Einführung in die Angewandte Bioinformatik: Multiples Alignment und Phylogenetik 04.06.2009

Prof. Dr. Sven Rahmann

Bisher: Paarweise Alignments

Optimales Alignment:

- Alignment mit höchstem Score unter allen Alignments
- arbeitet die Ähnlichkeiten zwischen zwei Sequenzen heraus
- Spaltentypen: Match/Mismatch, Insertion, Deletion mit Scores bzw. Kosten

Wichtig für:

- quantitative Bestimmung von Sequenzähnlichkeit
- Übertragen von Informationen (Struktur/Funktion) zwischen ähnlichen Seq.

Aber:

Proteinfamilien oder Domänen bestehen aus mehr als 2 Sequenzen

Erinnerung: Proteindomänen und Proteinfamilien (Pfam)

Domänen

sind wiederkehrende modulare Bausteine von Proteinen.
Durch verschiedene Kombinationen von Domänen
entstehen Proteine mit unterschiedlichen Eigenschaften.
Ziel: alle existierenden Domänen katalogisieren, analysieren

Proteinfamilien

Eine Domäne oder eine bestimmte Kombination von Domänen
kann eine bestimmte Familie von Proteinen charakterisieren.

Datenbanken zu Domänen

Pfam:	http://pfam.sanger.ac.uk/	protein families
SMART:	http://smart.embl-heidelberg.de/	simple modular architecture research tool

Modellierung von Proteindomänen

Wie kann man eine Domäne beschreiben?

- Aminosäuresequenz (Konsensus + Variationsmöglichkeiten)
  Sequenz angeben, evtl. mehrere Symbole pro Position
  (nicht sehr nützlich wg. Variationen)

- statistisches Modell (Hidden-Markov Model, HMM)

- multiples Alignment aus bekannten Beispiel-Sequenzen

Beschreibung durch HMM

Hidden Markov Model (HMM)

Stochastisches generatives Modell:

- wird aus gegebenen Beispiel-Sequenzen (Alignment) erstellt
- kann weitere ähnliche Sequenzen generieren
- kann benutzt werden, um zu prüfen, ob eine neue Sequenz zum Modell passt
  (Berechne Wahrscheinlichkeit, dass HMM diese Sequenz generiert)

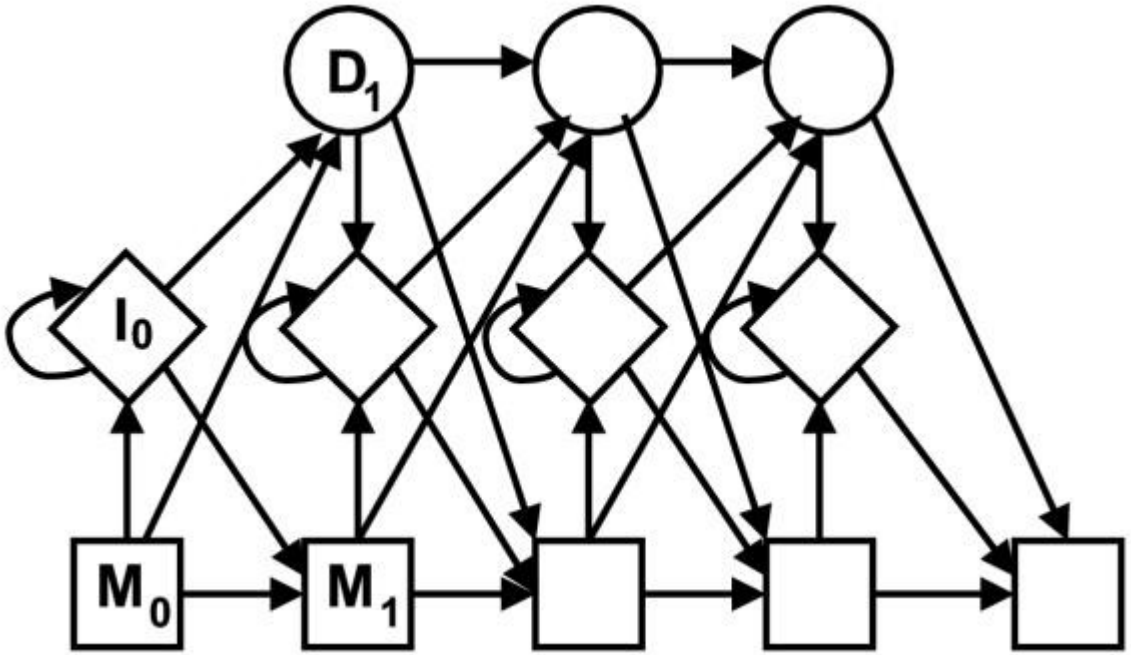
Modellparameter

Für jede Position wird angegeben:

