

HMM's – Hidden Markov Models Part II

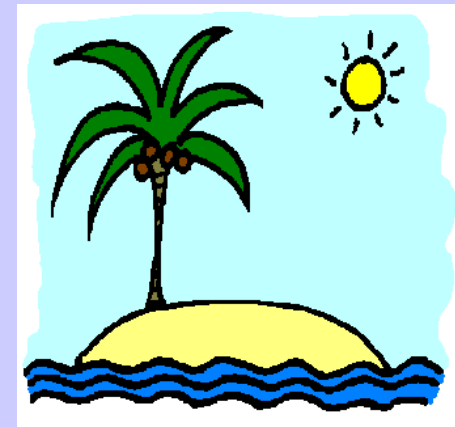
Sources:

- 1) <http://www.tau.ac.il/>
- 2) <http://evolution.genetics.washington.edu/phylip/doc/dnaml.html>
- 3) *Felsenstein J, Churchill GA. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol. 1996 Jan;13(1):93-104.*

CpG islands

In the human genome, for biological reasons, the occurrence of the dinucleotide CG is lower than would be expected from C & G independent probabilities.

- On the other hand, around promoters or the “start” regions of many genes, we see many more CG couplets (and in fact more C and G generally).



Markov chains in use

Let's answer one of the previous questions, using Markov Chains. For example: Given a sequence, is it a CpG island? Let's assume that we have the 2 sets of **transition probabilities**: in "standard" DNA, and transition probabilities in "CpG" DNA.

Markov chains in use - continued

+	A	C	G	T
A	0.18	0.274	0.426	0.12
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.3	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

CpG islands (+ model) “normal” DNA (- model)

Markov chains in use - continued

Thus, we have two models which we can compare statistically (likelihood ratio test):

$$S(x) = \log \frac{P(x | \text{model } +)}{P(x | \text{model } -)} = \sum_{i=1}^n \log \frac{P^+_{x_i - 1x_i}}{P^-_{x_i - 1x_i}}^*$$

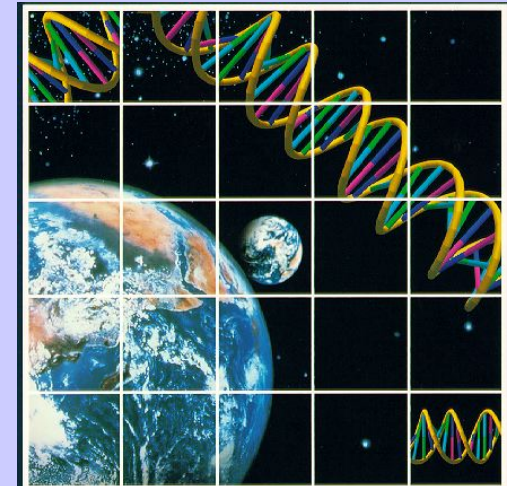
$$S(x) \sim \chi^2$$

*Note: $P_{x_0x_1} = P(x_1)$ – the probability of beginning with x_1 .

And now to HMM's

HMM's can answer a question we posed before that Markov chains cannot – how do we find CpG islands in a long un-annotated sequence?

- Instead of two Markov chain models (like before) we need one model that incorporates both the models from before. How do we do that in the previous example?



The solution...

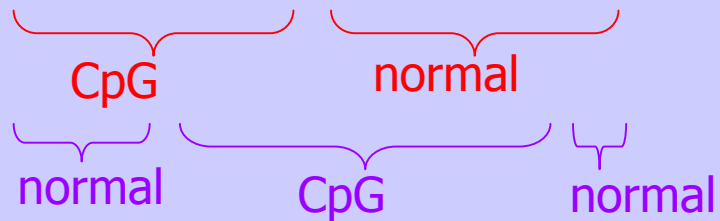
We **relable** the states. In order to integrate, we will define A^+ , G^+ , C^+ , T^+ (CpG areas) and A^- , G^- , C^- , T^- (“normal” DNA). Both A^+ and A^- emit the symbol A.

	A+	C+	G+	T+	A-	C-	G-	T-
A+								
C+								
G+								
T+								
A-								
C-								
G-								
T-								

Many state sequences emit a sequence of symbols:

A⁺G⁺C⁺T⁺G⁻C⁻C⁻T⁻ A⁻G⁻C⁺T⁺G⁺C⁺C⁺T⁻

A G C T G C C T



Most probable state path - Viterbi algorithm

We can compute the likelihood of each one of these state paths.

Look for the maximum likelihood state sequence.

This way, given a long unnotated sequence, one gets the ML state at the end:

A-G-C-G-T-T-T-C-G+C+A+G+A-C-G-T-C+G+T+

Viterbi algorithm does this with dynamic programming algorithm (an example – later).

Another example: Inferring phylogeny and rate of evolution from aligned sequences

Reminder: A **phylogeny** (or phylogenesis) is the origin and evolution of a set of organisms, usually of a species. A major task is to determine the ancestral relationships among known species (both living and extinct), and the most commonly used methods to infer phylogenies include maximum likelihood, and Bayesian.

Many methods assume equal rate of evolution at all sites – unrealistic

Rate4site – a method for deducing the rate at each sites. Assumes independence of the rate between sites.

Rate4site

- rate of evolution is not constant among amino acid sites
- rate variations correspond to different levels of purifying selection acting on these sites.
- purifying selection can be the result of geometrical constraints on the folding of the protein into its 3D structure, constraints at amino acid sites involved in enzymatic activity or in ligand binding or, alternatively, at amino acid sites that take part in protein-protein interactions.
- **Rate4Site** calculates the relative evolutionary rate at each site using a probabilistic-based evolutionary model. This allows taking into account the stochastic process underlying sequence evolution within protein families and the phylogenetic tree of the proteins in the family. The conservation score at a site corresponds to the site's evolutionary rate.

