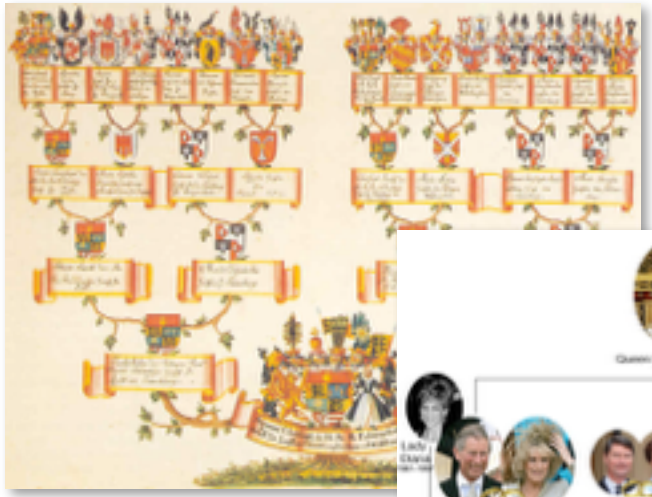
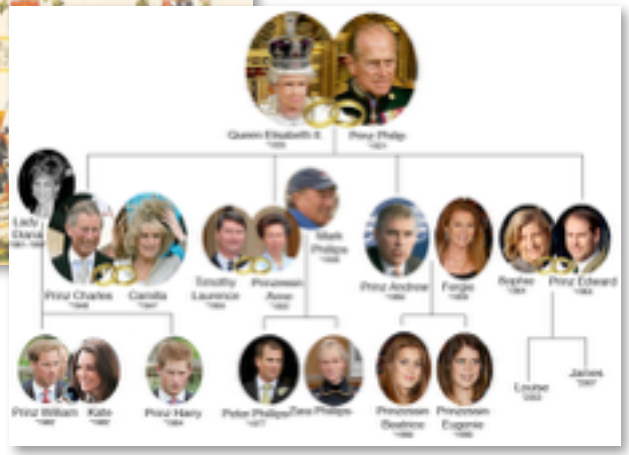
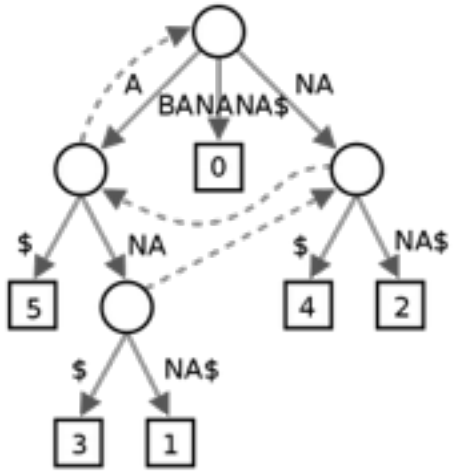
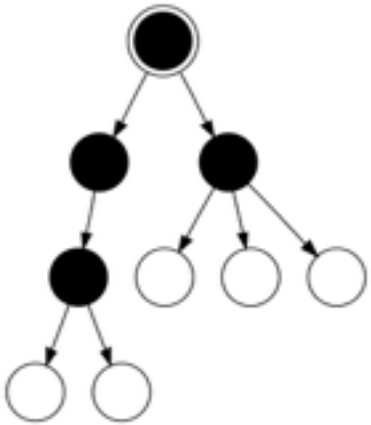
Bäume und Baumrekonstruktion

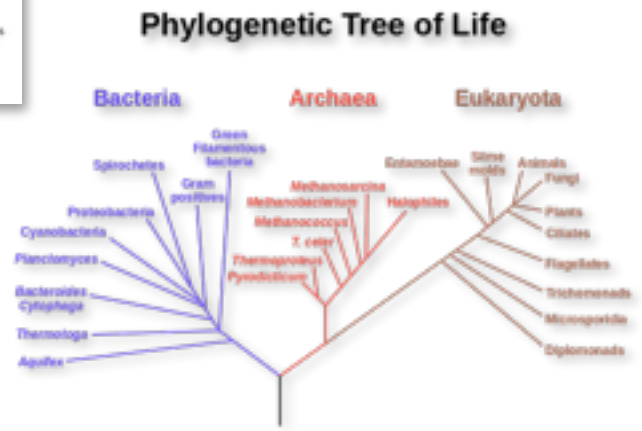
Ganz allgemein: Bäume repräsentieren die Beziehungen zwischen Dingen[1] und erzählen (häufig) Geschichten!







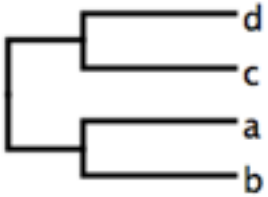




[1] e.g. between members of a family

Die Interpretation von Bäumen ist eigentlich einfach!

- Ein Aufspaltungsereignis im Baum spaltet **eine** parentale Entität in zwei Kind-Entitäten.
- Die Abfolge der Aufspaltungsereignisse im Baum bestimmt die Verwandtschaftschaftsverhältnisse der untersuchten Entitäten.
- Je näher an der Gegenwart ein Aufspaltungsereignis liegt, desto näher sind die daraus resultierenden Entitäten miteinander verwandt.



Die Rekonstruktion phylogenetischer Verwandtschaftsverhältnisse in Form eines Baumes als bioinformatisches Problem

Phylogenetik: Die Analyse evolutionärer Verwandtschaftsverhältnisse zwischen Gruppen von Organismen mittels der Analyse morphologischer Daten oder Molekularer Sequenzen.

Die Phylogenetik liefert Antworten auf folgende Fragen

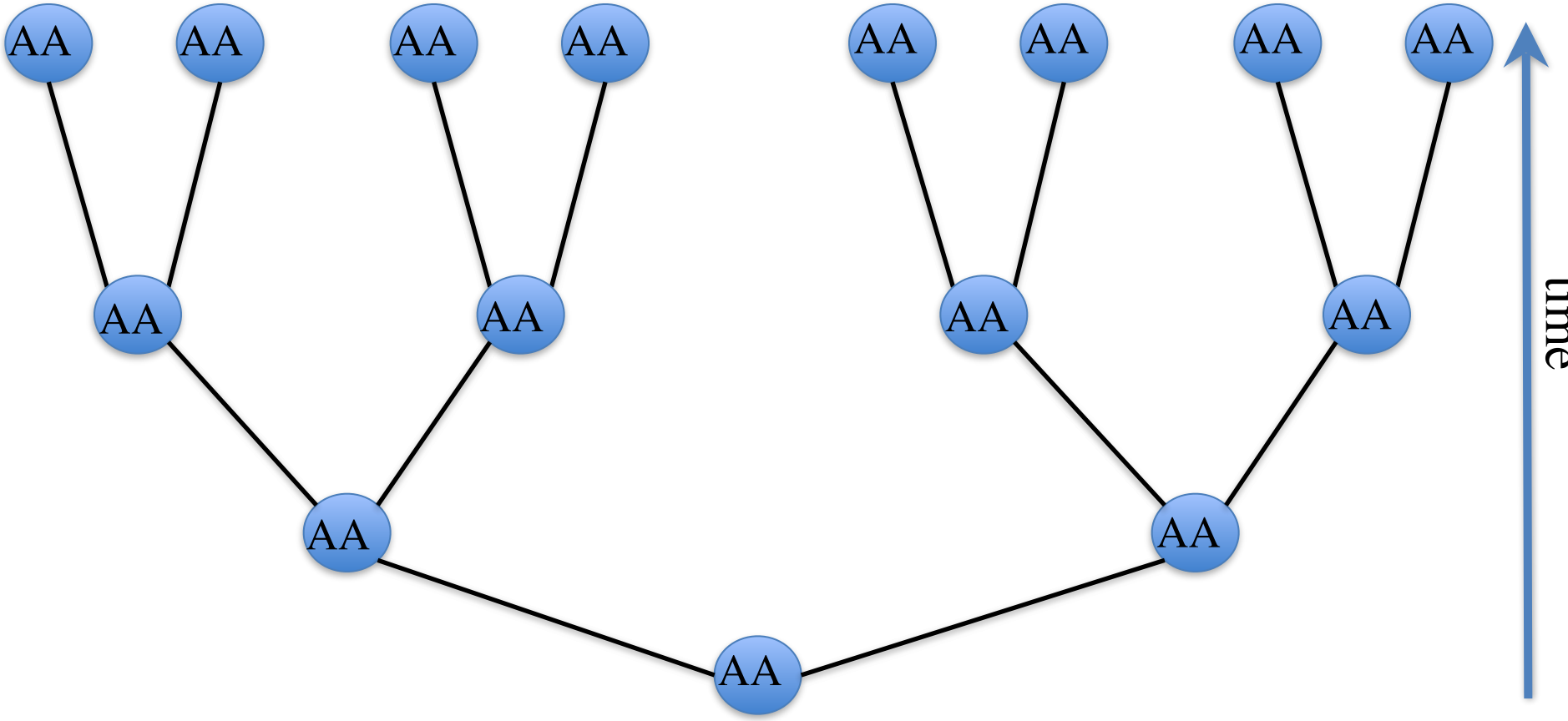
- ➢ Wie nahe sind die untersuchten Arten/Sequenzen miteinander verwandt?
- ➢ Wann in der evolutionären Geschichte sind bestimmte Ereignisse (z.B. Arbildung) passiert?

sie sagt aber nichts aus über

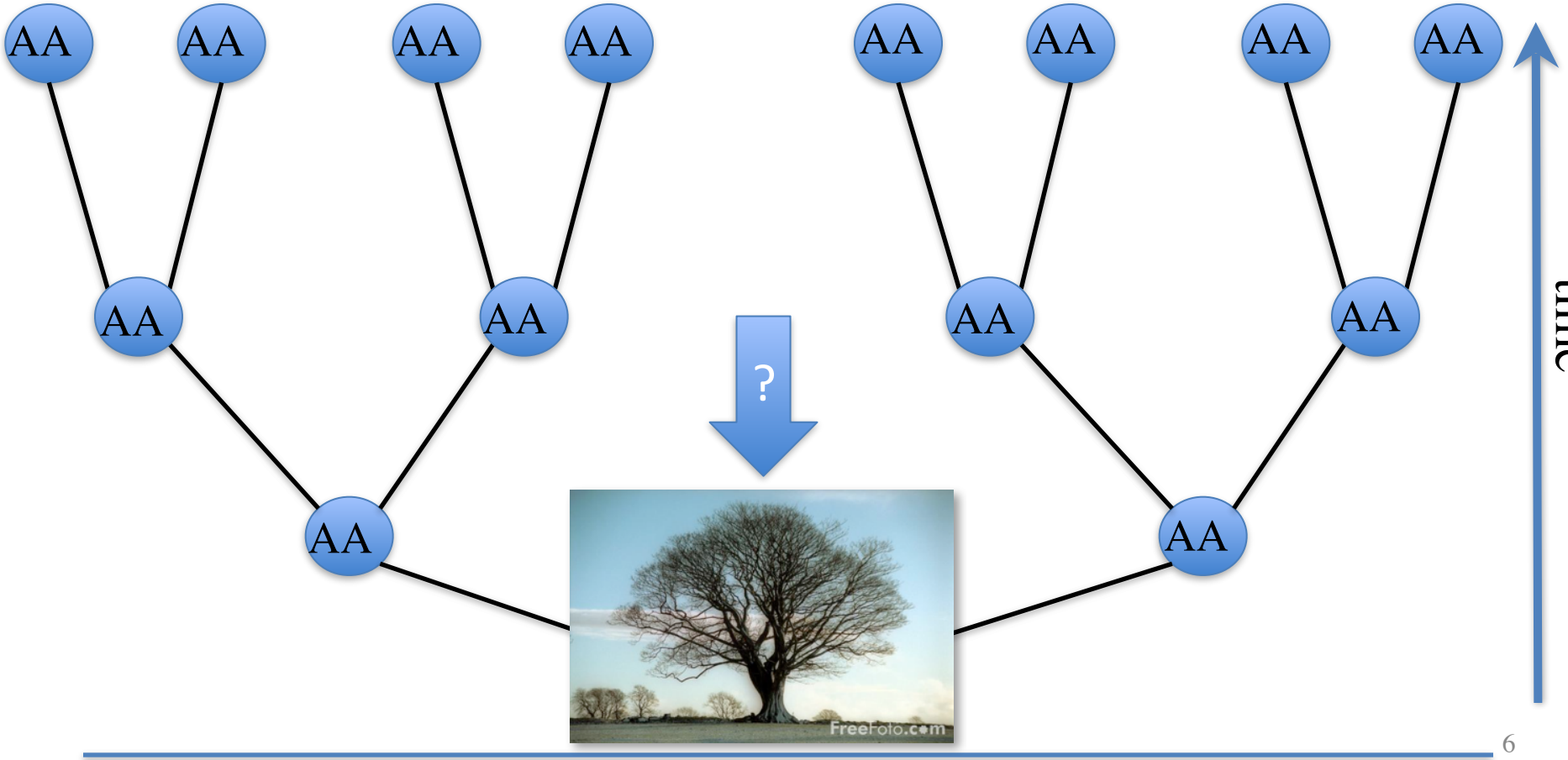
- ➢ Sind die untersuchten Arten/Sequenzen miteinander verwandt?[1]

[1] Bitte vergleichen mit der Einschränkung welche Sequenzen man alignieren darf!