- Wahrscheinlichkeit, Position auszulassen
- Aminosäure-Verteilung (Wahrscheinlichkeiten)
- Wahrscheinlichkeit, dahinter zusätzliche AS einzufügen
- Aminosäure-Verteilung der eingefügten AS

HMM (Profil-HMM)

Visuelle Vorstellung nach
Durbin, Eddy, Krogh, Mitchison:
Biological Sequence Analysis
Cambridge University Press



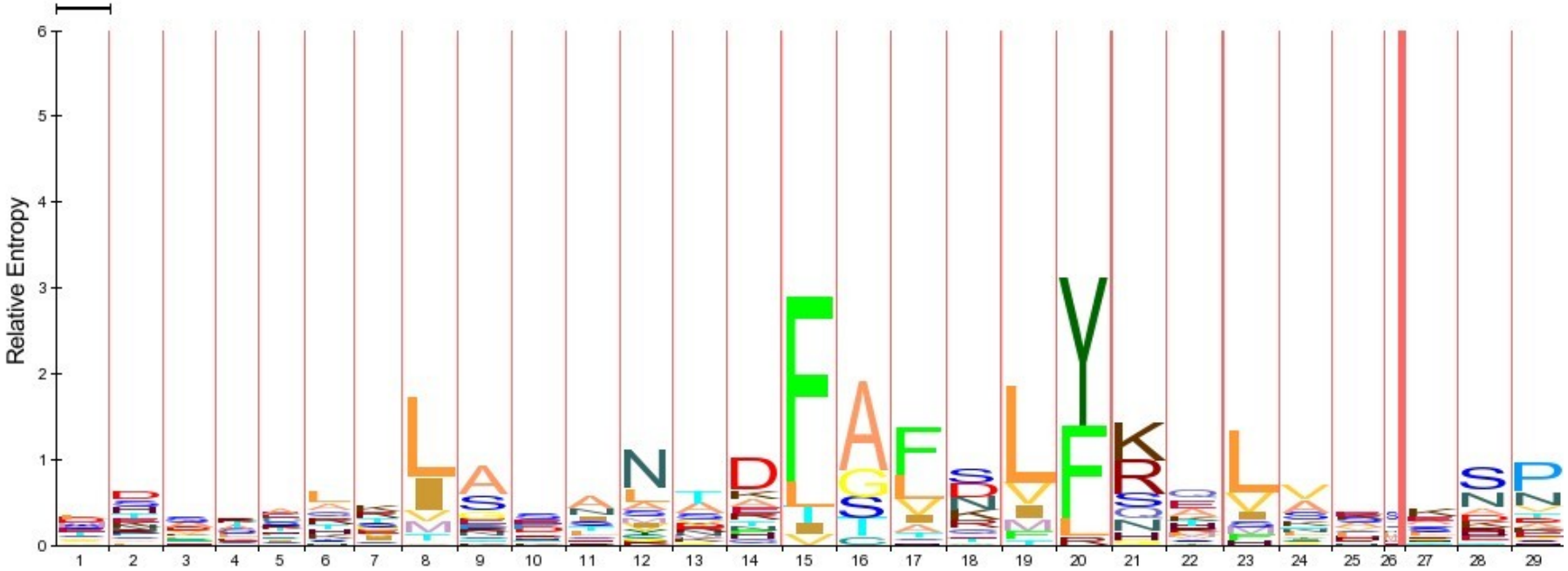
Modellparameter

Für jede Position wird angegeben:

- Wahrscheinlichkeit, Position auszulassen („deletion“ - D)
- Aminosäure-Verteilung (Wahrscheinlichkeiten in M0, M1, ...)
- Wahrscheinlichkeit, dahinter zusätzliche AS einzufügen („insertion“ - I)
- Aminosäure-Verteilung der eingefügten AS (Wahrscheinlichkeiten in I0, ...)

