Advanced Algorithms and Models for Computational Biology

-- a machine learning approach

Molecular Ecolution:

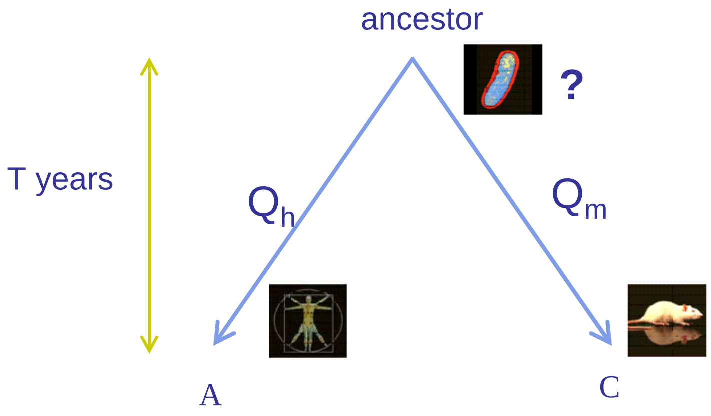
Phylogenetic trees

Eric Xing

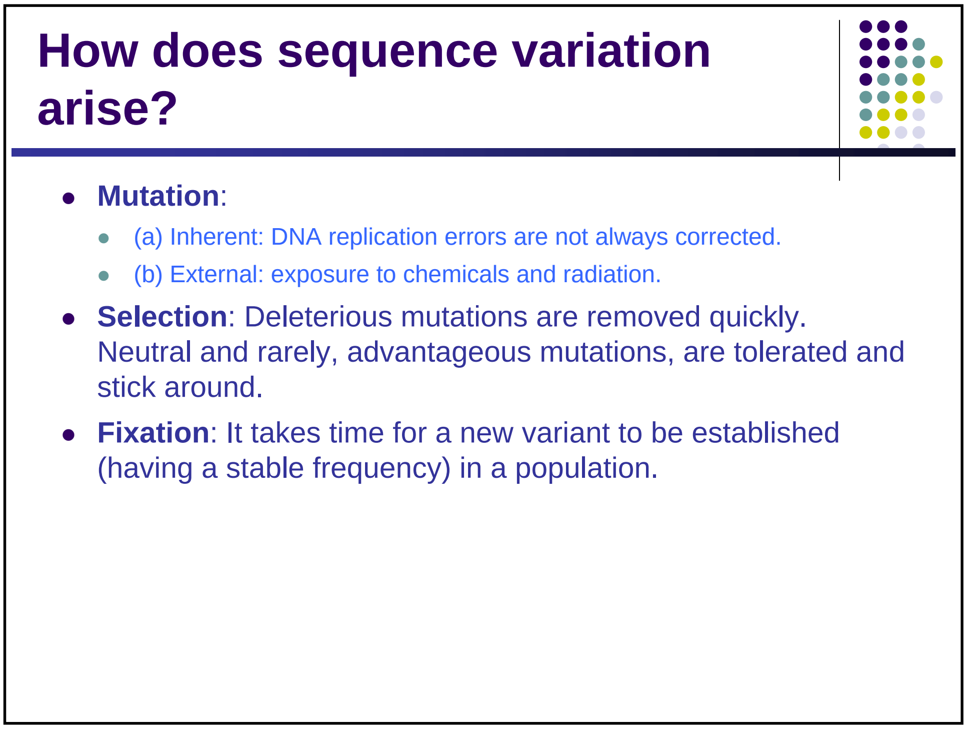
Lecture 21, April 5, 2006

