

# Bioinformatik

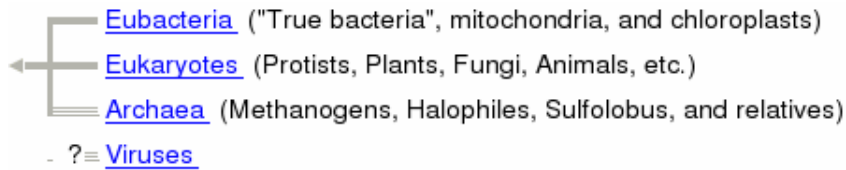
Distanzbasierte phylogenetische Algorithmen

Ulf Leser

Wissensmanagement in der  
Bioinformatik



# Klassifikation



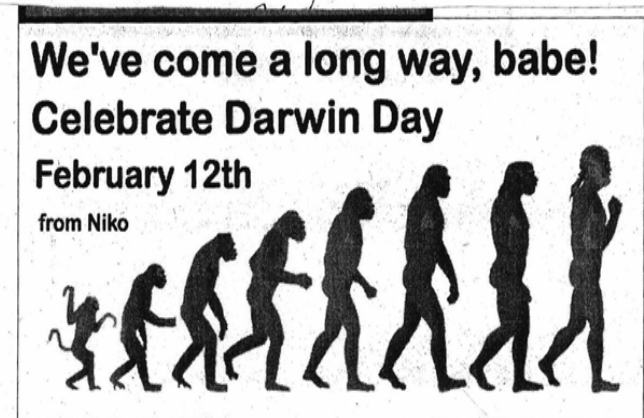
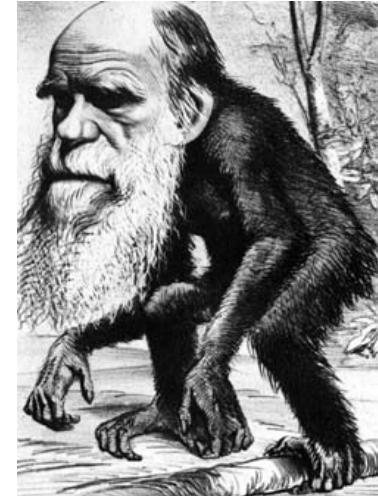
- Eukaryoten
- Tiere
- diverse Zwischenstufen
- Charniata (Schädelknochen)
- Vertebraten (Wirbeltier)
- Viele Zwischenstufen
- Mammals (Säugetiere)
- Eutheria (Placenta)
- Primaten (Affen)
- Catarrhini
- Hominidae (Mensch, Schimpanse, Orang-Utan, Gorilla)
- Homo (erectus, sapiens ...)
- Homo Sapiens

## Popular Groups on the Tree of Life

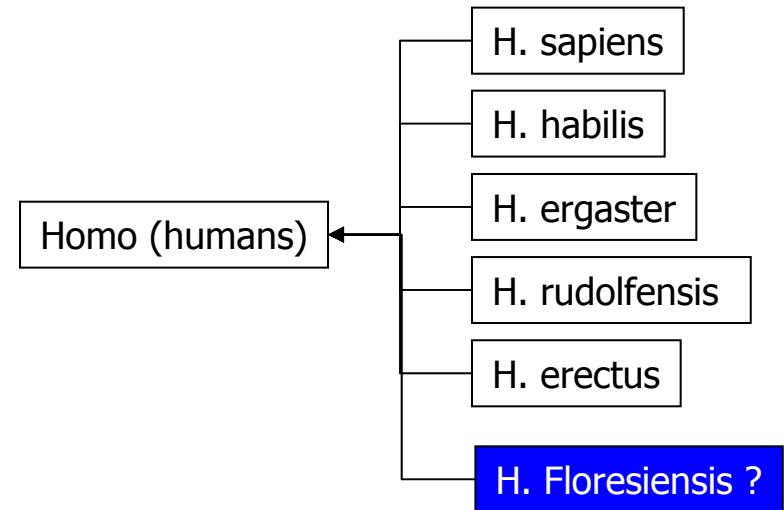


# Stammbäume

- Charles Darwin: „The origin of species“ (1859)
  - Arten sind nicht unveränderlich, sondern unterliegen im Laufe der Zeit einem **Wandel**
  - „Survival of the fittest“
  - Damals war noch unklar, was sich eigentlich wie wandelt
- **Stammbäume** (Abstammungsbäume)
  - Ergeben sich aus der Annahme der Evolution
  - Auch Stammbäume wurden lange aus morphologischen Eigenschaften abgeleitet
  - Jeder innere Knoten in einem Stammbaum hat einmal als Art existiert
  - Knoten im Baum (also Spezies) heißen **taxa**
- Was definiert eine Spezies?



# „Homo floresiensis“ (Hobbit?)

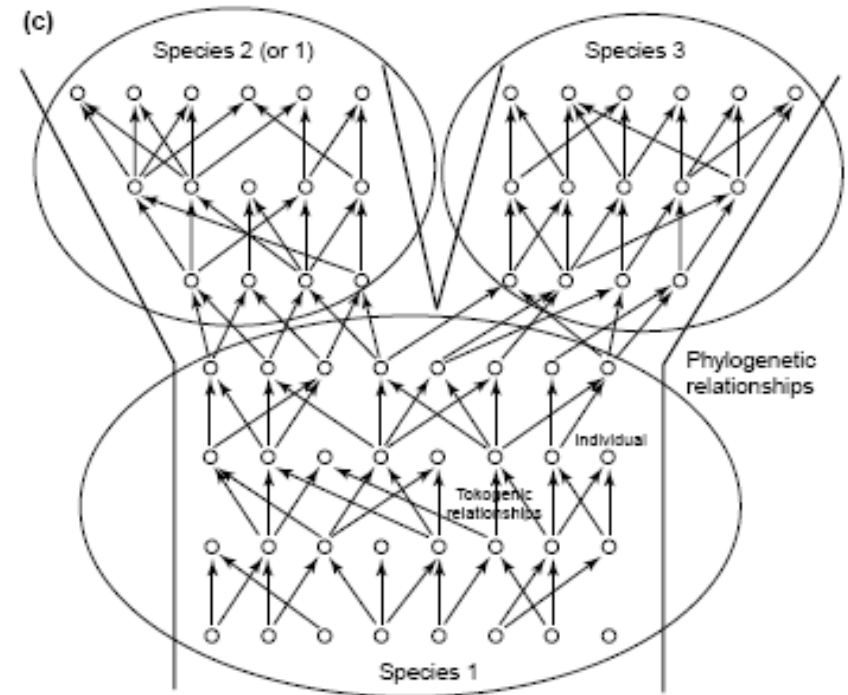


- Entdeckung 2003
- Lebte vermutlich vor 95 000 bis vor 12 000 Jahren auf der indonesischen Insel Flores
- Körperhöhe nur ca. 100 cm
- Gehirnvolumen nur 380 cm<sup>3</sup>
  - Homo erectus: 600-1200, h. sapiens: 1400
- Rückentwicklung? Vorläufer?

Quelle:  
Brown, P. *et al.* *Nature* 431, 1055-1061 (2004).

# Arten von Stammbäumen

- Individuelle Abstammung
  - Stammbäume, Ahnentafeln
  - **Kein Baum**: Zwei Eltern
  - Rekombination
- Speziesstammbäume
  - Das ist ein Baum, wenn **Spezies nicht verschmelzen** können
- **Gene Trees**
  - Geschichte eines Sequenzabschnitts
    - Nicht leicht zu definieren
  - Baumförmig, wenn Gene nicht verschmelzen
  - Aber: 2 Allele jedes Gens vorhanden (Besser: **Haplotyp Tree**)



TRENDS in Ecology & Evolution

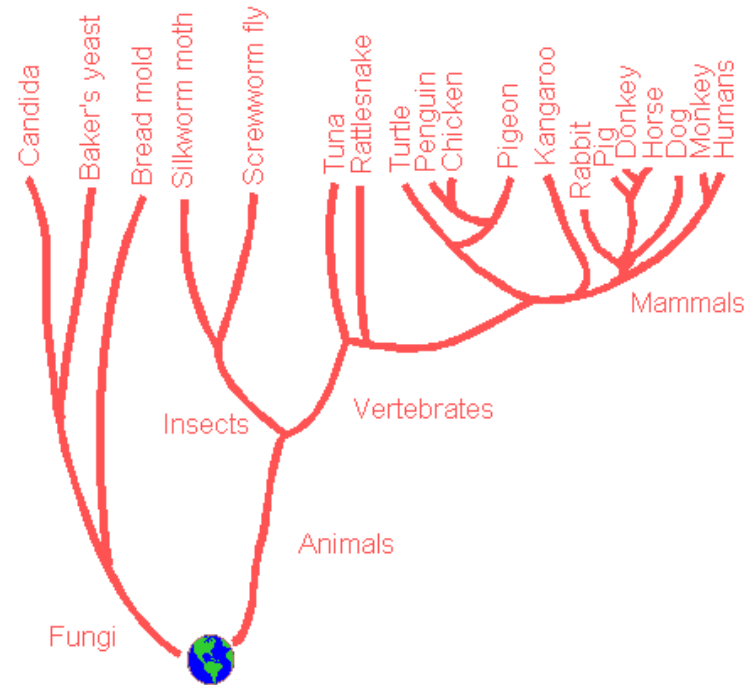
# Molecular Clock

---

- Häufige Annahme: **Molecular Clock**
  - Mutationen setzen sich bzgl. der Zeit immer mit gleicher Häufigkeit durch
  - Unabhängig von Teilbaum, Zeitpunkt, Ort und Art der Mutation
  - Anzahl der beobachteten Mutationen „proportional“ zur vergangenen Zeit
- Die ist hilfreich, aber falsch
  - Unterschiedliche Zeiten erzeugen unterschiedlichen Selektionsdruck
  - Unterschiedliche Sequenzabschnitte unterliegen unterschiedlichem Selektionsdruck
  - Unterschiedliche Spezies unterliegen unterschiedlicher Frequenz erfolgreicher Mutationen
  - ...