Methodology of Rate4site

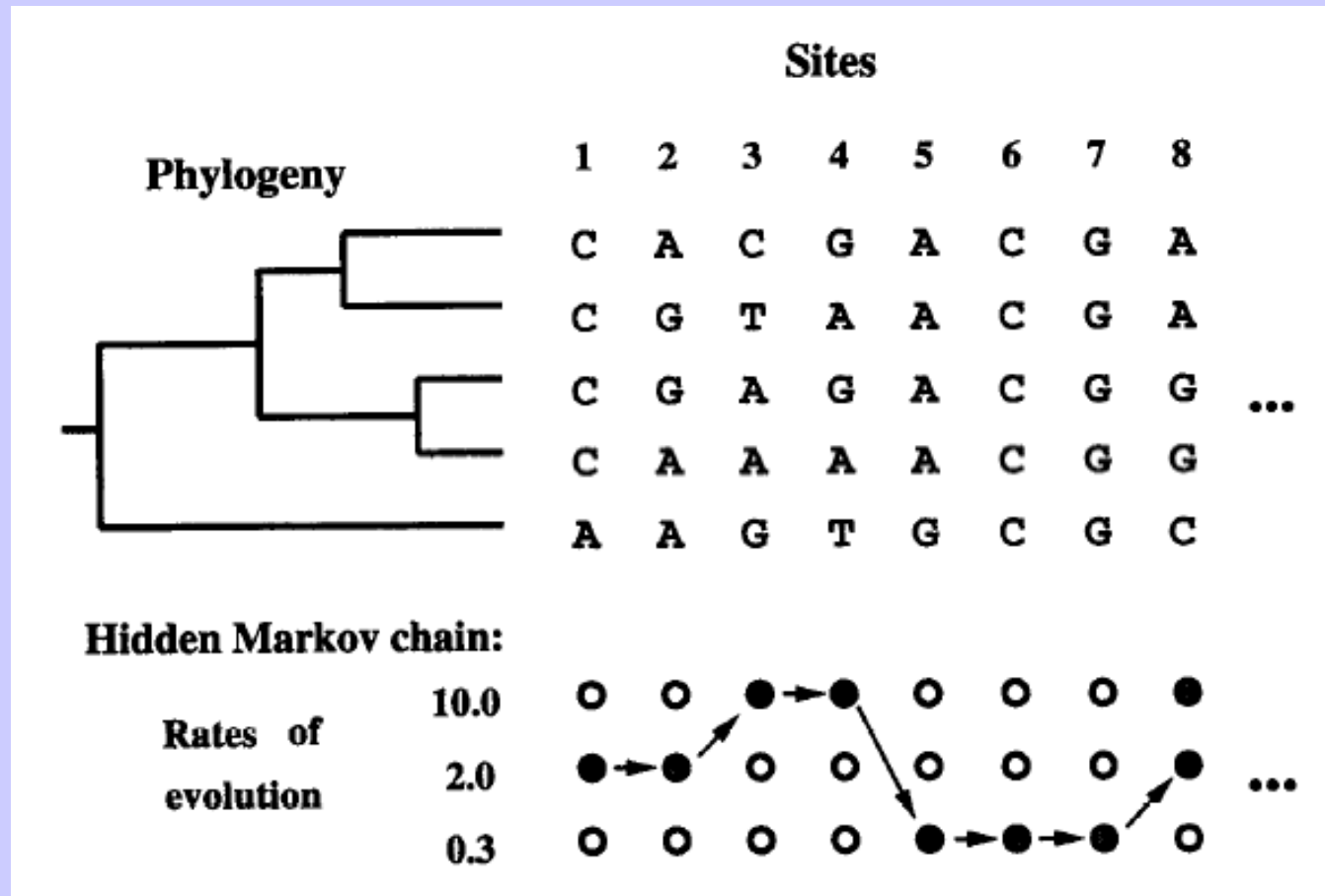
- The sole obligatory input to Rate4Site is an MSA file. The program then computes a phylogenetic tree that is consistent with the available MSA (the user can also input a pre-calculated tree). It then calculates the relative conservation score for each site in the MSA. This is carried out using either an empirical Bayesian method or a maximum likelihood method (Pupko et al., 2002).
- Differences between the two methods are explained in details in Mayrose et al (2004).
- References: Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. 2004. Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol Biol Evol* 21: 1781-1791.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions on proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18 Suppl: S71-S77.

Minimum requirements for (half-) realistic inference of phylogeny

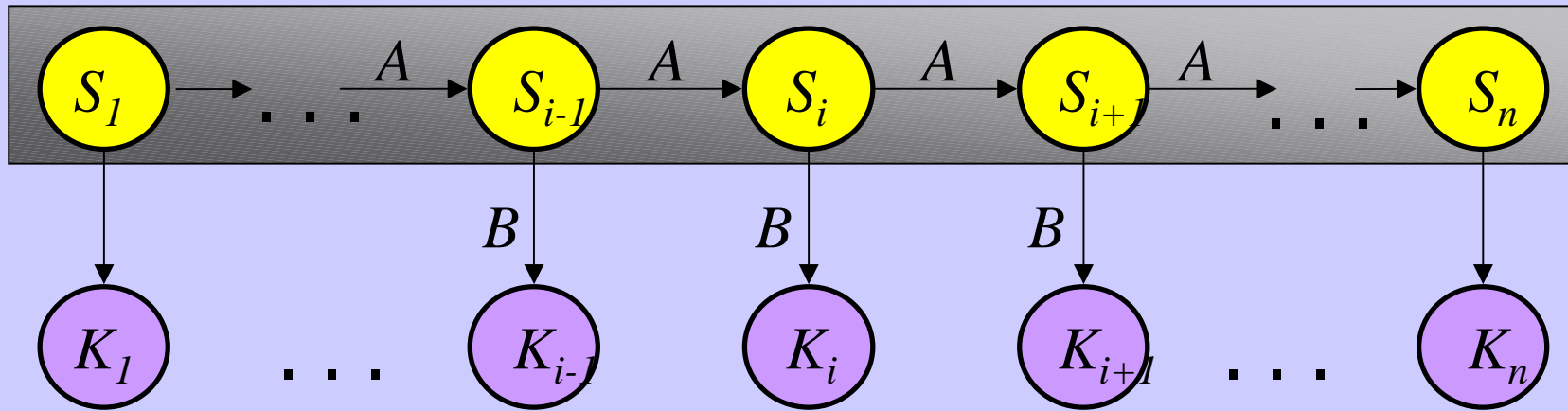
1. It must allow rates to differ among sites.
2. It must not assume that we know the relative rates of change at the individual sites, but must instead infer these from the data.
3. It must allow there to be some correlation between the rates of evolution at adjacent sites.

