

# Algorithmische Bioinformatik

## Distanzbasierte phylogenetische Algorithmen

Ulf Leser

Wissensmanagement in der  
Bioinformatik



# Inhalt dieser Vorlesung

---

- Ultrametrien
- Hierarchisches Clustering: UPGMA
- Additive Bäume und Neighbor Joining

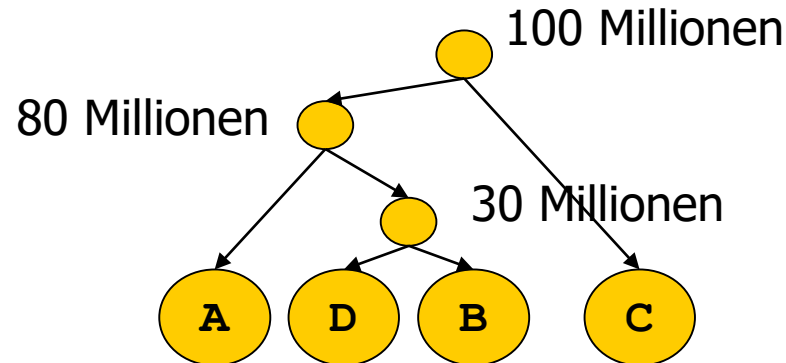
# Molecular Clock Assumption

---

- Häufige Annahme: **Molecular Clock**
  - Mutationen setzen sich bzgl. der Zeit immer mit gleicher Wsk durch
  - Unabhängig von Teilbaum, Zeitpunkt, Ort und Art der Mutation
- Die ist hilfreich, aber falsch
  - Zeiten erhöhter **Mutationshäufigkeit**: Sonneneruptionen, ...
  - Zeiten erhöhten **Selektionsdrucks**: Klimaverschiebungen, ...
  - Teilbäume, die schneller mutieren: Abhängig vom Anpassungsdruck
  - **Sequenzabschnitte**, die unterschiedlich schnell mutieren: Coding versus non-coding Regions, House-Keeping genes,
  - ...

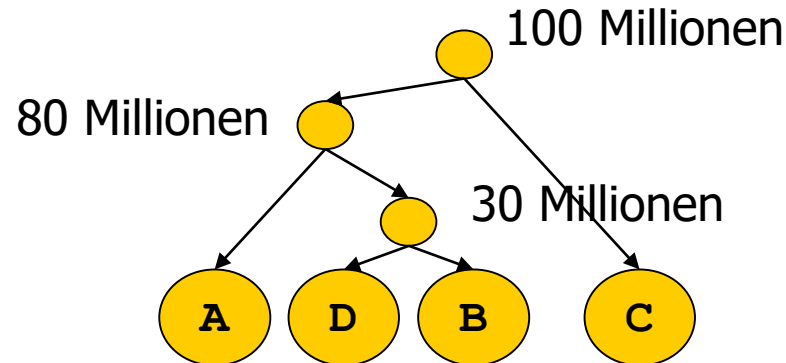
# Folgerungen

---



- Wenn die Molecular Clock Assumption gilt
  - Ist die Menge an Veränderungen auf einer Kante **proportional zu der verstrichenen Zeit**
  - Geschwister haben immer den gleichen Abstand zu allen Vorfahren
  - Der Editabstand zweier Knoten ist proportional zur **Summe der Editabstände beider Knoten** zum kleinsten gemeinsamen Vorfahr
    - Keine lineare Proportionalität – siehe PAM
- Damit kann man innere Knoten zeitlich verorten
  - Wenn man die Geschwindigkeit der Uhr kennt

# Ultrametrien



- Wenn man den Baum und die Zeitpunkte weiß, dann gilt
  - Zahlen auf Pfad von Wurzel zu Blatt nehmen strikt ab
  - Der **Zeitpunkt der Aufspaltung ist ein Abstandsmaß** für zwei Blätter
    - Für  $X, Y$  sei  $d(X,Y)$  das Label des letzten 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**

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

# Ultrametrische Bäume

---

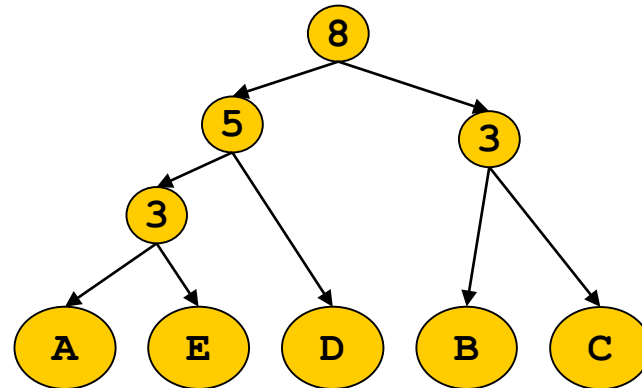
- Definition

*Sei  $T$  ein binärer, gewurzelter, ungeordneter Baum und  $D$  eine positive symmetrische Matrix mit  $n$  Zeilen und  $n$  Spalten und  $\forall i: D[i,i]=0$ .  $T$  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 **letzte gemeinsame Vorfahr** von  $i$  und  $j$  ist mit  $D[i,j]$  beschriftet*

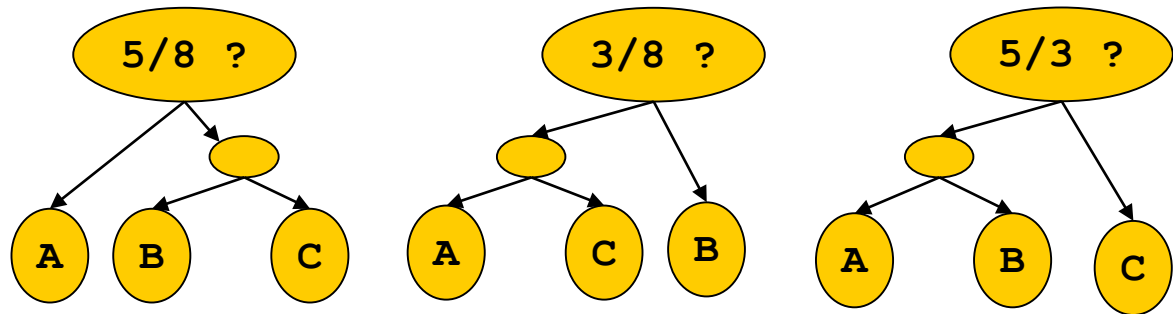
# Beispiel

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



Geht das immer?