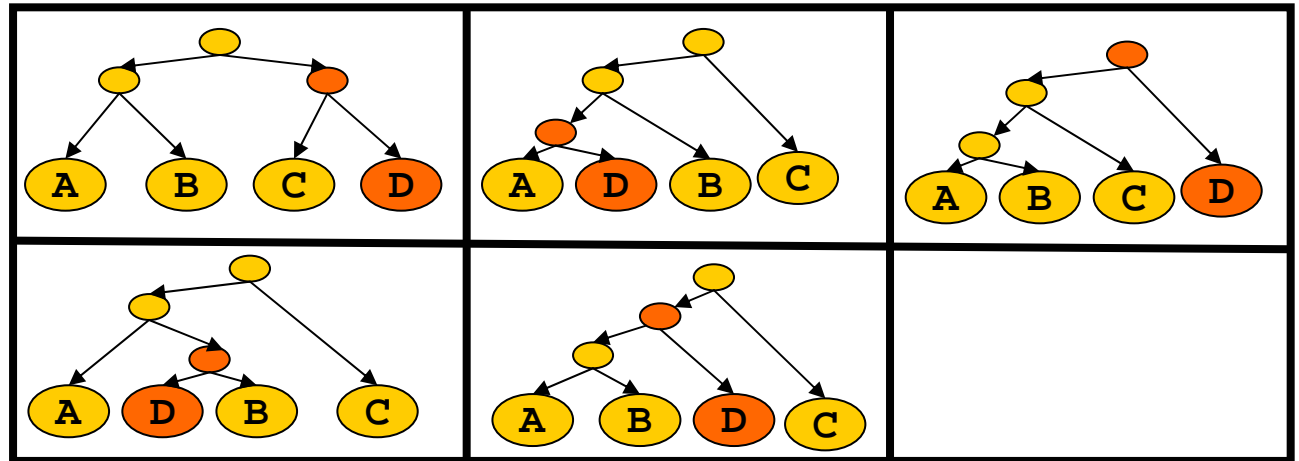
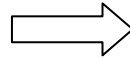
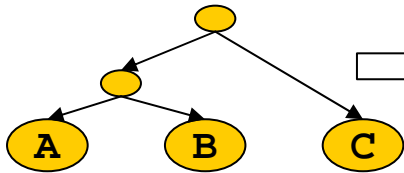
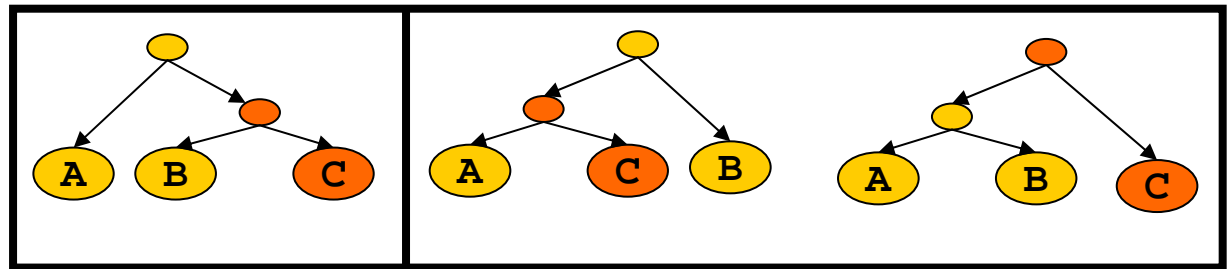
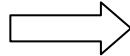
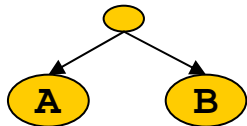
# Bäume

- Knoten = Arten
- Blätter = Lebende Arten
- Kanten
  - Länge kann (aber muss nicht) mit zeitlicher Entfernung korrelieren
  - Scaled trees
- Jeder Knoten hat exakt einen Vater
- Wurzel oder nicht
- Binäre Bäume
  - Unproblematisch (gleich)
- Reihenfolge der Kinder ist egal
- Viele Visualisierungsvarianten



# Wie schwierig wird das?

Wie viele binäre, ungeordnete Bäume für n Spezies gibt es?



# Ergebnis

- Sei  $t(n)$  die Zahl binärer Bäume mit  $n$  Blättern

$$\begin{aligned}
 t(n) &= t(2) * t(3) * t(4) * \dots * t(n-1) = \\
 &= 1 * 3 * 5 * \dots * (2(n-1) - 1) = \\
 &= \frac{(2n-3)!}{2 * 4 * 6 * \dots * (2n-4)} = \\
 &= \frac{(2n-3)!}{2 \binom{2}{2} * 2 \binom{4}{2} * 2 \binom{6}{2} * \dots * 2 \binom{2n-4}{2}} = \\
 &= \frac{(2n-3)!}{2 * (1) * 2 * (2) * 2 * (3) * \dots * 2(n-2)} = \\
 &= \frac{(2n-3)!}{2^{n-2} * (n-2)!}
 \end{aligned}$$

1	1
2	1
3	3
4	15
5	105
6	945
7	10.395
8	135.135
9	2.027.025
10	34.459.425
11	654.729.075
12	13.749.310.575
13	316.234.143.225
14	7.905.853.580.625
15	213.458.046.676.875
16	6.190.283.353.629.370
17	191.898.783.962.511.000
18	6.332.659.870.762.850.000
19	221.643.095.476.700.000.000
20	8.200.794.532.637.890.000.000

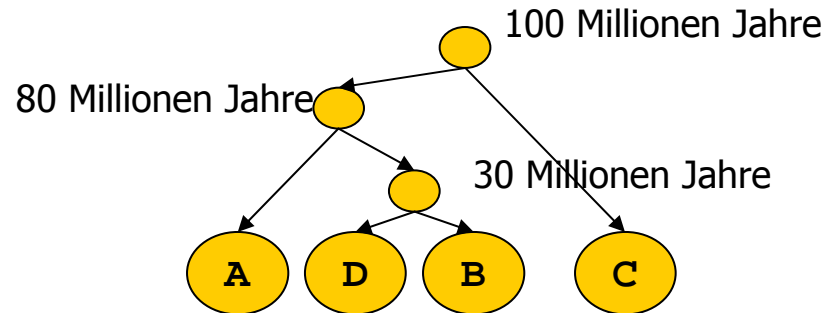
# Inhalt dieser Vorlesung

---

- Distanzbasierte Phylogenie
- Ultrametrien und UPGMA
- Additive Bäume und Neighbor Joining

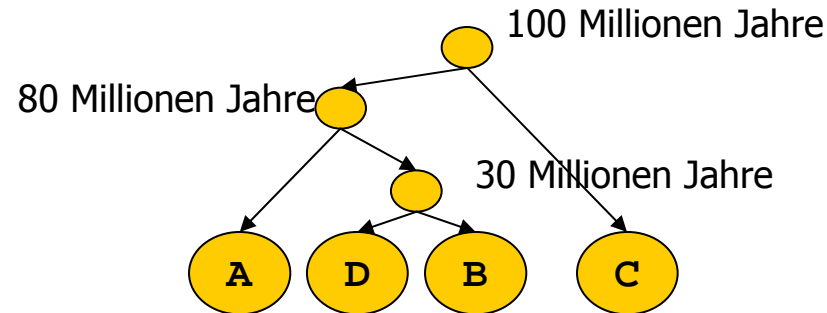
# Voraussetzungen

---



- An innere Knoten eines Stammbaum kann man den Zeitpunkt der Speziation schreiben („branch points“)
- Wenn die Molecular Clock Theory gilt
  - Ist die Menge an Veränderungen auf einer Kante **proportional zu der verstrichenen Zeit**
  - Damit proportional zur Länge der Kante
  - Haben alle Geschwister den gleichen Abstand zum Elternknoten
  - Ist der Editabstand zweier Knoten (=Evolution dazwischen) proportional zur Summe der **Editabstände beider Knoten zum kleinsten gemeinsamen Vater**

