Statistical Bioinformatics, Makerere
Lecture : Phylogenetics
Timo Koski

TK

02.08.2018

KTH Matematik

Phylogeny, Phylogenetics

Phylogeny refers to the evolutionary relationships among organisms; the patterns of lineage branching produced by the evolutionary history of the organisms being considered. Phylogenetics is the field of biology that deals with phylogeny.
A common practice today is to use molecular sequences as representatives of the organisms.

KTH Matematik

Three steps

How does a geneticist reconstruct molecular phylogenetic relationships ?
There are three steps (1)-(3)

(1) sequence selection and (multiple) alignment to determine the sites and site-by-site homologies and to detect DNA or amino acid differences.

(2) mathematical (probabilistic) modelling describing the evolution in time of the sequences.

(3) finding by a statistical method a tree topology and the branch lengths that best describe the sequences' phylogenies.

KTH Matematik

Site & Character

Each site obtained by a multiple alignment defines a *chararacter*, so that sequences having the same nucleotide (or amino acid) at that site exhibit the same *state of the character*.

KTH Matematik

Homology

A set S of organisms (taxa, species) in S can be represented by the values each species attains for each of the characters. Thereby the starting data after step (1) is a matrix-block, the columns of which recapitulate some aspect shared among the organisms by common descent. This attribute due to a common ancestor is called *homology*.

KTH Matematik

Multiple Alignment

S, species	character	character	character	character	character
Human	G	A	T	A	C
Gorilla	A	C	T	G	C
Orangutang	G	C	T	G	T
Chimpanzee	G	C	T	C	T
Gibbon	G	A	T	G	T

KTH VETENSKAP OCH KONST
KTH Matematik

Inference of Phylogenies (Step (3))

A phylogeny shows how organisms are related to the common ancestors. A phylogenetic tree is a way of coding data about evolutionary history. The task of molecular phylogenetics is to reconstruct the path of evolution from observation of molecular sequences of current species.

KTH VETENSKAP OCH KONST
KTH Matematik

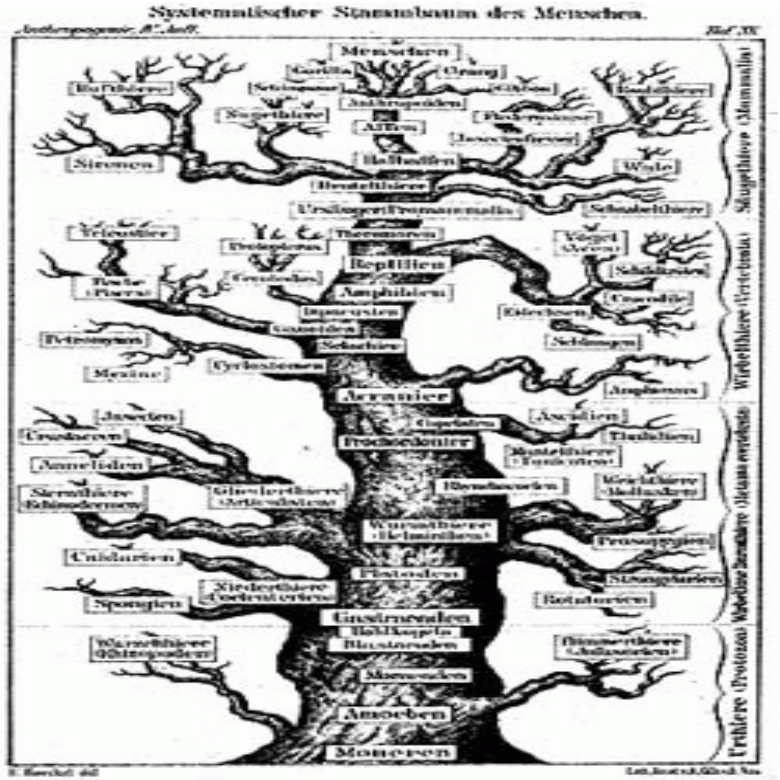
Inference of Phylogenies (Step (3))

The tree of life shows the path evolution took to get to the current diversity of life. It also shows how to ascertain the genealogy of disparate living organisms. This is evidence for evolution only in that we can construct such trees at all.

KTH Matematik

Inference of Phylogenies: trees

A tree of life from Ernst Haeckel's "Generelle Morphologie"[1], 1891.



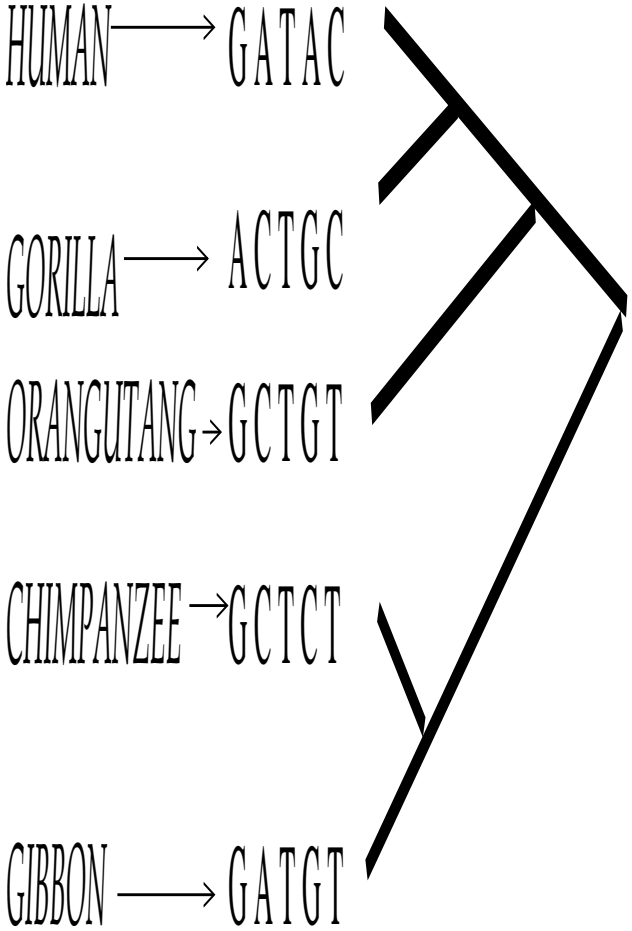
[1] The picture is avaiblable at
http://www.athro.com/evo/rhetoric.html

Inference of Phylogenies: trees

There are five hominoid species assigned to the leaves of the tree in the picture.

Inference of Phylogenies: trees

We think in fact of assigning each row in the matrix-block to a leaf.

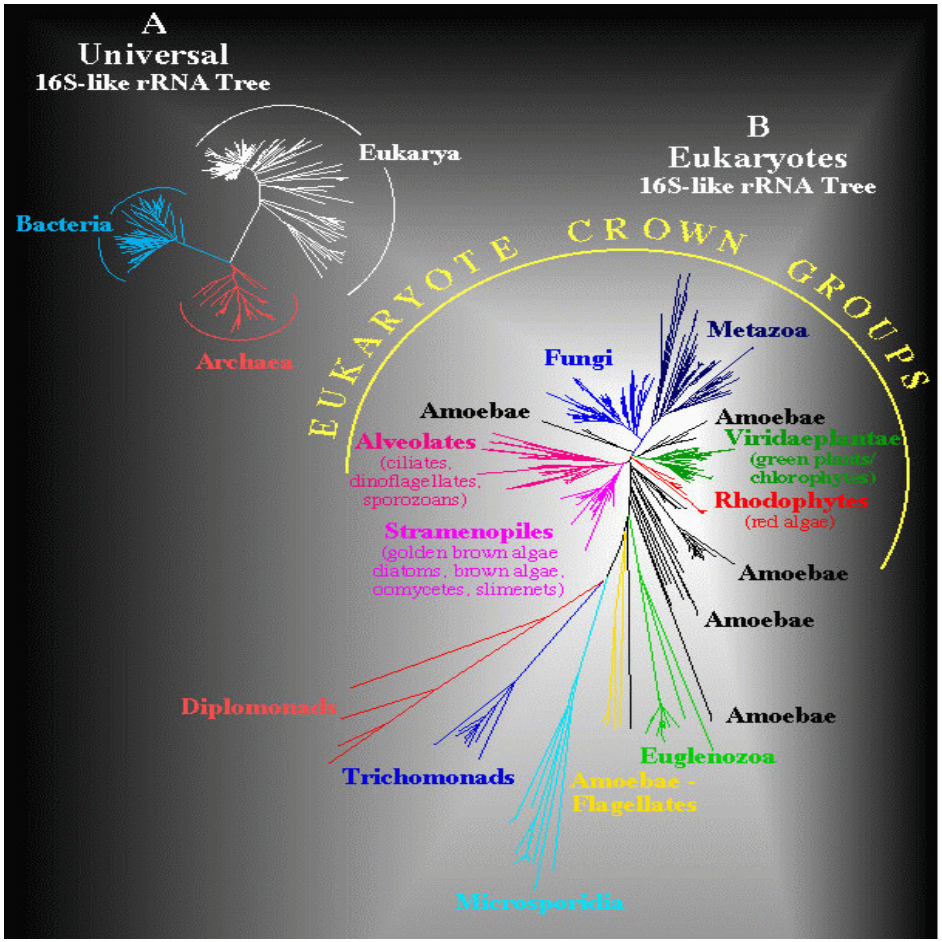


Phylogenetic Trees Based on 16s rDNA

Universal and Eukaryote Phylogenetic Trees: Structural similarites for 900 sites that can be unambiguously aligned in a data set of more than 800 eukaryotes and 100 prokaryotes were computed and converted to evolutionary distances using the Kimura two parameter model. The Neighbor-Joining method was used to infer the unrooted universal phylogeny shown in A. A similar computation was used to infer an unrooted phylogeny in B for diverse eukaryotes but this analysis is based upon comparisons of 1200 positions that can be unambiguously aligned. Evolutionary distances in these trees are proportional to length of line segments separating taxa.