Visualisierung durch HMM-Logos: Teil der Serpin-Domäne (Serin Protease Inhibitor)

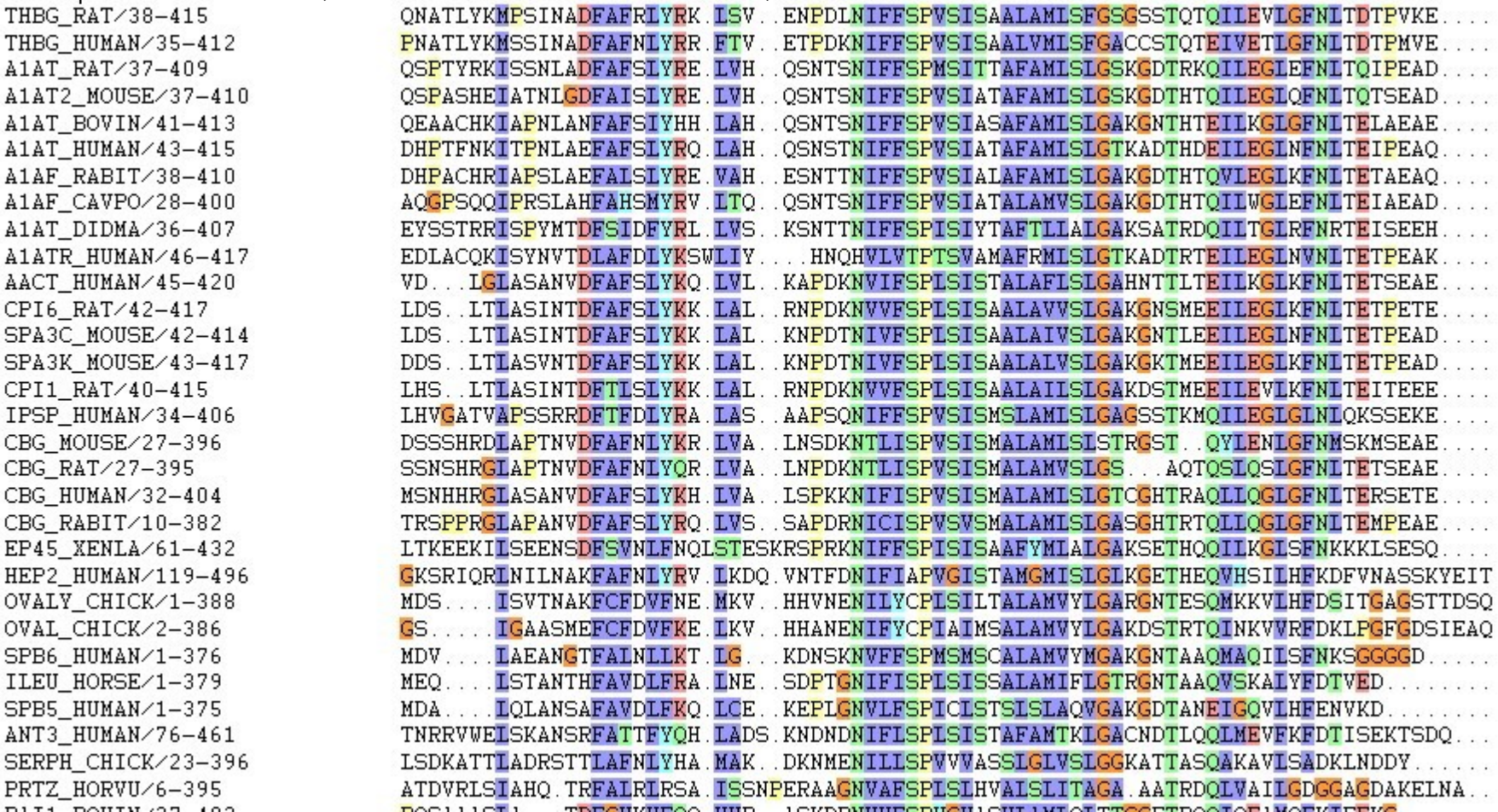
Höhe der Türme: Grad der Konserviertheit
Breite der Türme: Wahrscheinlichkeit, nicht ausgelassen zu werden
Breite der roten Balken: Insertionswahrscheinlichkeit zwischen zwei Positionen