# Ultrametrien



- Wenn man den Baum und die Zeitpunkte weiß, dann gilt
  - Alle Zahlen auf einem Pfad von der Wurzel zu einem beliebigen Blatt nehmen strikt ab
  - Der **Zeitpunkt der Aufspaltung ist ein Abstandsmaß** für zwei Arten
    - Für Knoten  $X, Y$  sei  $d(X, Y)$  das Label des kleinsten gemeinsamen Vorfahren
    - Im Beispiel:  $d(A, B) = 80$ ,  $d(B, C) = 100$ ,  $d(A, D) = 80$
  - Das ist eine Metrik
    - $d(X, X) = 0$ ,  $d(X, Y) > 0$ ,  $d(X, Y) = d(Y, X)$ , und  $d(X, Y) \leq d(X, Z) + d(Z, Y)$
  - Es ist sogar eine **Ultrametrik** (gleich)

# Alles ganz einfach?

---

- Wir sehen gleich
  - Schreibt man diese Abstände in eine Ähnlichkeitsmatrix, hat diese bestimmte Eigenschaften
  - Jede Matrix mit diesen Eigenschaften entspricht genau einem Baum und umgekehrt
  - Den Baum kann man aus der Matrix effizient berechnen
- Also ist das Phylogenie-Problem überhaupt kein Problem?
- Doch
  - **Molecular Clock Theory stimmt nicht**
  - Editabstand korreliert nicht direkt mit der vergangenen Zeit
  - Geschwisterknoten sammeln nicht die gleiche Menge an Mutationen

# Ultrametrische Bäume

---

- Definition

*Sei  $T$  ein binärer, gewurzelter Baum und  $D$  eine symmetrische Matrix mit  $n$  Zeilen und  $n$  Spalten.  $T$  heißt **ultrametrischer Baum** für  $D$  wenn gilt:*

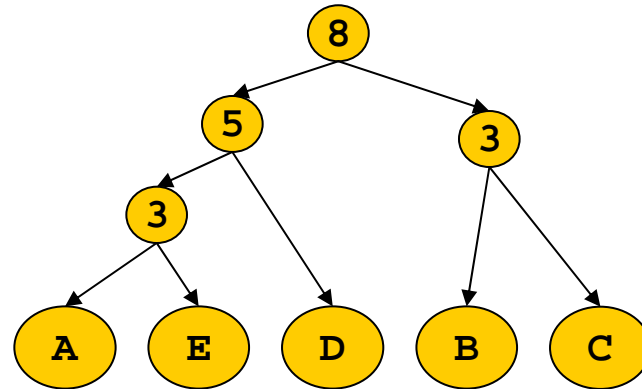
- *$T$  hat  $n$  Blätter, beschriftet mit den Zeilen von  $D$*
- *Jeder innere Knoten von  $T$  ist mit einem Wert aus  $D$  beschriftet*
- *Auf jedem Pfad von der Wurzel zu einem Blatt in  $T$  sind die Beschriftungen der inneren Knoten strikt abnehmend*
- *Für alle Blätter  $i, j$  mit  $i \neq j$  gilt: der **kleinste gemeinsame Vorfahr** von  $i$  und  $j$  ist mit  $D[i, j]$  beschriftet*

- Bemerkung

- Jeder Stammbaum ist ultrametrisch für die Abstandsmatrix mit den Speziationszeitpunkten als Abstandsmaß

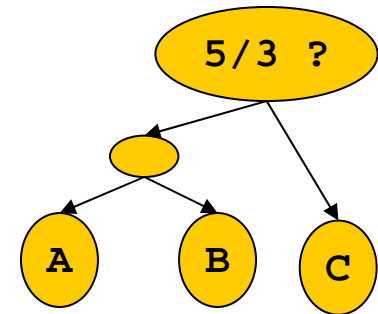
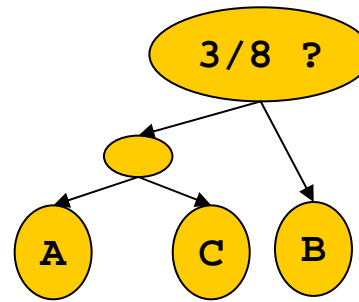
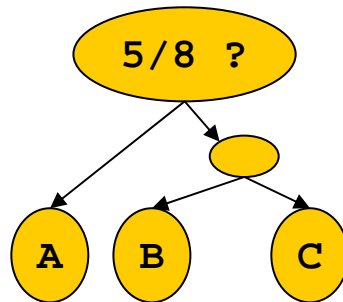
# Beispiel

	A	B	C	D	E
A		8	8	5	3
B			3	8	8
C				8	8
D					5
E					



Das geht nicht immer

	A	B	C
A		8	5
B			3
C			



# Überlegungen

---

- Das kann auch nicht immer gehen
  - Matrix hat  $(n^2-n)/2$  relevante Zellen
  - Baum hat nur  $n-1$  innere Knoten
  - Eine Matrix, zu der man einen ultrametrischen Baum konstruieren kann, muss also Duplikate enthalten
- Definition

*Eine symmetrische Matrix  $D$  mit  $n$  Spalten und Zeilen ist **ultrametrisch**, wenn für beliebige Zeilen  $i, j, k$  gilt, dass das Maximum von  $D[i,j]$ ,  $D[j,k]$  und  $D[i,k]$  genau zweimal vorkommt*
- Bemerkung
  - Also entweder
    - $D[i,j]=D[j,k]$  und  $D[i,j]>D[i,k]$
    - $D[i,j]=D[i,k]$  und  $D[i,j]>D[j,k]$
    - $D[j,k]=D[i,k]$  und  $D[j,k]>D[i,j]$

# Ultrametrik

---

- Definition

*Eine Ultrametrik ist eine Metrik für die gilt:*

$$d(a,c) \leq \max(d(a,b), d(b,c))$$

- Bemerkung

- Für Metriken muss nur gelten:  $d(a,c) \leq d(a,b) + d(b,c)$
- Jede Ultrametrik ist eine Metrik, aber nicht umgekehrt

# Von der Matrix zum Baum und zurück

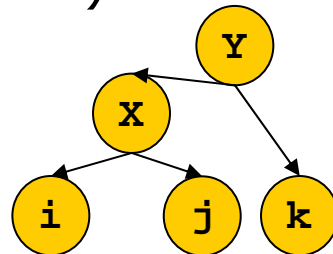
---

- Theorem

*Eine symmetrische Matrix  $D$  hat einen ultrametrischen Baum gdw.  $D$  ultrametrisch ist.*

- Beweis

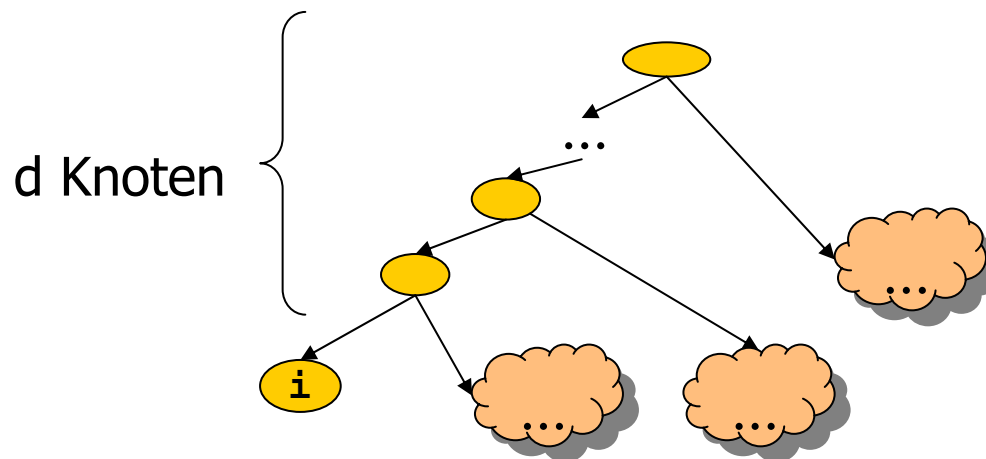
- (1) Nehmen wir erst an, dass zu  $D$  ein ultrametrischer Baum  $T$  existiert. Nehmen wir an, dass  $i, j, k$  wie folgt liegen (alle anderen Knoten können wir ignorieren)



- Dann gilt offensichtlich  $D[i,k]=D[j,k]=Y$  und  $D[i,k]>D[i,j]=X$
- Dito kann man für die zwei anderen möglichen Lagen von  $i, j, k$
- Das gilt für alle Tripel von Knoten
- Also ist  $D$  ultrametrisch