A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution (Churchill and Felsenstein 1996)



all correlation between sites will be assumed to be the consequence of the clustering (if any) of high and low rates at adjacent sites.

Reminder: HMM Formalism



$\{S, K, P, A, B\}$

$S: \{s_1 \dots s_N\}$ are the values for the hidden states

$K: \{k_1 \dots k_M\}$ are the values for the observations

The hidden states **emit/generate** the symbols (observations)

$\Pi = \{\pi_i\}$ are the initial state probabilities

$P = \{P_{ij}\}$ are the state transition probabilities

$B = \{b_{ik}\}$ are the emission probabilities

Motivation

- To find the path of rates that fits the data in the best way.
- We will represent a general path as follows:
 (c_1, c_2, \dots, c_n) (path of evolutionary rate *states*)
where $c_i = 10.0$ or 2.0 or 0.3 (in the example above).
- *Observations*: Multiple sequence alignments
- c_i corresponds to the rate at site i .
- n – is the length of the sequence.

Paths \rightarrow HMM

- Each path corresponds to a state path.
The “real” path is hidden.
- Before we had $A^+C^+G^+ \rightarrow A C G$
- Now we have c_1, c_2, c_3
corresponding to a multiple sequence alignment :

x_1	y_1	y_3
x_2	y_2	y_3
...
s^1	s^2	s^3

Motivation - notes

- Finding the path that best fits the data...how?
The path that makes the maximum contribution to the likelihood of the data.
- Instead of going over all possible paths (computationally hard) → dynamic programming.

Outline of the algorithm

1. Assume we have k categories of rate:

r_1, r_2, \dots, r_k

We know Π_{r_i} for $i=1\dots k$ (probabilities of occurrences of rate r_i)

e.g. $r_1:r_2:r_3:r_4 = 0:1:2.3:8.9$ for four different states

We know prior probabilities that a given site is in these k categories, e.g. $0.1:0.32:0.22:0.36$

Outline - continued

2. The likelihood of a given phylogeny T :

$$\begin{aligned} L &= \text{Prob}(D|T) \\ &= \sum_{c_1} \sum_{c_2} \cdots \sum_{c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ &\quad \times \text{Prob}(D|T, r_{c_1}, r_{c_2}, \dots, r_{c_n}). \end{aligned}$$

sum, over all assignments of rate categories, of the probability of the data D given that combination of rates, multiplied by the prior probability of that combination of rates.

n – the number of sites in sequence.

c_i denotes the category that a given rate combination assigns to site i , so that the rate assigned to site i is r_{c_i}

Outline - continued

3. Look for the combination which makes the largest contribution to the likelihood:

$$R = \max_{c_1, c_2, \dots, c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ \times \text{Prob}(D | T, r_{c_1}, r_{c_2}, \dots, r_{c_n}).$$

1. *Choosing rate categories*

How to choose the rates?

- estimate the values of relative rates r_i and probabilities f_i by ML, using the EM algorithm (significant increase in computation time)
- in practise: examine a few sets of rates and choose the one with ML

2. Computing the likelihood

Since each site evolves independently once the rate categories are determined, the likelihood

$$\begin{aligned} L &= \text{Prob}(D|T) \\ &= \sum_{c_1} \sum_{c_2} \cdots \sum_{c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ &\quad \times \text{Prob}(D|T, r_{c_1}, r_{c_2}, \dots, r_{c_n}). \end{aligned}$$

transfers into

$$\begin{aligned} L &= \sum_{c_1} \sum_{c_2} \cdots \sum_{c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ &\quad \times \prod_{i=1}^n \text{Prob}(D_i|T, r_{c_i}). \end{aligned}$$

2. Computing the likelihood (cont'd)

$$L = \sum_{c_1} \sum_{c_2} \dots \sum_{c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ \times \prod_{i=1}^n \text{Prob}(D_i | T, r_{c_i}).$$

Computational costs for computing likelihood are **tremendous**:
k categories, n sites:

number of combinations of categories is k^n

e.g. 1000 sites, 3 rates: $3^{1000} = 10^{477}$

cf. Number of particles in universe: 10^{80}

2. Computing the likelihood (cont'd)

Reminder: Hidden Markov Model specifies that each combination of rate categories c_1, c_2, \dots, c_n is the outcome of a stationary Markov chain, and thus its prior probability is simply the product of the prior probability of c_1 times a product of transition probabilities of that Markov chain:

$$\text{Prob}(c_1, c_2, \dots, c_n) = f_{c_1} P_{c_1, c_2} P_{c_2, c_3} \dots P_{c_{n-1}, c_n}.$$

Thus, L can be rewritten as:

or:

$$L = \sum_{c_1} f_{c_1} \left[\sum_{c_2} \sum_{c_3} \dots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \right. \\ \left. \times \text{Prob}(D | T, r_{c_1}, r_{c_2}, \dots, r_{c_n}) \right]$$

$$L = \sum_{c_1} f_{c_1} \left[\sum_{c_2} \sum_{c_3} \dots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \right. \\ \left. \times \prod_{i=1}^n \text{Prob}(D_i | T, r_{c_i}) \right].$$