KTH Matematik

Phylogenetic Trees Based on 16s rRNA



KTH Matematik

Evolution

Homologous sequences diverge from a common ancestor because mutations, insertions and deletions, duplications and transpositions of single nucleotides and of blocks of nucleotides occur. Some fraction of these are fixed into to the evolving population by selection and chance.

KTH Matematik

Evolution & Models

A completely general model of this process should take into account insertions and deletions, duplications and transpositions, but this course deals only with models of single site molecular substitutions. In this course modeled in terms of continuous-time Markov chains.

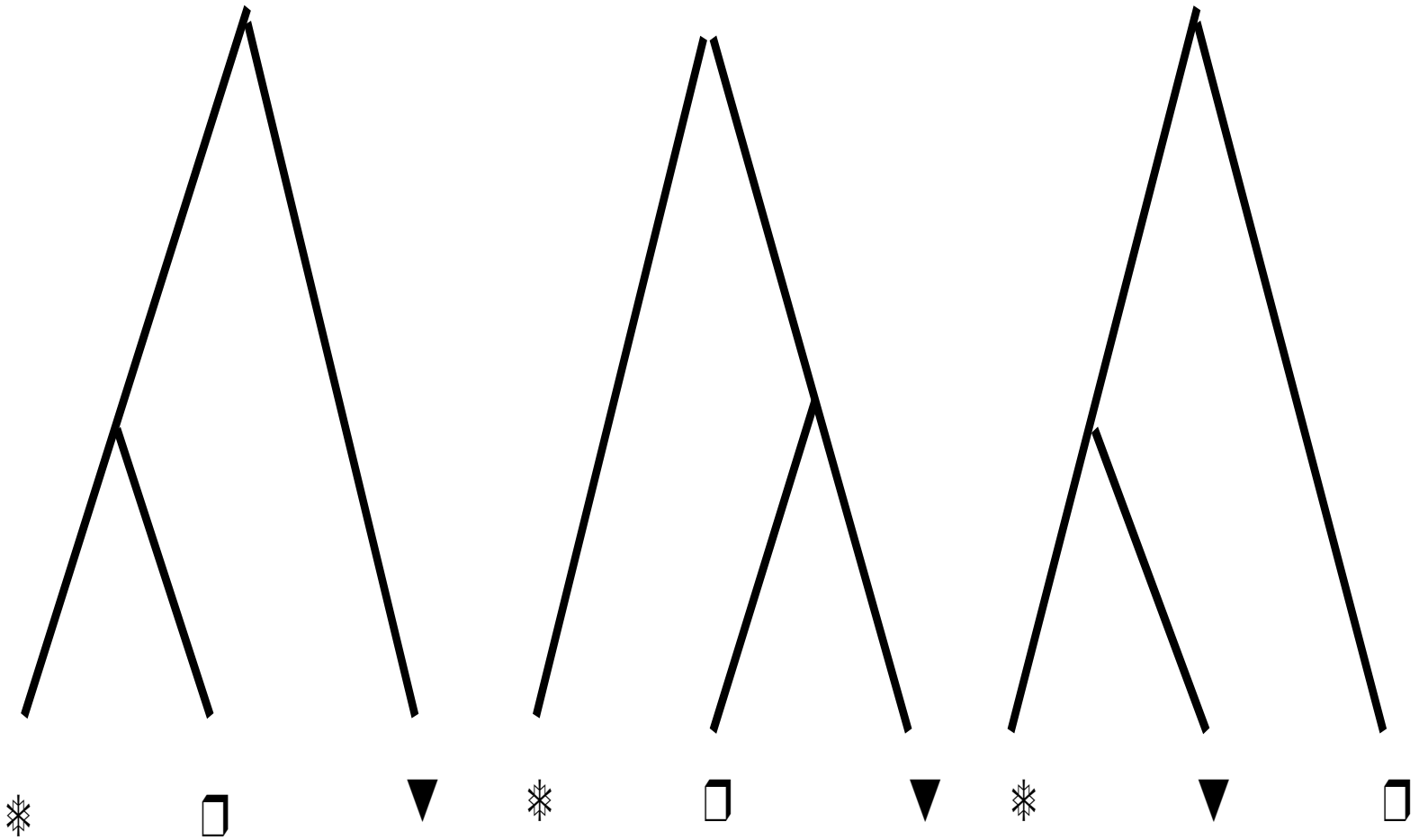
KTH Matematik

Topology

The pattern of branching connecting the species is called a *topology*. There are, as can be shown, 120 different (rooted) topologies for five species. The empirical data consists of fragments of genomes or of genes common to all of the respective species assigned to the leaves. The nodes at the branchings of the tree correspond to ancestral species not known to us.

KTH VETENSKAP OCH KONST
KTH Matematik

Topologies of phylogenetic rees



KTH Matematik

Inference of Phylogenies

There are several approaches to computational reconstruction of phylogeny. We take the task of reconstruction to mean that we infer the tree topology from sequence data by means of probabilistic models of nucleotide substitution.

KTH Matematik

Evolution & Probability

We strive first at computing the probability of one column or probability of a **site pattern** (the states of one character), e.g., the probability of the right-most column $p(CCTTT)$.

S, species	character	character	character	character	character
Human	G	A	T	A	C
Gorilla	A	C	T	G	C
Orangutang	G	C	T	G	T
Chimpanzee	G	C	T	C	T
Gibbon	G	A	T	G	T

KTH Matematik

Evolution & Probability

Then we multiply over all sites to get the probability of our molecular data

$$p(GAGGG)p(ACCCA)p(TTTTT)p(AGGCG)p(CCTTT)$$

S, species	character	character	character	character	character
Human	G	A	T	A	C
Gorilla	A	C	T	G	C
Orangutang	G	C	T	G	T
Chimpanzee	G	C	T	C	T
Gibbon	G	A	T	G	T

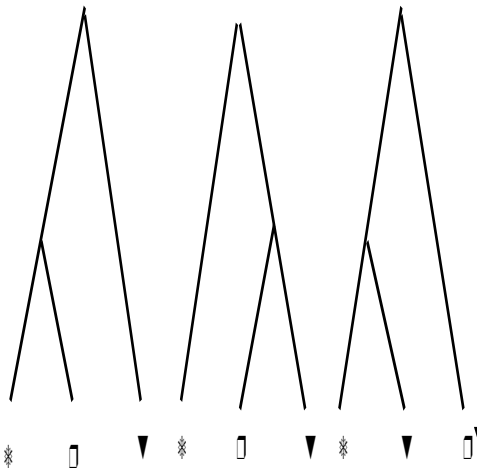
KTH VETENSKAP OCH KONST
KTH Matematik

Evolution & Probability

$$p(GAGGG)p(ACCCA)p(TTTTT)p(AGGCG)p(CCTTT)$$

The question is to find some way of computing these probabilities. We need a model.

Evolution & A Probability Model



We think of a Markov chain of mutations evolving as a Markov chain on the branches of the tree.

Neutral theory of evolution

The theory of evolution implies that extant life on earth has descended by divergence from a single common ancestor. Hence the nucleic acids and proteins will reflect those of the ancestor. The process of mutation in gene sequences is going on in a continuous fashion but leads seldom to what most of us in a conventional sense associate with (Darwinian) evolution: changes in the phenotype by adaptation to environment.

KTH VETENSKAP OCH KONST
KTH Matematik

Neutral theory of evolution

But if most molecular changes do not affect survival, natural selection ignores them. These selectively neutral molecules evolve, this is the hypothesis, like the ticking of a clock.

KTH VETENSKAP OCH KONST
KTH Matematik

Neutral theory of evolution and the molecular clock

With the clock having been set we can take a gene common to all of the species under study and examine the variations in a given molecule among species. By counting these variations we may try to estimate when a species branched off its family tree.

KTH VETENSKAP OCH KONST
KTH Matematik

The Molecular Clock Hypothesis

- All the mutations occur in the same rate in all the branches of the tree of life.
- The rate of the mutations is the same for all sites along the sequence.