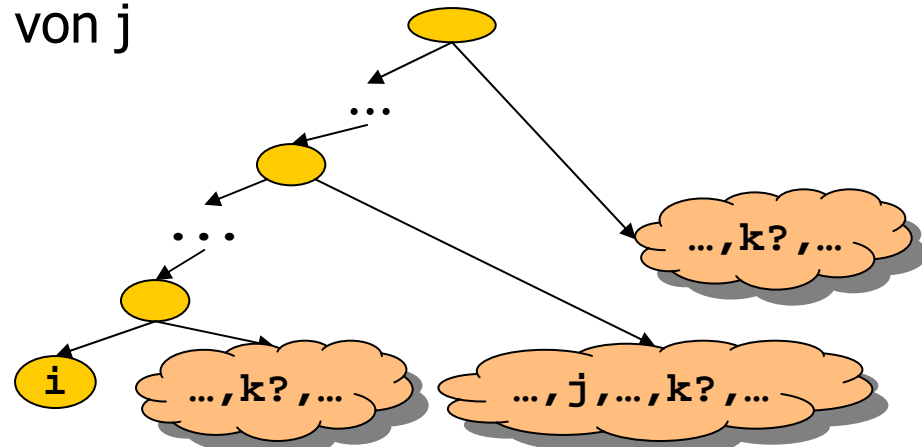
# Beweis Gegenrichtung

- (2) Nehmen wir an, dass  $D$  ultrametrisch ist. Wir konstruieren einen ultrametrischen Baum  $T$  aus  $D$ 
  - Betrachten wir eine beliebige Zeile  $i$ . Diese entspricht Blatt  $i$  in  $T$
  - $i$  hat diverse kleinste gemeinsame Vorfahren mit den  $n-1$  anderen Blättern
  - Diese Vorfahren müssen mit den Werten  $D[i,x]$ ,  $x \neq i$ , in aufsteigender Reihenfolge beschriftet werden.
  - Der Pfad von  $i$  zur Wurzel muss nicht  $n-1$  Knoten enthalten, denn die kleinsten gemeinsamen Vorfahren mit verschiedenen anderen Knoten sind oft identisch. Nehmen wir an, dass es auf dem Pfad  $d$  verschiedene Werte gibt ( $d \leq n-1$ )



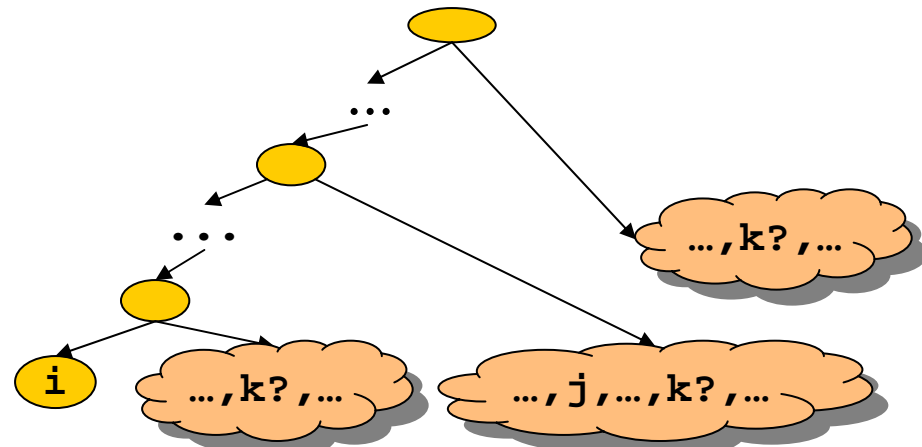
# Beweis Gegenrichtung -2-

- Die Menge aller Blätter (ohne  $i$ ) zerfällt damit **in  $d$  Klassen**
  - Alle Knoten einer Klasse befinden sich in einem Unterbaum bzgl. eines Knoten auf dem Pfad von  $i$  zur Wurzel
  - Alle Knoten einer Klasse haben den selben Abstand zu  $i$
- Betrachten wir ein Blatt  $j \neq i$  und ein beliebiges anderes Blatt  $k \neq i$ . Drei Möglichkeiten
  - $k$  liegt in der selben Klasse wie  $j$
  - $k$  liegt in einer Klasse „links“ von  $j$
  - $k$  liegt in einer Klasse „rechts“ von  $j$



# Beweis Gegenrichtung -3-

- Fall 1:  $j$  und  $k$  in der selben Klasse
  - Das gilt, wenn  $D[i,j]=D[i,k]$  und  $D[j,k]<D[i,j]$
- Fall 2:  $k$  liegt links von  $j$ 
  - Das gilt, wenn  $D[i,j]=D[k,j]$  und  $D[i,k]<D[i,j]$
- Fall 3:  $k$  liegt rechts von  $j$ 
  - Das gilt, wenn  $D[i,k]=D[j,k]$  und  $D[i,j]<D[j,k]$
- Damit können wir alle Knoten  $\neq i$  nach ihrem Abstand zu  $i$  in Klassen einsortieren
- Für jede der Klassen gehen wir rekursiv vor
- qed.



# Folgerung

---

- Der Beweis ist konstruktiv; man kann genau auf diese Weise einen ultrametrischen Baum bauen
- Theorem  
*Sei  $D$  eine ultrametrische Matrix. Dann gibt es **genau einen ultrametrischen Baum  $T$  für  $D$ .***
- Beweis
  - Durch Konstruktion
- Kann in  $O(n^2)$  implementiert werden
  - Später

# Einschub: Distanzmatrizen und Metriken

---

- Andere **Metriken** sind z.B. Manhattanndistanz oder der Euklidische Abstand
  - Allgemeiner: Minkowski Metriken

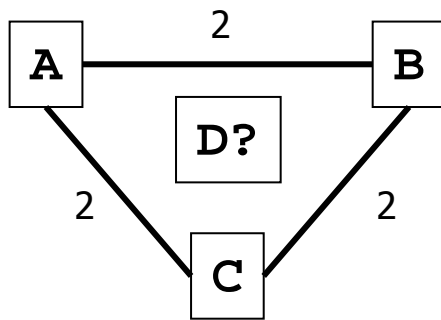
$$d_p(x, y) = \sqrt[p]{\sum |x_i - y_i|^p}$$

- Wenn wir Punkte und eine Metrik haben, können wir immer eine Distanzmatrix ausrechnen
- Aber geht das **auch immer umgekehrt?**
  - Kann man aus einer gegebenen Distanzmatrix und Metrik immer die Punkte zurückrechnen?

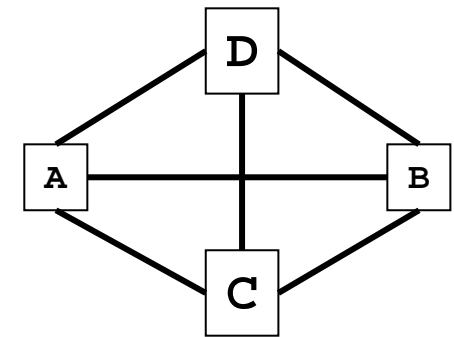
# Beispiel

- Euklidischer Abstand im zweidimensionalen Raum

	A	B	C	D
A		2	2	2
B			2	2
C				2
D				



- Hängt von der Metrik ab
- Hängt von der **Zahl der Dimensionen** ab



# Distanzbasierte Algorithmen

---

- Konstruktion des ultrametrischen Baumes basiert rein auf Distanzmaßen
  - Einen ultrametrischen Baum gibt es nicht für alle Matrizen
  - Es gibt weniger empfindliche Verfahren (gleich)
- Die Geschichte einzelner „Sites“ wird nicht berücksichtigt
- Solche Algorithmen zur Berechnung von Stammbäumen nennt man **distanzbasiert**
- Alternative: **Merkmalsbasierte Verfahren**
  - Beachten jedes einzelne Merkmal (Basen)
  - Vertreter: Perfect Phylogeny; Maximum Parsimony
  - Später

# UPGMA - Hierarchisches Clustering

---

- UPGMA
  - „Unweighted pair group method with arithmetic mean“
  - Anderer Name: [Hierarchisches Clustering](#)
- Sehr einfaches und allgemeines Verfahren, kann bei allen möglichen Problemen angewandt werden
- [Wenn eine Matrix ultrametrisch ist](#), dann findet UPGMA den dazugehörigen ultrametrischen Baum
  - UPGMA nimmt also die Molecular Clock an – alle Pfade von einem Blatt zur Wurzel haben am Ende die selbe Länge
- **Achtung:** UPGMA konstruiert immer einen Baum
  - Auch wenn die Matrix nicht ultrametrisch ist