2. Computing the likelihood (cont'd)

Let's define $L_{c_k}^{(k)}$ as the conditional likelihood of T for the data $D^{(k)}$ (= data set consisting of sites k through n) given that site k has rate category c_k

$$L = \sum_{c_1} f_{c_1} \left[\sum_{c_2} \sum_{c_3} \dots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \right. \\ \left. \times \prod_{i=1}^n \text{Prob}(D_i | T, r_{c_i}) \right].$$

$$[\dots] = L_{c_1}^{(1)}$$

Thus, L can be rewritten as:

$$L = \sum_{c_1} f_{c_1} L_{c_1}^{(1)}.$$

2. Computing the likelihood (cont'd)

Definition of likelihood L allows to reformulate:

$$L_{c_1}^{(1)} = \text{Prob}(D_1 | T, r_{c_1}) \sum_{c_2} \sum_{c_3} \cdots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \\ \times \prod_{i=2}^n \text{Prob}(D_i | T, r_{c_i}). \quad (7)$$

Because of Markov chain property:

$$L_{c_1}^{(1)} = \text{Prob}(D_1 | T, r_{c_1}) \sum_{c_2} P_{c_1, c_2} \\ \times \left[\sum_{c_3} \cdots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \right. \\ \left. \times \prod_{i=2}^n \text{Prob}(D_i | T, r_{c_i}) \right].$$

$$[\dots] = L_{c_2}^{(2)}$$

2. Computing the likelihood (cont'd)

$$L_{c_1}^{(1)} = \text{Prob}(D_1 | T, r_{c_1}) \sum_{c_2} P_{c_1, c_2} \\ \times \left[\sum_{c_3} \dots \sum_{c_n} \text{Prob}(c_2, \dots, c_n | c_1) \right. \\ \left. \times \prod_{i=2}^n \text{Prob}(D_i | T, r_{c_i}) \right].$$

$$[\dots] = L_{c_2}^{(2)}$$

Thus:

$$L_{c_1}^{(1)} = \text{Prob}(D_1 | T, r_{c_1}) \sum_{c_2} P_{c_1, c_2} L_{c_2}^{(2)}.$$

or generally:

$$L_{c_k}^{(k)} = \text{Prob}(D_k | T, r_{c_k}) \sum_{c_{k+1}} P_{c_k, c_{k+1}} L_{c_{k+1}}^{(k+1)}.$$

with:

$$L_{c_n}^{(n)} = \text{Prob}(D_n | T, r_{c_n}).$$

This recursion form can be easily derived

2. Computing the likelihood (cont'd)

General recursion form:

$$L_{c_k}^{(k)} = \text{Prob}(D_k | T, r_{c_k}) \sum_{c_{k+1}} P_{c_k, c_{k+1}} L_{c_{k+1}}^{(k+1)}.$$

$$L_{c_n}^{(n)} = \text{Prob}(D_n | T, r_{c_n}).$$

Pattern of computation reverses order of recursion:
proceed from last site n to the first one:

First compute likelihoods $\text{Prob}(D_k | T, r_{c_k})$ at each site for each possible rate category

Then determine $L_{c_n}^{(n)}$ and recursively $L_{c_{n-1}}^{(n-1)} \dots L_{c_1}^{(1)}$

Number of computations only $O(n \times k)$ (instead of $O(k^n)$!)

3. The most probable rates' combination (follows a version of the Viterbi algorithm (1967))

Ability to calculate the likelihood of phylogeny T allows to search for maximal likelihood phylogeny. Likelihood is computed by summing contributions from all possible combinations of rates. Most interesting is the combination that makes the maximal contribution to the likelihoods at the sites:

$$R = \max_{c_1, c_2, \dots, c_n} \text{Prob}(c_1, c_2, \dots, c_n) \\ \times \text{Prob}(D | T, r_{c_1}, r_{c_2}, \dots, r_{c_n}).$$

Define:

$$R_{c_k}^{(k)} = \max_{c_{k+1}, \dots, c_n} \{ \text{Prob}(c_{k+1}, \dots, c_n | c_k) \\ \times \text{Prob}[D^{(k)} | T, r_{c_k}, r_{c_{k+1}}, \dots, r_{c_n}] \}.$$

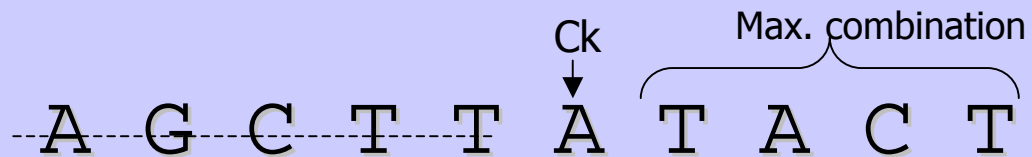
where for $k=n$:

$$R_{c_n}^{(n)} = \text{Prob}(D_n | T, r_{c_n}),$$

- $D^{(k)}$ – data of sites k till n
- D_k – data at site k

The most probable rates' combination - continued

This is the likelihood contribution of sites k through n , where site k has rate c_k , and the rest ($k+1$ through n) have the combination of rates that maximizes the contribution.



The most probable rates' combination - continued

We thus get the following recursive formula:

$$R_{c_k}^{(k)} = \text{Prob}(D_k | T, c_k) \max_{c_{k+1}} [P_{c_k, c_{k+1}} R_{c_{k+1}}^{(k+1)}].$$

Proof (the essentials)

$$R_{c_k}^{(k)} = \max_{c_{k+1}, \dots, c_n} \left\{ \text{Prob}(c_{k+1}, \dots, c_n \mid c_k) \times \text{Prob}(D^{(k)} \mid T, r_{c_k}, \dots, r_{c_n}) \right\} =$$

$$= \max_{c_{k+1}, \dots, c_n} \left\{ P_{c_k+1, c_k+2} \cdot \dots \cdot P_{c_n-1, c_n} \times$$

$$\times \text{Prob}[D_k \mid T, r_{c_k+1}, \dots, r_{c_n}] \times \text{Prob}[D^{(k+1)} \mid T, r_{c_k+1}, \dots, r_{c_n}]$$

1. Because the rate categories are the outcome of a Markov chain and $P(c_1, \dots, c_n) = f_{c_1} P_{c_1, c_2} P_{c_2, c_3} \dots P_{c_{n-1}, c_n}$
2. Due to the assumption that once the rate is set, the sites evolve independently

The most probable rates' combination - continued

Thus, we compute the above equation from sites n , $n-1$, $n-2$ up till 1, for each one of the rate categories.