Unsere Grundannahme: Die analysierten Entitäten (Sequenzen/ Arten) sind entlang eines Baumes evolviert.

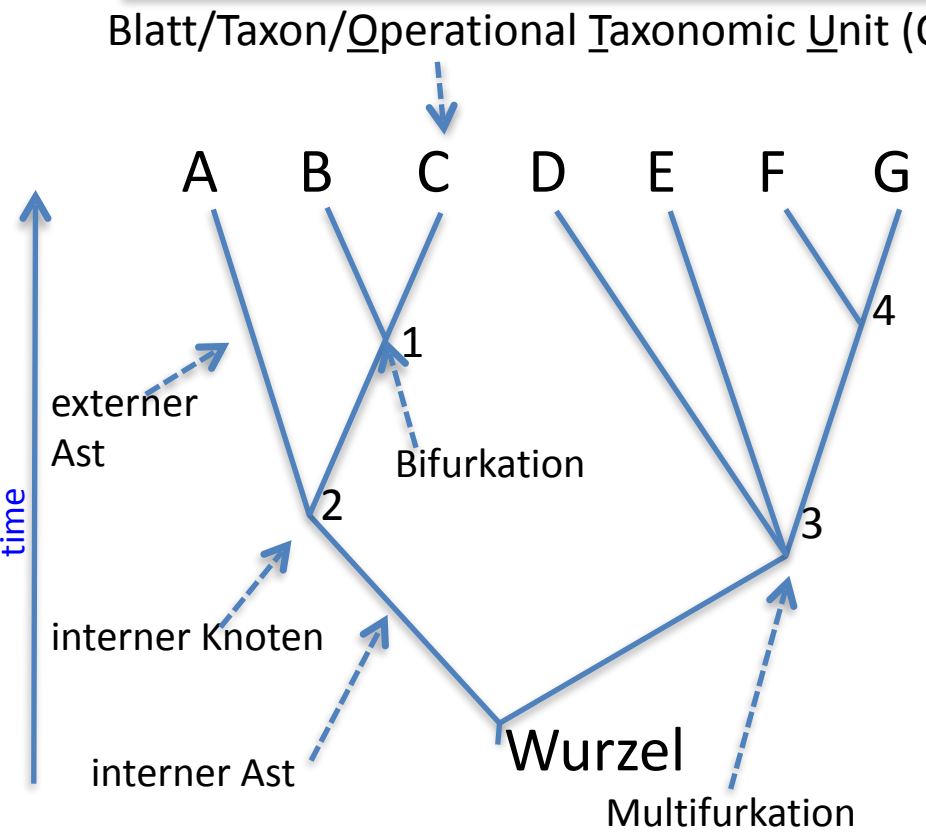


Wie schaffen wir es nun im Rückblick die Reihenfolge der Aufspaltungsereignisse zu rekonstruieren?

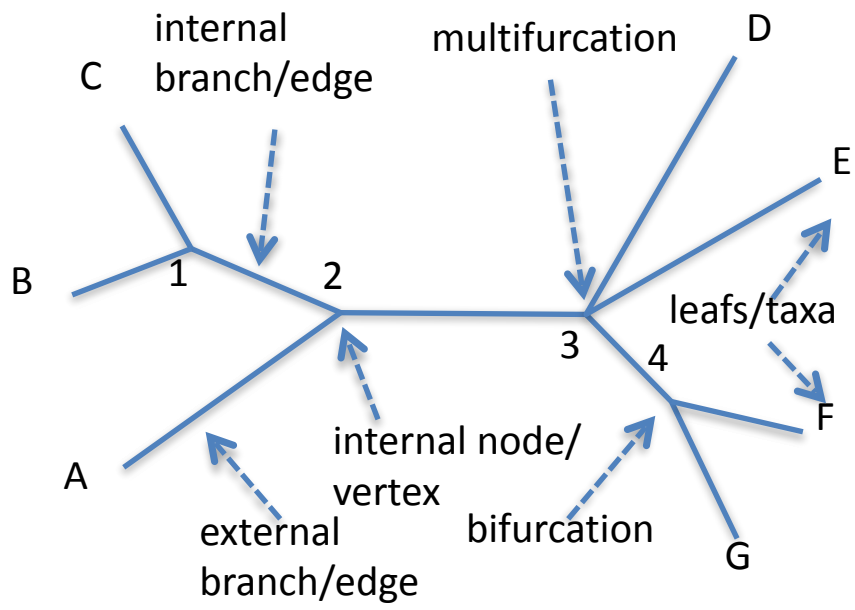


Phylogenetische Bäume: Terminologie und Konzepte

Notationen für Bäume



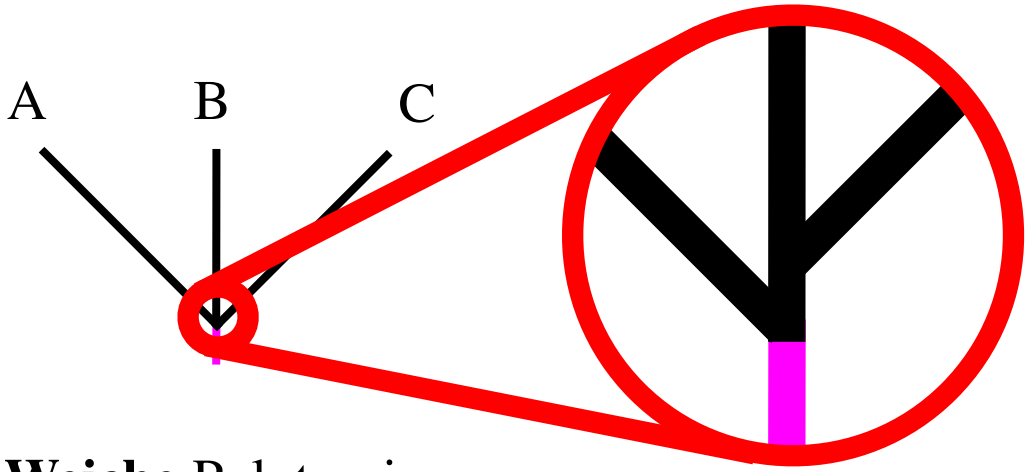
Gewurzelter Baum



Unrooted tree[1]

[1] Hier, die englischen Begriffe. Achtung, es fehlt die Wurzel und damit die Richtung der Zeit!

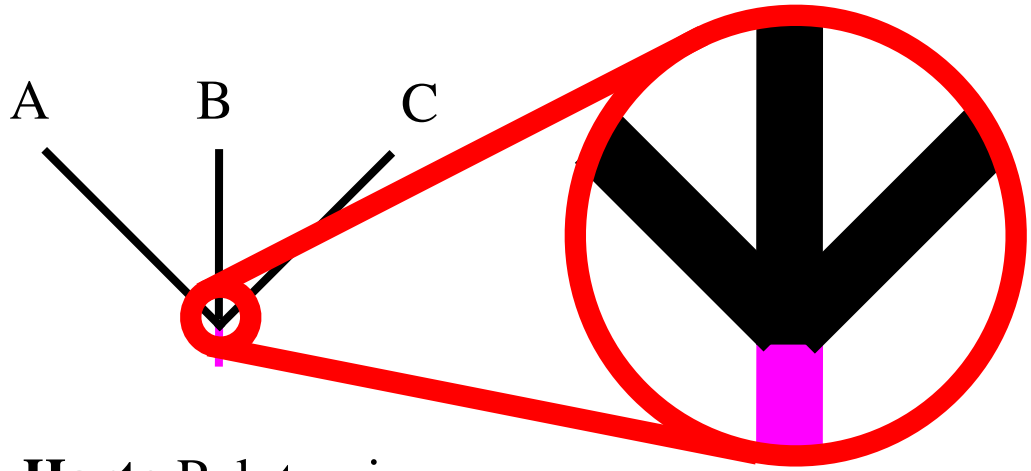
Wie interpretiert man Polytomien?



Weiche Polytomie

Es gibt nur Bifurkationen im Baum – Die internen Kanten sind so kurz, dass keine verwertbare evolutionäre Änderung beobachtbar ist.

Entsprechend kann das wahre Aufspaltungsmuster nicht rekonstruiert werden.



Harte Polytomie

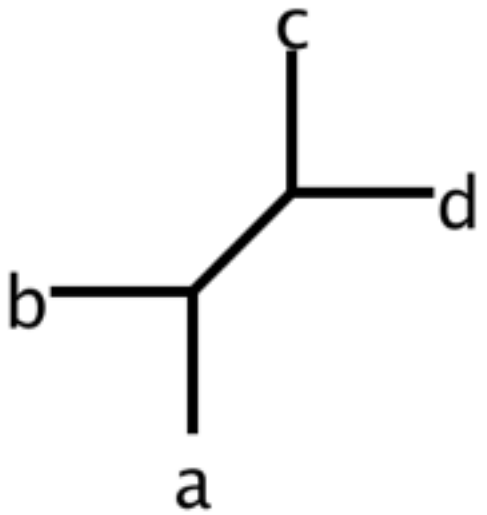
Aus einer anzestralen Linie sind gleichzeitig 3+ Linien entstanden.

Ungewurzelte Bäume
(Ein nicht ganz so intuitives Konzept[1])

Eine triviale Aussage: Der Baum hat keine Wurzel.

Eine nicht so triviale Konsequenz: Im Baum gibt es keine Aussage hinsichtlich der Richtung in die Zeit fließt.

Wir können also nicht zwischen Vor- und Nachfahre unterscheiden!



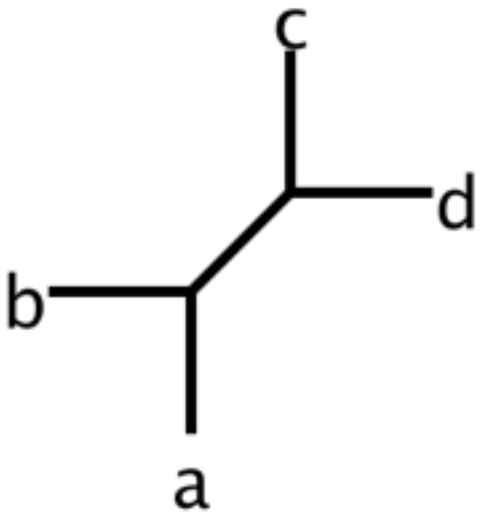
Achtung: In vielen Fällen wird bei der Verwendung und Interpretation von Bäumen aber implizit von einem gewurzelten Baum ausgegangen…

[1]1 Aber die meisten Programme zur Baumrekonstruktion liefern nur ungewurzelte Bäume!

Wir verdeutlichen das Problem:

Welches Taxon ist der nächste Verwandte von **d**?

