage	
	Complete linkage	
Objective criterion-based methods	Least-squares distance	
	Maximum parsimony	
	Minimum evolution	
	Maximum likelihood	

UPGMA

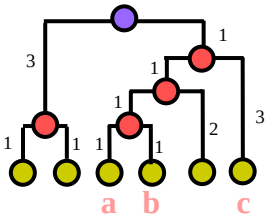
- Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA)
- First, construct a distance matrix:

A - GCTTGTCCGTTACGAT
B – ACTTGTCTGTTACGAT
C – ACTTGTCCGAAACGAT
D - ACTTGACCGTTTCCTT
E – AGATGACCGTTTCGAT
F - ACTACACCCTTATGAG

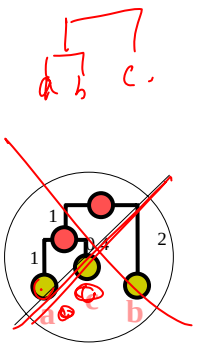
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html

Ultrametric Trees



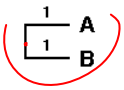
- Metric distances (for additive trees) must obey 4 rules:
  - Non-negativity: d(a,b) ≥ 0
  - Distinctness: d(a,b) = 0 if and only if a = b
  - Symmetry: d(a,b) = d(b,a)
  - Triangle Inequality: d(a,c) ≤ d(a,b) + d(b,c)
- Ultrametric must obey one additional rule:
  - Three point condition: d(a,b) ≤ max(d(a,c), d(b,c))



UPGMA

- First round

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8

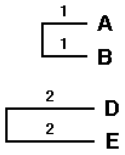
	B,A	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Choose the most similar pair, cluster them together and calculate the new distance matrix.

UPGMA

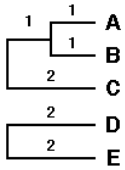
- Second round

	B,A	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



- Third round

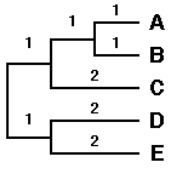
	B,A	C	,D E
C	4		
,D E	6	6	
F	8	8	8



UPGMA

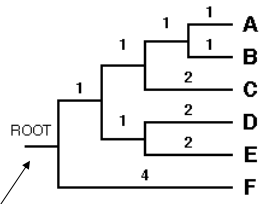
- Forth round

	C,AB	E,D
E,D	6	
F	8	8



- Fifth round

	DE,ABC
F	8

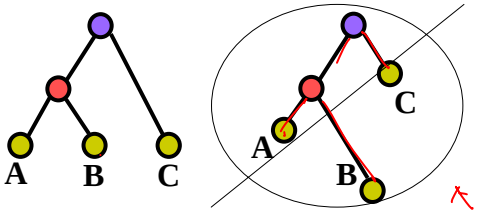


Note the this method identifies the root of the tree.

UPGMA assumes a molecular clock

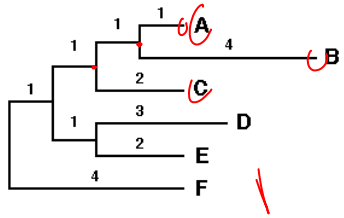
- The UPGMA clustering method is very sensitive to unequal evolutionary rates (assumes that the evolutionary rate is the same for all branches).
- Clustering works only if the data are ultrametric
- Ultrametric distances are defined by the satisfaction of the 'three-point condition'.

The three-point condition:

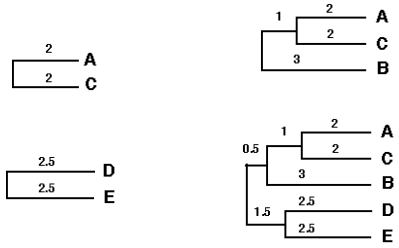


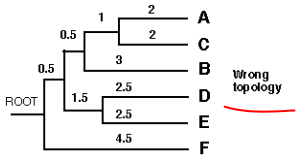
For any three taxa, the two greatest distances are equal.