At the end, we take the largest of the quantities

This is the quantity that maximizes the contribution to the likelihood.

$$f_{c1}R_{c1}^{(1)}$$

Backtracking the rate combination

We still don't know which combination c_1, \dots, c_n brought us here, so we backtrack: Start with rate category c_1 corresponding to maximum contribution. Then find value c_2 that is involved in maximum contribution...

Rate cat.	$R_{r1}^{(n)}$ r1	$R_{r2}^{(n)}$ r2	$R_{r3}^{(n)}$ r3
n			
n-1	$R_{r1}^{(n-1)}$	$R_{r2}^{(n-1)}$	$R_{r3}^{(n-1)}$
...
2	$R_{r1}^{(2)}$	$R_{r2}^{(2)}$	$R_{r3}^{(2)}$
1	$f_{r1}R_{r1}^{(1)}$	$f_{r2}R_{r2}^{(1)}$	$f_{r3}R_{r3}^{(1)}$

The diagram illustrates the backtracking process. Arrows point from the top row (n) down to the bottom row (1) across three columns (r1, r2, r3). Specifically, arrows point from the top row to the middle row (n-1) in each column, and from the middle row to the bottom row (1) in each column. A diagonal line is drawn from the top-left cell (n, r1) to the bottom-right cell (1, r3).

Implementation in DnaML: Assumptions

1. Each site in the sequence evolves independently.
2. Different lineages evolve independently.
3. Each site undergoes substitution at an expected rate which is chosen from a series of rates (each with a probability of occurrence) which we specify.
 1. All relevant sites are included in the sequence, not just those that have changed or those that are "phylogenetically informative".
 2. A substitution consists of one of two sorts of events:
 - (a) The first kind of event consists of the replacement of the existing base by a base drawn from a pool of purines or a pool of pyrimidines (depending on whether the base being replaced was a purine or a pyrimidine). It can lead either to no change or to a transition.
 - (b) The second kind of event consists of the replacement of the existing base by a base drawn at random from a pool of bases at known frequencies, independently of the identity of the base which is being replaced. This could lead either to a no change, to a transition or to a transversion.

Implementation in DnaML: Specification of evolutionary rates

For example, let's assume three possible rates of evolution, 1.0, 2.4, and 0.0, further prior probabilities of a site having these rates are 0.4, 0.3, and 0.3, further average patch length (number of consecutive sites with the same rate) is 2.0.

Average patch length: $1/(1-\lambda)$ with λ being the autocorrelation parameter
(in the above example $\lambda=0.5$)

Transition probability is then defined as:

(δ Kronecker symbol)

$$P_{ij} = \lambda\delta_{ij} + (1 - \lambda)f_j$$

λ close to 1: high degree of autocorrelation, $\lambda = 0$: no autocorrelation

Accordingly, the program will sum the likelihood over all possibilities, but giving less weight to those that (say) assign all sites to rate 2.4, or that fail to have consecutive sites that have the same rate.

Implementation in DnaML: Computation of tree topologies and optimal branch lengths

A) Search among tree topologies is done according to a stepwise addition followed by branch-swapping by nearest neighbor interchanges after each species is added

Optionally final round of branch swapping by subtree pruning and regrafting

Optionally multiple runs with different input orders of species

B) Derivatives of likelihood are obtained in recursive calculation along the sequence
These derivatives are used for estimating the branch length by use of the Newton-Raphson method.