- **a**
- **b**
- **c**
- **ab**
- **ac**
- **bc**
- **abc**

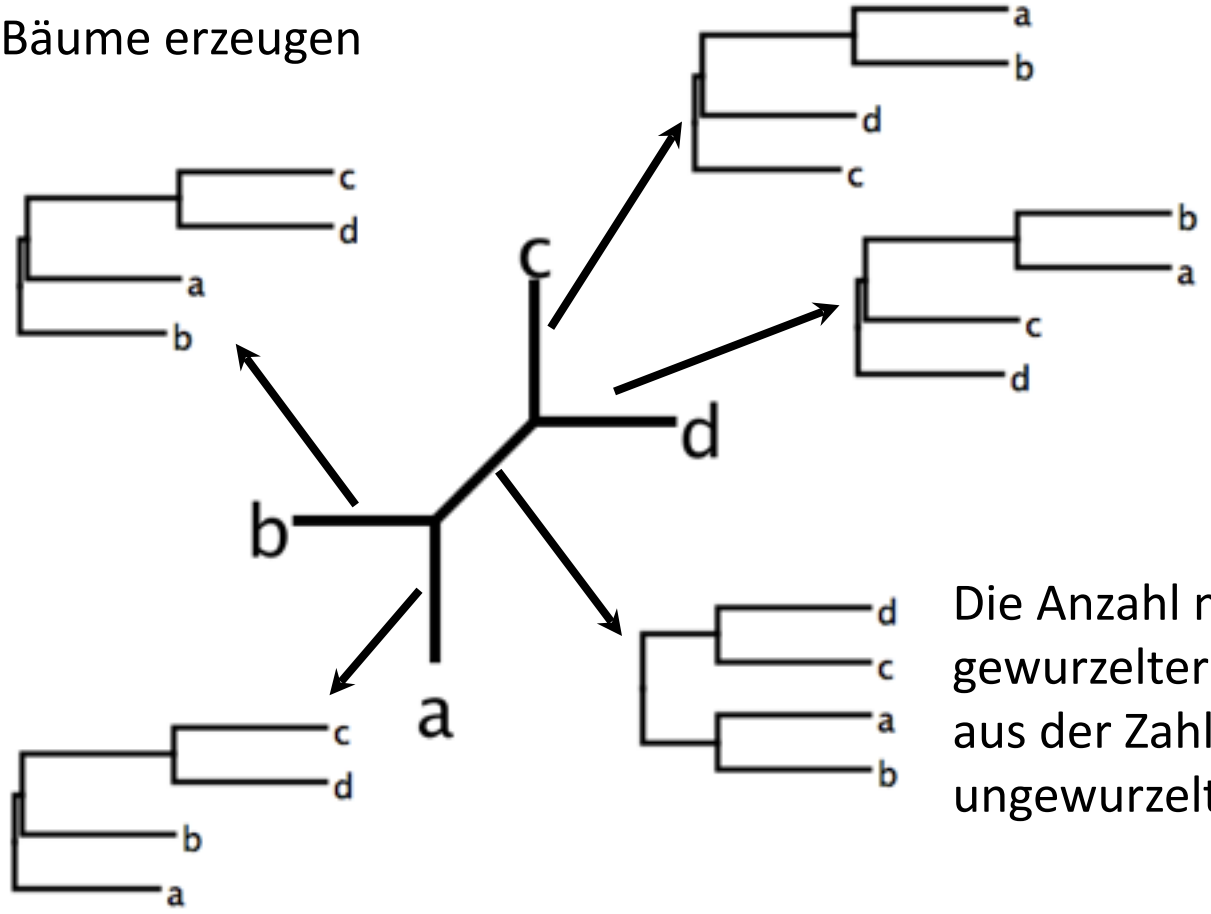


Wir können nicht entscheiden – Die Antwort hängt davon ab wo die Wurzel liegt!

Achtung: Unabhängig von der Position der Wurzel kann **d** nicht am nächsten verwandt zu **a** **oder** **b** sein.

Man kann ungewurzelte Bäume nachträglich wurzeln, z.B. mittels einer Außengruppe[1] (Bitte beachten, die Außengruppe selbst ist nicht dargestellt nur die resultierende Wurzel)

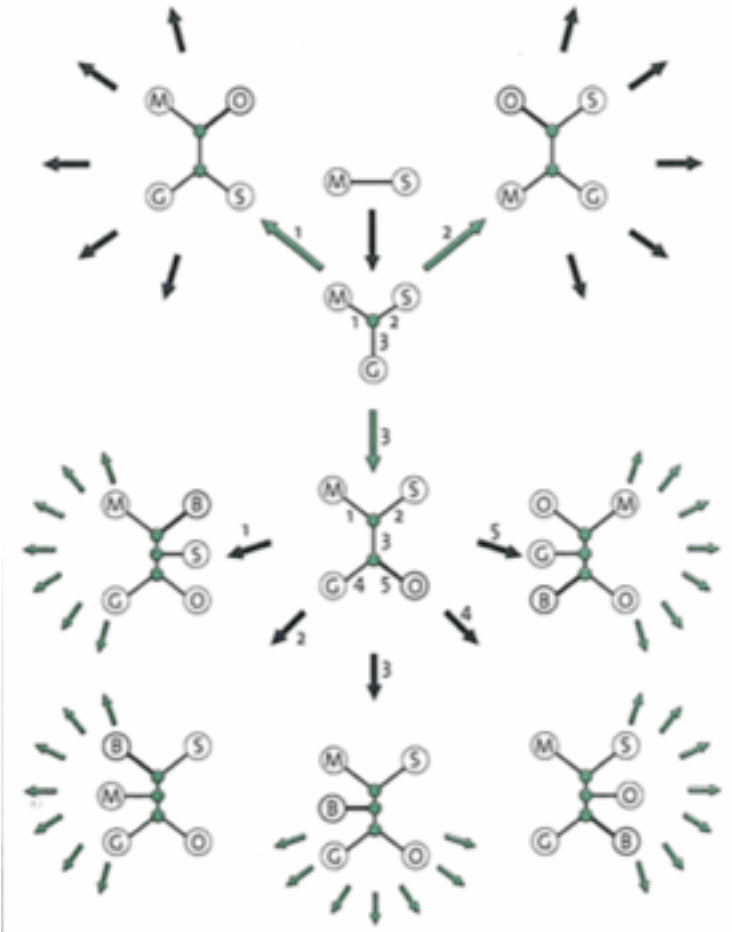
Ein ungewurzelter Baum kann viele gewurzelte Bäume erzeugen



Die Anzahl möglicher gewurzelter Bäume ergibt sich aus der Zahl der Äste im ungewurzelten Baum.

[1] Eine Außengruppe ist ein Taxon von dem man a priori **weiss**, dass es sich zuerst in der Phylogenie abgespalten hat.

Wie viele möglichen Bäume gibt es[1]?



$$b(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

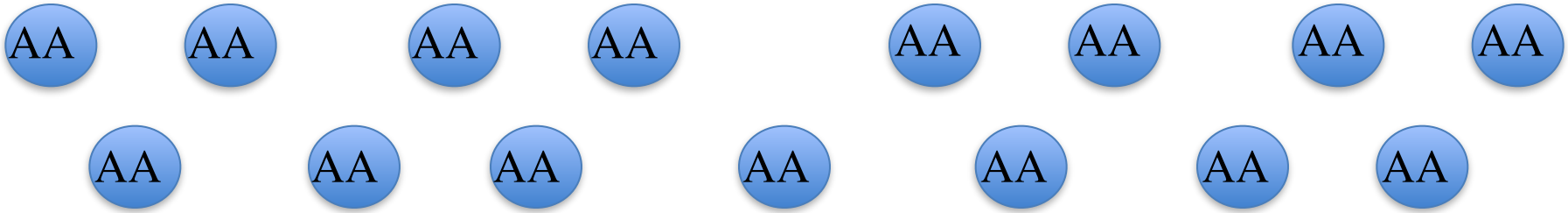
$$b(10) = 2027025$$

$$b(55) = 2.9 \times 10^{84}$$

$$b(100) = 1.7 \times 10^{182}$$

1 Dieses Beispiel gilt für ungewurzelte Bäume. Für gewurzelte Bäume erhöht man die Anzahl der Taxa um 1, die Wurzel.

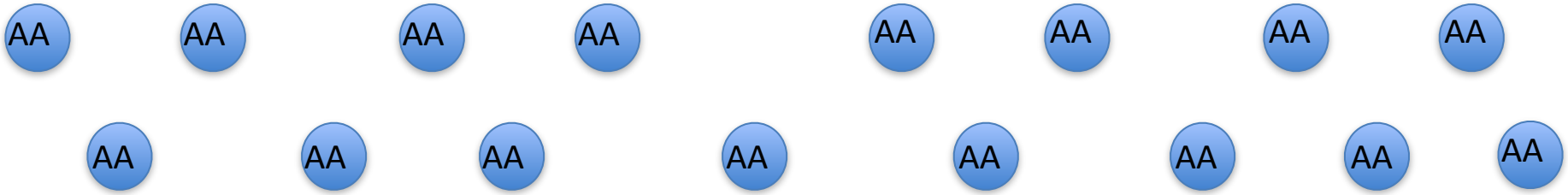
Rekonstruktion der Aufspaltungs-Reihenfolge der evolutionären Entitäten (Taxa).

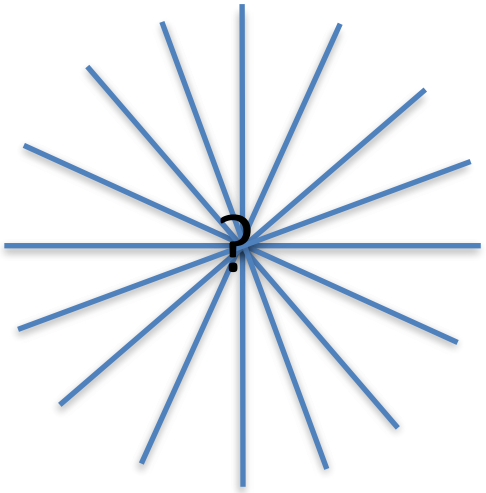


Ziel: Wir wollen die Taxa zuerst vereinigen, die sich zuletzt einen gemeinsamen Vorfahren geteilt haben. Wir wollen also im Rückblick Zeiträume rekonstruieren.

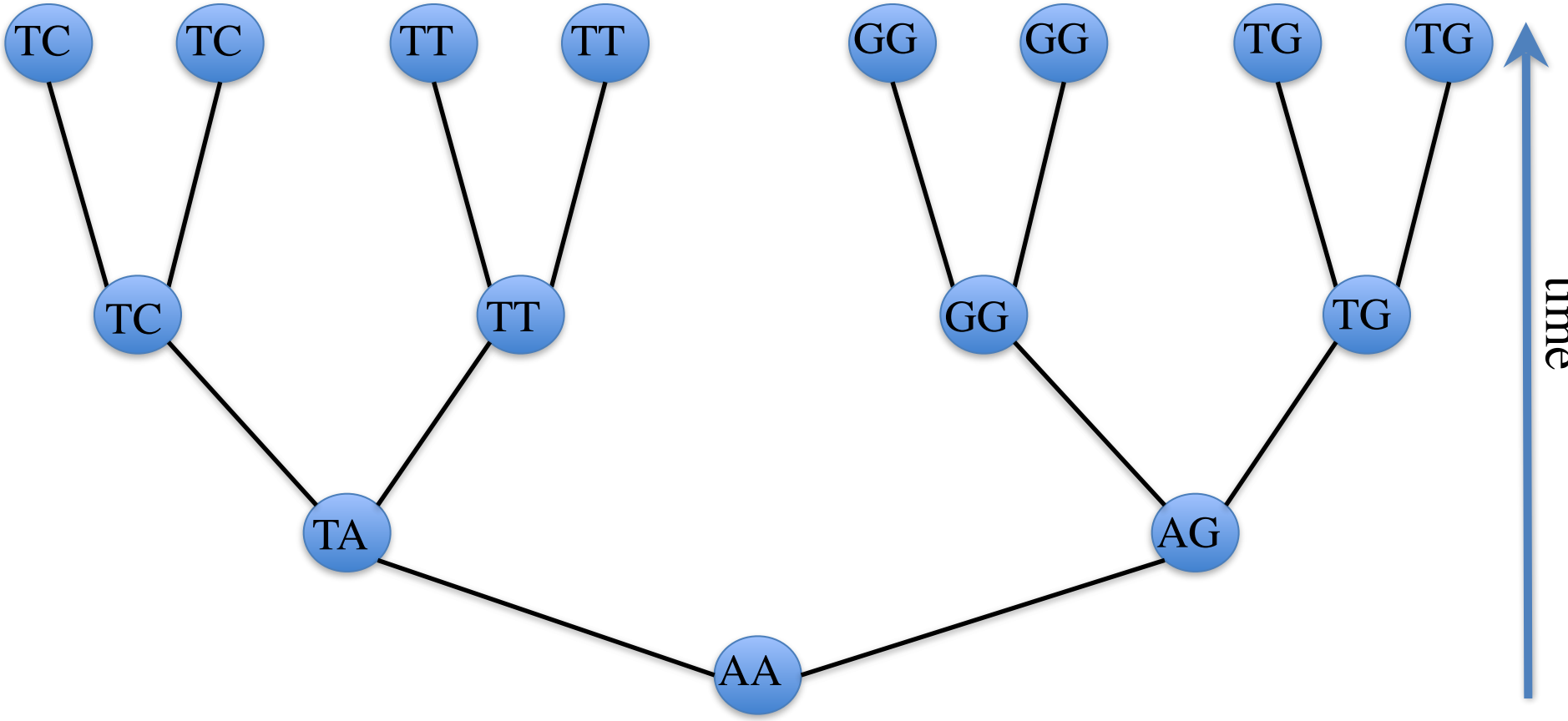
Ohne Veränderung der Charakteristika der einzelnen Taxa ist es allerdings unmöglich die Reihenfolge der Aufspaltungsereignisse zu rekonstruieren!