# UPGMA Verfahren

---

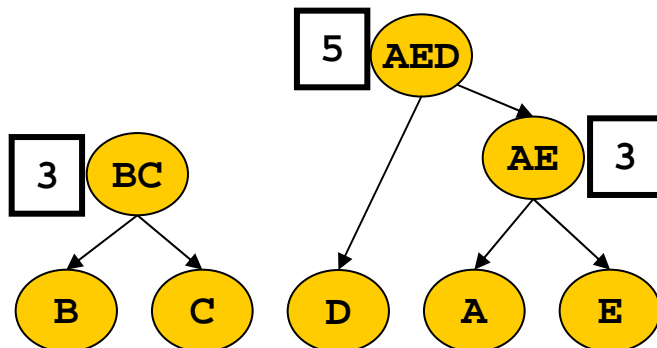
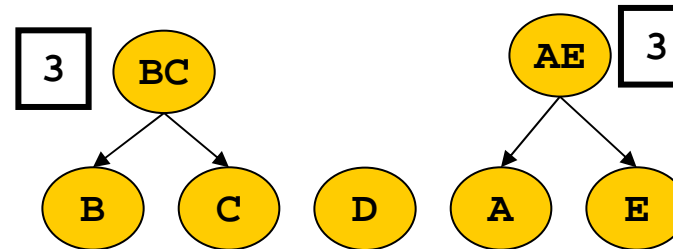
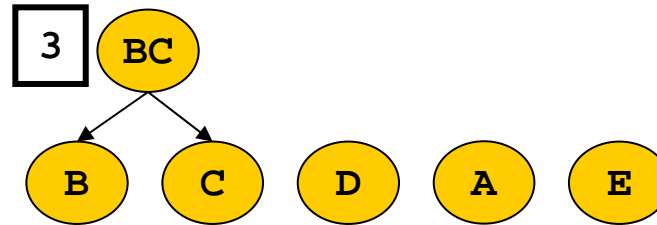
- Gegeben: Distanzmatrix D
- Erzeuge ein „Baumgerüst“ mit n Blättern
- Loop
  - Wähle den kleinsten Wert  $D[i,j]$  Wert der Matrix und verbinde die Knoten i und j durch einen neuen Knoten „ij“ mit Beschriftung  $D[i,j]$  und Kanten zu i und zu j
    - Anfangs sind i und j Blätter, später können es auch innere Knoten sein
  - Lösche Zeilen und Spalten i und j aus D
  - Füge in D eine Zeile und eine Spalte „ij“ hinzu mit  $D[ij,k] = (D[i,k]+D[j,k])/2$
- Bis D nur noch 2 Spalten/Zeilen hat

# Beispiel

	B	C	D	E
A	8	8	5	3
B		3	8	8
C			8	8
D				5

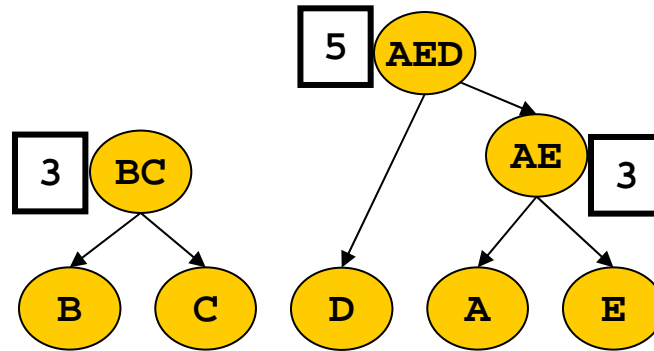
	BC	D	E
A	8	5	3
BC		8	8
D			5

	BC	D
AE	8	5
BC		8

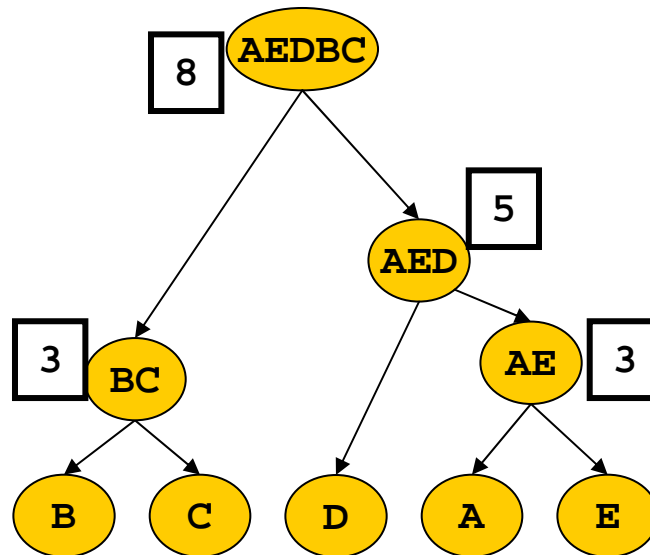


# Beispiel

	AE	BC	D
AE		8	5
BC			8



	BC
AED	8

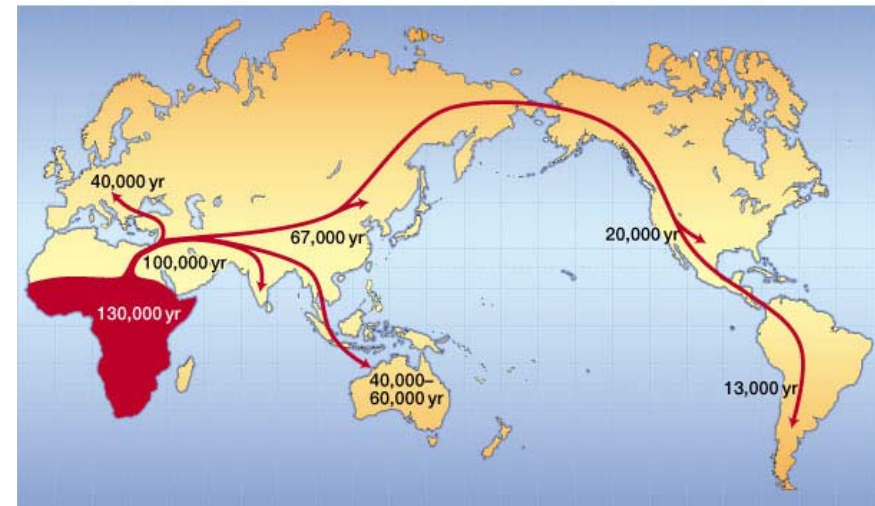
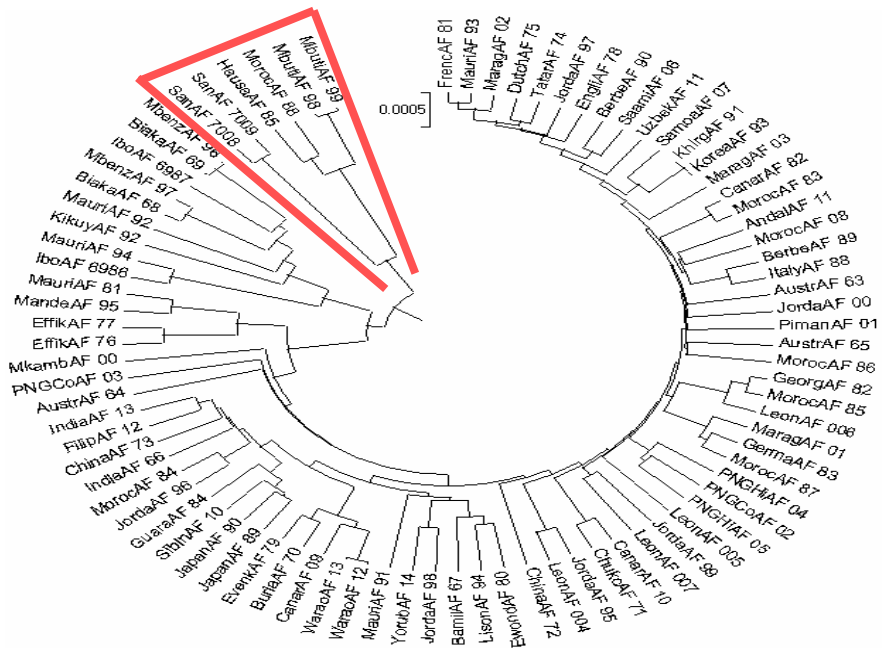


## Kontrolle

	B	C	D	E
A	8	8	5	3
B		3	8	8
C			8	8
D				5

# Anwendungsbeispiel

- Sequenzierung der **mitochondrialer DNA** (16 KB) von 86 geographisch verteilt lebenden Personen
- Ergebnis: Mitochondriale DNA scheint nach einer molekularen Uhr zu mutieren; Divergenz ist ca.  $1,7E-8$  pro Base und Jahr



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

# Komplexität

---

- n Durchläufe
  - In jedem Durchlauf wird die Matrix um eine Zeile/Spalte kleiner
- $n+2$  Berechnungen
  - 2 Kantenlängen
  - $O(n)$  neue Einträge in der Matrix
- Also  $O(n^2)$ ?
- So nicht
  - Es bleibt das Problem, den kleinsten Eintrag in der Matrix zu finden
  - Damit sind wir bei  $O(n \cdot (n \cdot \log(n) + n)) = O(n^2 \log(n))$
- Es gibt aber  $O(n^2)$  Algorithmen zur Rekonstruktion des Baumes aus einer ultrametrischen Matrix
  - Siehe Gusfield, Errata Webseite