	A	B	C
A		8	5
B			3
C			





# Überlegung

---

- 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

# Test auf Ultrametrien

---

- Definition

*Eine positive 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  $D[j,k]>D[i,j]$

- Eine Abstandsmatrix von Objekten ist also ultrametrisch, wenn das **Abstandsmaß eine Ultrametrik** ist

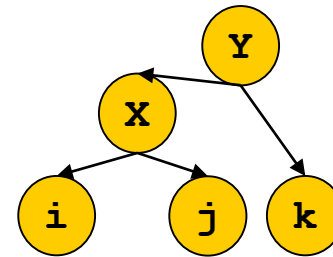
# Von der Matrix zum Baum und zurück

- Theorem

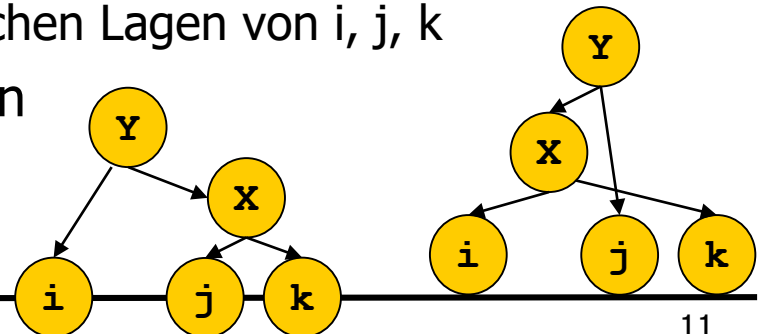
*Eine symmetrische Matrix  $D$  hat einen ultrametrischen Baum  $T$  gdw.  $D$  ultrametrisch ist.*

- Beweis

- (1) Nehmen wir erst an, dass zu  $D$  ein ultrametrischer Baum  $T$  existiert. Nehmen wir an, dass  $i, j, k$  in  $T$  so liegen:

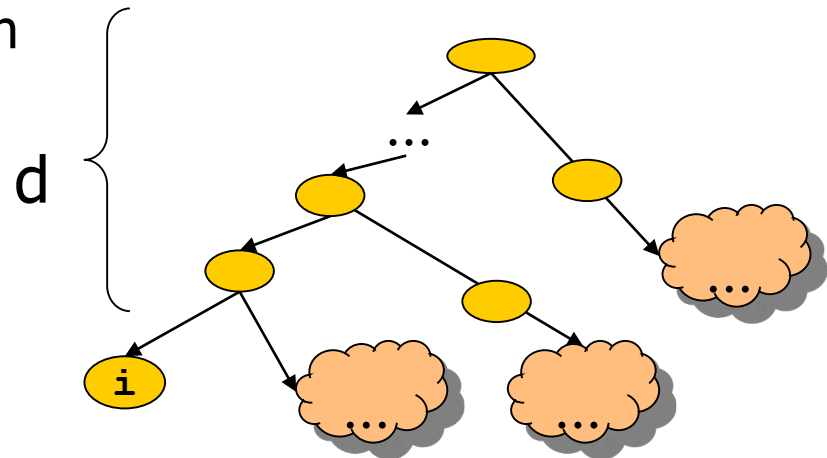


- Dann gilt offensichtlich  $D[i,k]=D[j,k]=Y$  und  $D[i,k]>D[i,j]=X$ 
  - Dito für die zwei anderen möglichen Lagen von  $i, j, k$
- Das gilt für alle Tripel von Knoten
- Also ist  $D$  ultrametrisch



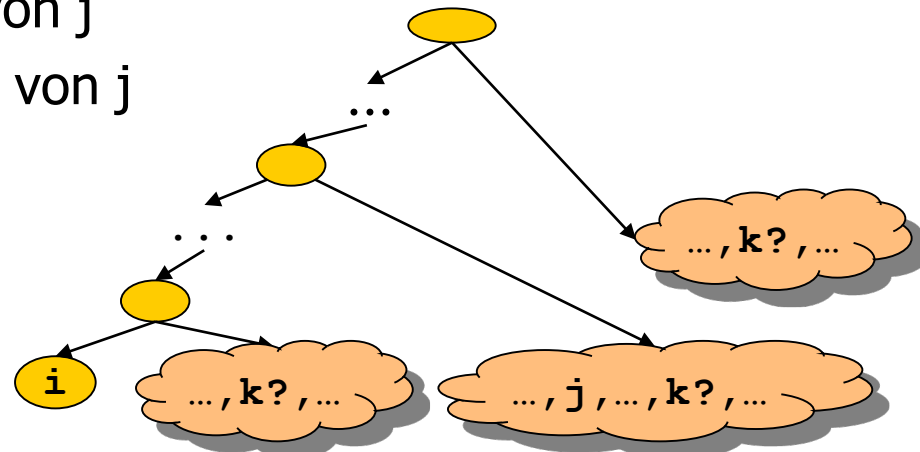
# Beweis Gegenrichtung

- (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 **letzte gemeinsame Vorfahren** mit allen  $n-1$  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 letzten 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 Blätter einer Klasse befinden sich in einem Unterbaum bzgl. genau eines Knoten auf dem Pfad von  $i$  zur Wurzel
  - Alle Blätter 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  $j$

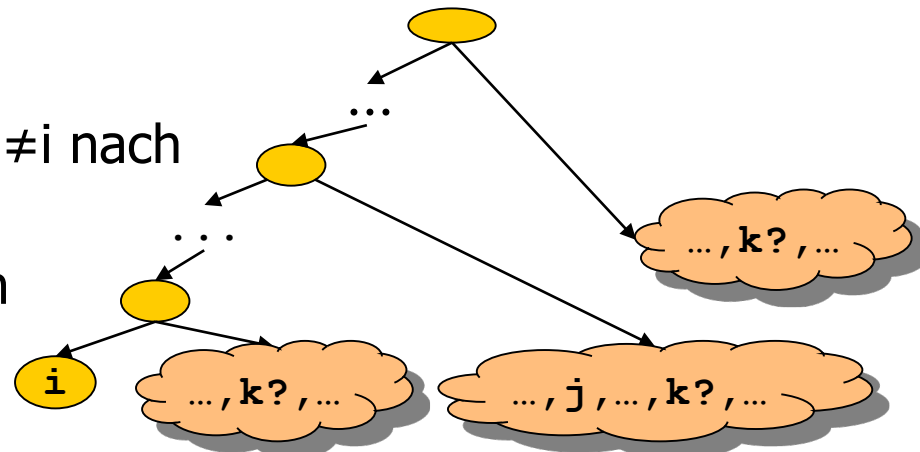


# Beweis Gegenrichtung -3-

- 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  $D[i,j]<D[j,k]$
- qed.