Merke: Wir können nicht Zeit selbst messen, sonder nur das was entlang dieser Zeit passiert ist!





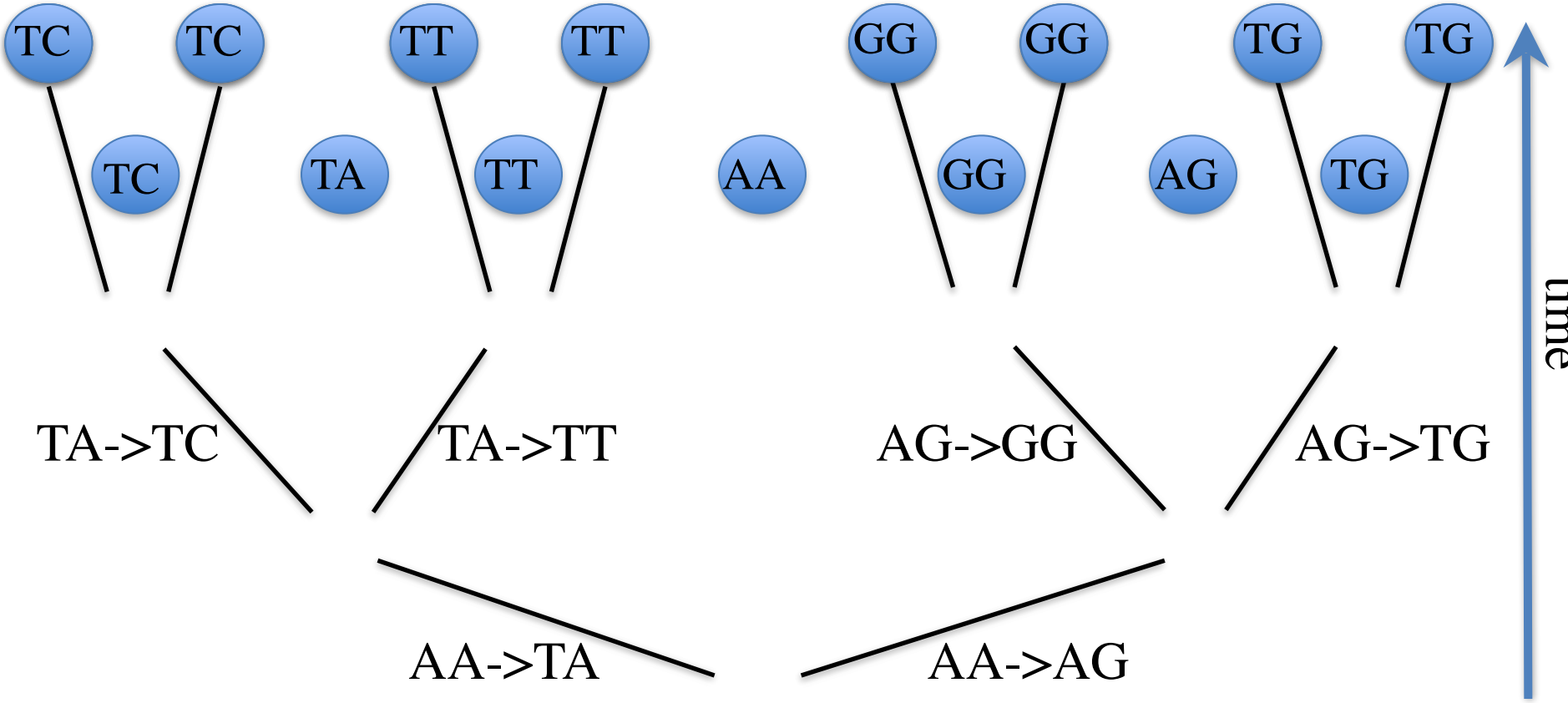
Um Aufspaltungsereignisse zu rekonstruieren brauchen wir Veränderung von Charakteristika.



Unterschiede zwischen Charakteren (Veränderungen[1]) liefern die Information für die Phylogenie-Rekonstruktion.

Dieser Baum erklärt die Daten mit **6** Substitutionen

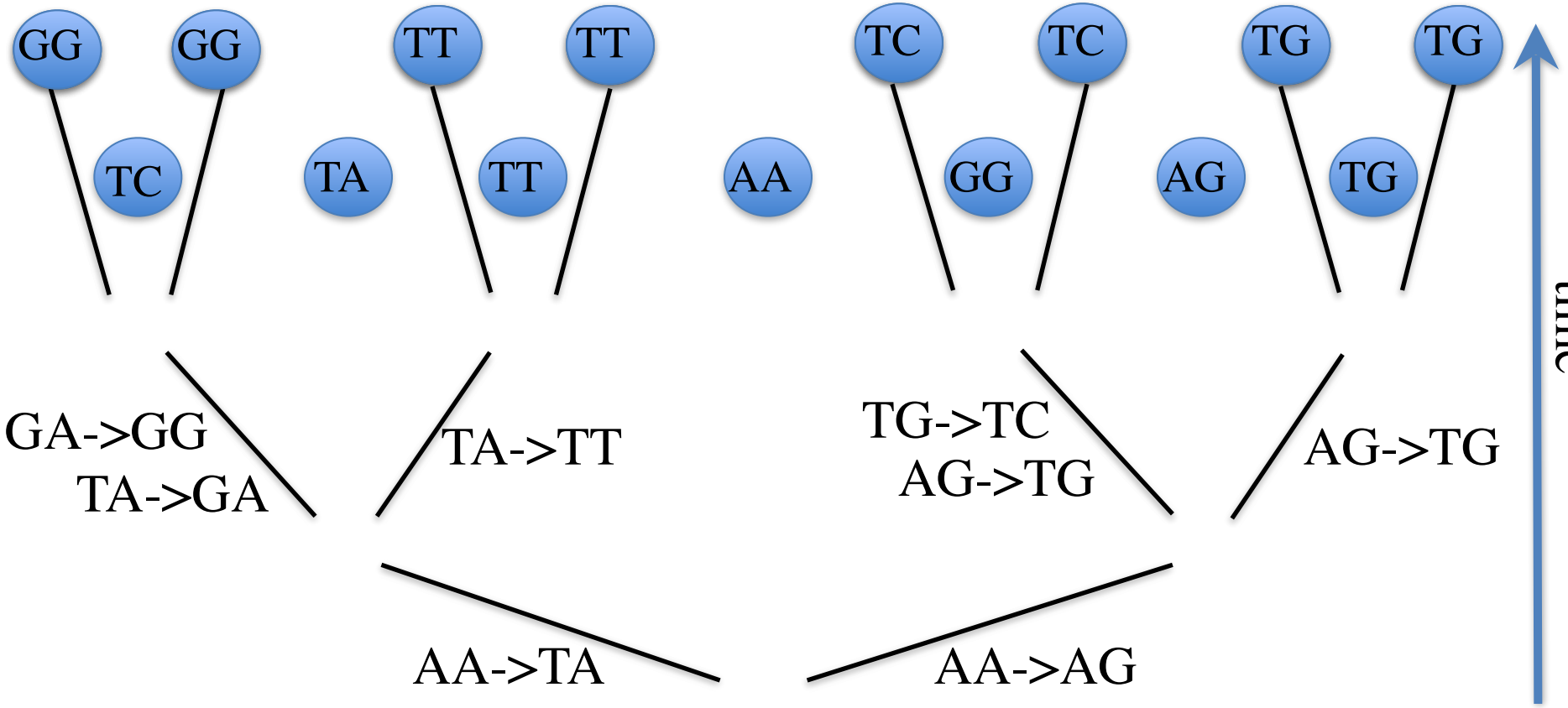
…aber es gibt mehr als eine Möglichkeit



[1] wir arbeiten in den Beispielen mit Substitutionen

Unterschiede zwischen Charakteren (Veränderungen) liefern die Information für die Phylogenie-Rekonstruktion.

Dieser Baum erklärt die Daten mit **8** Substitutionen!
Ist er besser oder schlechter als der vorherige Baum?



Wenn wir nicht-beobachtete Ereignisse rekonstruieren, verwenden wir meistens das (intuitive) Prinzip der maximalen Sparsamkeit[1]

William of Ockham, 1285-1347/49

Occam's ‚Razor' (Gesetz der Sparsamkeit) besagt:

Pluralitas non est ponenda sine necessitate.

Komplexität sollte nicht ohne Notwendigkeit angenommen werden

Das Prinzip besagt schlicht, von zwei gegenüberstehenden Hypothesen zur Erklärung von Beobachtungen sollte man die einfachere Erklärung bevorzugen[2].

1 engl. Maximum Parsimony

2 Achtung, hierbei handelt es sich nur um eine generelle Arbeitsanweisung ohne definierten Gültigkeitsbereich.

Mit dem Prinzip der maximalen Sparsamkeit haben wir die erste von drei Möglichkeiten zur Baum-Rekonstruktion.

Finde den Baum, der die Daten mit der geringsten Anzahl an Veränderungen erklärt

Data	Method	Evaluation Criterion
Characters (Alignment)	**Maximum Parsimony**	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	**Distance Methods**	

Rekonstruktion von phylogenetischen Bäumen mittels Maximum Parsimony: Die Datenmatrix ist in den meisten Fällen ein Alignment.

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	A	C	T	G	T	T
S2	C	G	C	A	C	T	G	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	G	A	C	T	G	C	T

Reconstruktion phylogenetischer Bäume: Maximum parsimony

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	A	C	T	G	T	T
S2	C	G	C	A	C	T	G	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	G	A	C	T	G	C	T

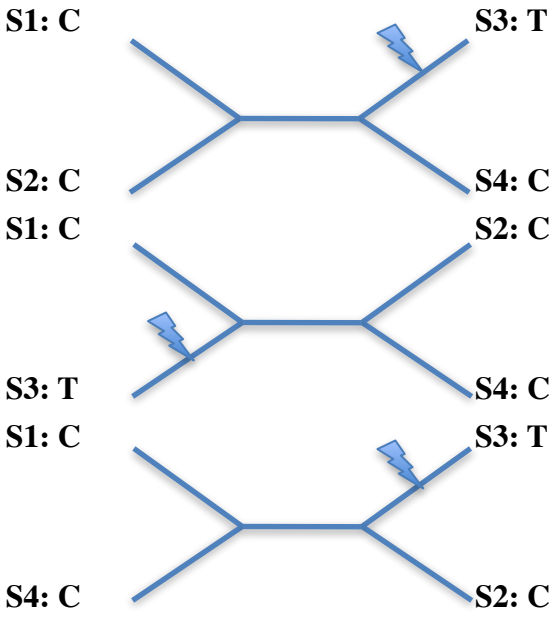
Prinzip: Finde für jede Alignment-Spalte den Baum der die niedrigste Anzahl von Veränderungen benötigt, um das Spalten-Muster[1] (engl. site pattern) zu erklären! Wähle schließlich den Baum, der die Anzahl der Veränderungen über das gesamte Alignment minimiert.

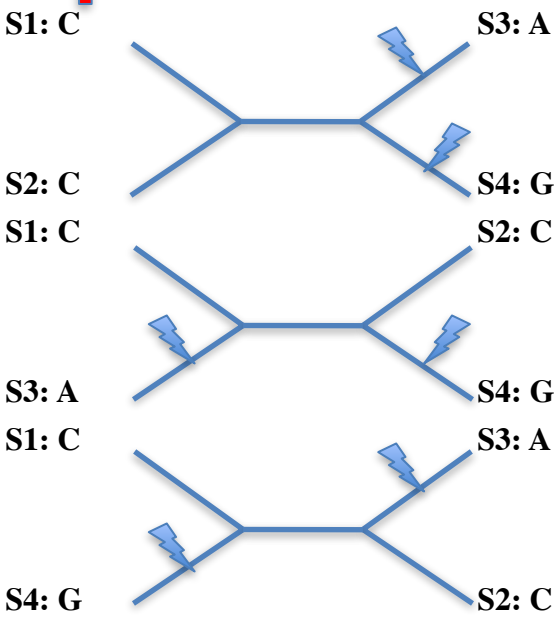
[1] Muster aus den Buchstaben des Alphabets in dieser Spalte.