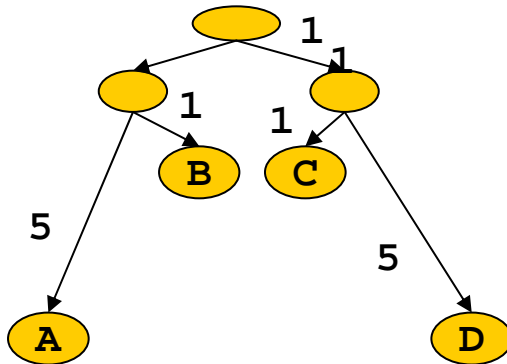
# UPGMA und Ultrametrien

---

- Theorem  
*Wenn eine Matrix ultrametrisch ist, berechnet UPGMA den dazu gehörenden ultrametrischen Baum.*
- Beweis
  - Literatur
- Aber: Wenn eine Matrix nicht ultrametrisch ist, berechnet UPGMA
  - ... auch einen Baum – aber wie gut ist der?
- Die **Molecular Clock Theory ist also Voraussetzung** für die (korrekte) Anwendung von UPGMA

# Ultrametrien und Sequenzabstände

- Realität
  - Matrix mit Editabständen der Sequenzen
- **Reale Daten** sind selten ultrametrisch
  - Spezies unterliegen spezifischem Selektionsdruck
- Die Zahl der Mutationen auf den Kanten zu Geschwistern sind nicht **gleich verteilt**
- Beispiel: Stammbaum, aber keine Ultrametrik



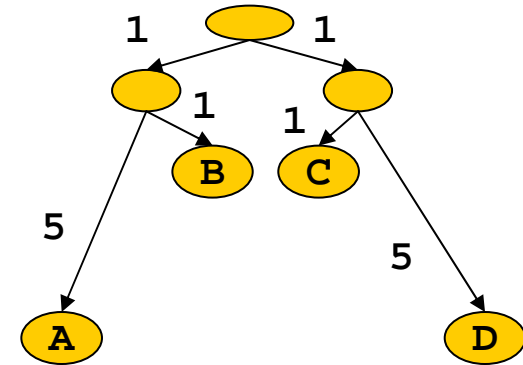
	A	B	C	D
A		6	8	12
B			4	8
C				6
D				

# Wo UPGMA irrt

Der echte Baum

	B	C	D
A	6	8	12
B		4	8
C			6
D			

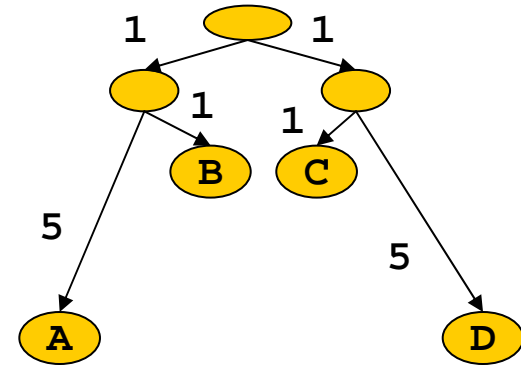
Was erzeugt UPGMA?



# Wo UPGMA irrt

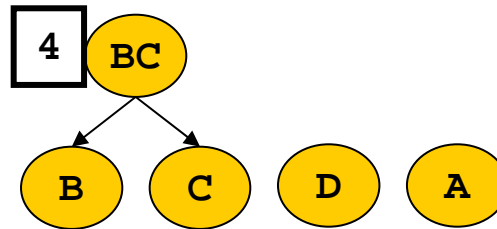
Der echte Baum

	B	C	D
A	6	8	12
B		4	8
C			6
D			



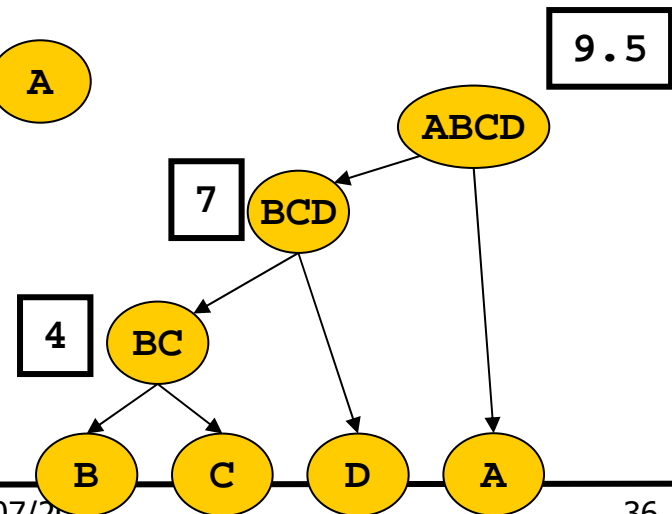
Was erzeugt UPGMA?

	B	C	D
A	6	8	12
B		4	8
C			6



	A	BC	D
A		7	12
BC			7

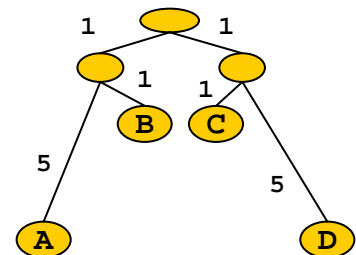
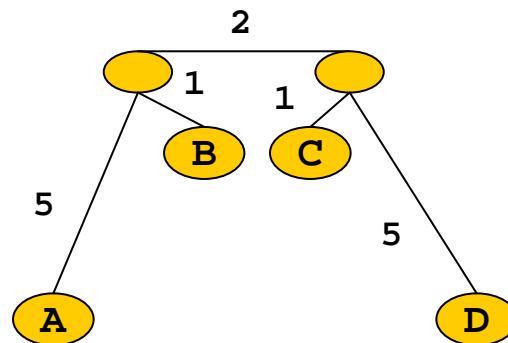
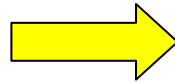
	A
BCD	9.5



# Additive Bäume

- Ultrametrien: Label der inneren Knoten
- Andere Betrachtung: Berechnung der **Kantenlabel**
- Problem  
Gegeben eine Ähnlichkeitsmatrix, finde einen binären Baum so, dass die **Summe der Kantenlabel** auf dem Pfad von jedem Knoten  $i$  zu jedem Knoten  $j$  gleich  $D[i,j]$  ist
- Bemerkung
  - Da es nur Abstände gibt, hat der Baum zunächst **keine Wurzel**

	B	C	D
A	6	8	12
B		4	8
C			6
D			



# Additive Bäume

---

- Definition

*Sei  $D$  eine positive symmetrische Matrix mit  $n$  Spalten und Zeilen und  $\forall i: D[i,i]=0$ . Ein **Baum  $T$**  heißt **additiver Baum** für  $D$  gdw*

- *$T$  hat  $n$  Blätter, beschriftet mit den Zeilen von  $D$*
- *Innere Knoten in  $T$  sind nicht beschriftet, Kanten sind beschriftet*
- *Für jedes Paar  $i,j$  ist  $D[i,j]$  gleich der **Summe der Kantenlabel** auf dem (eindeutigen) Pfad von  $i$  nach  $j$*

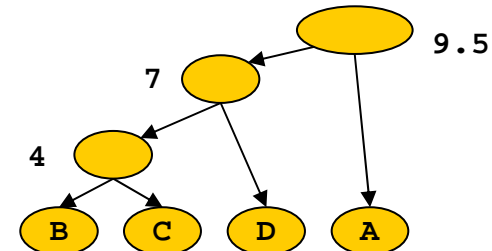
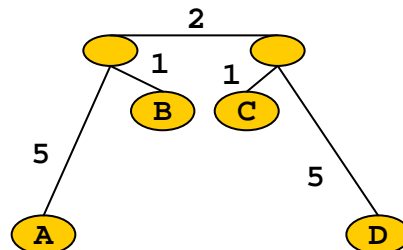
- Bemerkung

- Jede ultrametrische Matrix induziert einen additiven Baum

# Additive Bäume und Ultrametrien

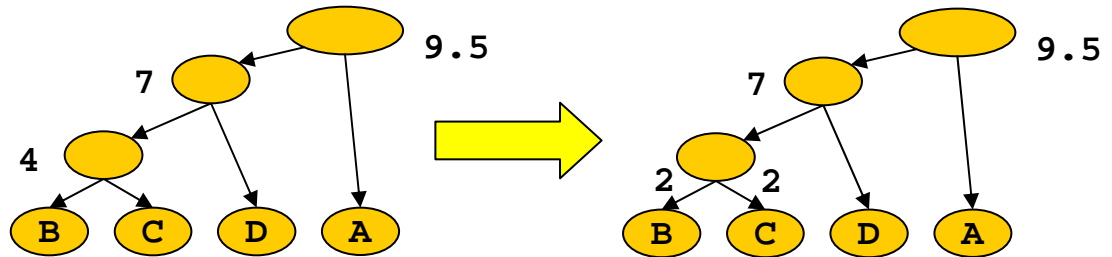
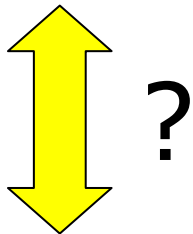
- Ultrametrien: Label auf den **inneren Knoten**
  - Die entsprechen den Speziationszeitpunkten
  - Kanten von inneren Knoten zu Kindern sind implizit alle gleich lang
  - Label sind unmittelbar in der Matrix enthalten
- Additive Bäume: Label auf den **Kanten**
  - Es werden keine Speziationszeitpunkte berechnet
  - Kanten von inneren Knoten zu Kindern können unterschiedlich lang sein
  - Die Werte in der Matrix sind keine Label im Baum
- Jeder ultrametrische Baum induziert einen additiven Baum, aber nicht umgekehrt
- Jede ultram. Matrix induziert einen add. Baum, aber nicht umgekehrt
- **Additive Bäume sind echt „mächtiger“ als Ultrametrien**