- Konstruktion

- $i$  beliebig wählen, alle Knoten  $\neq i$  nach Abstand zu  $i$  klassifizieren
- Strang von  $i$  zur Wurzel bauen
- Jede Klasse rekursiv lösen



# Folgerung

---

- Der Beweis ist konstruktiv; man kann auf diese Weise einen ultrametrischen Baum bauen
- Das wird immer derselbe sein, egal in welcher Reihenfolge man die Blätter wählt
- Theorem  
*Sei  $D$  eine ultrametrische Matrix. Dann gibt es **genau einen ultrametrischen Baum**  $T$  für  $D$ .*
- Beweis
  - Durch Konstruktion; Literatur

# Distanzbasierte Algorithmen

---

- Algorithmen zur Berechnung von Stammbäumen, die nur die Distanzmatrix benutzen, nennt man **distanzbasiert**
  - Die Geschichte einzelner „Sites“ (Basen, Sequenzabschnitte etc.) wird nicht berücksichtigt
  - Sequenzen an inneren Knoten können nicht rekonstruiert werden
- Alternative: **Merkmalsbasierte Verfahren**
  - Beachten jede einzelne Site (Basen)
  - Vertreter: Perfect Phylogeny, Maximum Parsimony, ...



# Inhalt dieser Vorlesung

---

- Ultrametrien
- Hierarchisches Clustering: UPGMA
- Additive Bäume und Neighbor Joining

# UPGMA - Hierarchisches Clustering

---

- UPGMA
  - „Unweighted pair group method with arithmetic mean“
  - Anderer Name: [Hierarchisches Clustering](#)
- Theorem

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

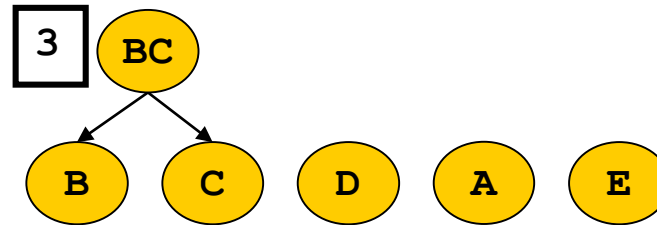
# UPGMA Algorithmus

---

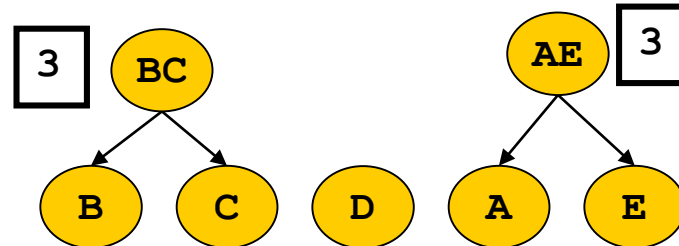
- Gegeben: Distanzmatrix D
- Erzeuge ein „Baumgerüst“ mit n Blättern
- Loop
  - Wähle den **kleinsten Wert**  $D[i,j]$  Wert der Matrix und verbinde die Knoten i und j durch einen neuen Knoten „ij“ mit Beschriftung  $D[i,j]$  und Kanten zu i und zu j
    - Anfangs sind i und j Blätter, später können 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

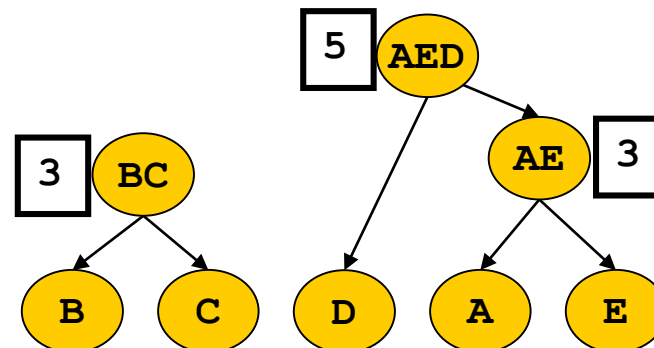
	B	C	D	E
A	8	8	5	3
B		3	8	8
C			8	8
D				5



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

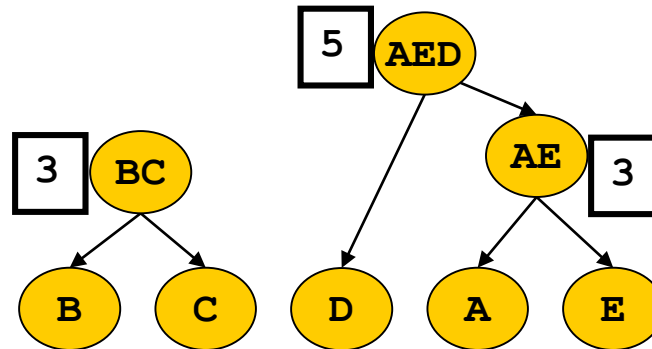


	BC	D
AE	8	5
BC		8

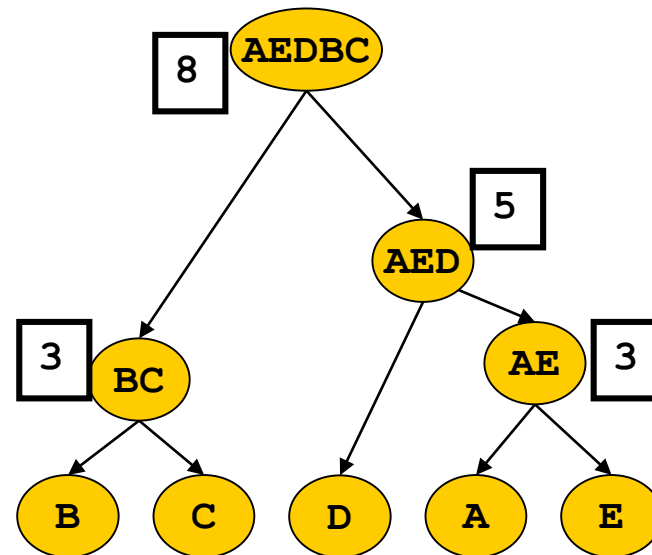


# Beispiel

	AE	BC	D
AE		8	5
BC			8



	BC
AED	8

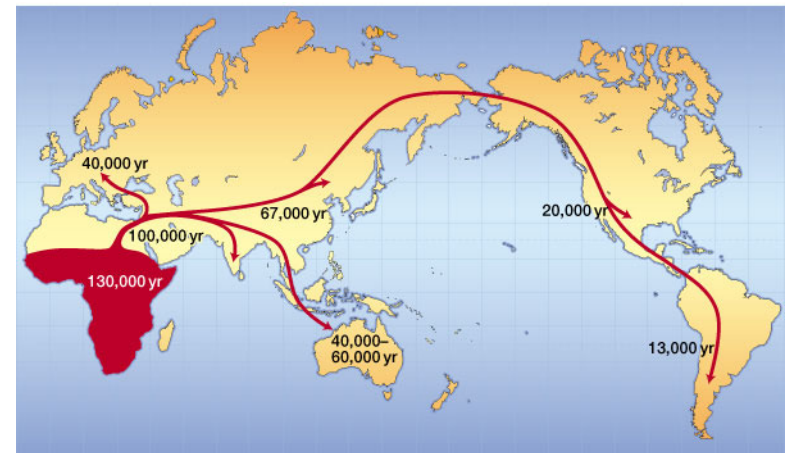
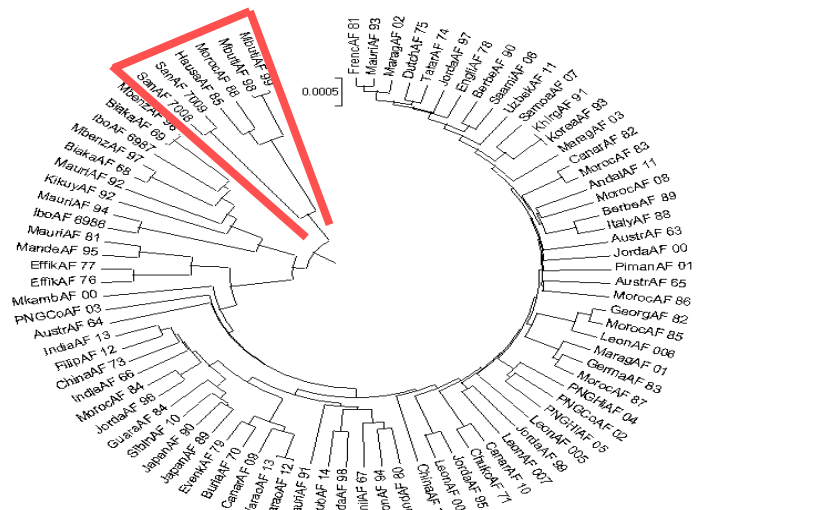