KTH Matematik

The Molecular Clock Hypothesis

The clock hypothesis is more precisely formulated with a simple statistical model, a continuous-time Markov chain, that describes events of substitution occurring at random times with constant rates

KTH VETENSKAP OCH KONST
KTH Matematik

The Molecular Clock Hypothesis

In (Zuckerkandl and Pauling 1965, Zuckerkandl et.al. 1971, Neyman 1970) the disturbances that may result in failure to restore DNA sequences correctly arise according to the probability of no disturbance (substitution) being

$$e^{-\mu t},$$

and the probability of at least one disturbance being

$$1 - e^{-\mu t}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

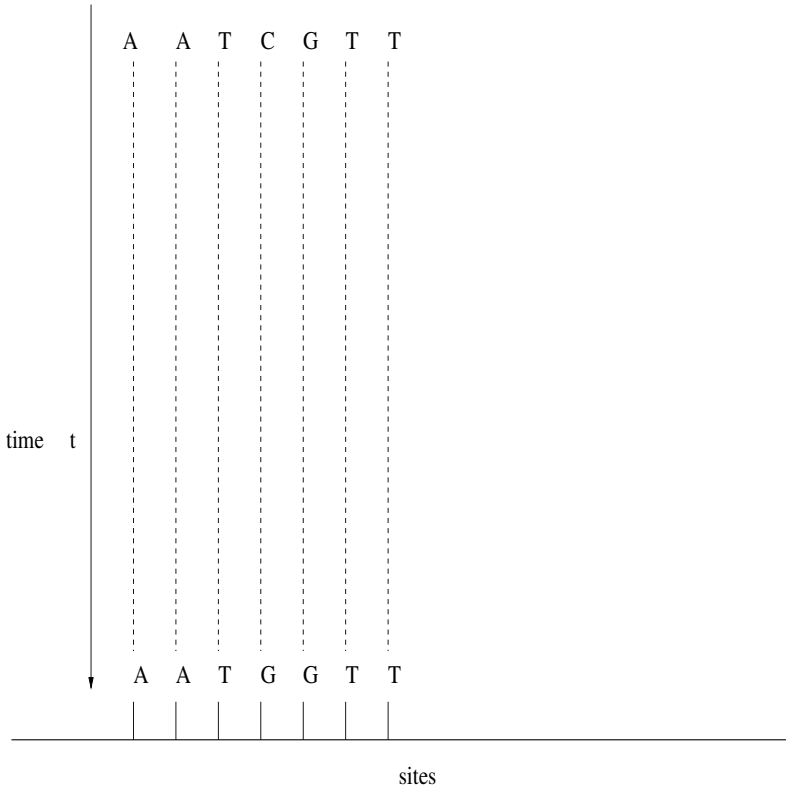
The Hypotheses

1) All nucleotide sites change independently.
2) All nucleotide sites vary equiprobably.
3) The substitution rate is constant over time and in different evolutionary lineages.
4) The base composition is at equilibrium.
5) The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.

KTH Matematik

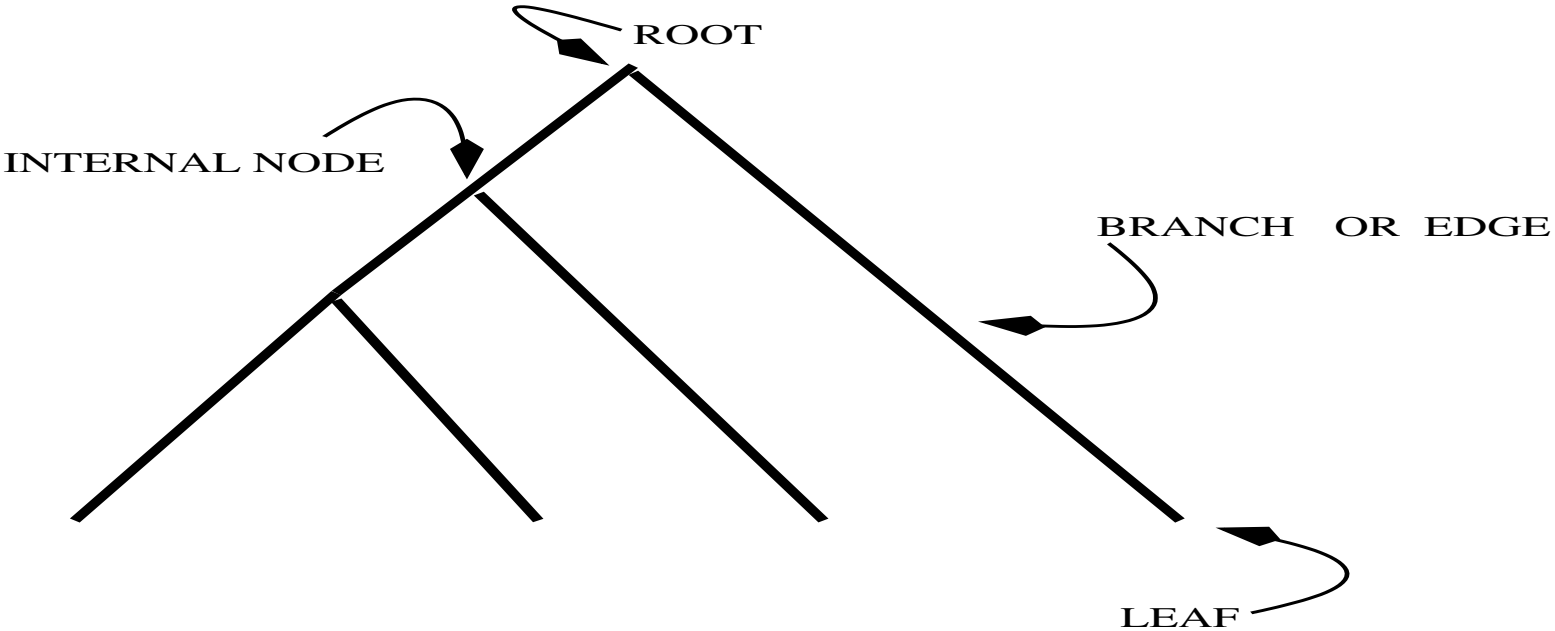
The Molecular Clock Hypothesis

We are thus dealing with a 'space-time' plane, in which a number of independent mutation process evolve in a parallel fashion.



Terminology on Trees

Each node of a tree represents a species, with edges or branches denoting immediate ancestor-descendant relationships. The leaves represent extant species, the internal nodes represent the ancestral species.



Binary Trees

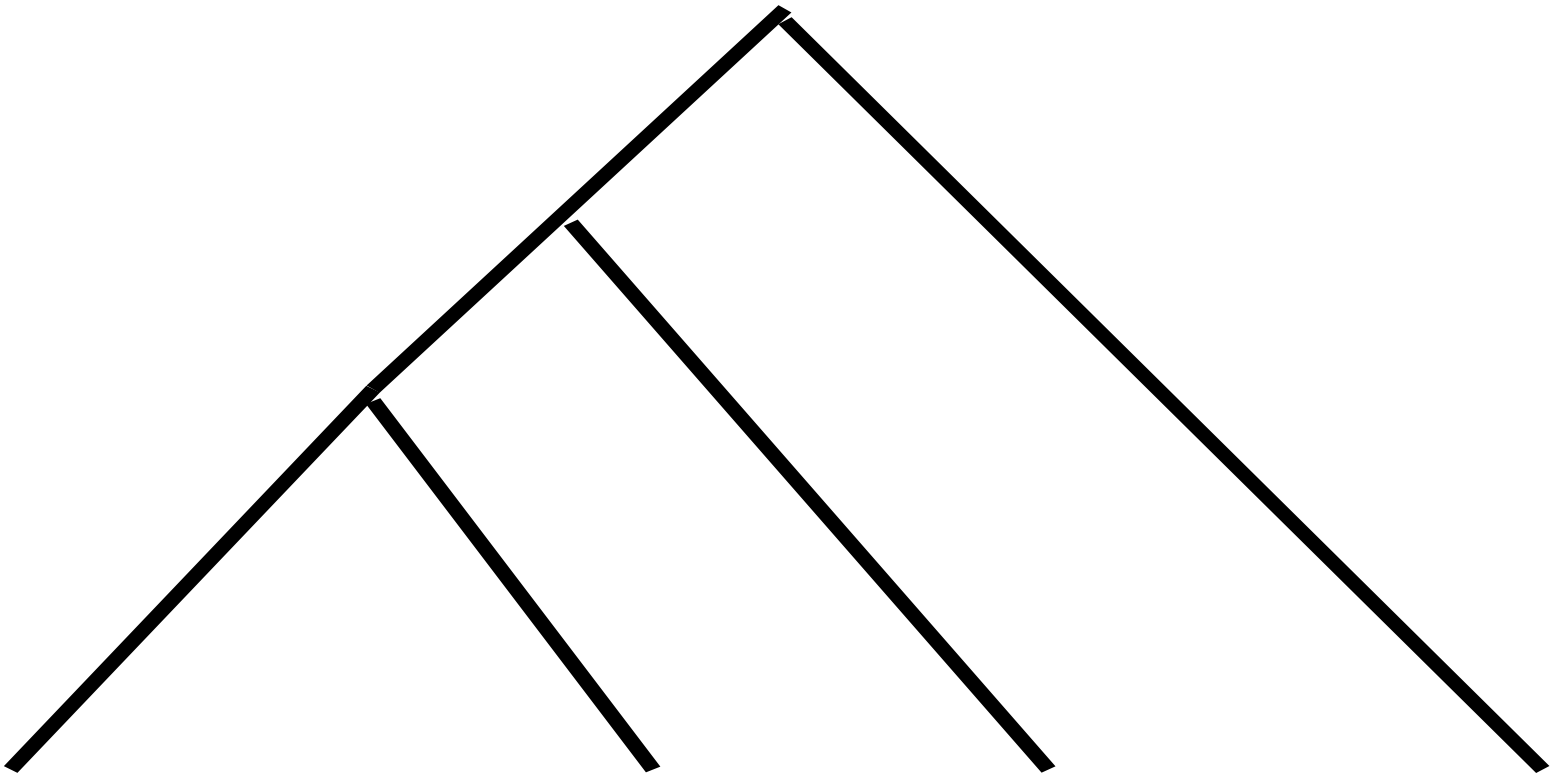
In a rooted tree the root represents the common ancestor, from which all other species under consideration are descended. Since we are primarily interested in speciation events, where the tree branches, we do not allow nodes of degree 2, except possibly in the root. Hence the trees under consideration here will be binary.