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

A pair of homologous bases

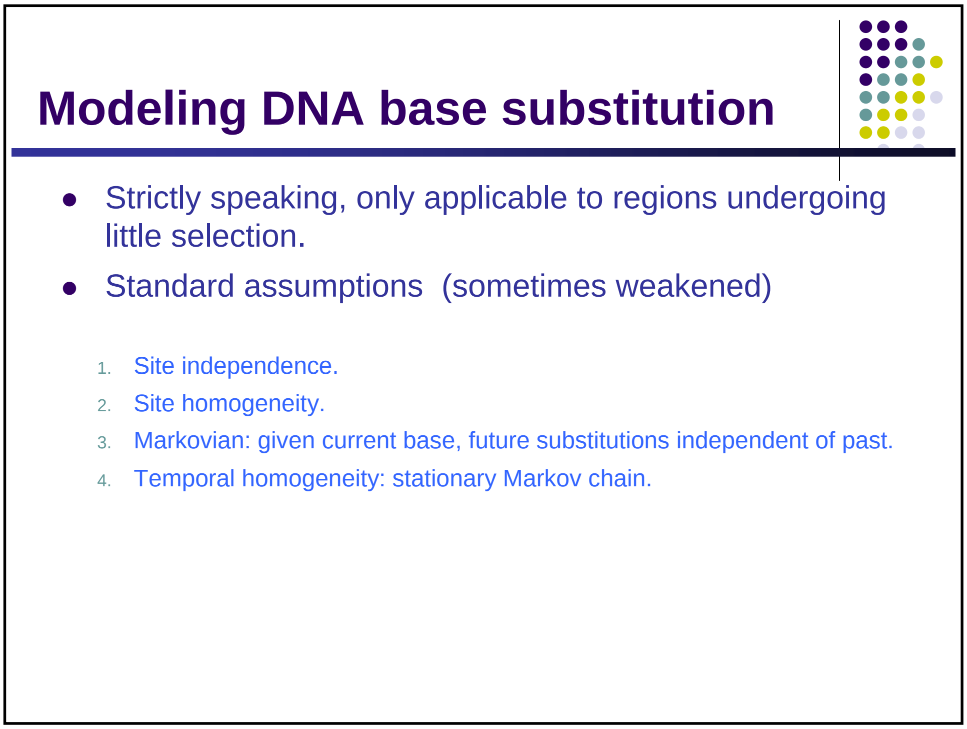


Typically, the ancestor is unknown.



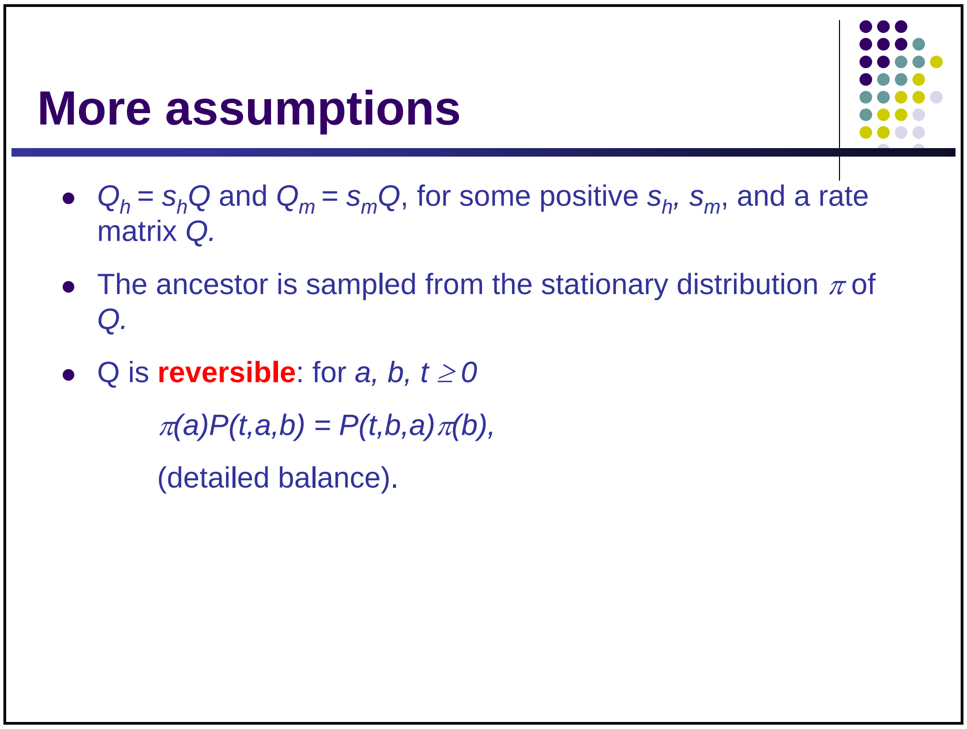
How does sequence variation arise?

- **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.