UPGMA fails when rates of evolution are not constant



	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8



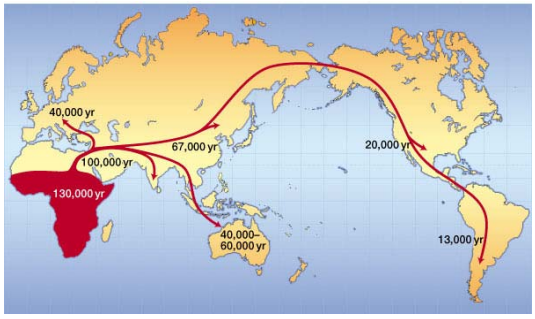


(Neighbor joining will get the right tree in this case.)

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html

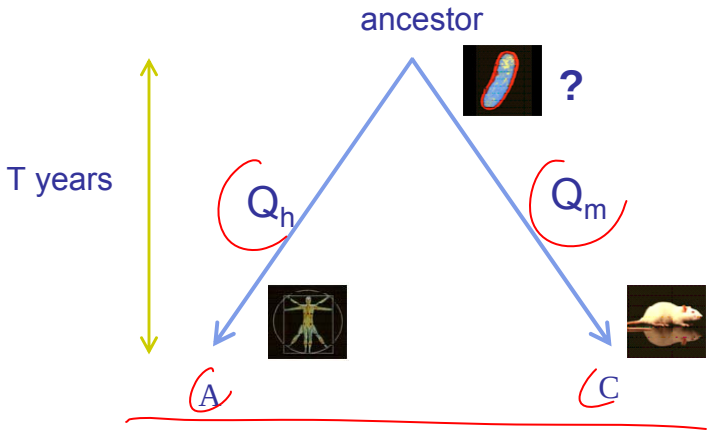
Phylogeny based upon the molecular clock

- Evidence for a human mitochondrial origin in Africa: African sequence diversity is twice as large as that of non-African
- Gyllensten and colleagues estimate that the divergence of Africans and non-Africans occurred 52,000 to 28,000 years ago.



Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. (2000) *Nature* 408: 708-713.

A pair of homologous bases



Typically, the ancestor is unknown.

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

- Strictly speaking, only applicable to regions undergoing little selection.
- Standard assumptions (sometimes weakened)

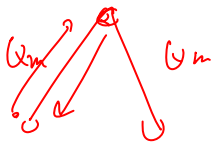
1. Site independence.
2. Site homogeneity.
3. Markovian: given current base, future substitutions independent of past.
4. Temporal homogeneity: stationary Markov chain.

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 π of Q.
- Q is **reversible**: for $a, b, t \geq 0$

  $\pi(a)P(t,a,b) = P(t,b,a)\pi(b)$,

  (detailed balance).



The stationary distribution

- A probability distribution π 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,

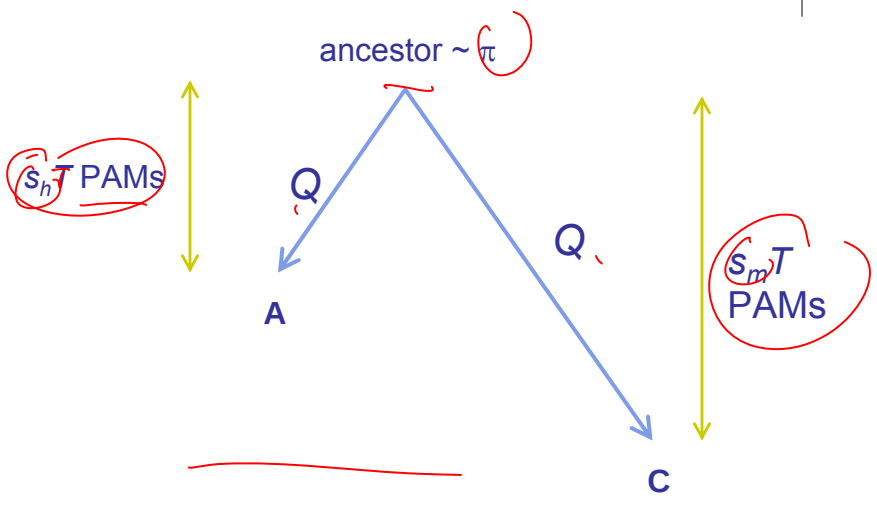
$$\sum_i \pi(i)\, P(i,j) = \pi(j).$$

- **Exercise**. Given any initial distribution, the distribution at time t of a chain with transition matrix P converges to π as $t \to \infty$. Thus, π 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.)