Parameter settings for estimating site specific rate variations

```

1          50          100          150          200          250          300
Tachyglies ATGTTGATT TACTGTTC TTAGAGACT OCTGTACCA ACCTGTGGG GCATGTGAC
Didelphis .....C..TC.GA.G.....A.T.C.A.C..TC.TA.G..C.G
Tarsius .....C..C.GA.A.....G.C.....T.G.....CA.G..A.G...
Rattus .....C..C.A.G.....G.....T.A.T.G.C.....A.A.G.....
Cepus .....C..C.GA.G.....G.....G.T.C.....C.A.G.....A
Oryzolog. ....C..TCA.G.A.G.....T.....C.....CA.G.....T
Homo .....C..C.C.GA.G.....T.....C.T.T.G.....C.A.G.....
Lemur .....A.T.T.C.....T.....C.A.G.....C.A.G.....T
11111122 11122211 21111122 22111122 21111122 21111122
111 2 2 2 111 2 22 2 111 111 2 2

61          110          160          210          260          310          360
Tachyglies GTCAATGAC TCGTGGGA GGCCTTGGC AGCCTGCTG TGTGTACCC CTGAGCCAG
Didelphis ...TG.CC.GA.CT...T.....A.....C..T.....A.CC...A.CC
Tarsius ...GC.A..TG..T.....G.....A.....A.....A.....G.....T.G...
Rattus ...CCG...T.S.T.....G.....T.....T.....T.....T.....
Cepus ...GG...G.T.....C.T.....G.....T.....T.....T.....
Oryzolog. ...GG.A...G..T.....G.....A.....T.....A.....A.....
Homo ...GC...G.T.....T.....G.....C.....T.....A.....T.....
Lemur ...AG.GA..G..T.....T.G.....T.....G.....A.....A.....T.....
22222222 21111111 11111122 11111122 11111122 21111111
2 2 111 11 111 1 11 11 111 111 111 111 111 1

121          170          220          270          320          370          420
Tachyglies AGTTTTCG AATCTTGG TCACTTTC AGCCGCGATG CTGTAAGG AAACGCAAG
Didelphis .....T..T.GAG..T...T.....TTC.T.GC.....C..TO...T.T...
Tarsius .....T..C...T..G.....GTC.T.GC.....T..A..C.T.T...
Rattus .....A..T..TAG..T..G.....TCT..TC.....A..C.....Y..C.T...
Cepus .....T..GCA..T..G.....TCT..T.....T..AA.C.T.T...
Oryzolog. ....G...T..G.....TCT..AC.....T..A..C.TC.T...
Homo .....T..G...T..G.....TCT..T.....C..C.T...
Lemur .....G...T..G.....TCT..TC.....T..G...C.T...
11111111 22211111 11111111 22222211 11111122 21122111
111 11 222111 1 1111 2 2 11 11 1122 1 221111

181          230          280          330          380          430          480
Tachyglies GTCAAGGAC ATGTGGGA GTCTGTACC TCTTGGGG ATGCTGTA GAACCTGAC
Didelphis ...TC.A.....T.....T.....T..A..AG.C..C.TT.G...
Tarsius .....C.....CAAG.....A..G..TA.T..C.G.A..GG.T.T.G...
Rattus .....G.....CAAG.....A.A..G.....AAT.....GC..T.G...
Cepus .....G.....CAAG.....AG..TA.T.C.G.A.....TC..T.T...
Oryzolog. ....G...T..CAAG.....T.T.G..A.T..G.CT...TC...G...
Homo .....G...T..CAAG.....A.....GCT.C..TA.T..C..GC...C...
Lemur .....G...T..CAAG.....CT.C..TA.T..A.CT..C..TC...G...
11111111 11112211 11122222 21112221 12222112 22221111
11 1 111 11 111 222 111 1 1 2 22 111

341          390          440          490          540          590          640
Tachyglies AACCTGAG GACCTTTC CAACCTGAC GACCTGACT GGCACAGT GGCATGAGC
Didelphis .....G.....T..T..AT...T...T...G.....T.....T.....
Tarsius .....C...T..T.....T.....T..G.....T.....AT.....
Rattus .....C.....T..TC.E...T...C.....T.....T.....T...
Cepus .....A.....T..TC.E...T...G.....T.....T.....T...
Oryzolog. ....A.....T..CA..T..G.....T.....T.....T...
Homo .....C.....T..CA..T..G.....T.....T.....T...
Lemur .....C.....T..CA..T..G.....T.....T.....T...
11111111 11111111 22211111 11111111 11111111 11111111
11111111 1111 11 1111 1 1 111 1111 1 111 11111

361          410          460          510          560          610          660
Tachyglies CCGAGATT TCAATGCT GGTAACTG CTGTGTGG TCTGTGGGG TCACTGTAC
Didelphis ...T...C...GATG...G..TA.C.A.T..GA.CT.G...TGA...T...
Tarsius ...T...C...GG.T...C..TA..A.T..GA.T..CT..G..A..C..G...
Rattus ...T...C...GG.T...C..TA..A.T..GA.T..CT..G..A..C..G...
Cepus ...T...C...GG.T...C.....G..T..S.....T..G..CA...
Oryzolog. ...T...C...GG.T...C.....G..T..S.....T..G..T..A...
Homo ...T...C...GG.T...C.....TOT..G.....A.....T...
Lemur ...TC...C...C.T...C.....G..T..G...TGA.....T...
11111111 11122211 11111211 11112222 21112222 21111111
111 111 1 11 1 1 1 1 2 2 22 11 222 21 11

421          470          520          570          620          670          720
Tachyglies GCGTGGGC ACAAGTACA GTGA
Didelphis .....A.....
Tarsius .....T.....
Rattus .....T.....
Cepus .....GA.T...A...
Oryzolog. ....T.....
Homo .....T.....
Lemur .....T.....
11111111 11111111 1121
111 111 1 111 1 11 1 21

```

1. Site-specific categories representing the three codon positions
best combination of parameters: rates 1.0:0.6:2.7
2. Two regional rates
best combination of parameters: rates 1.0:8.0

Inferred frequencies of the two regions 0.75:0.25
Inferred parameter $\lambda = 0.5454$

Inferred rates of evolution for beta-hemoglobin DNA sequences in 8 mammalian species