	B	C	D
A	6	8	12
B		4	8
C			6
D			

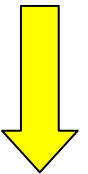
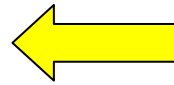
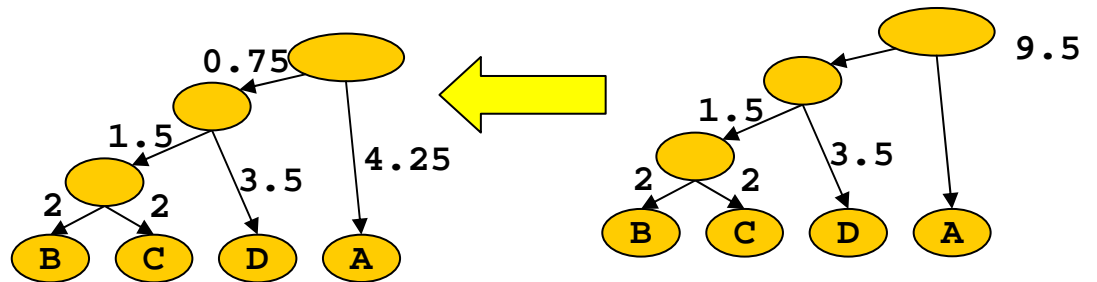
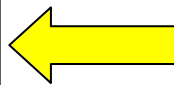


# Von ultrametrischen zu additiven Bäumen

	B	C	D
A	6	8	12
B		4	8
C			6
D			



	B	C	D
A	9.5	9.5	9.5
B		4	7
C			7
D			



# Matrizen und additive Bäume

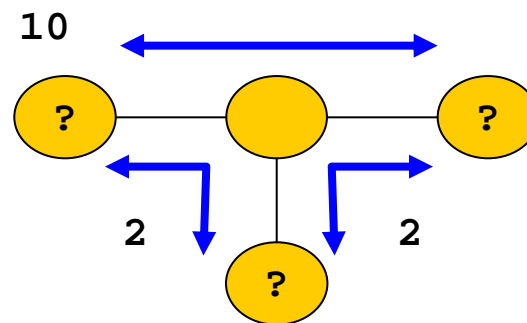
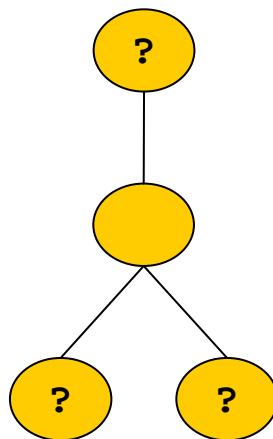
---

- Fragen
  - Existiert zu jeder Matrix ein additiver Baum?
  - Wie findet man einen additiven Baum zu einer gegebenen Matrix?

# Matrizen und additive Bäume

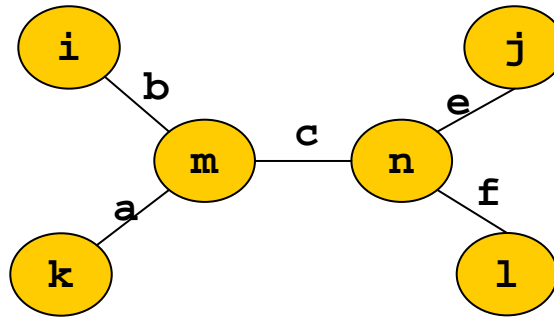
- Fragen
  - Existiert zu jeder Matrix ein additiver Baum?
  - Wie findet man einen additiven Baum zu einer gegebenen Matrix?
- Gegenbeispiel
  - Es gibt nur einen additiven Baum für drei Spezies

	A	B	C
A		10	2
B			2



# Beobachtung

- Betrachten wir beliebige vier Blätter  $i, j, k, l$ , eines Baumes mit Kantenlabeln



- In welchem Verhältnis stehen deren Abstände?
  - Die 6 Abstände setzen sich aus Summen von 5 Kantenlabeln zusammen
- Es gilt (zum Beispiel)
  - $D(i,k)+D(j,l) \leq D(i,j)+D(k,l) = D(i,l)+D(k,j)$
  - Denn:  $(a+b) + (e+f) \leq (b+c+e) + (a+c+f)$
- Aber die Knoten können auch anders angeordnet sein

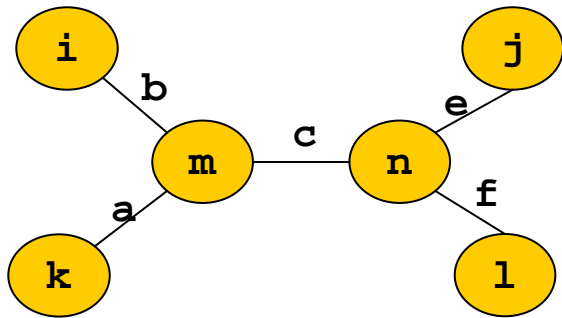
# 4-Punkt Bedingung

- Theorem

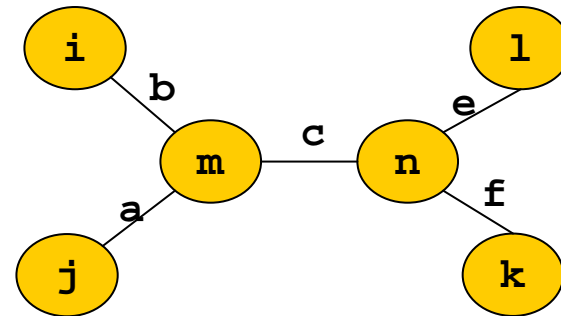
*Eine Matrix  $D$  hat einen additiven Baum gdw. für alle Quadrupel  $i, j, k, l$  die **4-Punkt Bedingung** gilt:*

$$D(i,k)+D(j,l) \leq \max( D(i,j)+D(k,l) , D(i,l)+D(k,j) )$$

- Beweis: Literatur



$$(a+b) + (e+f) \leq \max( (b+c+e) + (a+c+f), (b+c+f) + (a+c+e) )$$



$$(b+c+f) + (a+c+e) \leq \max( (a+b) + (e+f), (b+c+e) + a+c+f )$$

# Neighbor-Joining

---

- Matrizen und Algorithmen
  - Ultrametrische Matrizen – UPGMA
  - Additive Matrizen – Neighbor Joining
- Hierarchisches Clusterverfahren (wie UPGMA)
  - Erzeugt einen binären Baum ohne Wurzel
  - Grundaufbau wie UPGMA
    - Beginne mit so vielen Clustern wie Blättern
    - Wähle nach **bestimmtem Kriterium zwei Cluster**
    - Verschmelze die zwei Cluster und verbinde Knoten im Baum
    - Iteriere, bis nur noch ein Cluster vorhanden ist
- Unterschiede
  - UPGMA wählt Cluster zur Verschmelzung nur nach Nähe zueinander
  - Neighbor Joining wählt Cluster **nach der Nähe zueinander und dem Abstand zu anderen Clustern**

# Verfahren

---

- Bilde aus jeder Zeile einen Cluster
- Berechne für jeden Cluster  $i$  den Abstand  $u_i$  zu allen anderen Clustern

$$u_i = \sum_{k \neq i} \frac{D[i, k]}{n-2}$$

Nächster Nachbar

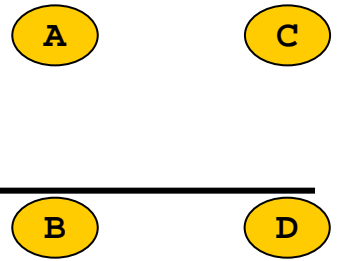
- Suche das Clusterpaar  $(i, j)$ , für das gilt  $D[i, j] - u_i - u_j = \min$
- Erzeuge Cluster  $ij$  mit Kanten zu  $i$  und  $j$  mit Kantenlängen

$$d(i, ij) = \frac{D[i, j] + u_i - u_j}{2} \quad d(j, ij) = \frac{D[i, j] + u_j - u_i}{2}$$

- Erzeuge neuen Clusterknoten mit Abständen zu anderen Clustern

$$D[k, ij] = \frac{D[i, k] + D[k, j] - D[i, j]}{2}$$

- Lösche Cluster  $i$  und  $j$
- Iteriere, solange mehr als ein Cluster existiert