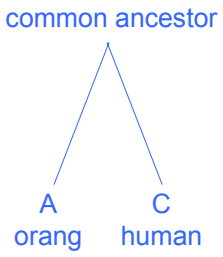
We often assume that the ancestor sequence is i.i.d π.

New picture from a statistical evolution model



The Jukes-Cantor adjustment

- Assume that the common ancestor has A, G, C or T with probability 1/4.

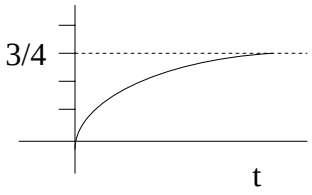


- Then the chance of the nt differing

$p_{\neq} = 3/4 \times (1 - e^{-8\alpha t})$

$= 3/4 \times (1 - e^{-4k/3})$, since $k = 2 \times 3\alpha t$

When $k = .01$, described as *1 PAM*



Joint probability of A and C

- Under the model in the previous slides, the joint probability is

$$p(\mathrm{A},\mathrm{C}) = \sum_a \pi(a)\, p(\mathrm{A} \mid s_h T, Q, a)\, p(\mathrm{C} \mid s_m T, Q, a)$$
$$= \sum_a \pi(\mathrm{A})\, p(a \mid s_h T, Q, \mathrm{A})\, p(\mathrm{C} \mid s_m T, Q, a)$$
$$= \pi(\mathrm{A})\, p(\mathrm{C} \mid s_h T + s_m T, Q, \mathrm{A})$$
$$= (t, \mathrm{A}, \mathrm{C})$$

  - where $t = s_h T + s_m T$ is the (evolutionary) distance between A and C. Note that s_h, s_m and T are not identifiable.

- The matrix $F(t)$ is symmetric. It is equally valid to view A as the ancestor of C or vice versa.

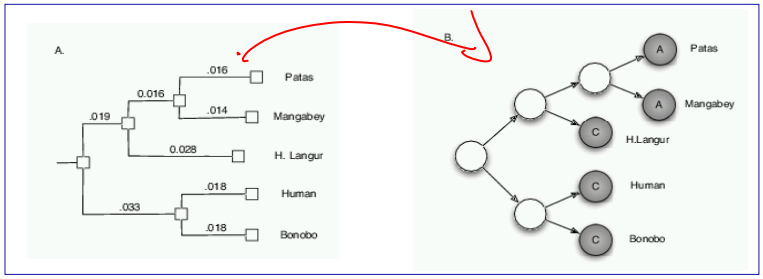
Estimating the evolutionary distance between two sequences

- Suppose two aligned protein sequences $a_1 \ldots a_n$ and $b_1 \ldots b_n$ are separated by t PAMs.
- Under a reversible substitution model that is IID across sites, the likelihood of t is

$$L(t) = p(a_1 \ldots a_n, b_1 \ldots b_n \mid \text{model})$$
$$= \prod_k (t, a_k, b_k)$$
$$= \prod_{a,b} (t, a, b)^{(a,b)}$$

  - *where* $c(a,b) = \#\{k : a_k = a,\ b_k = b\}$.
- Maximizing this quantity gives the maximum likelihood estimate of t. This generalizes the distance correction with Jukes-Cantor.