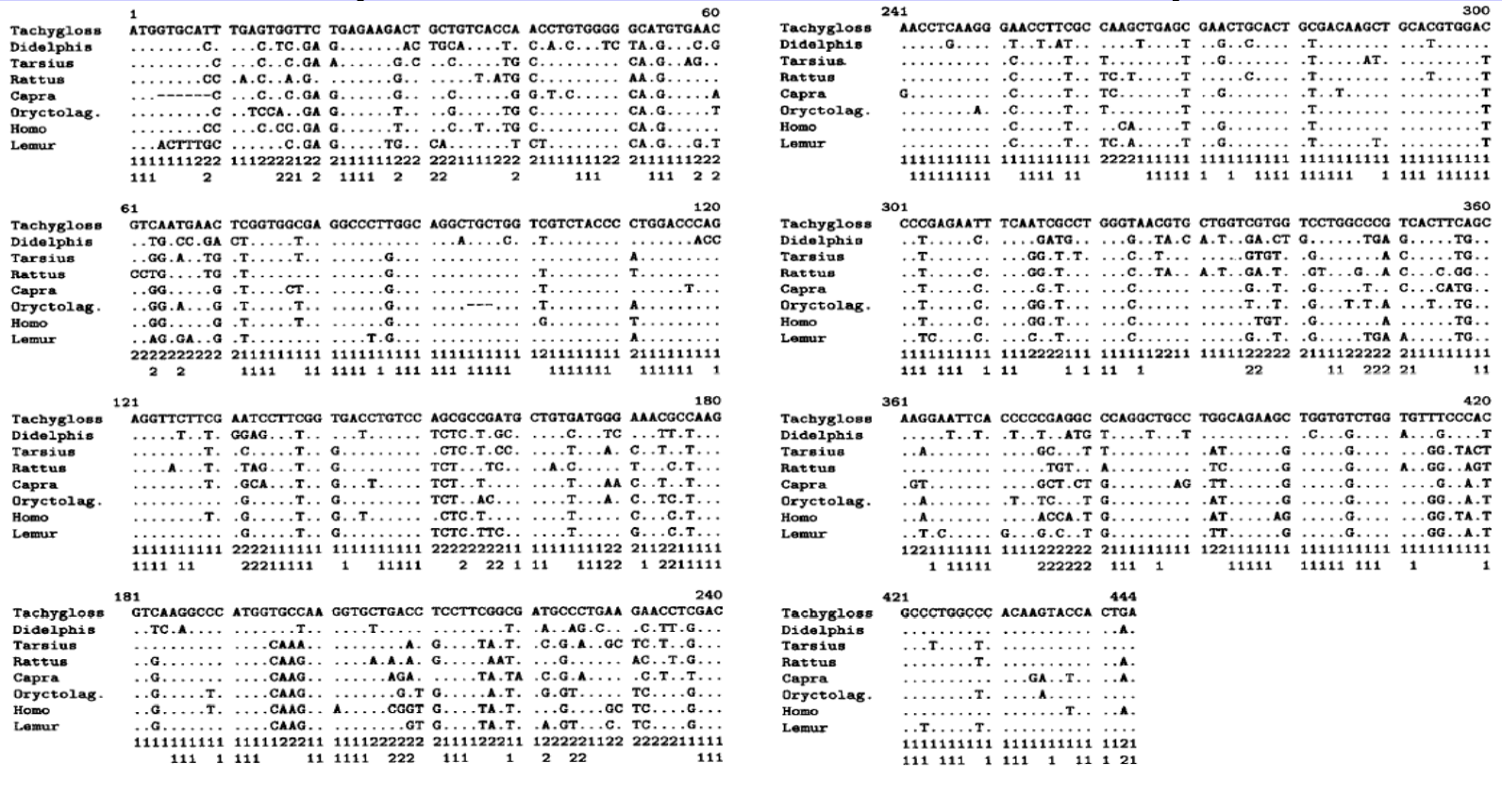


FIG. 3.—The β -hemoglobin coding sequences used in the data example. The dots are sites at which the sequence is the same as in *Tachyglossus*. The two rows of digits below each section of sequences are the regional rate categories inferred for each site. The first shows the single combination of regional rate assignments that contributes most to the likelihood. The second shows an assignment for each site provided that 95% or more of the likelihood is contributed by that rate being assigned to that site (otherwise no assignment is shown). Category 1 has the lower rate.

Inferred rates of evolution for beta-hemoglobin DNA sequences in 8 mammalian species

- Conserved codon for heme-binding histidines
- α -helical regions of proteins (4 out of 8 highlighted)

1		241		300	
Tachygloss	ATGGTGCATT TGAGTGGTTC TGAGAAGACT GCTGTCACCA ACCTGTGGGG GCATGTGAAC	Tachygloss	AACCTCAAGG GAACCTTCGC CAAGCTGAGC GAACTGCAC T GCGACAAGCT GCACGTGGAC		
DidelphisC.....C.TC.GA G.....AC TGCA...T. G.A.C...TC TA.G...C.G	DidelphisG.....T.T.AT...T...T...T...G...C.....T.....T.....T.....		
TarsiusC.....C.C.GA A.....G.C...C.....TG C.....CA.G...AG..	TarsiusC.....T...T.....T.....I...G.....T.....AT.....T.....		
RattusCC.A.C.A.G.....G.....T.ATG C.....AA.G.....	RattusC.....T...TC.T...T.....C.....T.....T.....T.....T.....		
CapraC.....C.C.GA G.....G.....G.T.C.....CA.G.....A	Capra	G.....C.....T...TC.....T...G.....T...T.....T.....T.....		
Oryctolag.C.....TCCA..GA G.....T.....G.....TG C.....CA.G.....T	Oryctolag.A.....C.....T...T.....T.....T.....T.....T.....T.....T.....		
HomoCC.....C.CC.GA G.....T.....C.T...TG C.....CA.G.....	HomoC.....T...CA.....T...G.....T.....T.....T.....T.....		
LemurACTTGC.....C.GA G.....TG...CA.....T CT.....CA.G...G.T	LemurC.....T...TC.A.....T...G.....T.....T.....T.....T.....		
1111111222	111222122 2111111222 2221111222 211111122 2111111222	1111111111	1111111111 2221111111 1111111111 1111111111 1111111111		
111	2 221 2 1111 2 22 2 111 111 2 2	1111111111	1111 11 11111 1 1 1111 111111 1 111 111111		
6		301		360	
Tachygloss	GTCAATGAAC TCGGTGGCGA GGCCTTGGC AGGGTGTGG TCGTCTACCC CTGGACCCG	Tachygloss	CCCGAGAATT TCAATGCGCT GGGTAACGTG CTGGTCGTGG TCCTGGCCCG TCACCTCAGC		
Didelphis	..TG.CC.GA CT.....T.....A.....C...T.....A.....CCG	Didelphis	..T.....C.....GATG...G...TA.C A.T...GA.CT G.....TGA G.....TG..		
Tarsius	..GG.A.TG T.....T.....G.....A.....A.....A.....	Tarsius	..T.....G...GG.T.T...C.T...GTGT...G.....A C.....TG..		
Rattus	CCTG...TG T.....G.....T.....T.....T.....T.....	Rattus	..T.....C.....GG.T...C.TA..A.T...GA.T...GT...G.A C...C.GG..		
Capra	..GG.....G T.....CT.....G.....T.....T.....T.....	Capra	..T.....C.....G.T...C.....TA..A.T...G.T...G.....T...C...CATG..		
Oryctolag.	..GG.A..G T.....T.....G.....T.....A.....A.....	Oryctolag.	..T.....C.....GG.T...C.....T...T...G...T.T.A...T...TG..		
Homo	..GG.....G T.....T.....G.....G.....T.....T.....	Homo	..T.....C.....GG.T...C.....T...T...G.....TGT...G.....A...TG..		
Lemur	..AG.GA..G T.....T.....G.....A.....A.....	Lemur	..TC.....C.....C.T...C.....G.....T...G.....TGA A.....TG..		
2222222222	2111111111 1111111111 1111111111 1211111111 2111111111	1111111111	1112222111 1111112211 1111122211 1111122222 2111122222		
2	2 1111 11 1111 1 111 11 11111 111111 111111 1	111 111 1 11	1 1 11 1 22 11 222 21 11		
12		361		420	
Tachygloss	AGGTTCTTCG AATCCTTCGG TGACCTGTCC AGCGCGGATG CTGTGATGGG AAACGCCAAG	Tachygloss	AAGGAATTC A CCCCAGGCG CCAGGCTGCC TGGCAGAAGC TGGTGTCTGG TGTTCGCC		
DidelphisT.T.GGAG...T.....T.....TCTC.T.CC...C...TC...TT.T...	DidelphisT.T...T.T...ATG T...T...T...C...G...A...G...T...		
TarsiusT...C...T...G.....TCTC.T.CC...T...A.C.T.T...	Tarsius	..A.....G...GC...T T.....AT.....G.....G.....GG.TACT		
RattusA...T...TAG...T...G.....TCT...TC...A.C...T...C.T...	RattusT...G...T...A.....TC.....G.....G...A...GG...AGT		
CapraT...GCA...T...G...T...TCT...T...T...AA C...T...T...	Capra	..GT.....G...GCT.CT G.....AG...TT.....G.....G...A...T		
Oryctolag.G...T...T...G...T...TCT...AC...T...A.C...TC.T...	Oryctolag.	..A.....T...TC...T G.....AT.....G.....G...GG...A...T		
HomoT...G...T...G...T...TCTC.T.CC...T...C...C.T...	Homo	..A.....A...ACCA.T G.....AT...AG...G.....G...GG...TA.T		
LemurG...T...T...G...T...TCTC.TTC...T...T...G...C.T...	Lemur	..T.C...G...G.C.T G.....TT.....G.....G...GG...A...T		
1111111111	2221111111 1111111111 2222222211 1111111122 2112211111	1221111111	1111222222 2111111111 1221111111 1111111111 1111111111		
1111 11	22211111 1 11111 2 22 1 11 1122 1 2211111	1 11111	222222 111 1 1111 1111 111 1		
181		421		444	
Tachygloss	GTCAAGGCC ATGGTGCCAA GGTGCTGACC TCCTTCGGCG ATGCCCTGAA GAACCTCGAC	Tachygloss	GCCTGGGCC ACAAGTACCA CTGA		
Didelphis	..TC.A.....T.....T.....T.....A.AG.C...C.TT.G...	DidelphisA.....		
TarsiusCAAA.....A.G...TA.T...C.G.A...GC TC.T...G...	TarsiusT...T.....		
RattusG.....CAAG.....A.A.A.G...AAT...G.....AC...T.G...	RattusT.....A.....		
CapraG.....CAAG.....AGA...TA.TA.C.G.A...C.T.T...	CapraGA...T...A.....		
Oryctolag.G...T...CAAG.....G.T G...A.T...G.GT...TC...G...	Oryctolag.T.....A.....		
HomoG...T...CAAG...A...CGGT G...TA.T...G...GC TC...G...	HomoT.....T...A.....		
LemurG.....CAAG.....GT G...TA.T...A.GT...C.TC...G...	LemurT.....T.....		
1111111111	1111122211 111122222 2111122211 122221122 2222211111	1111111111	1111111111 1121		
111 1 111	11 1111 222 111 1 2 22 111	111 111 1 111	1 11 1 21		