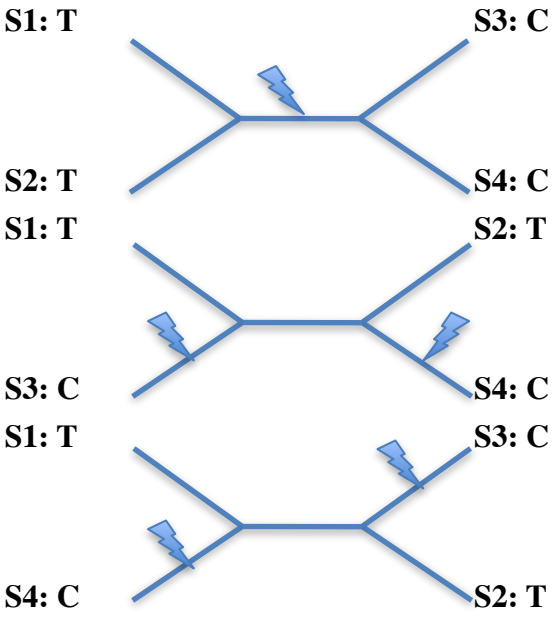
Rekonstruktion des Maximum Parsimony Baumes: Evaluiere für jede informative Spalte alle möglichen Bäume

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	A	C	T	G	T	T
S2	C	G	C	A	C	T	G	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	G	A	C	T	G	C	T

??[2]





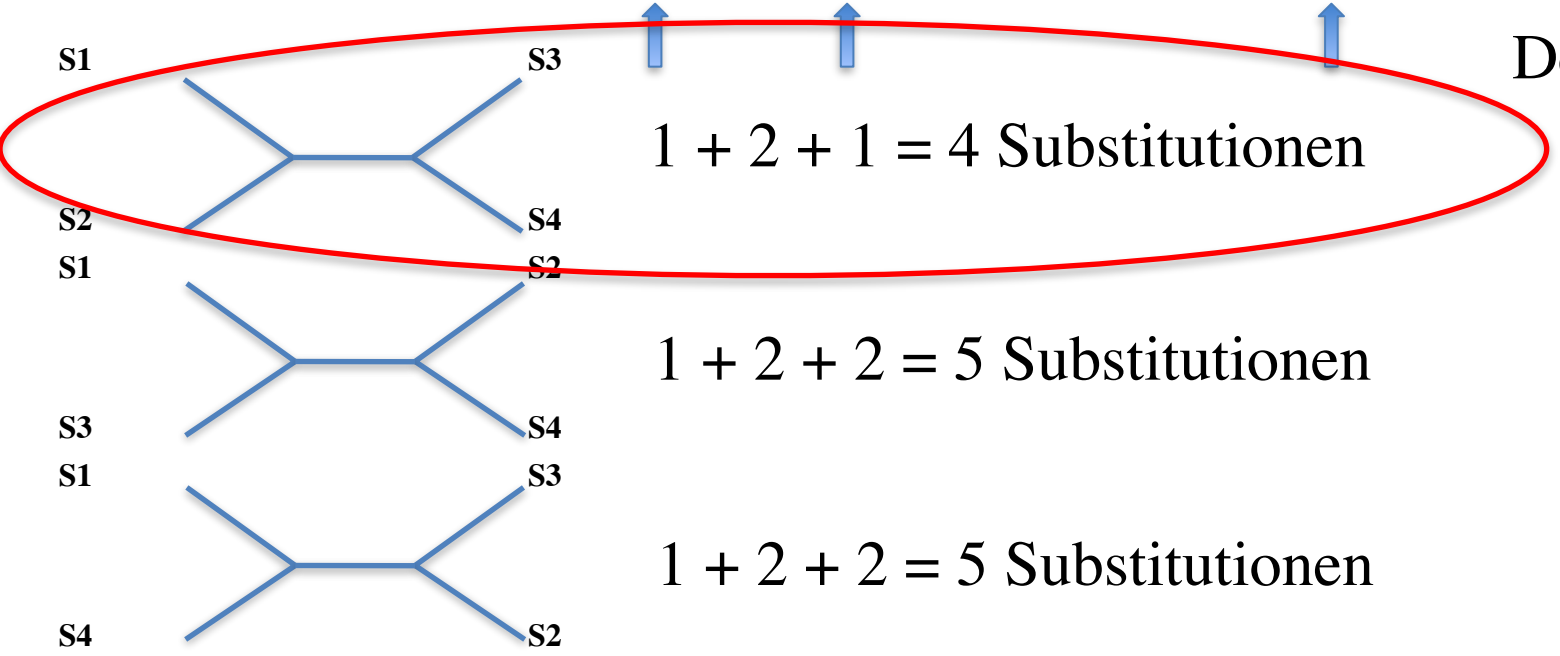


substitution

[2] konstante Spalten sind uninformativ!

Der Maximum Parsimony Baum ist der, der die Anzahl der Veränderungen über das gesamte Alignment minimiert

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	A	C	T	G	T	T
S2	C	G	C	A	C	T	G	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	G	A	C	T	G	C	T



Der MP Baum!

1 + 2 + 1 = 4 Substitutionen

1 + 2 + 2 = 5 Substitutionen

1 + 2 + 2 = 5 Substitutionen

Einige zusammenfassende Aspekte zum Thema "Maximum Parsimony"

1. Das Parsimonie-Konzept wird in der Regel als Modell-frei betrachtet!

2. Tatsächlich wird nichts modelliert aber der Algorithmus macht eine sehr starke Annahme: Veränderungen sind selten und Rückmutationen geschehen nicht. Diese Annahme selbst ist allerdings ein implizites ‚Modell'.

3. Annahme 2 trifft in der Regel auf morphologische Daten zu. Allerdings wird sie von biologischen Sequenzen-Daten häufig verletzt[1]

4. MP ist eigentlich eine Methode zur Baum-Evaluierung, nicht zur Baum-Rekonstruktion...

[1] vgl. Vorlesungsinhalte zur Modellierung von Sequenzevolution

Distanzen und deren Verwendung zur Baum-Rekonstruktion

Finde den Baum der die geringste Anzahl an Veränderungen erfordert

Data	Method	Evaluation Criterion
Characters (Alignment)	Maximum Parsimony	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	Distance Methods	

Rekonstruieren den Baum der den besten 'Fit' an eine paarweise Distanzmatrix liefert

Rekonstruktion von phylogenetischen Bäumen: **Distanz**

Die paarweise Distanz zwischen Sequenzen approximiert die Zeit, die diese Sequenzen getrennt evolvieren

seq 1	a	g	c	t	t	a	c	c	t	g	t	t	a	c	t
seq 2	c	g	t	a	a	a	t	t	t	c	c	c	g	a	t
seq 3	c	g	c	a	a	g	t	t	t	c	c	c	g	a	t
seq 4	c	a	c	t	t	a	t	t	a	g	t	c	a	a	c

	Seq 1	Seq 2	Seq 3
Seq 2	11		
Seq 3			
Seq 4			

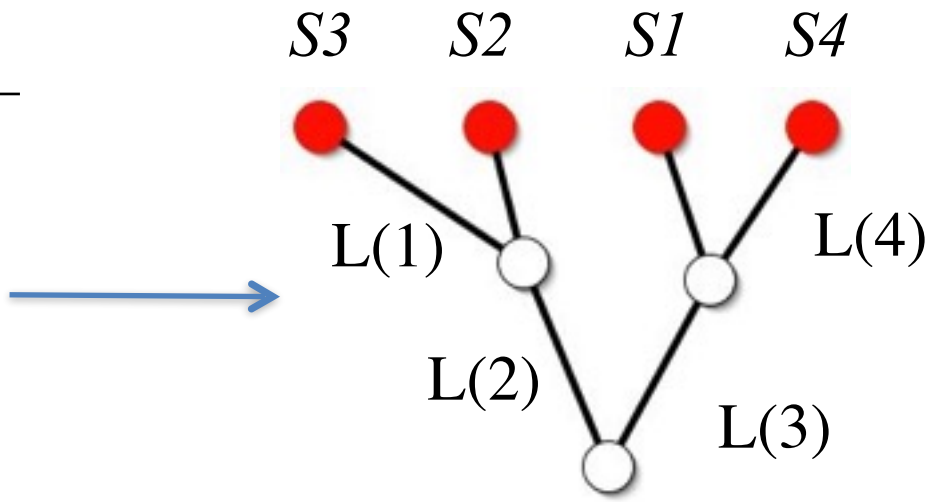
Phylogenie-Rekonstruktion
Distanz-basierte Methoden

seq 1	a	g	c	t	t	a	c	c	t	g	t	t	a	c	t
seq 2	c	g	t	a	a	a	t	t	t	c	c	c	g	a	t
seq 3	c	g	c	a	a	g	t	t	t	c	c	c	g	a	t
seq 4	c	a	c	t	t	a	t	t	a	g	t	c	a	a	c

	Seq 1	Seq 2	Seq 3
Seq 2	11		
Seq 3	11	2	
Seq 4	8	10	9

Phylogenie-Rekonstruktion
Distanz-basierte Methoden

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9

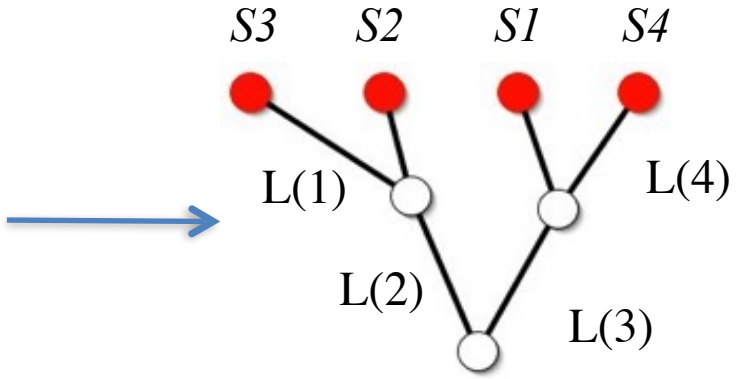


Erstelle einen Baum *T* mit Astlängen L(b) so dass die Summe der Astlängen zwischen zwei beliebigen Blättern möglichst nahe an die gemessenen paarweisen Distanzen kommen.

$$D_{measured}(S3, S4) \approx L(1) + L(2) + L(3) + L(4)$$

Distanz-basierte Baumrekonstruktion entspricht einer hierarchischen Clusteranalyse
- Metrik -

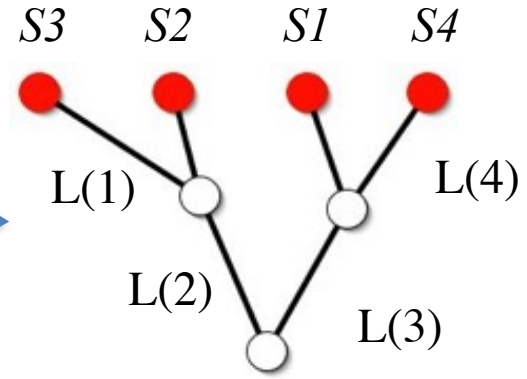
	S1	S2	S3
S2	**11**		
S3	**11**	**2**	
S4	**8**	**10**	**9**



Eine **Metrik** ist eine **Funktion**, die je zwei Elementen des Raums einen nicht negativen reellen Wert zuordnet, der den **Abstand** der beiden Elemente voneinander repräsentiert.

Hierarchische Clusteranalyse
- Metrik -

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9



Sei X eine beliebige Menge. Eine Abbildung

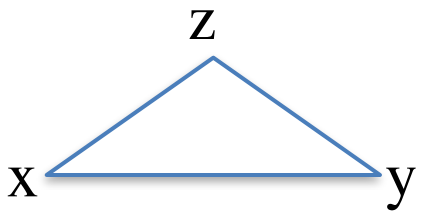
$$d: X \times X \rightarrow \mathbb{R}$$

heißt Metrik auf X, wenn für beliebige Elemente x, y und z von X die folgenden Axiome erfüllt sind:

positive Definiertheit: $d(x,y) \geq 0$ und $d(x,y) = 0 \Leftrightarrow x = y$,

Symmetrie: $d(x,y) = d(y,x)$,

Dreiecks-Ungleichung: $d(x,y) \leq d(x,z) + d(z,y)$.



Typical clustering algorithms: UPGMA (Unweighted Pair Group Methods using arithmetic Averages)

The algorithm uses a pair-wise distance matrix to cluster the elements in a rooted tree. At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters ***X*** and ***Y*** is taken to be the average of all distances between pairs of objects ***x*** in ***X*** and ***y*** in ***Y***, that is, the mean distance between elements of each cluster.