If you use HMM-Logos in your publication, please cite
"Schuster-Boeckler B, Schultz J, Rahmann S: HMM Logos for visualization of protein families. BMC Bioinformatics 2004, 5:7"
The paper is "open access": http://www.biomedcentral.com/1471-2105/5/7

Beschreibung durch Multiples Alignment

Serpin-Domäne (Serin Protease Inhibitor)



Multiple Alignments

Motivation

- Möchte alle Mitglieder einer Proteinfamilie auf einmal betrachten, Gemeinsamkeiten und Unterschiede auf einen Blick sehen.
- Homologe (evolutionär sich entsprechende) Positionen besser sichtbar.
- Drei optimale paarweise Alignments von drei Sequenzen evtl. inkonsistent: a mit b aligniert, b mit c, aber a nicht mit c.
  In einem multiplen Alignment unmöglich: Alle Sequenzen gleichzeitig aligniert!

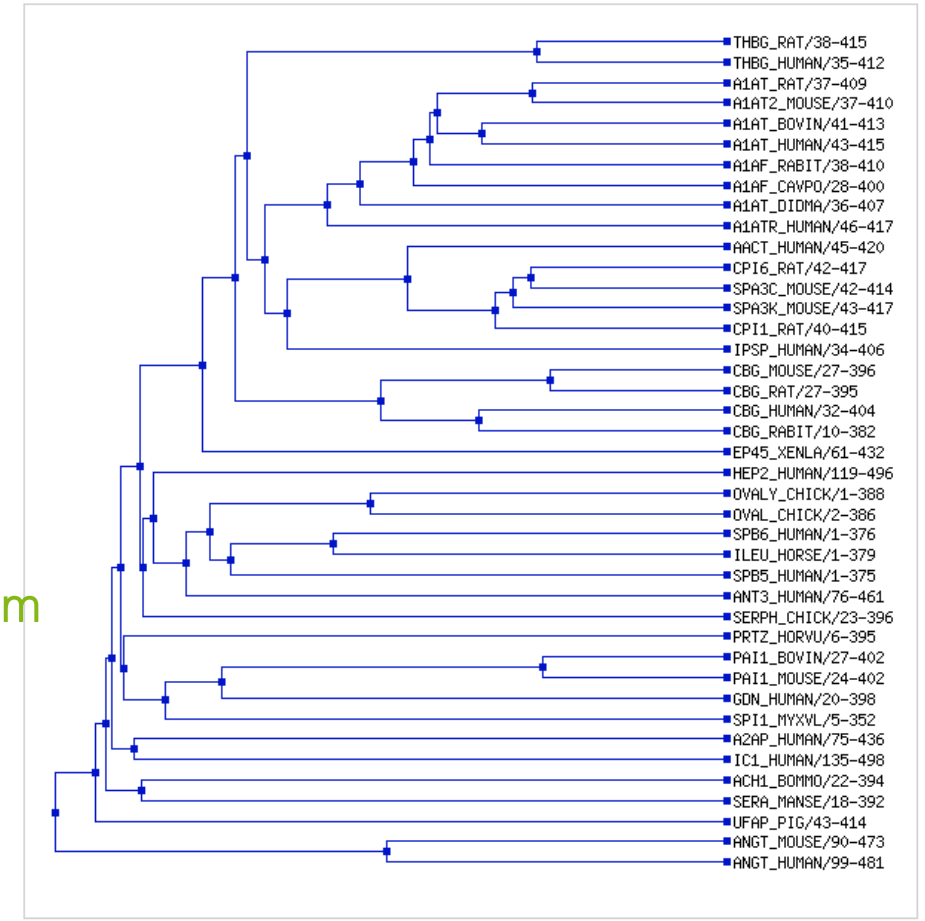
Multiples Alignment :=
Alignment von mindestens 2, i.d.R. mindestens 3 Sequenzen.
Jede Zeile entspricht (durch Weglassen der Gaps) einer der Sequenzen.
Bei k Sequenzen kann jede Spalte 0 bis k-1 Gap-Zeichen enthalten.

Wann sind multiple Alignments sinnvoll?