## Kontrolle

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

# Anwendungsbeispiel

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

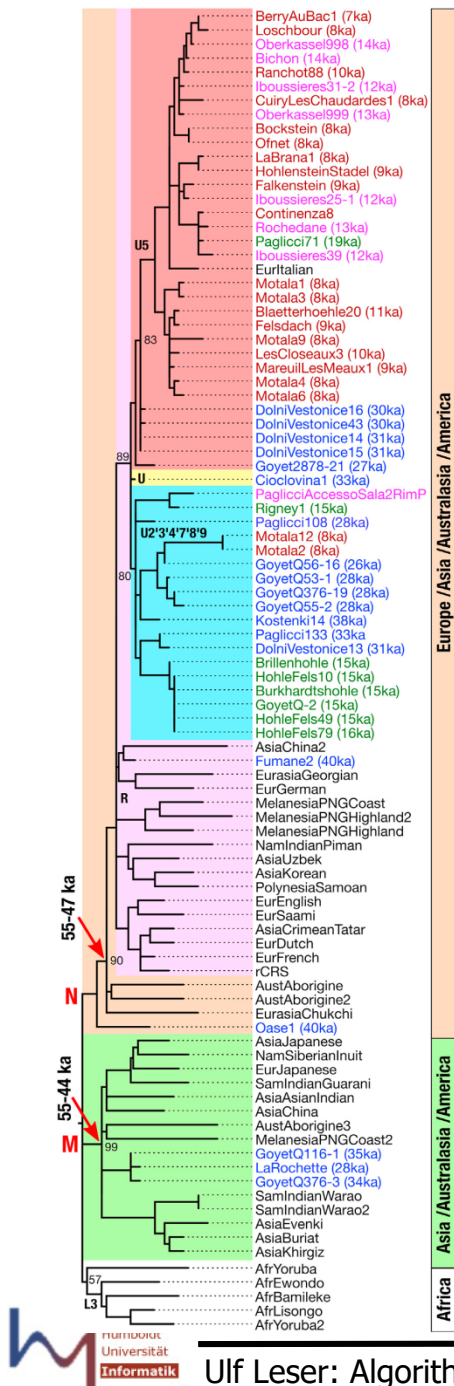


Quelle:  
<http://www.genpat.uu.se/mtDB/sequences.html>  
Methode: UPGMA

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

# Aktuelle Variante

- 55 Proben aus Europa, bis zu 35.000 Jahre alt
- Vergleich mit 311 aktuellen und 66 prähistorischen Proben Weltweit
- Bayesian phylogenetic inference
- Resultat: Es gab nur eine Einwanderungswelle nach Europa
- Posth et al., Pleistocene Mitochondrial Genomes Suggest a **Single Major Dispersal** of Non-Africans and a Late Glacial Population Turnover in Europe, Current Biology (2016)



# Komplexität

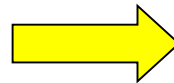
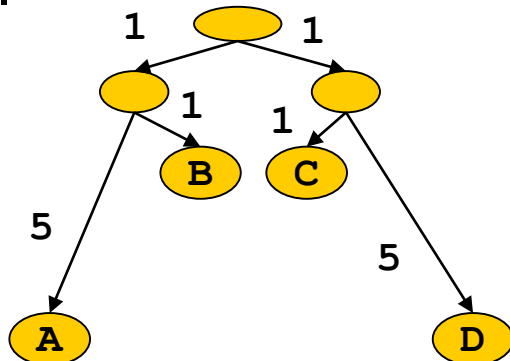
---

- n Durchläufe
  - In jedem Durchlauf wird die Matrix um eine Zeile/Spalte kleiner
- Pro Durchlauf:  $O(n)$  Berechnungen
  - 2 Spalten / Zeilen löschen (nicht durch Matrix kopieren!)
  - $O(n)$  neue Einträge in der Matrix
- Also  $O(n^2)$ ?
- Wir müssen noch kleinsten Eintrag in der Matrix zu finden
- Mit Priority Queue:  $O(n \cdot (n \cdot \log(n) + n)) = O(n^2 \log(n))$
- Es gibt  $O(n^2)$  Algorithmen zur Rekonstruktion des Baumes aus einer ultrametrischen Matrix
  - Siehe Gusfield, Errata Webseite



# Ultrametrien und Sequenzabstände

- **Reale Abstandsmatrizen** sind selten ultrametrisch
  - Molecular Clock Assumption gilt idR nicht
  - Editabständen der Sequenzen sind nur Annäherungen an wahren evolutionären Abstand (siehe auch Jukes-Cantor Modell)
  - Sequenzen haben Fehler
- In der Realität sind Mutationen auf den Kanten zu Geschwistern nicht **gleich verteilt**
  - Wir heben die MCA zum Ziel auf
- Beispiel



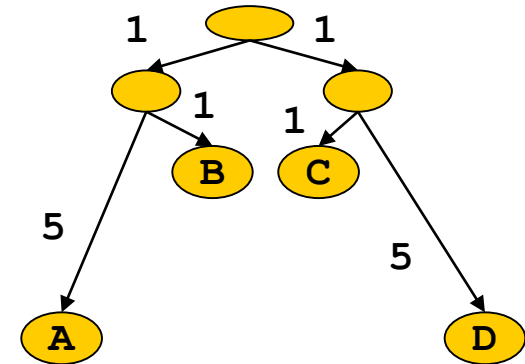
	A	B	C	D
A		6	8	12
B			4	8
C				6
D				

# Wo UPGMA irrt

Der echte Baum

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

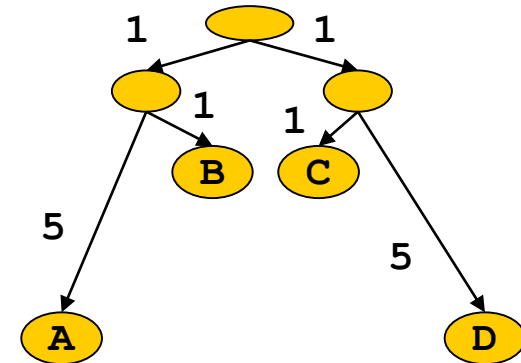
Was erzeugt UPGMA?