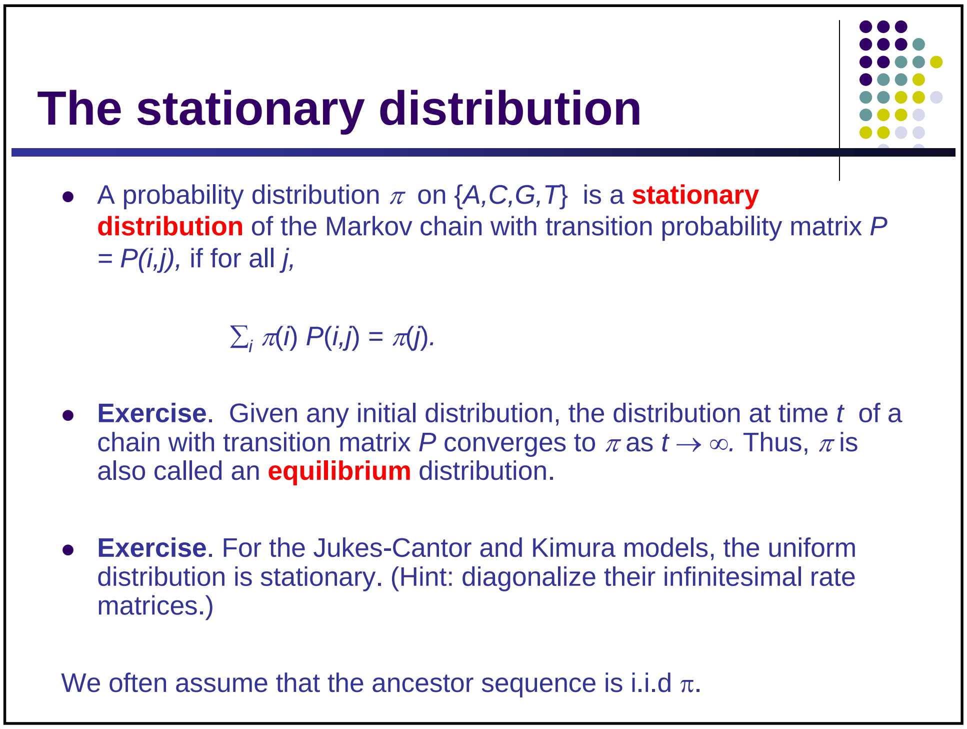


More assumptions

- $Q_h = s_h Q$ 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,

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

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

Phylogeny methods

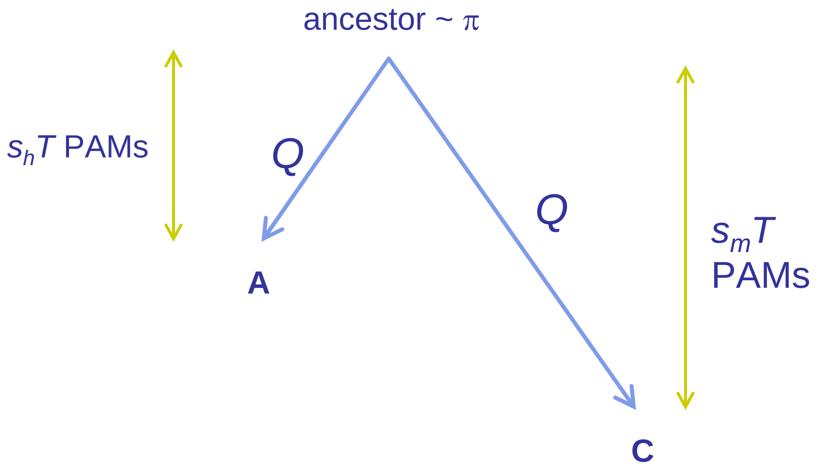
Basic principles:

- 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

New picture



Joint probability of A and C

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

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

  - where t = s_hT+ s_mT 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.

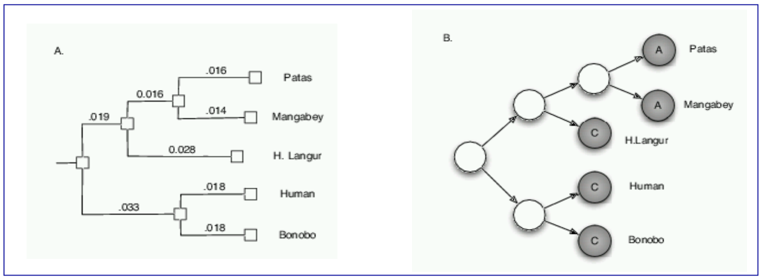
- Under a reversible substitution model that is IID across sites, the likelihood of *t* is

$$\begin{aligned} L(t) &= p(a_1 \ldots a_n, b_1 \ldots b_n \mid \mathrm{model}) \\ &= \prod_k F(t, a_k, b_k) \\ &= \prod_{a,b} F(t, a, b)^{c(a,b)} \end{aligned}$$

  - *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.