Globales multiples Alignment nur sinnvoll, wenn
- Sequenzen global ähnlich,
- alle Sequenzen evolutionär verwandt

Lokales multiples Alignment sinnvoll:
- hinreichend lange gemeinsame ähnliche Teilsequenz aller Sequenzen (Domäne?)
- diese Bereiche evolutionär verwandt

Verwandtschaftsbeziehungen werden durch phylogenetischen Baum beschrieben.
Beispiel: Serpine in Pfam



Berechnung multipler Alignments

Ziel:

Biologisch / evolutionär korrektes multiples Alignment
(evolutionär voneinander abstammende Aminosäuren,
oder solche mit gemeinsamem Vorfahren, stehen untereinander)

Formulierung als Optimierungsproblem:

Definiere (wie schon auf paarweisen Alignments) eine Scorefunktion.
Finde das multiple Alignment, das den Score maximiert.

Problematisch:

Es gibt vermutlich kein Scoring-Verfahren,
das immer das evolutionär korrekte Alignment
zu dem mit der höchsten Score macht (Beispiel).

Scoring von multiplen Alignments

Sum-of-pairs Score:

Multiples Alignment aus k Sequenzen enthält ~ $k^2/2$ paarweise Alignments.
Summe aller paarweisen Scores ergibt den Score des multiplen Alignments.

Tree score:

Annahme: Zwischen Sequenzen bestehen evolutionäre Verwandtschaften, gegeben durch phylogenetischen Baum.
Summiere paarweise Scores von im Baum benachbarten Sequenzen.

gewichteter Sum-of-pairs Score

wie Sum-of-pairs Score, aber jedes Paar erhält individuelles Gewicht

Optimierungsprobleme beim multiplen Alignment

Sum-of-pairs Problem

Gegeben *k* Sequenzen, Scorematrix, Gapkosten.
Finde multiples Alignment, das (gewichteten) sum-of-pairs Score maximiert.

Tree Alignment Problem

Gegeben zusätzlich ein Baum mit den *k* Sequenzen an den Blättern.
Finde Belegung der inneren Knoten mit Sequenzen (gemeinsame Vorfahren)
und multiples Alignment, das den Tree Score maximiert.

Verallgemeinertes Tree Alignment Problem

Gegeben *k* Sequenzen, Scorematrix, Gapkosten (kein Baum!),
finde Baumtopologie, Belegung der inneren Knoten mit Sequenzen
und multiplea Alignment, das den Tree Score maximiert

Schwierigkeit des multiplen Alignment - Problems

Für alle drei Varianten des Problems
- Sum-of-pairs Problem
- Tree Alignment Problem
- Verallgemeinertes Tree Alignment Problem

gilt:

Es gibt exakte Algorithmen,
aber deren Zeitbedarf ist exponentiell in der Anzahl der Sequenzen k.
Diese sind nur praktikabel für 3 – 7 Sequenzen.

Die Probleme sind NP-schwer
(mindestens genauso schwierig wie alle anderen Probleme in NP;
NP: nichtdeterministisch in Polynomialzeit lösbar).

Heuristiken

Zwei Probleme bisher:
- Score-Maximierung eines Alignments biologische Korrektheit
- Exakte Score-Maximierung dauert zu lange

Verwendung von Heuristiken (gr. heuriskein, „(auf-)finden“, „entdecken“),
(Kunst, mit begrenztem Wissen und wenig Zeit zu guten Lösungen zu kommen)

Laufzeit-Heuristik:
- löst das Problem in jedem Fall exakt.
- versucht, durch „Abkürzungen“ die Lösung möglichst schnell zu finden.

(Qualitäts-)Heuristik:
- garantierte (schnelle) Laufzeit.
- gefundene Lösung nicht notwendig optimal.

Heuristiken für multiples Alignment

Center-star-Methode:

- Wähle Sequenz (mit geringster evolutionärer Abstandssumme zu den anderen).
- Aligniere jede andere Sequenz paarweise daran.
- Setze k-1 paarweise Alignments zu einem multiplen Alignment zusammen.
- Nachteil: Es werden nur k-1 der möglichen paarweisen Alignments betrachtet.