# Wo UPGMA irrt

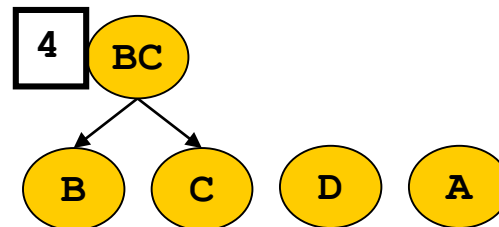
Der echte Baum

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



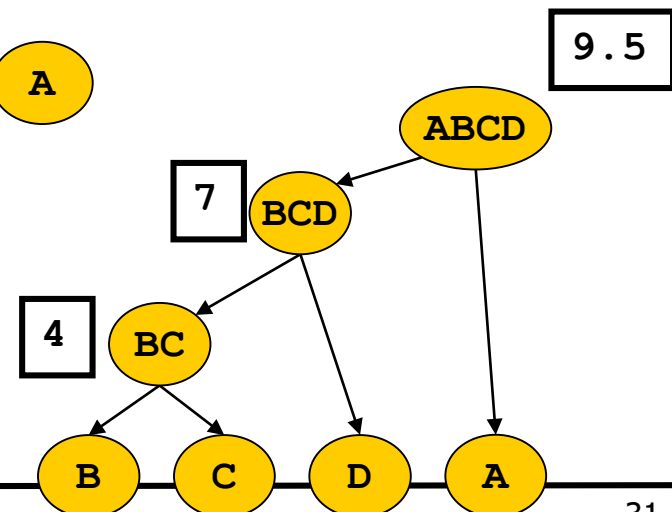
Was erzeugt UPGMA?

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



	A	BC	D
A		7	12
BC			7

	A
BCD	9.5



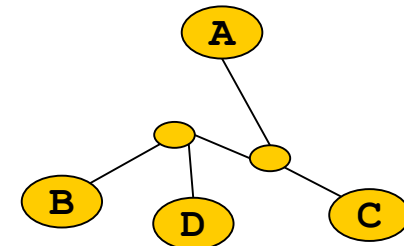
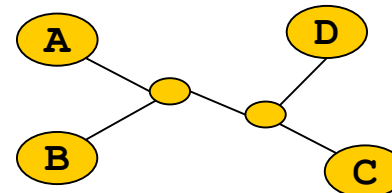
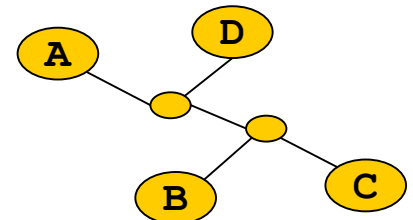
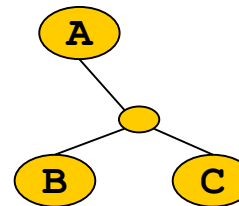
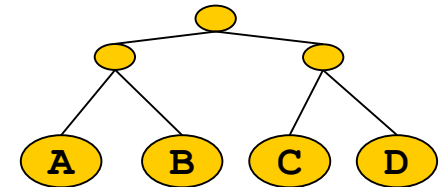
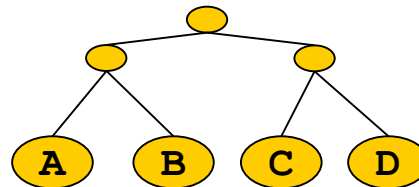
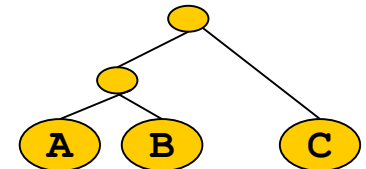
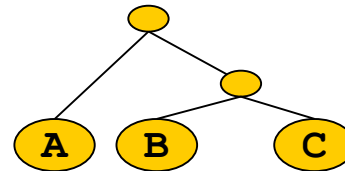
# Inhalt dieser Vorlesung

---

- Ultrametrien
- Hierarchisches Clustering: UPGMA
- Additive Bäume und Neighbor Joining

# Einschub: Ungewurzelte Bäume

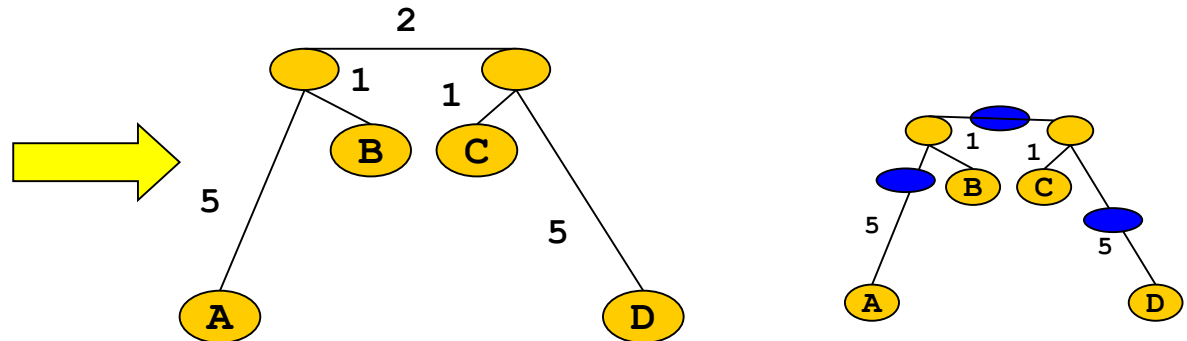
- Ein binärer, ungerichteter, **gewurzelter Baum** ist ein zyklensfreier Graph, in dem **ein Knoten Grad zwei** hat (Wurzel) und alle anderen Grad drei (innere Knoten) oder Grad eins (Blätter)
- Ein binärer, ungerichteter, **ungewurzelter Baum** ist ein zyklensfreier Graph, in dem alle Knoten Grad drei oder Grad eins haben
  - Einfach Wurzel rausnehmen



# Additive Bäume

- Problem: Gegeben  $D$ , 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
- Da es nur Abstände zwischen Blättern gibt, kann die Wurzel nicht identifiziert werden – Berechnung eines **ungewurzelten Baums**
  - Beide Kanten zur Wurzel sind in allen Blatttdistanzen entweder **zusammen oder gar nicht**
  - Können nicht durch Differenzbildung identifiziert werden