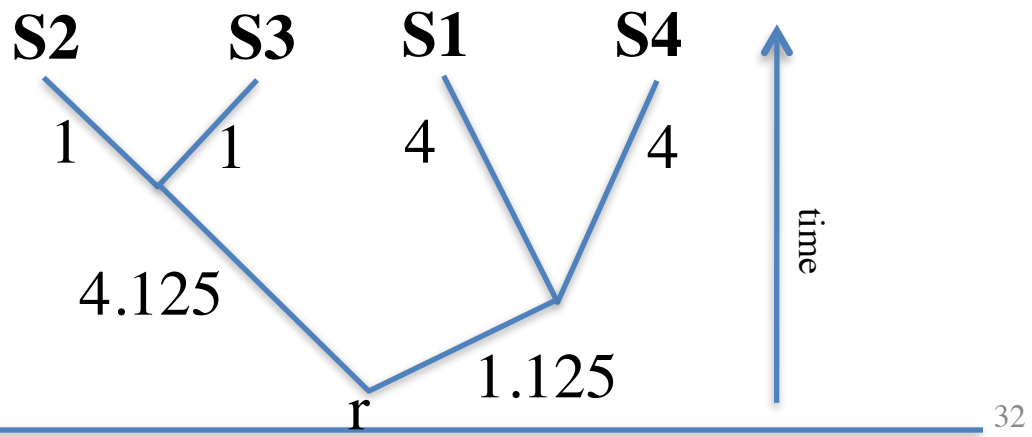
a) Matrix of pair-wise distances

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9

b) Formula to compute the distances between clusters

$$d(X,Y) = \frac{1}{|X| \times |Y|} \sum_{x \in X} \sum_{y \in Y} d(x,y)$$

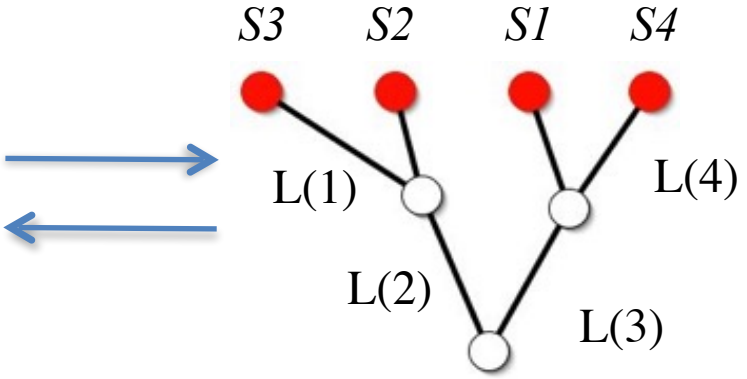
d((S2,S3),(S1,S4))=
¼*(11+11+10+9)
=10.25



Our **assumptions** so far (make sure to understand them!)

1) The data evolved along a clock-like evolving tree.
2) The pair-wise evolutionary distances were fairly accurately estimated from the data.
3) It is for this reason that we can use the distance matrix for tree reconstruction.
4) Any algorithm will typically use only a distance matrix as input and will produce a tree ignoring the previous 3 points!

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9



Hence we need to test whether a given distance matrix can be represented on a (ultrametric[1]) tree.

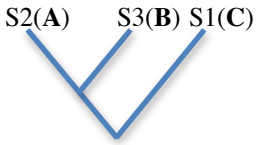
[1] this is a tree in which the distance between all leafs and the root are the same. It is sometimes referred to as 'clock-like' tree.

Theorem: The ultrametric inequality

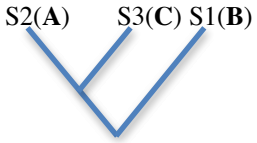
A distance matrix $(d_{i,j})$, $i,j=1....n$, is representable as a clock-like tree, if and only if

$$d(A,B) \le \max\{d(A,C), d(B,C)\}$$
for all triples (A,B,C)

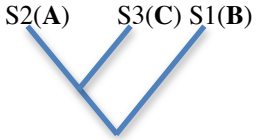
	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9



2 <= max(11,11)



11 <= max(11,2)



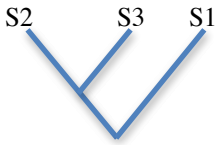
11 <= max(11,2)

Theorem: The ultrametric inequality

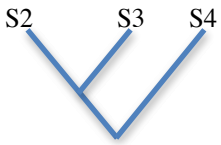
A distance matrix $(d_{i,j})$, $i,j=1....n$, is representable as a clock-like tree, if and only if

$$d(A,B) \le \max\{d(A,C), d(B,C)\}$$
for all triples (A,B,C)

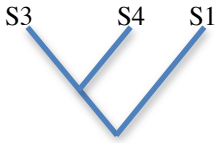
	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9



2 <= max(11,11); 11 <= max(11,2); 11 <= max(11,2)



2 <= max(9,10); 9 <= max(2,10); 10 <= max(2,9)



11 <= max(8,9); 9 <= max(11,8); 8 <= max(11,9)

Theorem: The ultrametric inequality

A distance matrix $(d_{i,j})$, $i,j=1....n$, is representable as a clock-like tree, if and only if

$$d(A,B) \leq \max\{d(A,C), d(B,C)\}$$
for all triples (A,B,C)

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9

Not suitable for UPGMA!

(S1,S2,S3):
11 <= max(11,2); 11 <= max(11,2); 2 <= max(11,11)

(S2,S3,S4):
2 <= max(9,10); 9 <= max(2,10); 10 <= max(2,9)

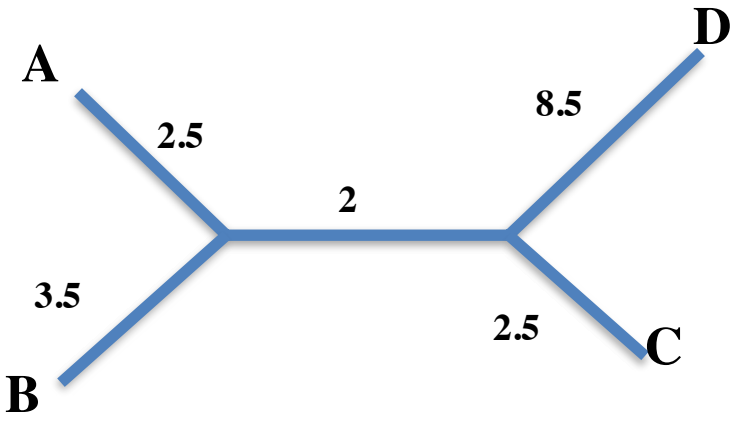
(S1,S2,S4):
11 <= max(8,9); 8 <= max(11,9); 9 <= max(11,8)

Is my distance matrix representable as a tree?

Theorem: Four-Point-Condition

A distance matrix $(d_{i,j})$, $i,j=1....n$, is representable as a tree, if and only if

$$d(A,B)+d(C,D)\leq \max\{d(A,C)+d(B,D),\, d(A,D)+d(B,C)\}$$
$$\text{for all } A,B,C,D\in\{1,2,\ldots,n\}$$



Theorem: Four-Point-Condition

A distance matrix $(d_{i,j})$, $i,j=1....n$, is representable as a tree, if and only if

$$d(A,B)+d(C,D) \le \max\{d(A,C)+d(B,D), d(A,D)+d(B,C)\}$$
$$\text{for all } A,B,C,D \in \{1,2,...,n\}$$

	S1	S2	S3
S2	11		
S3	11	2	
S4	8	10	9

Not representable as a tree!

(S1,S2,S3,S4):
11+9 <= max(11+10,8+2)

(S1,S2,S3,S4):
8+2 <= max(11+10,11+9)

(S1,S2,S3,S4):
11+10 <= max(11+9,8+2)

Bascially, we have three different means to reconstruct phylogenetic trees from sequence data

Find tree that requires the least number of changes

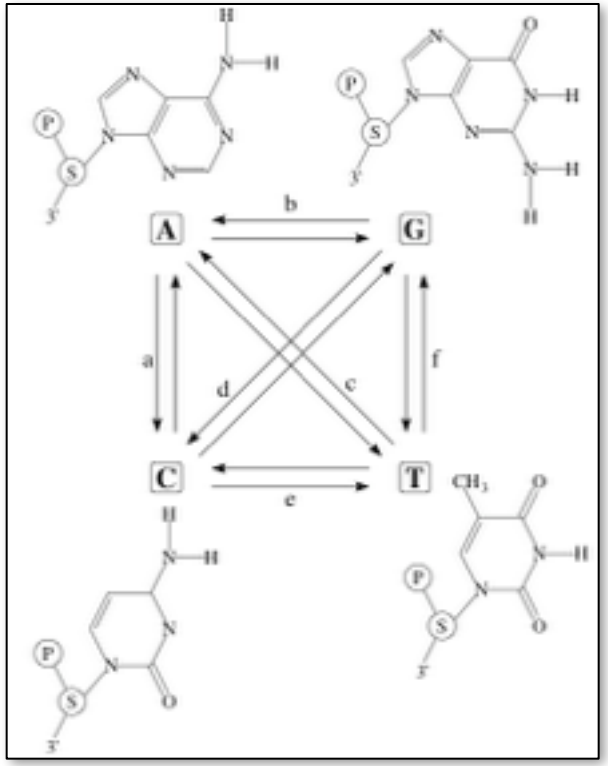
Data	Method	Evaluation Criterion
Characters (Alignment)	Maximum Parsimony	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	Distance Methods	

Find the tree that most likely gave rise to the data

Reconstruct the best fitting tree from a pair-wise distance matrix

Modeling sequence evolution

Evolutionary models are often described using a substitution rate matrix Q and character frequencies Π.[1]



Jukes and Cantor (1969) came up with the simplest substitution model for DNA sequences(JC69): all substitution rates are the same, and all nucleotides occur with the same frequency

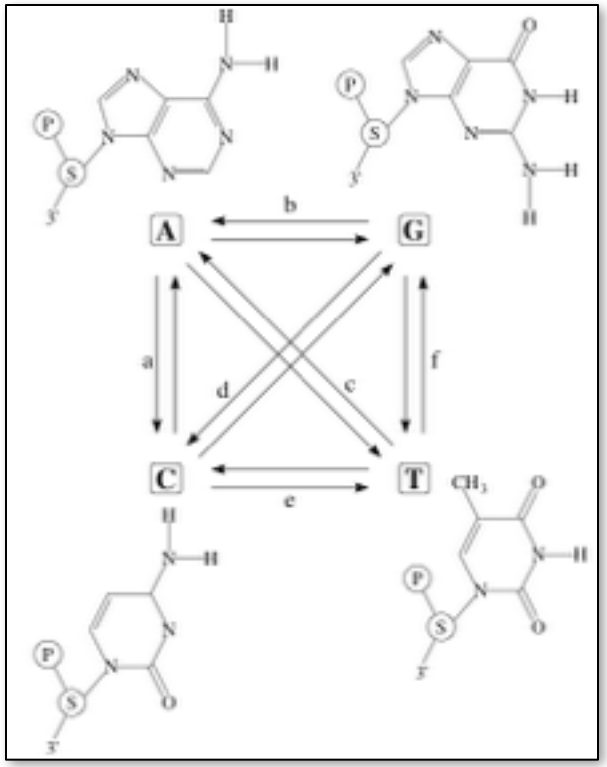
$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & a & a \\ a & - & a & a \\ a & a & - & a \\ a & a & a & - \end{pmatrix} \end{array}$$

$$\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$$

[1] remember the lecture “Modeling Sequence Evolution”

Modeling sequence evolution

Evolutionary models are often described using a substitution rate matrix Q and character frequencies Π.



$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

From Q and Π we reconstruct a substitution probability matrix P where $P_{ij}(t)$ is the probability of changing i to j in time t.

Computing probabilities for observing a given character pairing i,j after time t

Substitution matrix

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Probability matrix

$$P(t) = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} \\ \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} \\ \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} \\ \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}-e^{-4\alpha t} & \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} \end{pmatrix}$$

Now, we can define the probability to observe any site pattern in a pairwise alignment given time *t*

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4at}$$

$$P_{diff}(t) = \frac{3}{4}\left(1 - e^{-4\alpha t}\right)$$

With the likelihood function, we can now compute the likelihood that sequence *S* changes to sequence *S'* in time *t*

S: GGTCCTGACAGAAATAAAC

S': GATCCTGAGAGAAATAAAC

$$L(t \mid s \rightarrow s') = \prod_{i=1}^{m} \left(\pi_{s_i} \times P_{s_i s'_i}(t) \right)$$

m: alignment length

S_i: character at position *i* in sequence *S*

S'_i: character at position *i* in sequence *S'*

This value denotes the probability to observe the site pattern (alignment column) at position ***i*** in the alignment.

More precisely, it is the probability that nucleotide $\boldsymbol{S_i}$ has been substituted by nucleotide $\boldsymbol{S'_i}$ after time ***t***.