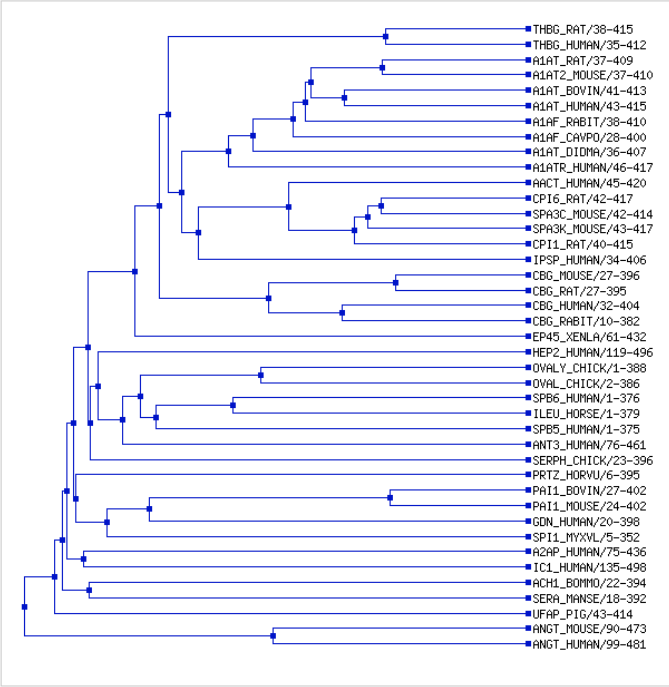
Divide-and-conquer Alignment:

- Suche etwa in der Mitte jeder Sequenz gut konservierte Stellen,
- an der man relativ sicher ist, dass alle diese Stellen eine Alignmentspalte bilden.
- Teile Problem in linkes und rechtes Problem auf, verfahre dort genauso.
- Nachteil: funktioniert nicht gut, wenn es keine solchen Stellen gibt.

Heuristiken für multiples Alignment

Progressives Alignment:

- Beginne mit zwei nah verwandten Sequenzen
- Aligniere in jedem Schritt eine weitere Sequenz optimal paarweise an bestehendes multiples Alignment.
- Reihenfolge durch evolutionäre Distanzen der Sequenzen bestimmt (Baum).
- Nachteil: falsche frühe Entscheidungen können nicht rückgängig gemacht werden.
- Wichtigstes Beispiel: **Clustal**



Clustal (http://www.clustal.org)

- schnelle Heuristik (und Software-Paket) für multiple Alignments

Webserver zum Berechnen von Alignments:

- am EBI: http://www.ebi.ac.uk/Tools/clustalw2/index.html
- am SIB: http://www.ch.embnet.org/index.html

Idee und Verfahren

- Berechne eine Folge von paarweisen Alignments
- Nimm dazu einen Baum zur Hilfe („guide tree“)
  (Art phylogenetischer Baum, aber eher ein Hilfsmittel zur Berechnung).
- Bereits existierende Teil-Alignments werden dabei nicht mehr verändert.

1. Berechne Distanzen zwischen allen Sequenzpaaren
2. Berechne aus den Distanzwerten einen Baum (mehr dazu später)
3. Aligniere Sequenzen und existierende Alignments
   in der Reihenfolge, die der Baum vorgibt (bottom-up).

Abhängigkeit von Baum und Alignment

Baum bildet Verwandtschaftsverhältnisse der Sequenzen ab.
Wichtiges Hilfsmittel beim Berechnen des multiplen Alignments.
Woher bekommt man einen solchen Baum („guide tree“) ?

Man schätzt evolutionäre Distanzen (z.B. in PAM-Einheiten)
zwischen den Sequenzen.
Dazu braucht man aber das Alignment.

Henne-Ei-Problem:
- Berechnung des Baums benötigt Alignment.
- Berechnung des Alignments benötigt Baum.

Abhängigkeit von Baum und Alignment

Lösung des Henne-Ei-Problems:

Beginne mit (groben) Schätzungen für Distanzen,
z.B. aus paarweisen Alignments (unterschätzen wahren Distanzen).
Berechne ersten Baum.
Erstelle erstes multiples Alignment.

Schätze daraus neue Distanzen, neuen Baum, neues Alignment.

Iteriere so lange, bis sich nichts mehr ändert,
oder eine Maximalzahl an Iterations-Schritten gemacht wurde.