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



# Formal

---

- Definition

*Sei  $D$  eine positive symmetrische Matrix mit  $n$  Spalten und Zeilen und  $\forall i: D[i,i]=0$ . Ein binärer, ungeordneter und ungewurzelter **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

- Wenn eine Matrix einen additiven Baum besitzt, so nennen wir die **Matrix additiv**

# Additive Bäume versus Ultrametrien

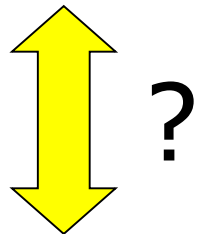
---

- Ultrametrien: Label auf den inneren Knoten
  - Abstände korrelieren mit Speziationszeitpunkten (per Annahme)
  - Kanten zu Kindern sind immer gleich lang
  - Zeitpunkt des Vorfahren (Speziation) ist der Abstand
  - Label sind als Werte in der Matrix enthalten
- Additive Bäume: Label auf den Kanten
  - Es werden keine Speziationszeitpunkte berechnet
  - Kanten zu Kindern können unterschiedlich lang sein
  - Die Werte in der Matrix sind i.A. keine Label im Baum
- Jede ultrametrische Matrix hat einen additiven Baum, aber nicht umgekehrt

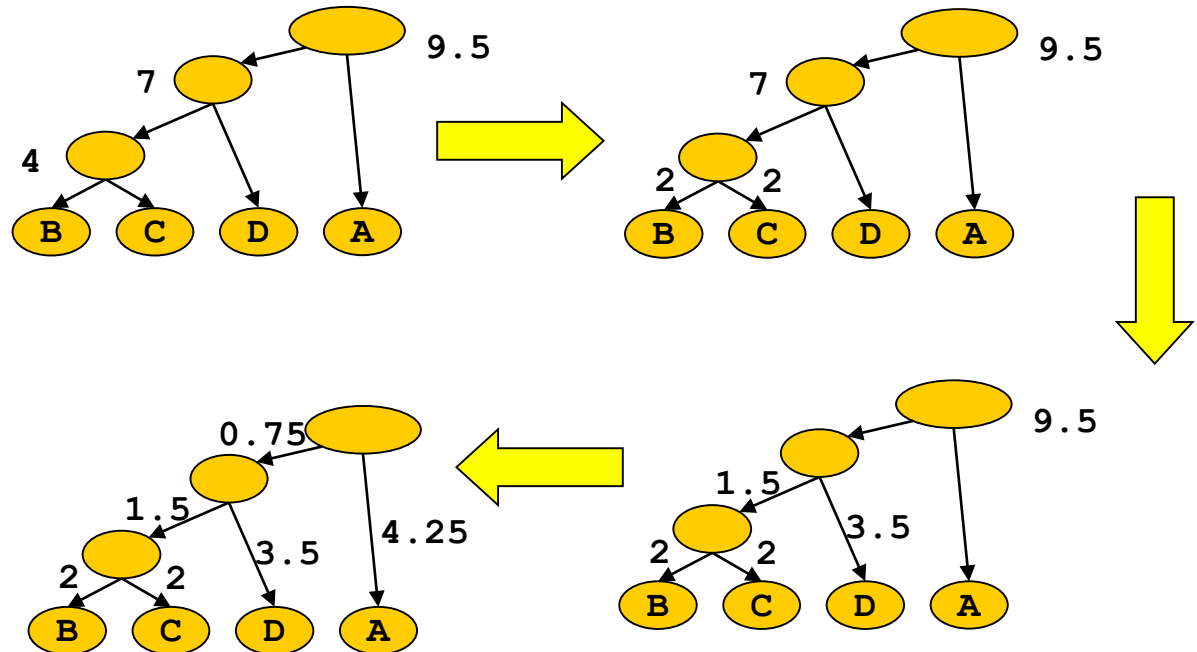


# Von ultrametrischen zu additiven Bäumen

	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

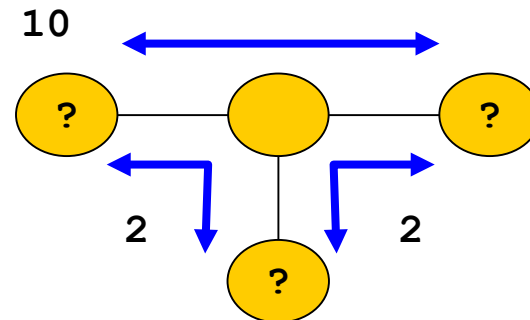
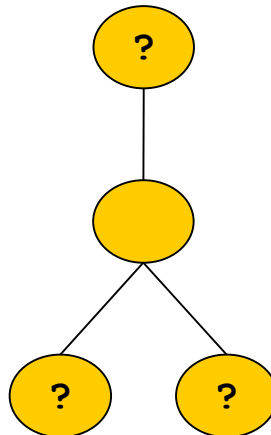
---

- Fragen
  - Existiert zu jeder Matrix ein additiver Baum?
  - Wann ist eine Matrix additiv?
  - Wie findet man einen additiven Baum zu einer additiven Matrix?

# Matrizen und additive Bäume

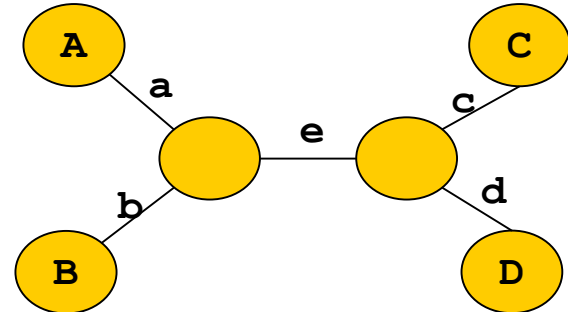
- Fragen
  - Existiert zu jeder Matrix ein additiver Baum?
- Gegenbeispiel
  - Es gibt nur einen ungewurzelten Baum für drei Spezies

	A	B	C
A		10	2
B			2



# Beobachtung

- Fragen
  - Wann ist eine Matrix additiv?
- Betrachten wir beliebige vier Blätter A, B, C, D
- In welchem Verhältnis stehen deren Abstände?
  - Die 6 Abstände setzen sich aus 5 Kantenlabeln zusammen
- Beobachtung
  - $D(A,B) + D(C,D) < D(A,D) + D(B,C) = D(A,C) + D(B,D)$
  - Denn:  $(a+b) + (c+d) \leq (a+e+d) + (b+e+c) = (a+e+c) + (b+e+d)$
- Aber: Die Knoten können auch anders angeordnet sein