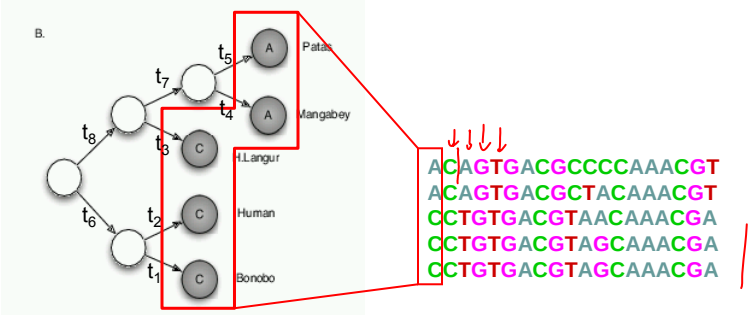
Phylogeny



- 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

Likelihood methods

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



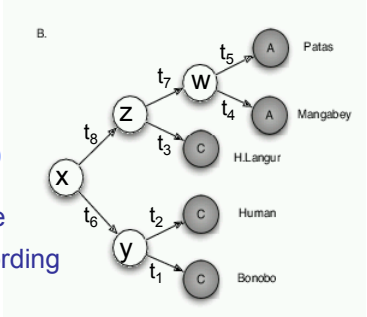
  - Since the sites evolve independently on the same tree,

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

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:

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

- 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:

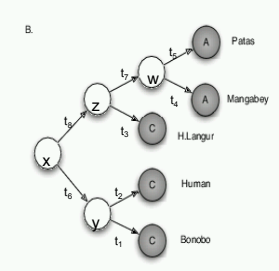
$$P(D^{(i)} \mid T) = \sum_x \sum_y \sum_z \sum_w P(A, A, C, C, C, x, y, z, w \mid T) =$$

$$\sum_x P(x)$$

$$\left(\sum_y P(y \mid x, t_6) \quad P(\mathrm{C} \mid y, t_1)\, P(\mathrm{C} \mid y, t_2) \right)$$

$$\left(\sum_z P(z \mid x, t_8) \quad P(\mathrm{C} \mid z, t_3) \right.$$

$$\left. \left(\sum_w P(w \mid z, t_7) P(\mathrm{A} \mid w, t_4) P(\mathrm{A} \mid w, t_5) \right) \right)$$



Felsenstein's Pruning Algorithm

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

Initialization:

Set k = 2N – 1

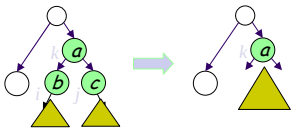
Recursion: Compute $P(L_k \mid a)$ for all $a \in \Sigma$

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 $P(L_i \mid b)$, $P(L_j \mid c)$ for all b and c, 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 = $P(x_1, x_2, \ldots, x_N \mid T, t) = \sum_a P(L_{2N-1} \mid a) P(a)$

- This algorithm can easily handle Ambiguity and error in the sequences (how?)

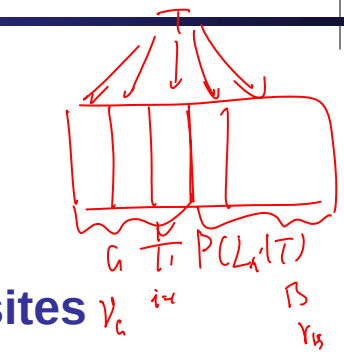
Finding the ML tree

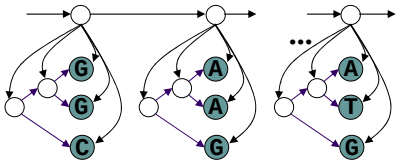
- So far I have just talked about the computation of the likelihood for one tree with branch lengths known.
- To find a ML tree, we must search the space of tree topologies, and for each one examined, we need to optimize the branch lengths to maximize the likelihood.

Bayesian phylogeny methods

- Bayesian inference has been applied to inferring phylogenies (Rannala and Yang, 1996;Mau and Larget, 1997; Li, Pearl and Doss, 2000).
  - All use a prior distribution on trees. The prior has enough influence on the result that its reasonableness should be a major concern. In particular, the depth of the tree may be seriously affected by the distribution of depths in the prior.
  - All use Markov Chain Monte Carlo (MCMC) methods. They sample from the posterior distribution.
  - When these methods make sense they not only get you a point estimate of the phylogeny, they get you a distribution of possible phylogenies.