The degree of a node is the number of branches pointing to it and leaving it, with direction taken in the sense of moving away from the root.

KTH Matematik

Binary Trees: A Formal Definition

A tree $\mathcal{T}$ is *binary*, if for every internal node v has degree 3.



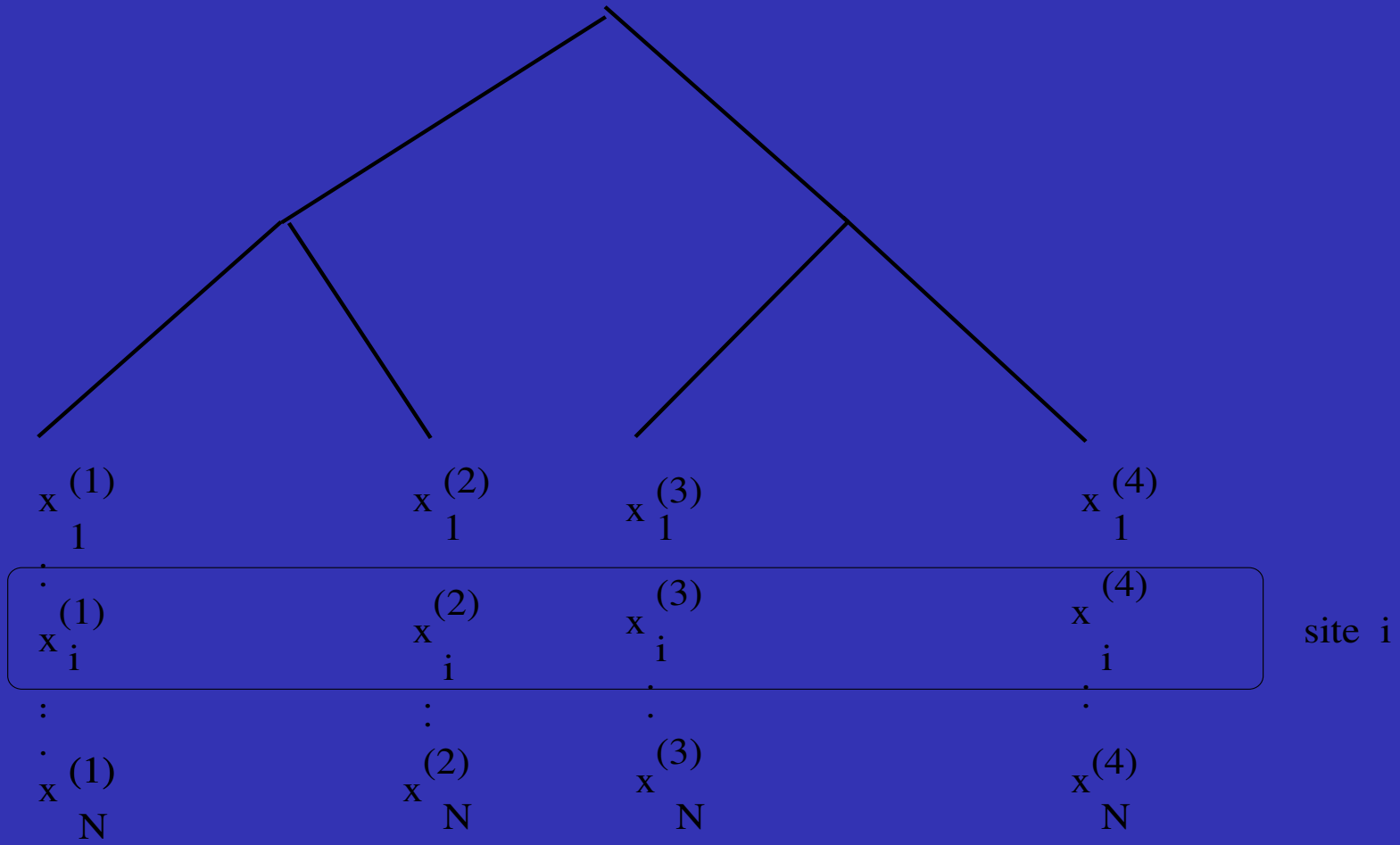
KTH Matematik

A rooted binary phylogenetic tree $\mathcal{T}$

We view the edges E as being directed away from the root r and then regard $\mathcal{T}$ as describing the evolution of the extant species from a common hypothetical ancestral species at the root r. The other interior nodes correspond to further hypothetical ancestral species.