Phylogenetik: Berechnung phylogenetischer Bäume

Phylogenetik (phylum = Stamm):
Rekonstruktion von evolutionären Stammbäumen
(= phylogenetische Bäume, **Phylogenien**)
aus unterscheidbaren **Merkmalen**, insbes. aus DNA- / Proteinsequenzen,
und/oder aus Merkmalen berechneten **Distanzen**

- für Spezies [**Spezies-Bäume**]
- für einzelne Gen- / Protein-Familien [**Gen-Bäume**]

Tree Of Life - Projekt

Langfristiges Ziel:
Stammbaum aller existierenden (und ausgestorbenen) Arten.
Tree Of Life Web Project: http://www.tolweb.org

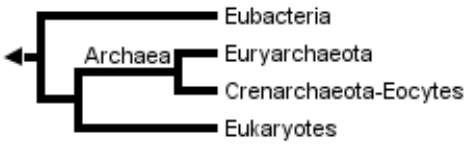
Tree Of Life Web Project

http://www.tolweb.org

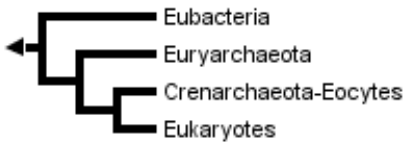
Kontroverse über Verzweigungen nahe der Wurzel

Eubacteria ("True bacteria", mitochondria, and chloroplasts)
Eukaryotes (Protists, Plants, Fungi, Animals, etc.)
Archaea (Methanogens, Halophiles, Sulfolobus, and relatives)
?≡ Viruses

The "archaea tree":

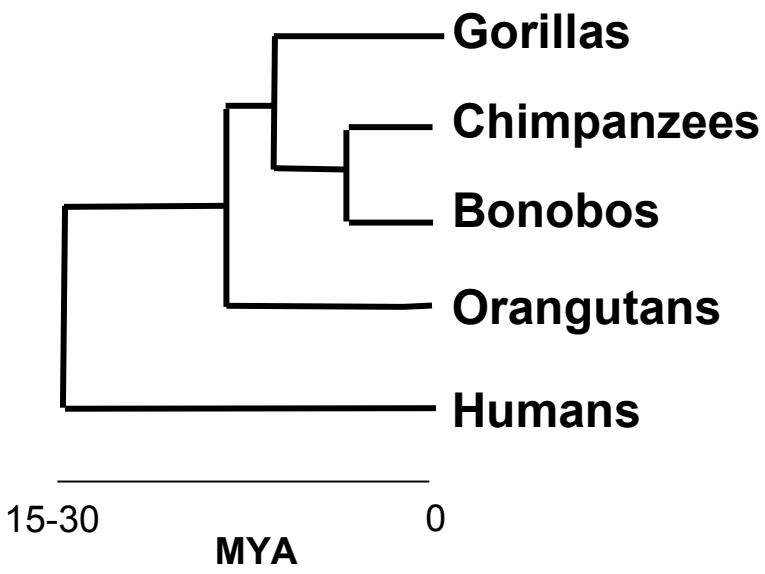


The "eocyte tree":

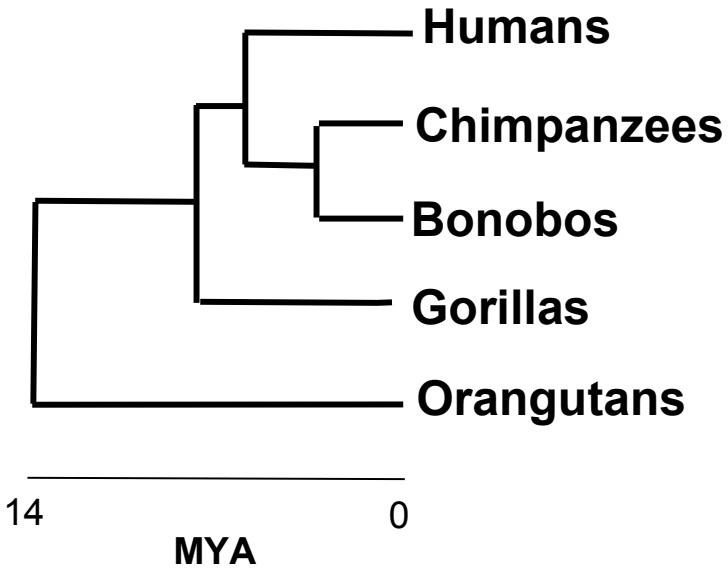


Aufgaben der Phylogenetik (Speziesbäume)

Bestimmung der Verwandtschaftsverhältnisse zwischen Arten (Spezies) und der Ereignisse, die bei den Arten seit der Speziation stattfanden.