Modeling rate variation among sites



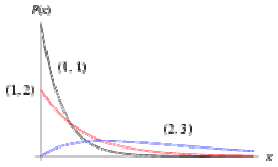


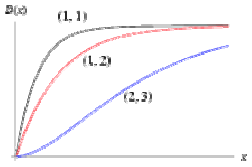
A model of variation in evolutionary rates among sites

- The basic idea is that the rate at each site is drawn independently from a distribution of rates. The most widely used choice is the Gamma distribution, which has density function:

$$f(r) = \frac{\lambda^{\alpha} r^{\alpha-1} e^{-\lambda r}}{\Gamma(\alpha)} = \frac{r^{\alpha-1} e^{-r/\theta}}{\Gamma(\alpha)\theta^{\alpha}}$$

- Gamma distributions (α,θ)





Unrealistic aspects of the model:

- There is no reason, aside from mathematical convenience, to assume that the Gamma is the right distribution.
- A common variation is to assume there is a separate probability f_0 of having rate 0.
- Rates at different sites appear to be correlated, which this model does not allow.
- Rates are not constant throughout evolution, they change with time.

Rates varying among sites

- If $L^{(i)}(r_i)$ is the likelihood of the tree for site i given that the rate of evolution at site i is r_i, we can integrate this over a gamma density:

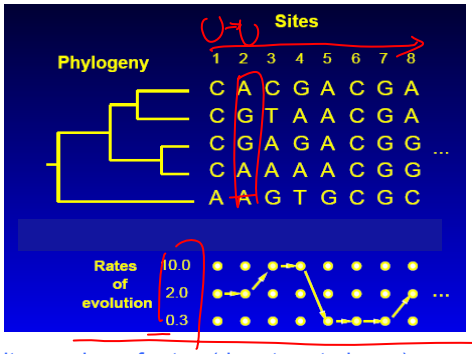
$$L^{(i)} = \int_0^\infty f(r_i;\alpha) L^{(i)}(r_i) dr_i$$

  - so that the overall likelihood is

$$L = \prod_{i=1}^{m} \left[\int_0^\infty f(r_i;\alpha) L^{(i)}(r_i) dr_i \right]$$

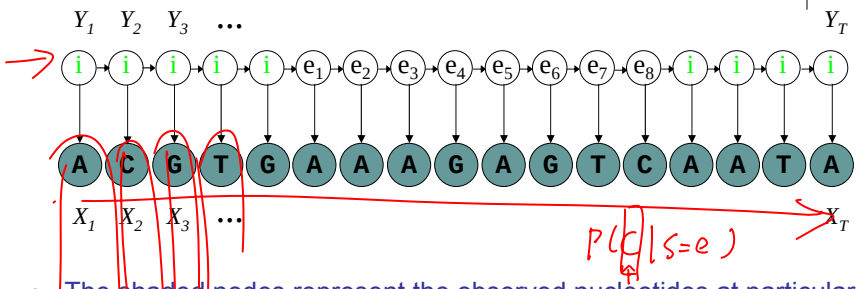
- Unfortunately these integrals cannot be evaluated for trees with more than a few tips as the quantities $L^{(i)}(r_i)$ becomes complicated.

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 appropriately 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)

- Observation space
  - **Alphabetic set:** $\mathbb{C} = \{c_1, c_2, \cdots, c_K\}$
  - **Euclidean space:** $\mathbb{R}^d$
- Index set of hidden states

  $\mathbb{I} = \{1,2,\cdots,M\}$
- Transition probabilities between any two states

  $p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j}$,

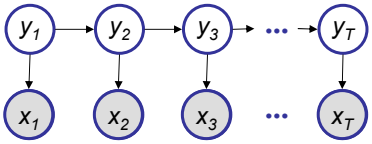
  or $p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \forall i \in \mathbb{I}.$
- Start probabilities

  $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M)$.
- Emission probabilities associated with each state

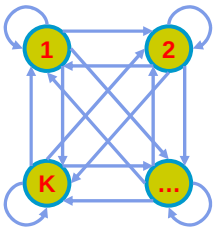
  $p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in \mathbb{I}.$

  or in general:

  $p(x_t \mid y_t^i = 1) \sim \text{f}(\cdot \mid \theta_i), \forall i \in \mathbb{I}.$