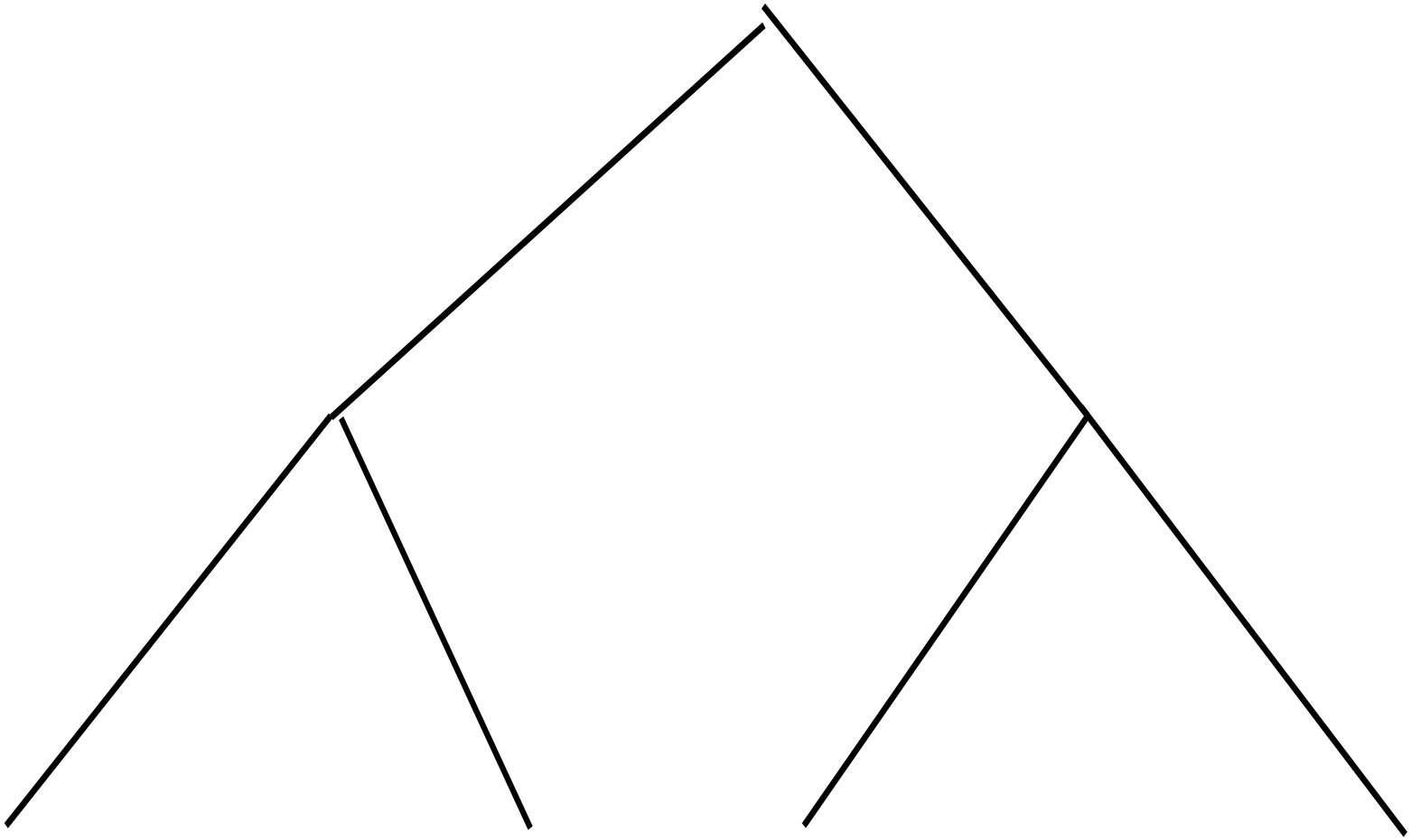
KTH VETENSKAP OCH KONST
KTH Matematik

Site patterns assigned to $\mathcal{T}$



A rooted binary phylogenetic tree $\mathcal{T}$

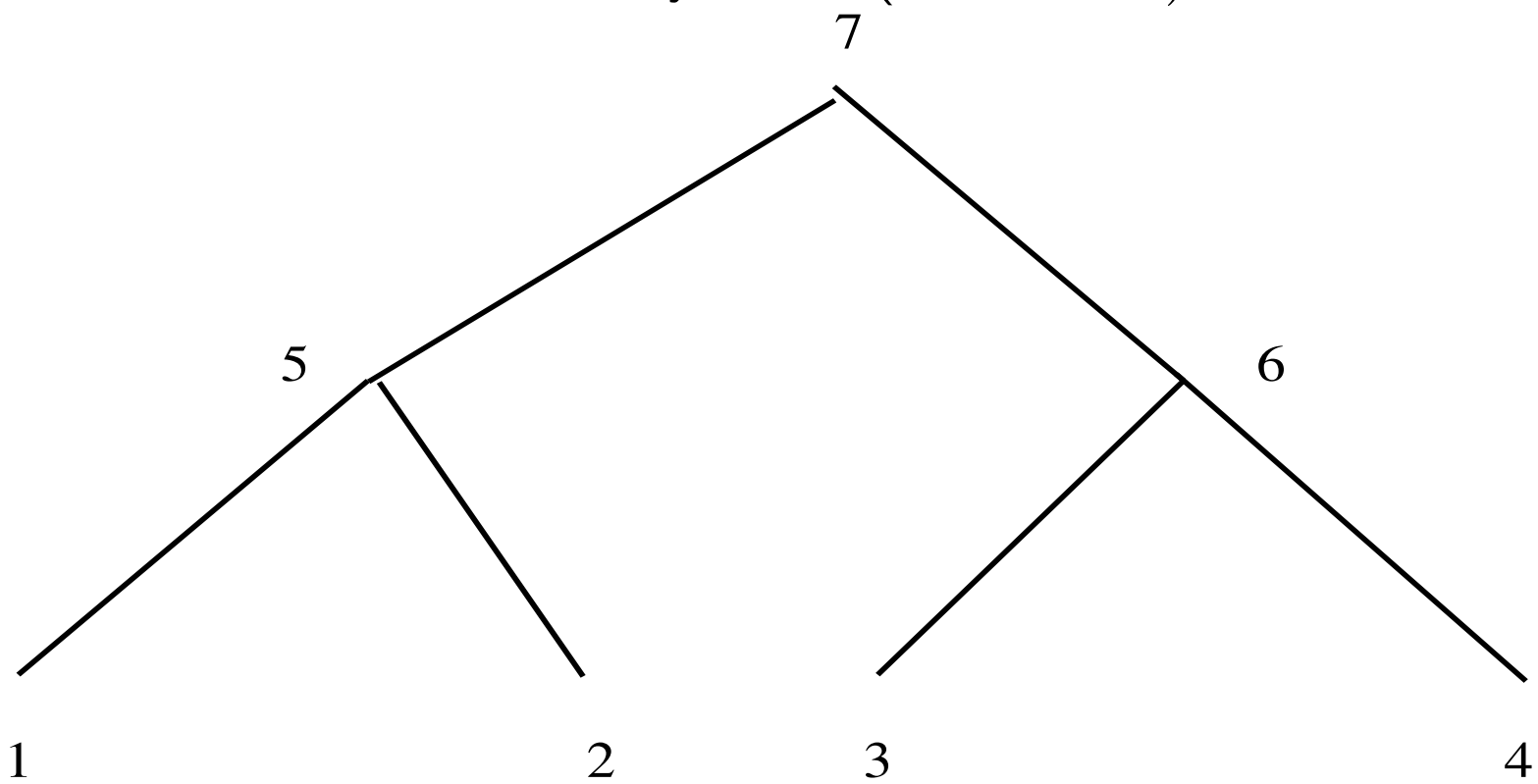
Let $\mathcal{T}$ be a binary rooted tree as in the figure



KTH Matematik

A rooted binary phylogenetic tree $\mathcal{T}$

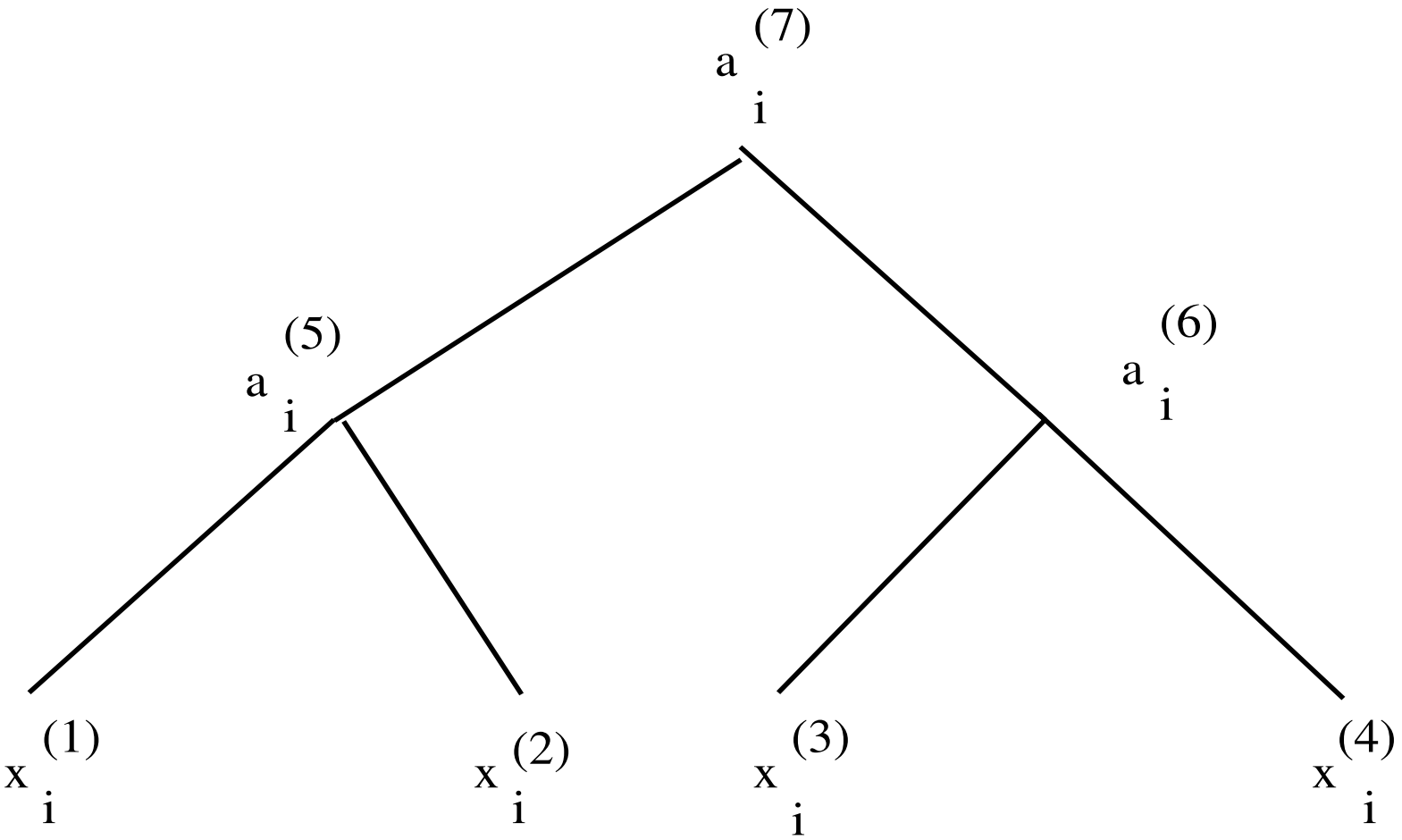
The leaves are $\mathbf{l} = \{1, 2, 3, 4\}$, the internal nodes are labelled by $\{5, 6\}$, and the root r is labelled by $r = 7$ $(= 2 \cdot 4 - 1)$.



KTH Matematik

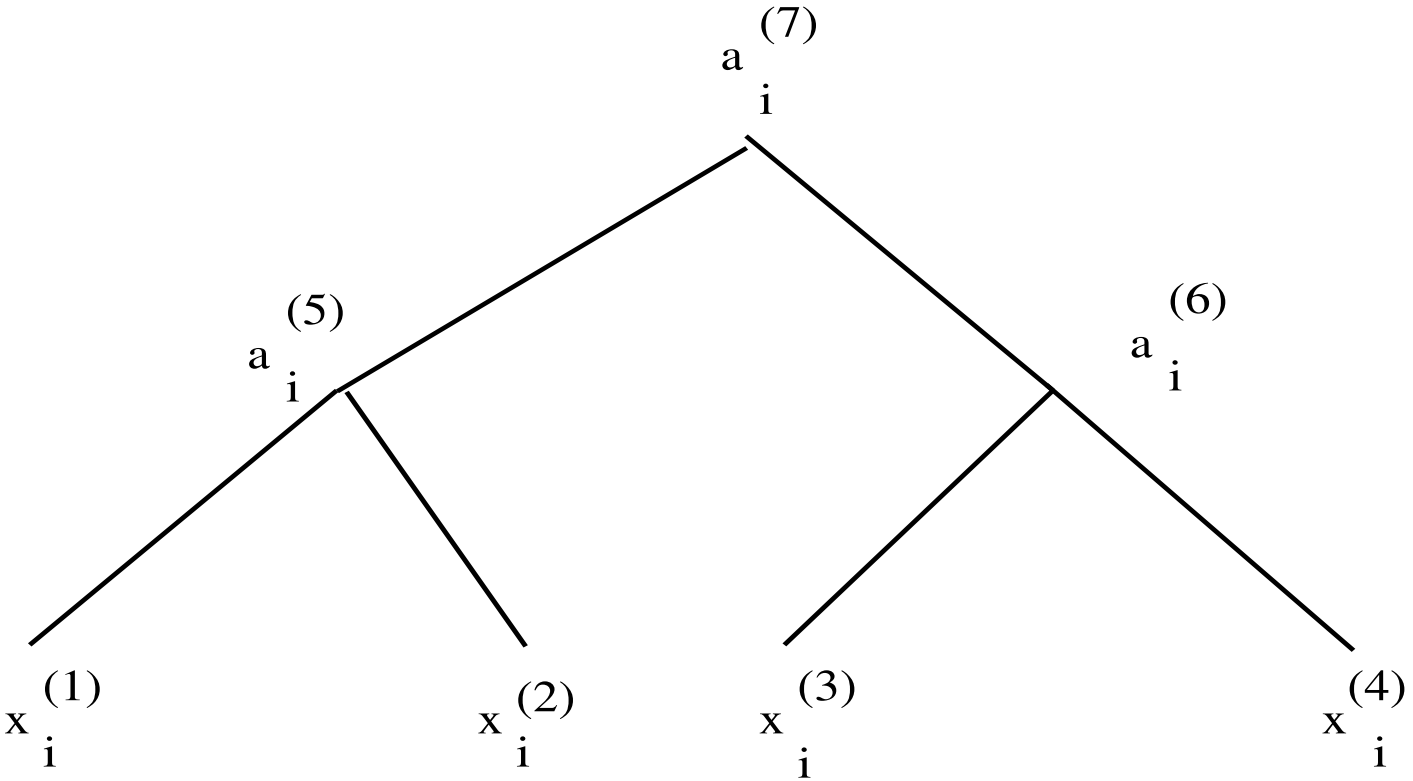
A rooted binary phylogenetic tree $\mathcal{T}$