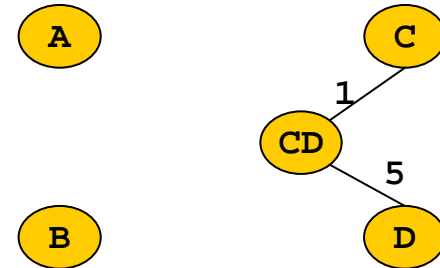


# Beispiel (hier scheiterte UPGMA)

NJ-Abstände der Clusterpaare

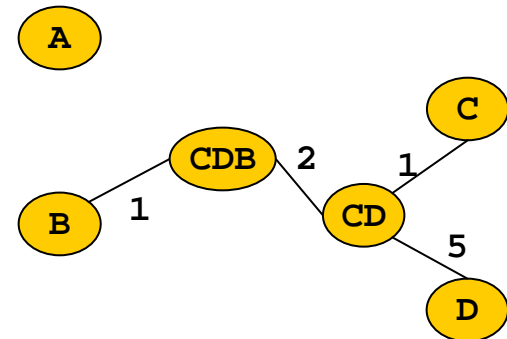
	A	B	C	D
A		6	8	12
B			4	8
C				6
$u_i$	13	9	9	13

	B	C	D
A	-16	-14	-14
B		-14	-14
C			-16

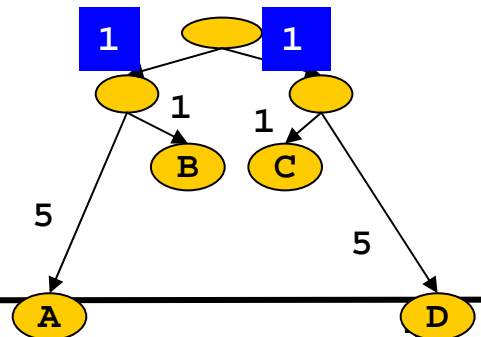
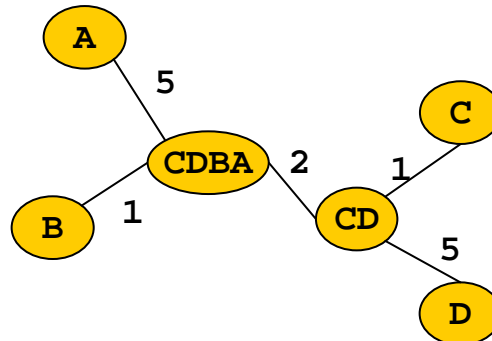


	A	B	CD
A		6	7
B			3
$u_i$	13	9	10

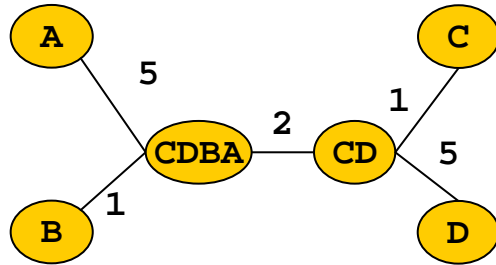
	B	CD
A	-16	-16
B		-16



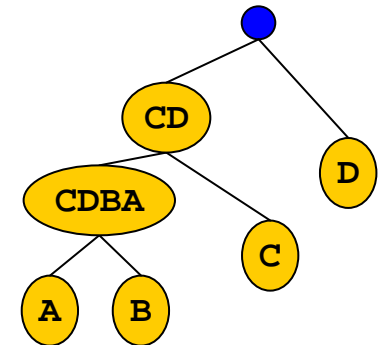
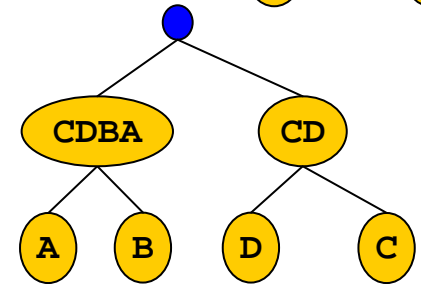
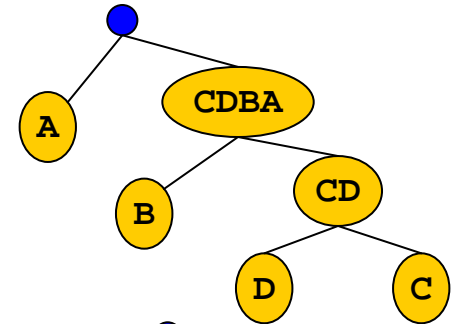
	A	BCD
A		5



# Rooting eines Baumes

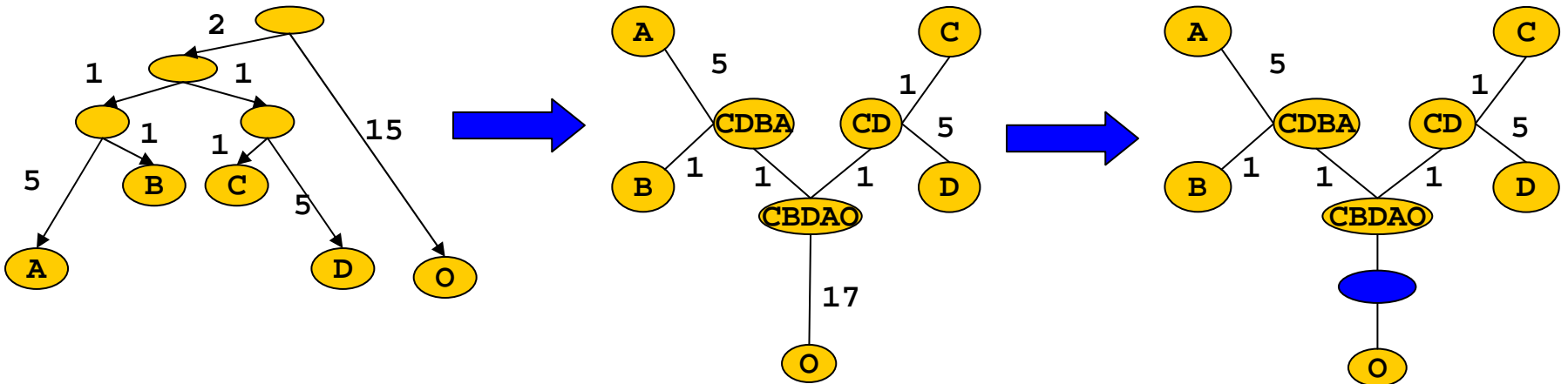


- NJ berechnet nur Abstände zueinander
- Eine **zeitliche Reihenfolge** der Knoten kommt nicht heraus
- Ein Wurzelknoten kann prinzipiell auf jeder Kante platziert werden
- Wie findet man die echte Wurzel?
  - Externe Datierung einzelner Knoten, z.B. durch Fossilienfunde und C14-Methode
  - Benutzung einer **Outgroup**



# Outgroups

- Eine **Outgroup** ist ein Taxa, das weiter von allen anderen Taxa entfernt ist als diese untereinander
  - Beispiel: Menschen, Mäuse, Ratten, Schweine – Storch
- Was passiert mit der Outgroup?
  - NJ ordnet sie im Baum ein
  - Offensichtlich muss die Kante, die zu der Outgroup führt, den **Wurzelknoten** enthalten
  - Damit wird der ganze Baum zeitlich angeordnet



# Wenn die Daten nun ...

---

- Weder ultrametrisch noch additiv sind?
- Approximieren; z.B. durch
  - Gegeben Abstandsmatrix  $D$ . Wähle eine Baumtopologie  $T$  und berechne die tatsächlichen Abstände im Baum  $d(i,j)$  so, dass der Fehler minimiert wird

$$error(t) = \min_d \left( \sum_{i=1}^n \sum_{j \neq i} (D[i, j] - d(i, j))^2 \right)$$

- Für gegebene Topologie  $T$  ist das effizient lösbar, aber ...
  - Man muss alle Topologien ausprobieren
  - Damit ist das Problem NP-vollständig
- Oder man versucht **Heuristiken**: UPGMA, Neighbor Joining
- Oder ganz anders aufgebaute Methoden: Maximum Parsimony, Maximum Likelihood, ...

# Literatur

---

- Eine gute Einleitung zu den Problemen (sehr praktisch orientiert)
  - Baldauf, S. L. (2003). "Phylogeny for the faint of heart: a tutorial." *Trends Genet* **19**(6): 345-51.
- Schöne, ausführliche Übersicht (weniger über die Algorithmen)
  - Morrison „Phylogenetic Tree Building“, Int J of Parasitology, 1996
- Ultrametrien und Additivität
  - Gusfield (Kapitel 17 )
- UPGMA und Neighbor Joining
  - Oft beschrieben, z.B. im Script von Ron Shamir, University Tel Aviv
  - <http://www.math.tau.ac.il/~rshamir/algmb/98/scribe/ps/lec09.ps>