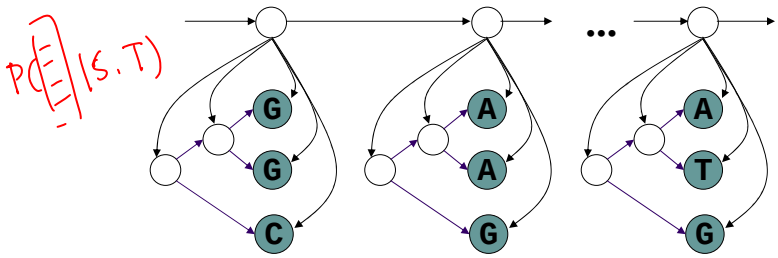


Graphical model



State automata

Hidden Markov Phylogeny



- Replacing the standard emission model with a tree
  - A process not visible to us (.hidden") assigns rates to sites. It is a Markov process working along the sequence.
  - For example it might have transition probability Prob ($j|i$) of changing to rate j in the next site, given that it is at rate i in this site.
- These are the most widely used models allowing rate variation to be correlated along the sequence.

The Forward Algorithm

- We can compute α_t^k for all k, t, using dynamic programming!

Initialization:

$$\alpha_1^k = P(x_1 \mid y_1^k = 1)\pi_k$$

$$\alpha_1^k = P(x_1, y_1^k = 1)$$
$$= P(x_1 \mid y_1^k = 1)P(y_1^k = 1)$$
$$= P(x_1 \mid y_1^k = 1)\pi_k$$

Iteration:

$$\alpha_t^k = P(x_t \mid y_t^k = 1)\sum_i \alpha_{t-1}^i a_{i,k}$$

Termination:

$$P(\mathbf{x}) = \sum_k \alpha_T^k$$

The Backward Algorithm

- We can compute β_t^k for all k, t, using dynamic programming!

Initialization:

$$\beta_T^k = 1, \ \forall k$$

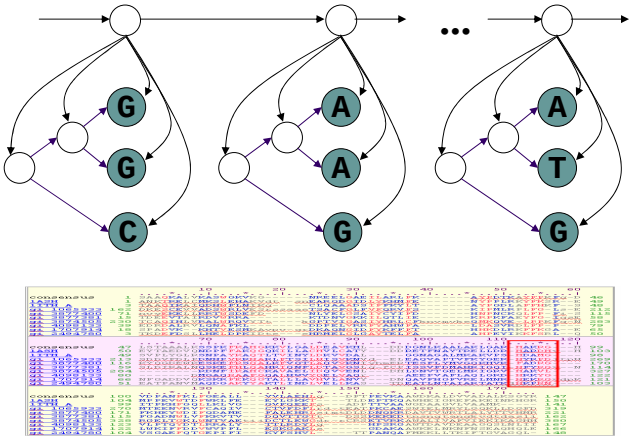
Iteration:

$$\beta_t^k = \sum_i a_{k,i} P(x_{t+1} \mid y_{t+1}^k = 1)\beta_{t+1}^i$$

Termination:

$$P(\mathbf{x}) = \sum_k \alpha_1^k \beta_1^k$$

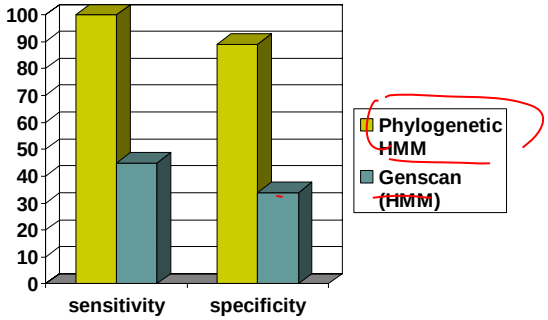
Hidden Markov Phylogeny



- this yields a gene finder that exploits evolutionary constraints

A Comparison of comparative genomic gene-finding and isolated gene-finding

- 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%.



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.

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
- **Itai Yanai:** for some of the slides modified from his lectures at Weizmann Institute of Science