Label the nodes as in the figure



A rooted binary phylogenetic tree $\mathcal{T}$

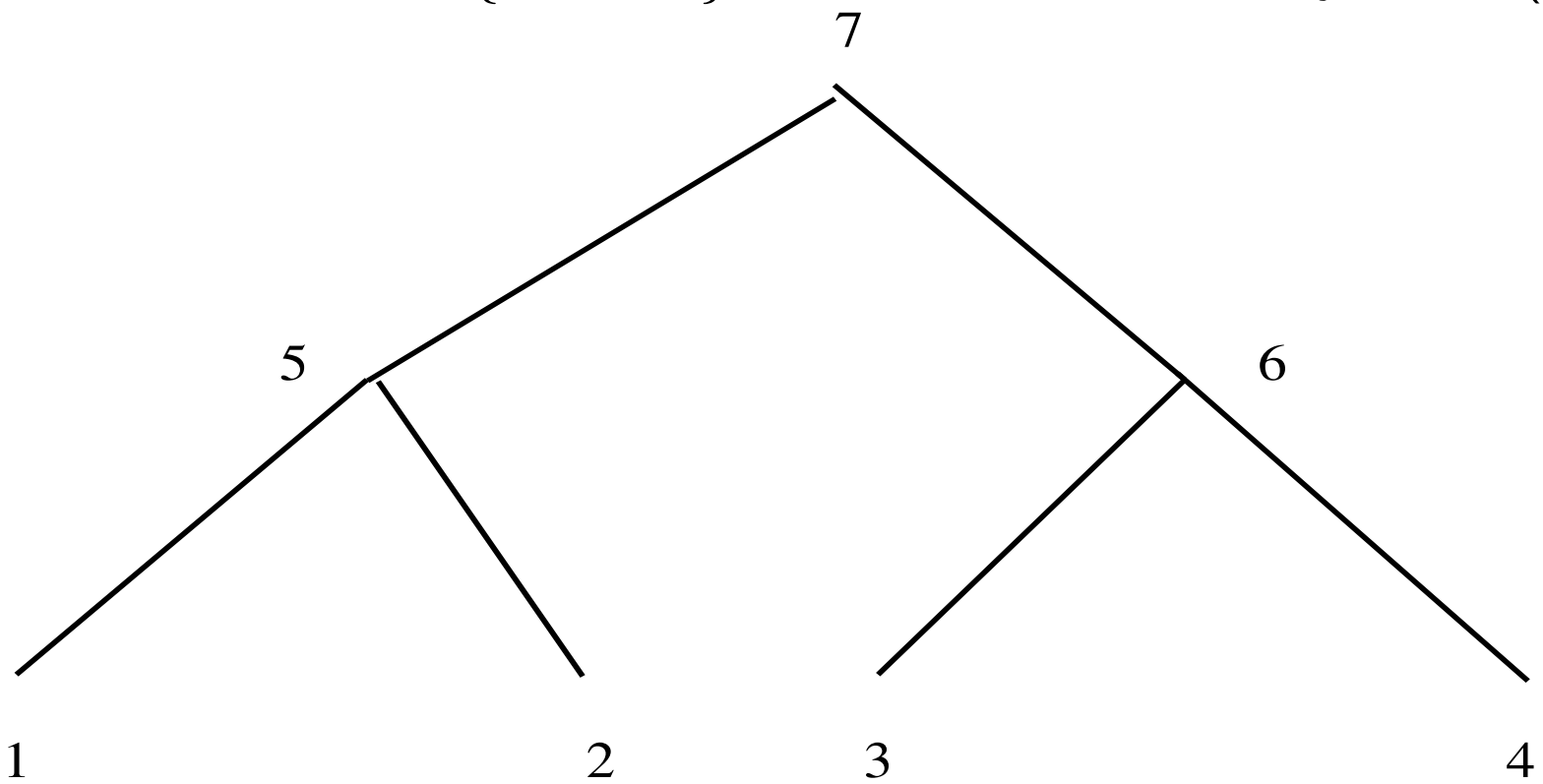
$x_i^{(l)}$ is the base at site i for the extant species l, $a_i^{(k)}$ is the base at site i for the ancestral species k.



KTH Matematik

A rooted binary phylogenetic tree $\mathcal{T}$

The leaves are $\mathbf{l} = \{1, 2, 3, 4\}$, the root r is labelled by $r = 7$ $(= 2 \cdot 4 - 1)$.



A rooted binary phylogenetic tree $\mathcal{T}$

The joint probability of the bases (ancestral and extant) at site i in the tree is computed as

$$P\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)}, a_i^{(5)}, a_i^{(6)}, a_i^{(7)} | \mathcal{T}\right) =$$

$$\pi_{a_i^{(7)}} \cdot P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot P\left(x_i^{(1)} | a_i^{(5)}\right) \cdot P\left(x_i^{(2)} | a_i^{(5)}\right) P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot P\left(x_i^{(3)} | a_i^{(6)}\right) \cdot P\left(x_i^{(4)} | a_i^{(6)}\right).$$

π is the equilibrium probability.

KTH VETENSKAP OCH KONST
KTH Matematik

A Markov Chain on a rooted binary phylogenetic tree $\mathcal{T}$

Note the assumption inherent in this: the probabilities are of the form

$$P(\text{ child } \mid \text{ parent})$$

one for each branch of the tree, all are multiplied together. This implies a Markov property on the binary tree under consideration.

KTH Matematik

A rooted binary phylogenetic tree $\mathcal{T}$

We eliminate the unknown ancestral sequences $\{\mathbf{a}^{(5)}, \mathbf{a}^{(6)}, \mathbf{a}^{(7)}\}$, by marginalization. The subscript $_{\text{av}}$ refers to the resulting probability of a site pattern:

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \mathcal{T}\right)$$

$$= \sum_{a_i^{(7)} \in \mathcal{X}} \sum_{a_i^{(6)} \in \mathcal{X}} \sum_{a_i^{(5)} \in \mathcal{X}} \pi_{a_i^{(7)}} \cdot \Phi$$

$$\Phi = P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot P\left(x_i^{(1)} | a_i^{(5)}\right) \cdot P\left(x_i^{(2)} | a_i^{(5)}\right) P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot P\left(x_i^{(3)} | a_i^{(6)}\right) \cdot P\left(x_i^{(4)} | a_i^{(6)}\right).$$

KTH Matematik

Peeling

Apply the distributive law of summation to move each of the summations as deep into the expression as possible.

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \mathcal{T}\right)$$

$$= \sum_{a_i^{(7)} \in \mathcal{X}} \pi_{a_i^{(7)}} \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot \left[P\left(x_i^{(1)} | a_i^{(5)}\right)\right] \cdot \left[P\left(x_i^{(2)} | a_i^{(5)}\right)\right] \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot \left[P\left(x_i^{(3)} | a_i^{(6)}\right)\right] \cdot \left[P\left(x_i^{(4)} | a_i^{(6)}\right)\right] \right\}.$$

KTH VETENSKAP OCH KONST

KTH Matematik

Peeling

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \mathcal{T}\right)$$

$$= \sum_{a_i^{(7)} \in \mathcal{X}} \pi_{a_i^{(7)}} \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} P\left(a_i^{(5)} | a_i^{(7)}\right) \cdot [P\left(x_i^{(1)} | a_i^{(5)}\right)] \cdot [P\left(x_i^{(2)} | a_i^{(5)}\right)] \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}\right) \cdot [P\left(x_i^{(3)} | a_i^{(6)}\right)] \cdot [P\left(x_i^{(4)} | a_i^{(6)}\right)] \right\}.$$

Then the pattern of the parentheses, $\{[][]\}\{[][]\}$ has an exact relationship to the topology of the tree. There is one probability for each subtree.