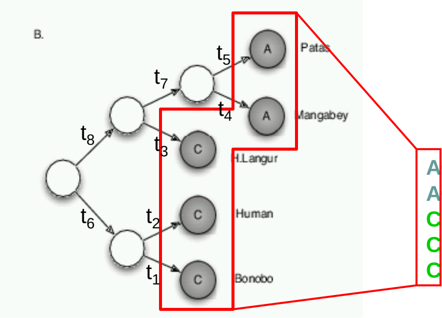
Phylogeny



- 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.



ACAGTGACGCCCCAAACGT
ACAGTGACGCTACAAACGT
CCTGTGACGTAACAAACGA
CCTGTGACGTAGCAAACGA
CCTGTGACGTAGCAAACGA

- Since the sites evolve independently on the same tree,

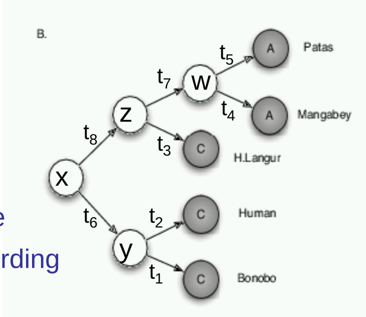
$$L = P(D \mid T) = \prod_{i=1}^{m} P(D^{(i)} \mid T)$$

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:



$$\begin{aligned}
&P(A,A,C,C,C,x,y,z,w \mid T) = \\
&\qquad P(x) \quad P(y \mid x, t_6) \quad P(A \mid y, t_1)\, P(C \mid y, t_2) \\
&\qquad\qquad\quad P(z \mid x, t_8) \quad P(C \mid y, t_3) \\
&\qquad\qquad\qquad\qquad\quad P(w \mid z, t_7)\, P(C \mid y, t_4)\, P(C \mid y, t_5)
\end{aligned}$$

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

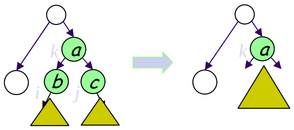
Getting a recursive algorithm

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

$$\begin{aligned}
&P(D^{(i)} \mid T) = \sum_x \sum_y \sum_z \sum_w P(A,A,C,C,C,x,y,z,w \mid T) = \\
&\qquad \sum_x P(x) \\
&\quad \Big(\sum_y P(y \mid x, t_6) \quad P(A \mid y, t_1)\, P(C \mid y, t_2) \Big) \\
&\quad \Big(\sum_z P(z \mid x, t_8) \quad P(C \mid z, t_3) \\
&\qquad\qquad\qquad \Big(\sum_w P(w \mid z, t_7) P(C \mid w, t_4) P(C \mid w, t_5) \Big)\Big)
\end{aligned}$$

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

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

Finding the ML tree

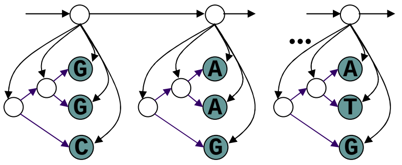
- So far I have just talked about the computation of the likelihood for one tree with branch lengths known.

- 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

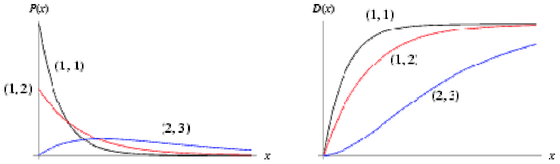


A model of variation in evolutionary rates among sites

- 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:

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

- 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 f0 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:

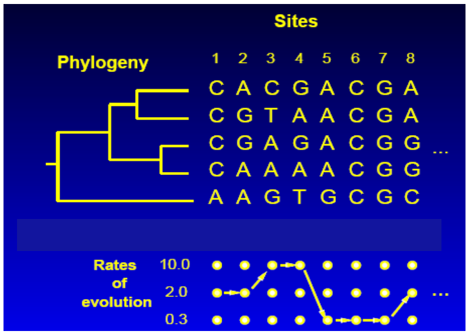
$$L^{(i)} = \int_0^\infty f(r_i;\alpha)L^{(i)}(r_i)dr_i$$

  - 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.