# 4-Punkt Bedingung

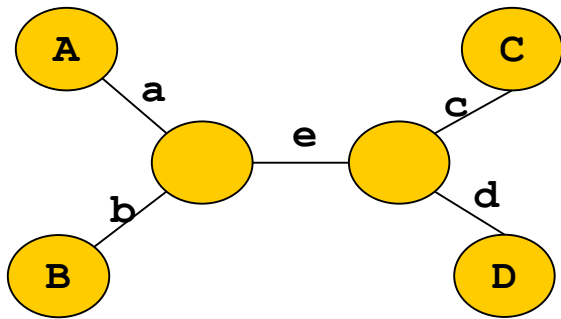
- Theorem

*Eine Matrix  $D$  ist additiv gdw. für alle Quadrupel  $A, B, C, D$  die **4-Punkt Bedingung** gilt:*

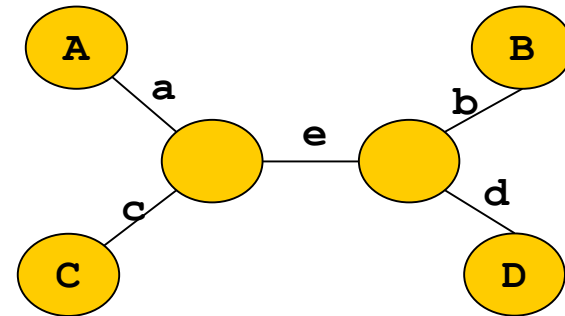
$$D(A, B) + D(C, D) \leq \max( D(A, D) + D(B, C), D(A, C) + D(B, D) )$$

- Beweis

- $\Rightarrow$  Es gibt nur eine Topologie – alle Lablings ausprobieren
- $\Leftarrow$  Literatur



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



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

# Rekonstruktion

---

- Fragen
  - Wie findet man einen additiven Baum zu einer additiven Matrix?
- Verfahren von Warnow
  - T. Warnow. Tree compatibility and inferring evolutionary history. Journal of Algorithmics, 16:388–407, 1994.
  - Läuft in  $O(n^2)$
  - Startet mit beliebigem Paar und **fügt iterativ Blätter** zu einem wachsenden Baum
  - Über die 4-Punkt Bedingung kann die Position des neuen Blatts immer genau bestimmt werden
- Aber: Echte Matrizen sind **praktisch nie additiv**
  - Dann funktioniert Warnow's Algorithmus nicht – findet keine Lösung

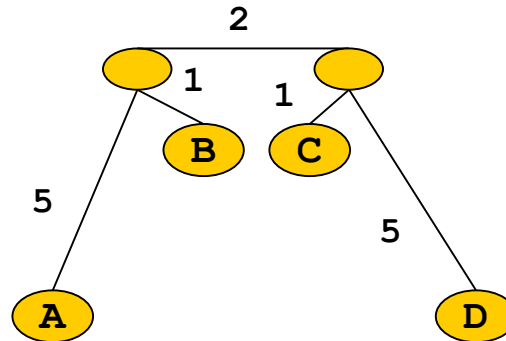
# Neighbor-Joining

---

- Findet **additiven Baum zu einer (fast) additiven Matrix**
- Auch ein hierarchisches Clusterverfahren
  - Erzeugt einen binären Baum ohne Wurzel
  - Beginne mit so vielen Clustern wie Blättern
  - Wähle zwei Cluster, die garantiert Nachbarn im Baum werden
    - Nach einem **bestimmtem Kriterium**
  - Verschmelze diese Cluster und verbinde Knoten im Baum
  - Iteriere, bis nur noch ein Cluster vorhanden ist
- Unterschiede zu UPGMA
  - UPGMA wählt Cluster zur Verschmelzung nur nach Nähe zueinander
  - Neighbor Joining wählt Cluster **nach der Nähe zueinander und dem Abstand** zu anderen Clustern aus
- Beweise sehr schön in [VSL02]

# Intuition

---



- B,C
  - Nahe beieinander
  - Aber auch relativ nahe an allen anderen
  - Kein starkes Signal, dass die sofort zusammengehören
- A,B
  - Nicht ganz so nahe beieinander
  - Aber A ist weit weg von allen anderen – B ist noch am nächsten



# Verfahren

---

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

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

- $n-2$ : Für durchschnitt. Abstand zu allen Knoten außer dem gewählten Paar (das sind  $n-2$  viele)
- Suche das Clusterpaar  $(i, j)$ , für das gilt

$$D[i, j] - u_i - u_j = \min$$

- Möglich nahe beieinander
- Möglichst weit weg von allen anderen

Wählt auf jeden Fall ein Nachbarpaar

# Verfahren 2

---

- 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  $ij$**  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

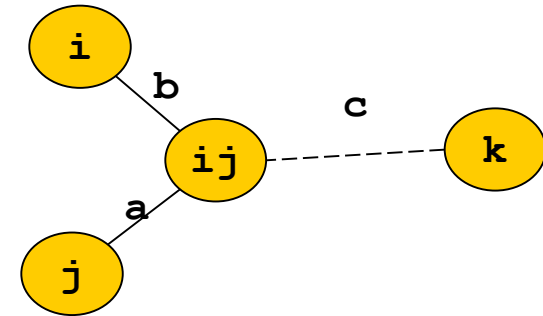
# Intuition

---

- Kantenbeschriftungen

- Wir kennen  $x=a+b$   
(aus der Matrix)
- Wir kennen  $y=a+c$   
(aus den mittleren Abständen;  $u_j$ )
- Genauso  $z=b+c$   
(aus den mittleren Abständen;  $u_i$ )
- Es ergibt sich  $c = (y+z-x)/2 = (a+c+b+c-a-b)/2 = (2c)/2 = c$

- $c$  ist der neue Wert in der Matrix – Abstand von  $k$  zu  $ij$



# Beispiel (hier scheiterte UPGMA)

NJ-Abstände der Clusterpaare

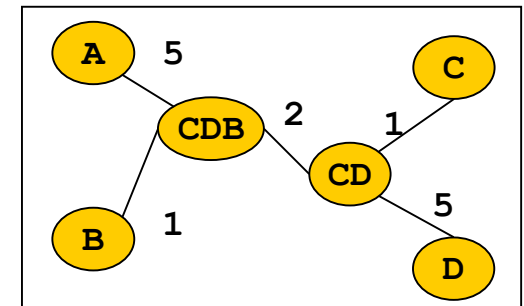
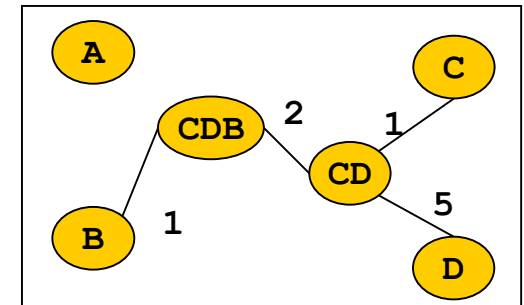
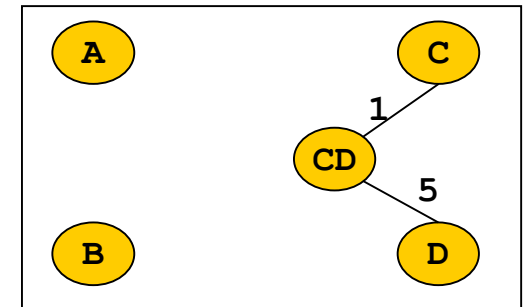
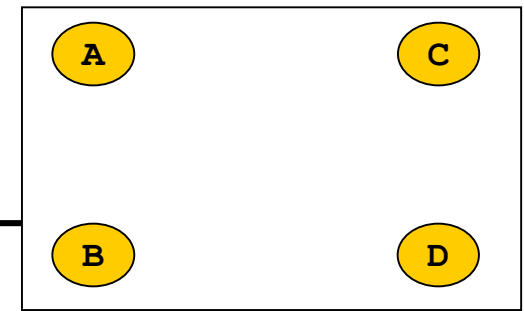
	A	B	C	D
A		6	8	12
B			4	8
C				6
$u_i$	13	9	9	13