KTH VETENSKAP OCH KONST

KTH Matematik

Peeling

The expression can be evaluated by working outwards from the innermost parentheses. This corresponds to starting from the leaves and working towards the root, also known as *post-order traversal*.

KTH Matematik

Post-order traversal

Define $L_s(v)$ as the probability of everything that is observed from node v upwards (away from root), when the base at node v is $s \in \mathcal{X}$.

$$L_{a_i^{(7)}}(7) = \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} P\left(a_i^{(5)} | a_i^{(7)}\right) L_{a_i^{(5)}}(5) \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}\right) L_{a_i^{(6)}}(6) \right\}.$$

KTH Matematik

Post-order traversal

The conditional likelihoods $L_{a_i^{(6)}}(6)$ and $L_{a_i^{(5)}}(5)$ have been computed at previous steps of the same algorithm. At the leaves l we take

$$L_s(l) = \begin{cases} 1 & s = x_i^{(l)} \\ 0 & \text{otherwise.} \end{cases}$$

Post-order traversal

The final step of the peeling algorithm is

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \mathcal{T}\right) = \sum_{a_i^{(7)} \in \mathcal{X}} \pi\left(a_i^{(7)}\right) \cdot L_{a_i^{(7)}}(7).$$

Probability of $\mathcal{T}$

The joint probability of sequences on the leaves of the tree is

$$\prod_{i=1}^{N} P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} | \mathcal{T}\right),$$

which means that we treat the site patterns as outcomes of independent random variables.

KTH Matematik

Substitution probabilities

The substitution probabilities

$$P\left(a_i^{(j)}|a_i^{(\tau(j))}\right) \quad \text{and} \quad P\left(x_i^{(j)}|a_i^{(\tau(j))}\right),$$

are expressed for time-reversible models:

$$\mathbf{P}_v = e^{u_v \mathbf{Q}_*},$$

with

$$u_v = \mu \cdot t_v$$

as the Markov transition matrix on the branch $(\tau(v), v)$. v is a node, $\tau(v)$ is the parent node of v.

KTH VETENSKAP OCH KONST
KTH Matematik

Substitution probabilities: **F 81**

In **F 81** we get thus

$$\mathbf{P}_v = e^{-u_v}\mathbf{I} + \mathbf{A}\left(1 - e^{-u_v}\right),$$

as has been shown in earlier lectures.

KTH Matematik

Substitution probabilities: **F 81**

We have (suppressing the site index)

$$P\left(a^{(v)}|a^{(\tau(v))}\right) = e^{-u_v}\delta_{a^{(\tau(v)},a^{(v)}} + \left(1-e^{-u_v}\right)\pi_{a^{(v)}},$$

or

$$P\left(x^{(v)}|a^{(\tau(v))}\right) = e^{-u_v}\delta_{a^{(\tau(v))},x^{(v)}} + \left(1-e^{-u_v}\right)\pi_{x^{(v)}},$$

where u_v is the *branch length* which depends on the unknown time between the substitution of the nucleotide at the ancestral node $\tau(v)$ with nucleotide at node v.

KTH VETENSKAP OCH KONST
KTH Matematik

Substitution probabilities: **F 81**

We shall incorporate the weights in the notation, as follows, e.g.,

$$P\left(x^{(v)}|a^{(\tau(v))};u_v\right) = e^{-u_v}\delta_{a^{(\tau(v))},x^{(v)}} + \left(1-e^{-u_v}\right)\pi_{x^{(v)}}.$$

KTH Matematik

Insert Branch Lengths

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)} \mid \underline{u}, \mathcal{T}\right)$$

$$= \sum_{a_i^{(7)} \in \mathcal{X}} \pi_{a_i^{(7)}} \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} P\left(a_i^{(5)} | a_i^{(7)}; u_5\right) \cdot \left[P\left(x_i^{(1)} | a_i^{(5)}; u_1\right)\right] \cdot \left[P\left(x_i^{(2)} | a_i^{(5)}; u_2\right)\right] \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}; u_6\right) \cdot \left[P\left(x_i^{(3)} | a_i^{(6)}; u_3\right)\right] \cdot \left[P\left(x_i^{(4)} | a_i^{(6)}; u_4\right)\right] \right\}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Substitution probabilities: **F 81**

$$= \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} \sum_{a_i^{(7)} \in \mathcal{X}} \frac{\pi_{a_i^{(7)}} P\left(a_i^{(5)} | a_i^{(7)}; u_5\right) \cdot \left[P\left(x_i^{(1)} | a_i^{(5)}; u_1\right)\right] \cdot \left[P\left(x_i^{(2)} | a_i^{(5)}; u_2\right)\right]}{} \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}; u_6\right) \cdot \left[P\left(x_i^{(3)} | a_i^{(6)}; u_3\right)\right] \cdot \left[P\left(x_i^{(4)} | a_i^{(6)}; u_4\right)\right] \right\}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Reversibility $\Rightarrow$

$$\pi_{a_i^{(7)}} P\left(a_i^{(5)} | a_i^{(7)}; u_5\right) = \pi_{a_i^{(5)}} P\left(a_i^{(7)} | a_i^{(5)}; u_7\right)$$

where we have

$$u_5 = u_7$$

Insertion $\Rightarrow$

$$= \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} \sum_{a_i^{(7)} \in \mathcal{X}} \frac{\pi_{a_i^{(5)}} P\left(a_i^{(7)} | a_i^{(5)}; u_7\right) \cdot [P\left(x_i^{(1)} | a_i^{(5)}; u_1\right)] \cdot [P\left(x_i^{(2)} | a_i^{(5)}; u_2\right)]}{} \right\}$$

$$\cdot \left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(a_i^{(6)} | a_i^{(7)}; u_6\right) \cdot [P\left(x_i^{(3)} | a_i^{(6)}; u_3\right)] \cdot [P\left(x_i^{(4)} | a_i^{(6)}; u_4\right)] \right\}.$$

KTH VETENSKAP OCH KONST

KTH Matematik

Chapman- Kolmogorov $\Rightarrow$

$$\sum_{a_i^{(7)} \in \mathcal{X}} P\left(a_i^{(7)} | a_i^{(5)}; u_7\right) P\left(a_i^{(6)} | a_i^{(7)}; u_6\right) =$$

$$= P\left(a_i^{(6)} | a_i^{(5)}; u_7 + u_6\right) = P\left(a_i^{(6)} | a_i^{(5)}; u_5 + u_6\right)$$

KTH Matematik

$\Rightarrow$ Pulley Principle

$$= \left\{ \sum_{a_i^{(5)} \in \mathcal{X}} \pi_{a_i^{(5)}} \cdot [P\left(x_i^{(1)} | a_i^{(5)}; u_1\right) P\left(x_i^{(2)} | a_i^{(5)}; u_2\right) \right\}$$

$$\left\{ \sum_{a_i^{(6)} \in \mathcal{X}} P\left(x_i^{(3)} | a_i^{(6)}; u_3\right) \cdot P\left(x_i^{(4)} | a_i^{(6)}; u_4\right) P\left(a_i^{(6)} | a_i^{(5)}; u_5 + u_6\right) \right\}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Pulley Principle

Likelihood depends only on the sum $u_5 + u_6$, and the root can thus be placed anywhere between the nodes 5 and 6 without changing the value of the likelihood. Continuing in this manner we can convince ourselves of correctness of the Pulley Principle, namely, that the root can be placed anywhere in the tree without changing the likelihood.

KTH Matematik

Probabilistic Phylogeny Reconstruction

The sequences at leaves

$$\mathbf{D} = \left\{ \mathbf{x}^{(k)} \right\}_{k=1}^{L}.$$

The probabilistic phylogeny reconstruction on basis of the data $\mathbf{D}$ is to rank binary rooted phylogenetic trees $\mathcal{T}i$ either by MAP or MLE.

KTH VETENSKAP OCH KONST
KTH Matematik

MAP = Maximum A Posterior

Maximize as a function of $\mathcal{T}_i$

$$P(\mathcal{T}_i|\mathbf{D}) = \frac{P(\mathbf{D}|\mathcal{T}_i) \cdot P(\mathcal{T}_i)}{\sum_{i=1}^{\Psi_r(L)} P(\mathbf{D}|\mathcal{T}_i) \cdot P(\mathcal{T}_i)},$$

where $P(\mathbf{D}|\mathcal{T}_i)$ is a likelihood of the tree $\mathcal{T}_i$ and $P(\mathcal{T}_i)$ is a prior probability of this tree, and the summation is over all trees with L leaves, i.e., there can be as many as $\Psi_r(L)$ terms.

KTH VETENSKAP OCH KONST
KTH Matematik

Prior on Trees

The uniform prior is, of course,

$$P(\mathcal{T}_i) = \frac{1}{\Psi_r(L)}.$$

$\Psi_r(L)$ is the number of rooted binary trees with L leaves. This is given explicitly below.

KTH Matematik

MLE= Maximum Likelihood Estimate

Maximizing the likelihood $P(\mathbf{D}|\mathcal{T}_i)$ as a function of $\mathcal{T}_i$.
A maximum likelihood tree topology $\widehat{\mathcal{T}}^{\text{MLE}}$ is a tree topology such that

$$P\left(\mathbf{D}|\widehat{\mathcal{T}}^{\text{MLE}}\right) \geq P(\mathbf{D}|\mathcal{T}_i)),$$

for all binary rooted phylogenetic trees $\mathcal{T}$ with L leaves.