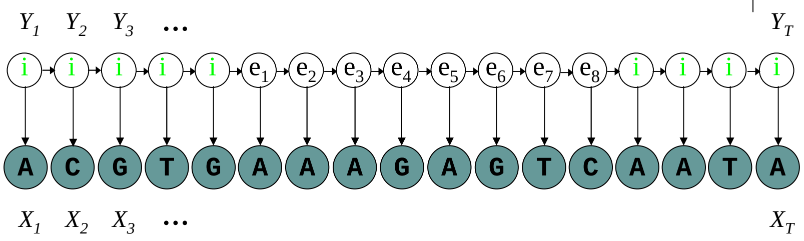
Modeling rate variation among sites



1	2	3	4	5	6	7	8
C	A	C	G	A	C	G	A
C	G	T	A	A	C	G	A
C	G	A	G	A	C	G	G
C	A	A	A	A	C	G	G
A	A	G	T	G	C	G	C

- 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


Y_1
Y_2
Y_3
…
Y_T
i i i i i e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8 i i i i
A C G T G A A A G A G T C A A T A
X_1
X_2
X_3
…
X_T

- 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

  $\mathrm{I} = \{1, 2, \cdots, M\}$

- Transition probabilities between any two states

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

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

- Start probabilities

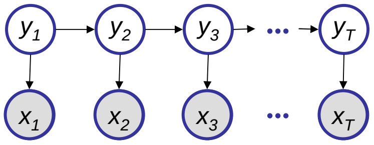
  $p(y_1) \sim \mathrm{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$

- Emission probabilities associated with each state

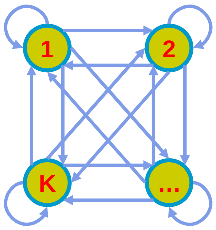
  $p(x_t \mid y_t^i = 1) \sim \mathrm{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in \mathrm{I}.$

  or in general:

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

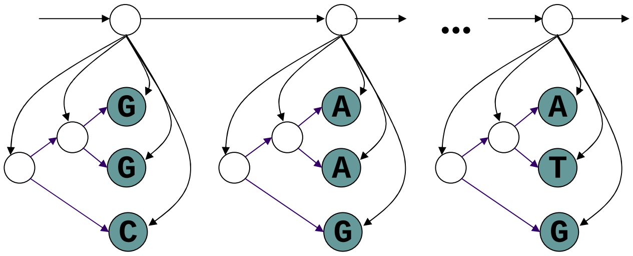


Graphical model



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:

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

$$\begin{aligned}\alpha_1^k &= P(x_1, y_1^k = 1)\\ &= P(x_1 \mid y_1^k = 1)P(y_1^k = 1)\\ &= P(x_1 \mid y_1^k = 1)\pi_k\end{aligned}$$

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:

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

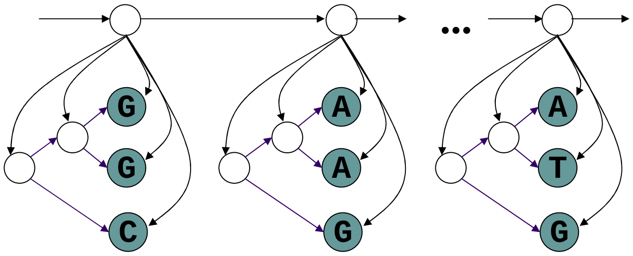
Iteration:

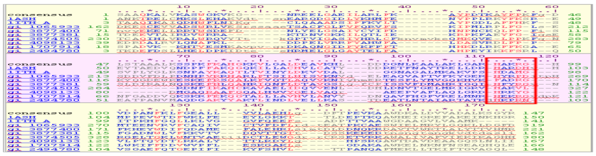
$$\beta_t^k = \sum\nolimits_i a_{k,i} P(x_{t+1} \mid y_{t+1}^k = 1) \beta_{t+1}^i$$

Termination:

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

Hidden Markov Phylogeny

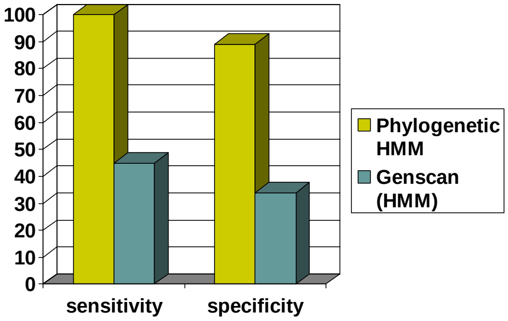




- this yields a gene finder that exploits evolutionary constraints

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