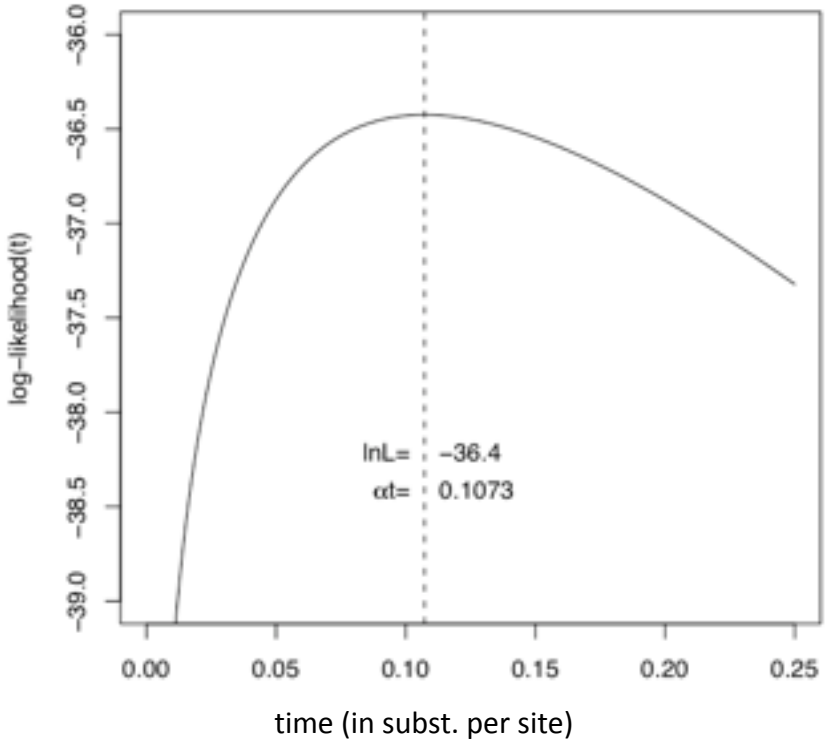
We can now compute the likelihood that S has changed to S' for any given time interval *t* and identify the time for which L(t|S->S')* is maximal.

S: GGTCCTGACAGAAATAAAC

S': GATCCTGAGAGAAATAAAC

$$L\left(t \mid s \rightarrow s'\right) = \prod_{i=1}^{m}\left(\pi_{s_i} \times P_{s_i s_i'}(t)\right)$$

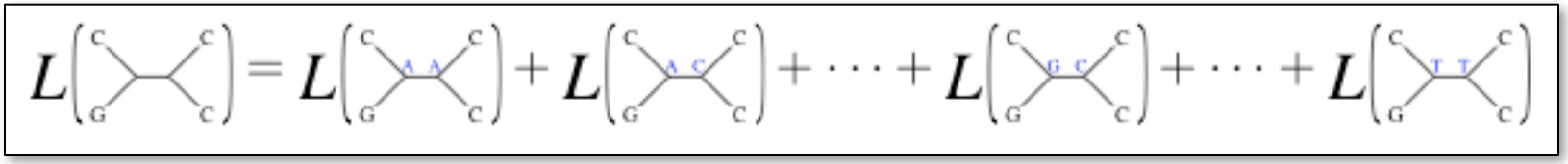
Log-Likelihood surface under JC69



*Note, since the products of probabilities quickly become very small, the likelihood is typically computed and given in log-scale (log-likelihood).

Tree likelihoods

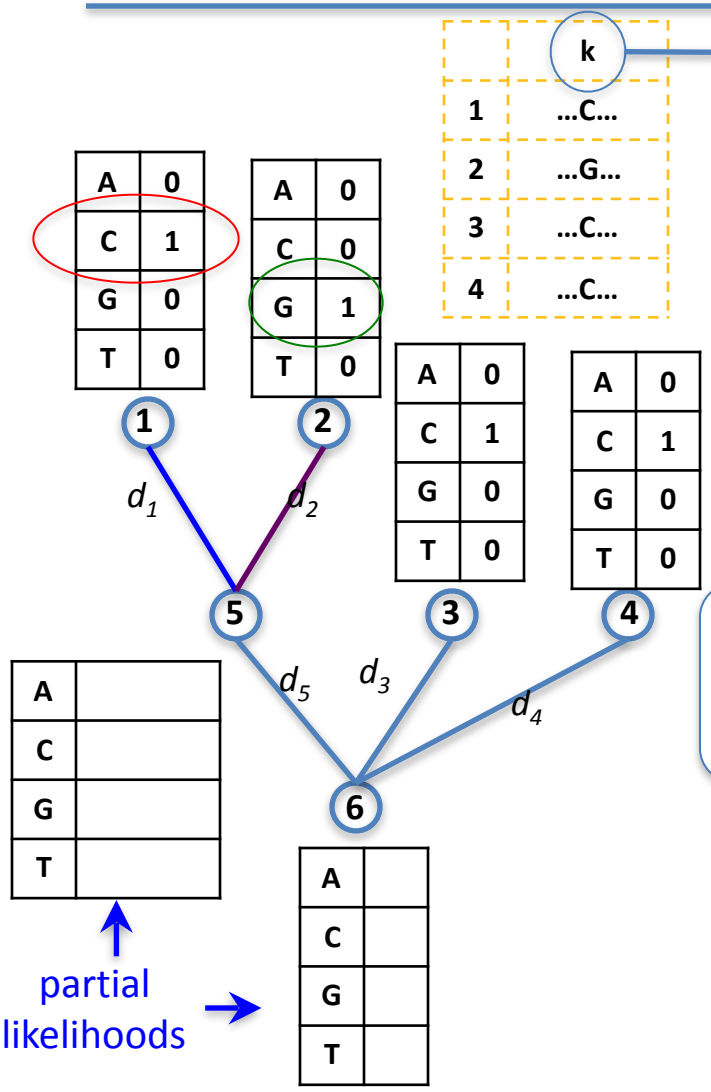
Given a tree with branch lengths and sequences for all nodes, the computation of likelihood values is straightforward. Usually no sequences are available for the inner nodes (ancestral sequences). Hence we have to evaluate every possible labeling at the inner nodes!



for every column in the alignment.

But there is a quicker way…

Calculating tree likelihoods



For a single site pattern *k* and a given tree we compute:

$$L_5(i) = [L(C) \times P_{iC}(d_1)] \times [L(G) \times P_{iG}(d_2)], \forall\, i \in \{A, C, G, T\}$$ [1]

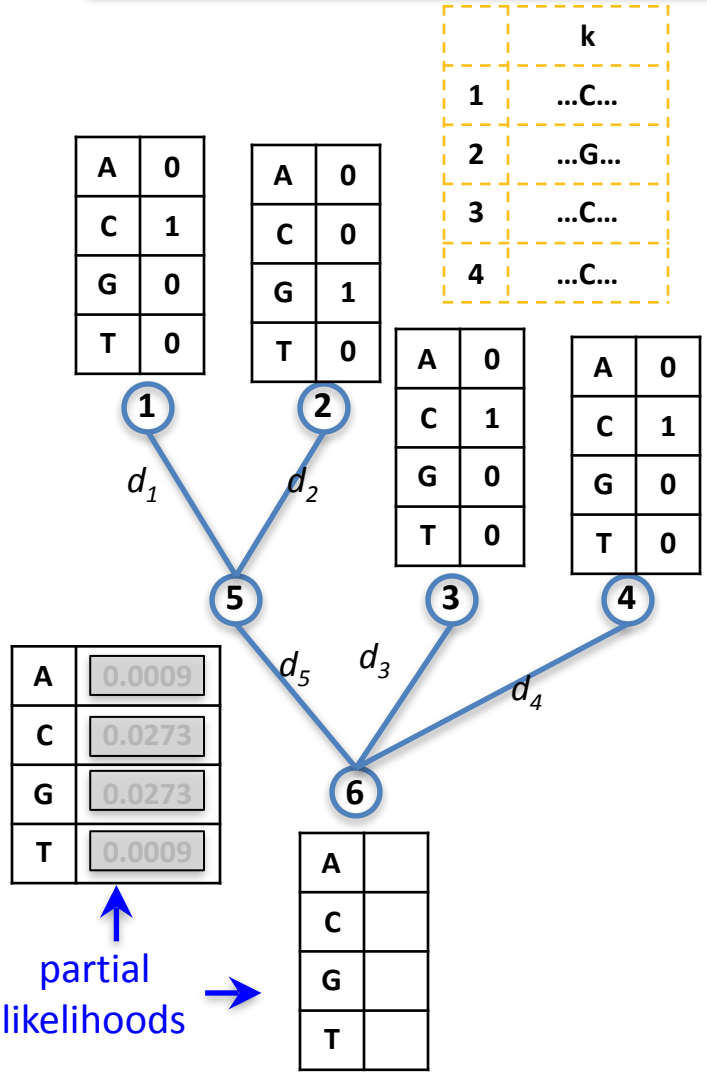
Note, these probabilities do not change as the bases are observed in the data!

Note, these probabilities DO change with *i and d,* and are specified by the substitution model!

with $d_x = 0.1\ \forall\ x \in \{1,..,5\}$, and $P_{ij}(0.1) = \begin{cases} 0.91 \text{ if } i = j \\ 0.03 \text{ for each } i \neq j \end{cases}$

[1] here you compute the partial likelihood $L_5(i)$ for each ancestral nucleotide *i* in node 5 of the tree, given the data in *k* and the model

Calculating tree likelihoods



For a single site pattern *k* and a given tree:

$$L_5(i) = [L(C) \times P_{iC}(d_1)] \times [L(G) \times P_{iG}(d_2)], \forall\, i \in \{A, C, G, T\}$$

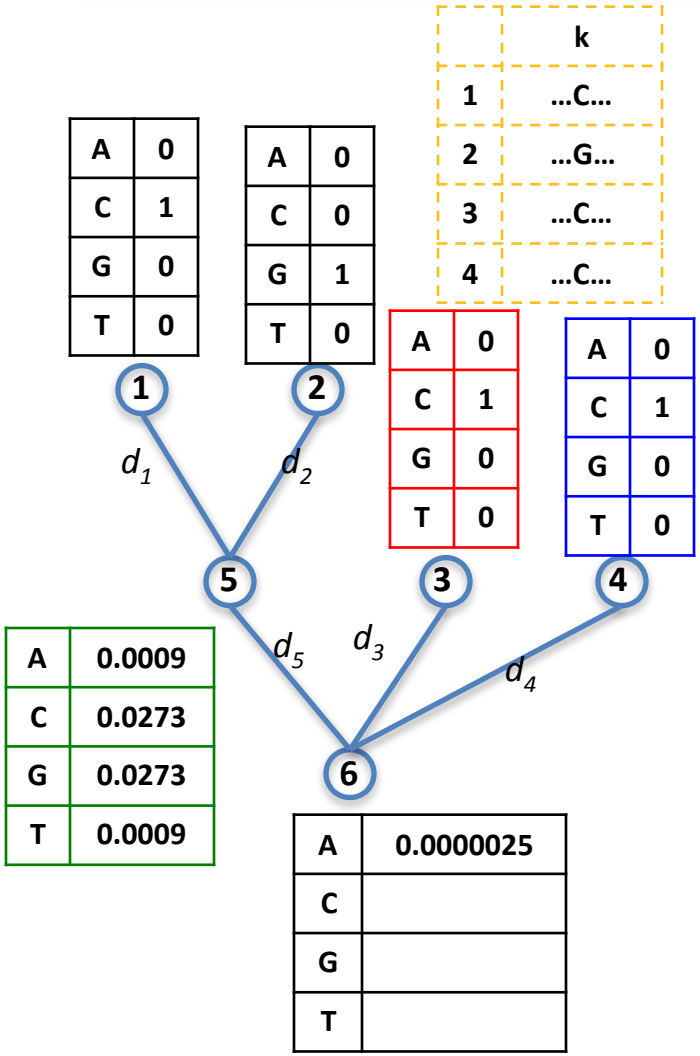
$$L_5(A) = [1 \times P_{AC}(0.1)] \times [1 \times P_{AG}(0.1)]$$
$$= 1 \times 0.03 \times 1 \times 0.03$$
$$= 0.0009$$

$$L_5(C) = [1 \times P_{CC}(0.1)] \times [1 \times P_{CG}(0.1)]$$
$$= 1 \times 0.91 \times 1 \times 0.03$$
$$= 0.0273$$

$L_5(G)$ and $L_5(T)$ are computed analogously

with $d_x = 0.1\ \forall\ x \in \{1,..,5\}$, and $P_{ij}(0.1) = \begin{cases} 0.91 \text{ if } i = j \\ 0.03 \text{ for each } i \neq j \end{cases}$