KTH Matematik

Refinement of the Preceding

The likelihood of a tree involves further quantities, like the parameters in the **Q**-matrix of the substitution model, the branching times, and possibly the gamma parameter for the heterogeneity of the sites.

$$P\left(\mathbf{D}|\mathcal{T},\theta\right).$$

KTH Matematik

Refinement of the Preceding

A peculiar statistical estimation problem; some of the the parameters in θ will depend on $\mathcal{T}$. The maximum likelihood estimate of topology can be found by maximizing

$$P\left(\mathbf{D}|\mathcal{T},\widehat{\theta}^{\mathrm{ML}}\right)$$

as a function of $\mathcal{T}$, where $\widehat{\theta}^{\mathrm{ML}}$, the maximum likelihood estimate of θ. It is not certain that a local or global maximum of the likelihood function will be in practice found by this two-level procedure.

KTH VETENSKAP OCH KONST
KTH Matematik

Algorithms

There are several algorithms and software (e.g., PHYLIP, PAML) for finding

$$\mathcal{T}^{\mathrm{ML}}, \widehat{\theta}^{\mathrm{ML}}$$

`http://evolution.genetics.washington.edu/phylip/software.html`

KTH Matematik

Counting the number of trees:

One problem in all of this is the size of the parameter space: The number of binary roooted phylogenetic trees with $L \geq 2$ leaves is

$$\Psi_r(L) := \prod_{j=2}^{L} (2j-3) = 1 \cdot 3 \cdot \cdots \cdot (2L-3).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Counting the number of rooted trees:

$$\Psi_r(4) = 1 \cdot 3 \cdot 5 = 15, \Psi_r(8) = 1 \cdot 3 \cdot 5 \cdot 7 \cdot \cdot 9 \cdot 11 \cdot 13 = 135135,$$

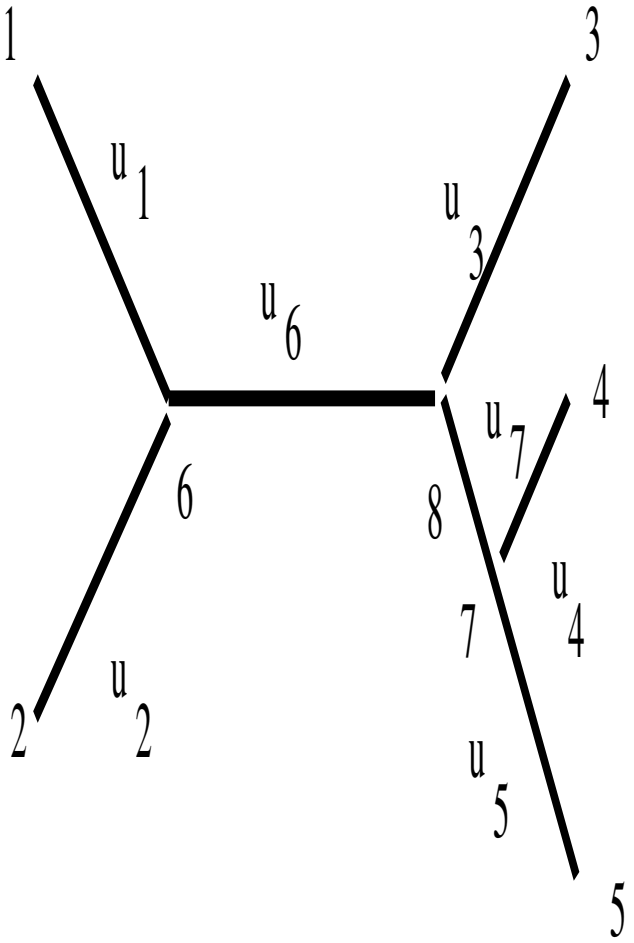
$$\Psi_r(10) = 135135 \cdot 15 \cdot 17 = 34459425$$

KTH VETENSKAP OCH KONST

KTH Matematik

An Unrooted Tree

By the Pulley principle we have placed the root in the branch $(8, 7)$.



ML for Branch Weights

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)}, x_i^{(5)} \mid \underline{u}, \mathcal{T}\right)$$

$$= \sum_{a_i^{(0)}} \sum_{a_i^{(8)}} \sum_{a_i^{(7)}} \pi_{a_i^{(0)}} P\left(a_i^{(8)} | a_i^{(0)}; u_8\right) L_{a_i^{(8)}}(8) P\left(a_i^{(7)} | a_i^{(0)}; v_6\right) L_{a_i^{(7)}}(7)$$

$$= \sum_{a_i^{(0)}} \pi_{a_i^{(0)}} L_{a_i^{(0)}}(0) \sum_{a_i^{(7)}} P\left(a_i^{(7)} | a_i^{(0)}; v_6\right) L_{a_i^{(7)}}(7)$$

KTH VETENSKAP OCH KONST
KTH Matematik

ML for Branch Weights

$$\sum_{a_i^{(0)}} \pi_{a_i^{(0)}} L_{a_i^{(0)}}(0) \sum_{a_i^{(7)}} P\left(a_i^{(7)} | a_i^{(0)}; v_6\right) L_{a_i^{(7)}}(7)$$

We insert

$$P\left(a^{(7)} | a^{(0)}; u_7\right) = e^{-u_7} \delta_{a^{(0)}, a^{(7)}} + \left(1 - e^{-u_7}\right) \pi_{a(7)}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

ML for Branch Weights

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)}, x_i^{(4)}, x_i^{(5)} \mid \underline{u}, \mathcal{T}\right) =$$

$$= e^{-u_7} \sum_s \pi_{a_i^{(s)}} L_s(8) L_s(7)$$

$$+ \left(1 - e^{-u_7}\right) \sum_{a_i^{(8)}} \pi_{a(8)} L_{a_i^{(8)}}(8) \sum_{a_i^{(7)}} \pi_{a(7)} L_{a_i^{(7)}}(7).$$

KTH VETENSKAP OCH KONST
KTH Matematik

ML for Branch Weights

$$P_{\text{av}}\left(x_i^{(1)}, x_i^{(2)}, x^{(3)}, x_i^{(4)}, x_i^{(4)}, x_i^{(5)} \mid \underline{u}, \mathcal{T}\right) =$$

$$e^{-u_7} A + \left(1 - e^{-u_7}\right) B = Aq + Bp$$

ML for Branch Weights: multiply over sites

$$\prod_{i=1}^{N} (A_i q + B_i p)$$

KTH Matematik

ML for Branch Weights

$$\prod_{i=1}^{N} (A_i q + B_i p)$$

is differentiated w.r.t. p and the derivative is set equal to zero. Then we get

$$p^{k+1} = \frac{1}{N} \sum_{i=1}^{N} \frac{B_i p^{(k)}}{A_i q^{(k)} + B_i p^{(k)}}$$

KTH VETENSKAP OCH KONST
KTH Matematik

ML for Branch Weights

Iterate for u_7 and then move to next u_v.

KTH Matematik

Searching for Topologies

Build a tree by adding species to it successively, starting by a two species tree. Try each branch to add a new species, and then place to give highest likelihood. Do rearrangements, if needed.
Then try adding next species.

KTH VETENSKAP OCH KONST
KTH Matematik

Inference of Phylogenies: PAML

PAML is a program package for phylogenetic analyses of DNA or protein sequences using maximum likelihood.
Possible uses of the programs are

- Estimation of branch lengths in a phylogenetic tree and parameters in the evolutionary model such as the transition/transversion rate ratio, the shape parameter of the gamma distribution for variable evolutionary rates among sites, and rate parameters for different genes;
- Test of hypotheses concerning sequence evolution, such as rate constancy and independence among nucleotide or amino acid sites, rate constancy among lineages (the molecular clock), and homogeneity of evolutionary process in multiple genes;
- Calculation of substitution rates at sites;
- Phylogenetic tree reconstruction by maximum likelihood and Bayesian methods (??).

KTH VETENSKAP OCH KONST
KTH Matematik

Inference of Phylogenies: PAML

A summary of the types of analyses performed by different programs in the package is given below.

- baseml: ML analysis of nucleotide sequences: estimation of tree topology, branch lengths, and substitution parameters under a variety of nucleotide substitution models (JC69, K80, F81, F84, HKY85, TN93, REV); constant or gamma rates for sites; molecular clock (rate constancy among lineages) or no clock, among-gene and within-gene variation of substitution rates; models for combined analyses of multiple sequence data sets; calculation of substitution rates at sites; reconstruction of ancestral nucleotides.
- basemlg: ML analysis of nucleotide sequences under the model of gamma rates among sites. The (continuous) gamma model is used with one of the following substitution models: JC69, K80, F81, F84, HKY85, TN93, and REV.

KTH VETENSKAP OCH KONST
KTH Matematik

Inference of Phylogenies: MrBayes

MrBayes (Huelsenbeck and Ronquist 2001) is a program for doing Bayesian phylogenetic analysis. The program uses a Markov Chain Monte Carlo (MCMC) method for exploring parameter space in a stepwise fashion. Each new step is either accepted or rejected based on the change in likelihood. The posterior probability for each parameter is based on the frequency with which the parameter values are observed.

`http://morphbank.ebc.uu.se/mrbayes/`

KTH VETENSKAP OCH KONST
KTH Matematik

End of Last Lecture of TAMS 23

Scientific Name: Cucumaria

Location: Ross Sea, Antarctica

KTH Matematik