All other regions: non-helical have markedly fewer high rates than low
 => Helical regions are under less constraint than the non-helical ones

*An example: beta-hemoglobin DNA sequences
in 8 mammalian species*

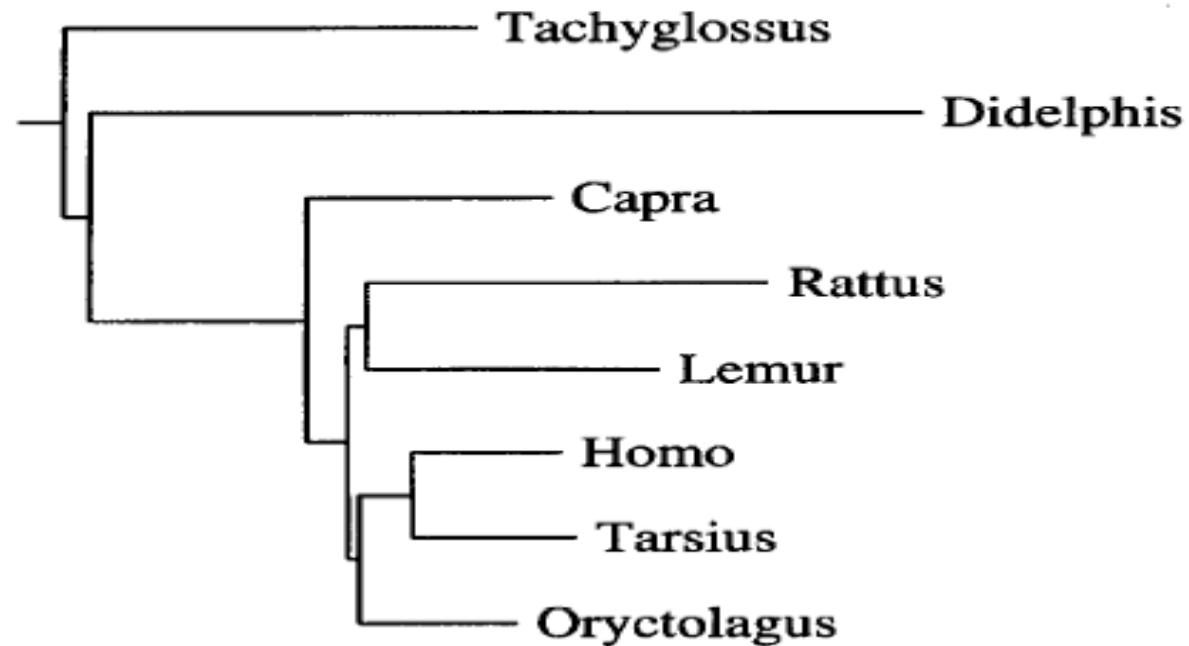


FIG. 2.—The phylogeny estimated for the eight hemoglobin β DNA sequences. The shorter branches are not statistically significant.