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

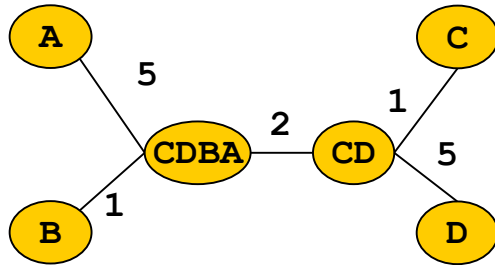
	A	B	CD
A		6	7
B			3
$u_i$	13	9	10

	B	CD
A	-16	-16
B		-16

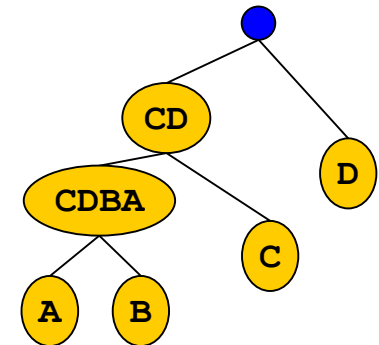
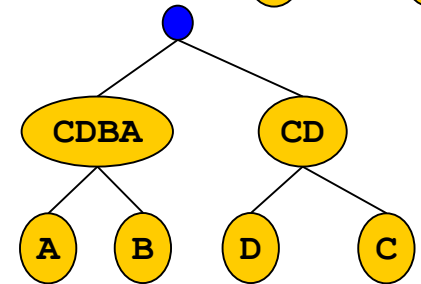
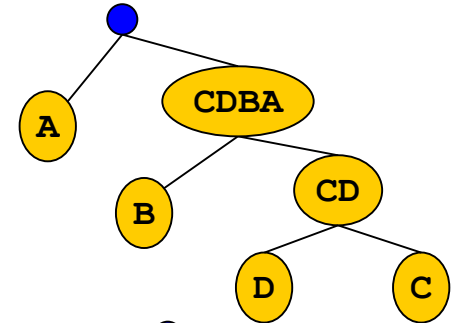
	A	BCD
A		5



# Rooting eines Baumes

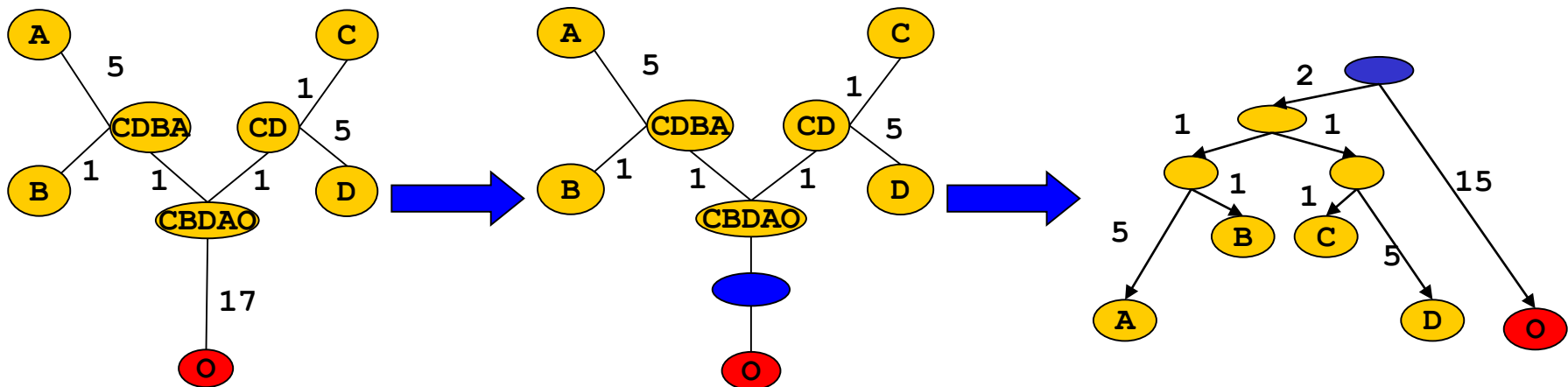


- NJ berechnet keine **zeitliche Reihenfolge** der Knoten
- Ein Wurzelknoten kann prinzipiell auf jeder Kante platziert werden
- Wie findet man die echte Wurzel?
  - Externe Datierung einzelner Knoten, z.B. durch Fossilien und C14-Methode
  - Benutzung einer **Outgroup**



# Outgroups

- Eine **Outgroup** ist ein Taxon, 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?
- Man lebt mit dem Fehler ...
- oder man formuliert ein Fehlerminimierungsproblem ...
  - Gegeben D. Finde Topologie T und Abstände im Baum  $d(i,j)$  so, dass der **folgende Fehler minimiert** wird

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

- Für gegebene Topologie T ist das effizient lösbar, aber ...
  - Man muss alle Topologien ausprobieren
  - Das Problem ist NP-schwierig
- oder man nimmt ganz andere Methoden (nächste Stunde)

# Literatur

---

- Gute Einführung (sehr praktisch orientiert)
  - Baldauf, S. L. (2003). "Phylogeny for the faint of heart: a tutorial." *Trends Genet* **19**(6): 345-51.
- 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
  - Sehr gut: Vingron, Stoye, Luz: Algorithms for Phylogenetic Reconstructions, Lecture Notes, 2002/2003



# Selbsttest

---

- Was besagt die „Molecular Clock Assumption“? Was folgt aus hier? Ist sie realistisch?
- Was ist ein additiver Baum?
- Beweisen Sie, dass es zu jeder Distanzmatrix genau dann einen ultrametrischen Baum gibt, wenn sie ultrametrisch ist
- Erklären Sie die 4-Punkt Bedingung für additive Matrizen an einem Beispiel
- Führen sie ein hierarchisches Clustering mit UPGMA auf der folgenden Matrix durch: ...
- Zeigen Sie, dass jede Ultrametric eine Metrik ist, aber nicht umgekehrt
- Geben Sie die WC-Komplexität von Neighbor Joining an