Calculating tree likelihoods



For a single site pattern *k* and a given tree:

$$L_6(i) = \prod_{v=\{3,4,5\}} \left[\sum_{j=\{A,C,G,T\}} L_v(j) \times P_{ij}(d_v) \right], \forall\, i \in \{A,C,G,T\} \quad *$$

Likelihood of nucleotide **A** at node **6** in column **k**!

$$L_6(A) =$$

$$[L_5(A) \times P_{AA}(0.1)] \times [L_3(C) \times P_{AC}(0.1)] \times [L_4(C) \times P_{AC}(0.1)] +$$

$$[L_5(C) \times P_{AC}(0.1)] \times [L_3(C) \times P_{AC}(0.1)] \times [L_4(C) \times P_{AC}(0.1)] +$$

$$[L_5(G) \times P_{AC}(0.1)] \times [L_3(C) \times P_{AC}(0.1)] \times [L_4(C) \times P_{AC}(0.1)] +$$

$$[L_5(T) \times P_{AC}(0.1)] \times [L_3(C) \times P_{AC}(0.1)] \times [L_4(C) \times P_{AC}(0.1)]$$

$$= [(0.0009 \times 0.91) + (0.0273 \times 0.03) + (0.0273 \times 0.03) + (0.0009 \times 0.03)] \times 0.03 \times 0.03$$

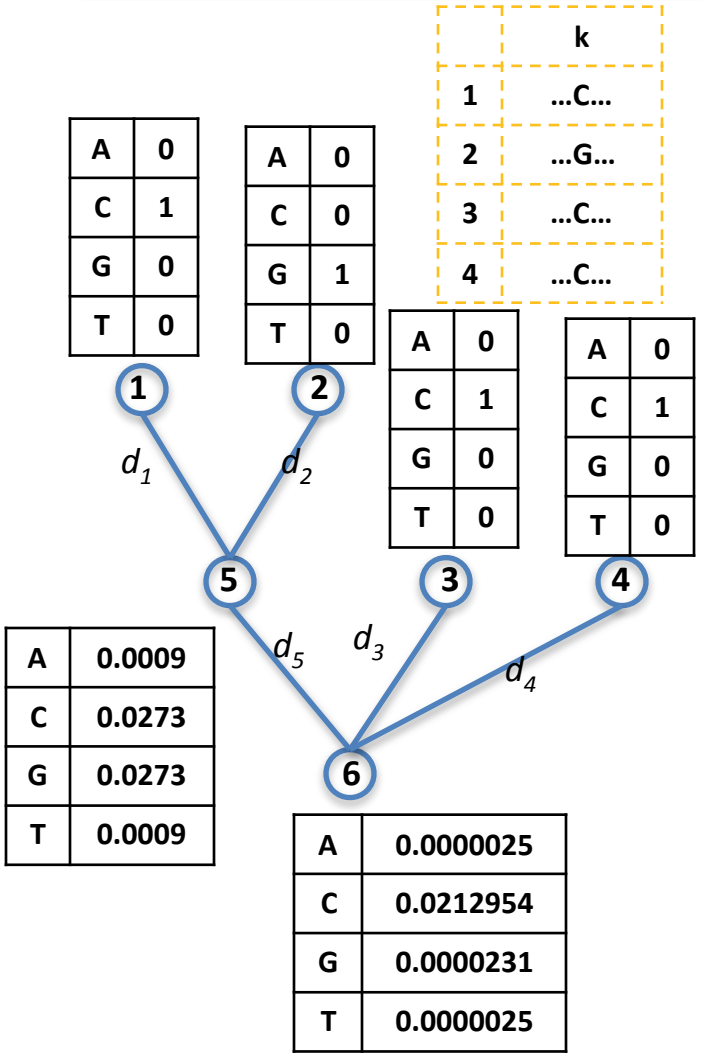
$$= 0.002727 \times 0.03 \times 0.03$$

$$= 0.0000025$$

with $d_x = 0.1 \;\forall\; x \in \{1,..,5\}$, and $P_{ij}(0.1) = \begin{cases} 0.91 \text{ if } i = j \\ 0.03 \text{ for each } i \neq j \end{cases}$

* Note, the v represents the nodes for which the partial likelihoods have already been computed. The sum indicates that you sum over all possible internal labels. Note, that for leaf nodes the probability of the observed nucleotide is 1 and that of the other nucleotides is 0! Hence, for nodes 3 and 4 there is no need to compute a sum!

Calculating tree likelihoods



For a single site pattern *k* and a given tree:

$$L_5(i) = [L(C) \times P_{iC}(d_1)] \times [L(G) \times P_{iG}(d_2)], \forall\, i \in \{A, C, G, T\}$$

$$L_6(i) = \prod_{v=\{3,4,5\}} \left[\sum_{j=\{A,C,G,T\}} L_v(j) \times P_{ij}(d_v) \right], \forall\, i \in \{A, C, G, T\} \quad *$$

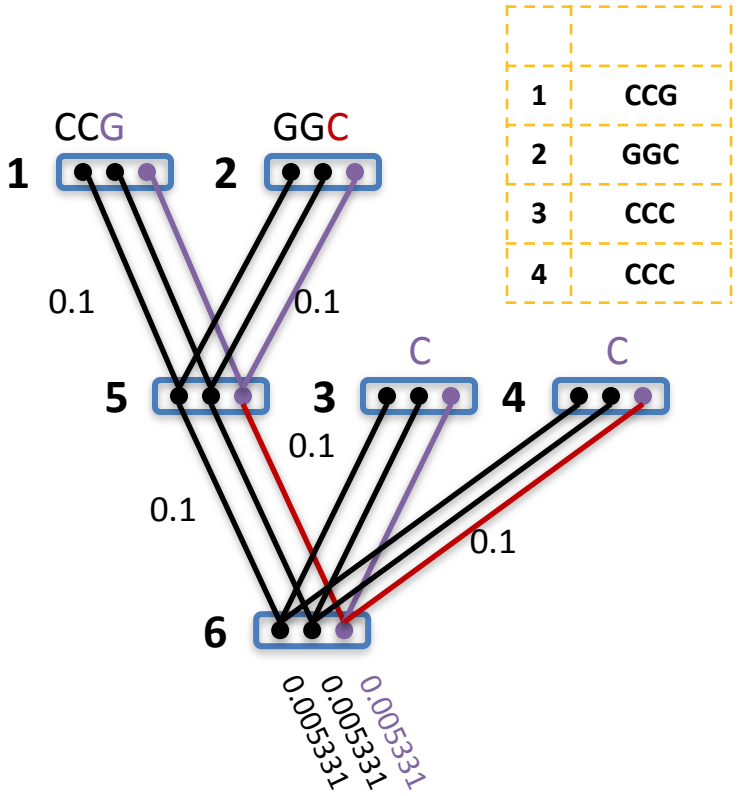
$$L^{(k)} = \sum_{i=\{A,C,G,T\}} \pi_i \times L_6(i) = 0.005331; \text{ mit } \pi_i = 0.25 \forall i \in \{A, G, C, T\}$$

This is the **site likelihood** of the pattern *k* given the tree

$$\text{with } d_x = 0.1 \;\forall\; x \in \{1,..,5\}, \text{ and } P_{ij}(0.1) = \begin{cases} 0.91 \text{ if } i = j \\ 0.03 \text{ for each } i \neq j \end{cases}$$

* Note, the v represents the nodes for which the partial likelihoods have already been computed. The sum indicates that you sum over all possible internal labels.

Calculating tree likelihoods



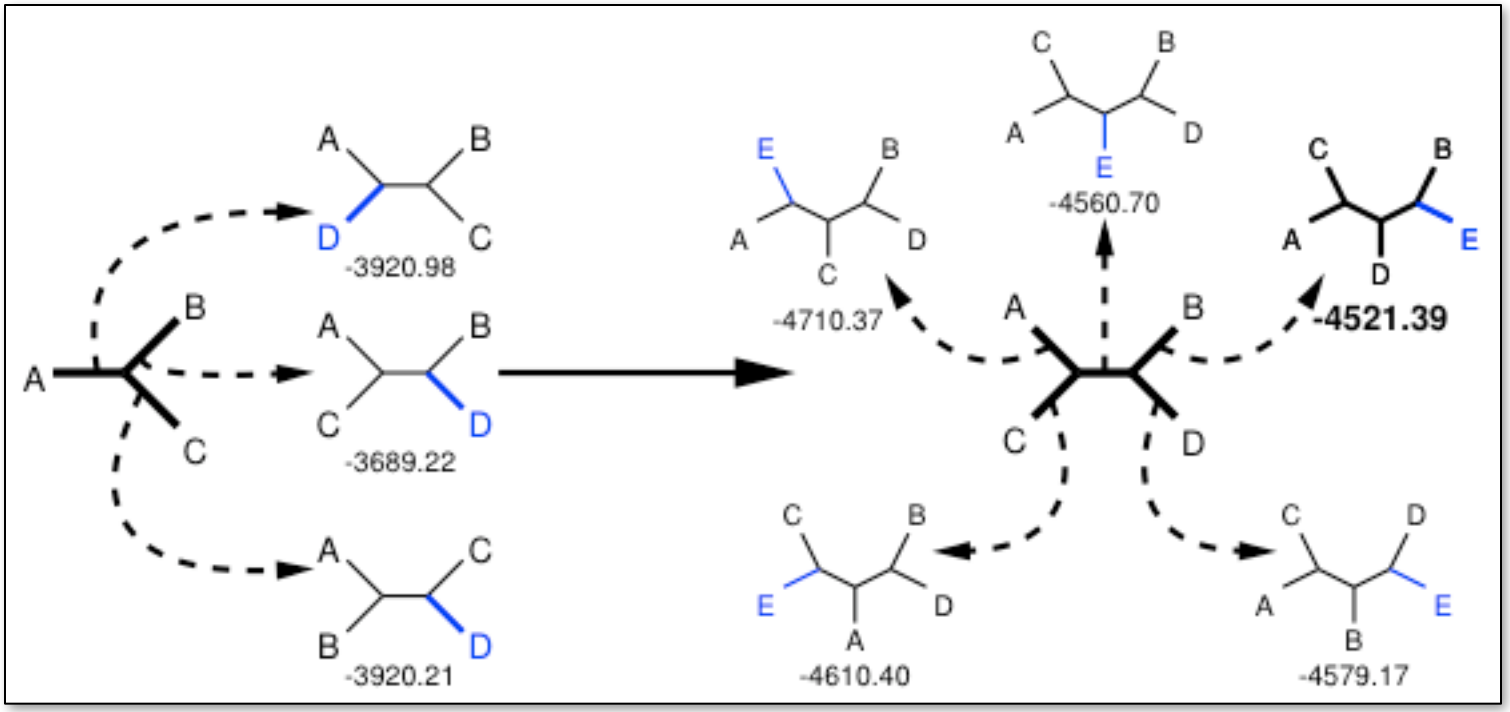
For an alignment of four sequences and length m=3 the likelihood is then

$$L(T) = \prod_{k=1}^{m} L^{(k)} = 0.005331^2 \times 0.005331 = 0.000000152$$

or the log-likelihood is

$$\ln L(T) = \sum_{k=1}^{m} \ln L^{(k)} = -15.7$$

Building and evaluating a tree is simple: e.g. stepwise insertion starting from a 4-taxon tree



Searching tree space is a bit more complicated…

The missing bit: Tree evaluation using Bayes theorem

So far we have computed

$$P(D \mid \mathrm{T}, \Theta)$$

i.e. the likelihood of the data *D* given the tree *T* and the parameter vector Θ.

However, what we are interested in most of the times is the likelihood of *T* and Θ given *D*, i.e.

$$P(\mathrm{T}, \Theta \mid D)$$

The missing bit: Tree evaluation using Bayes theorem

So far we have computed

$$P(D \mid \mathrm{T}, \Theta)$$

i.e. the likelihood of the data *D* given the tree *T* and the parameter vector Θ.

However, what we are interested in most of the times is the likelihood of *T* and Θ given *D*, which is given by Bayes' theorem

$$P(\mathrm{T}, \Theta \mid D) = \frac{P(D \mid \mathrm{T}, \Theta) * P(\mathrm{T}, \Theta)}{P(D)}$$

$P(D)$: total probability of the data considering all hypotheses. This is the problematic bit!

$P(\mathrm{T}, \Theta)$: **prior information** on the probability of a given hypothesis (T, Θ)

Finding the best tree is highly problematic!

1. **Exhaustive Search:** evaluates every possible tree and hence an optimal solution is guaranteed. Limit: 10-12 taxa

2. **Branch and Bound:** excludes parts from the tree space from the search where the optimal tree cannot be found. Guarantees to find the optimal tree.

3. **Heuristics:** Can be applied to large taxon sets but does not guarantee an optimal solution

To be told on another day….