Klassische Sichtweise,
basierend auf äußeren Merkmalen



Moderne Sichtweise,
basierend auf DNA-Vergleich

Aufgaben der Phylogenetik (Genbäume)

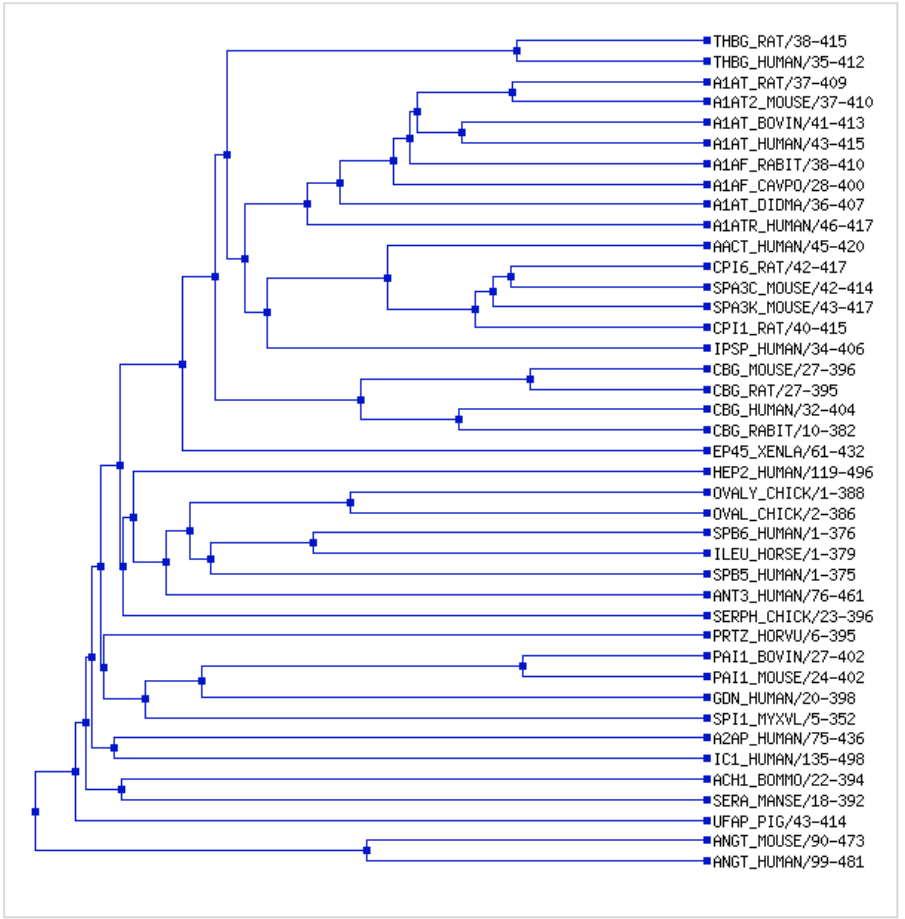
Bestimmung der Verwandtschaftsverhältnisse zwischen Genen in einer Gen- / Proteinfamilie / Domäne.

Bekanntes Beispiel:
Serpin-Familie in der pfam-Datenbank

Dies ist ein Gen-Baum (gene tree), kein Arten-Baum (species tree):

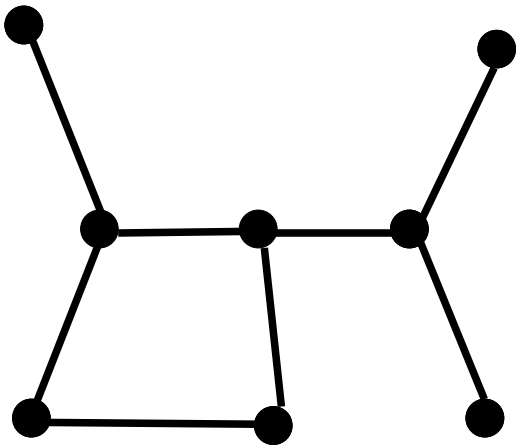
An den Blättern (rechts) stehen einzelne Gene/Proteine/Sequenzen, keine Spezies.

Evolutionäre Ereignisse (innere Knoten):
- Genduplikation
- Speziation



Grundlagen: Graphen

Ein Graph besteht aus:
- **Knoten** (vertices *V*, nodes) und
- **Kanten** (edges *E*),
  die zwischen den Knoten verlaufen.



Graphen können **gerichtet** oder **ungerichtet** sein
(Kanten haben eine Richtung oder nicht).

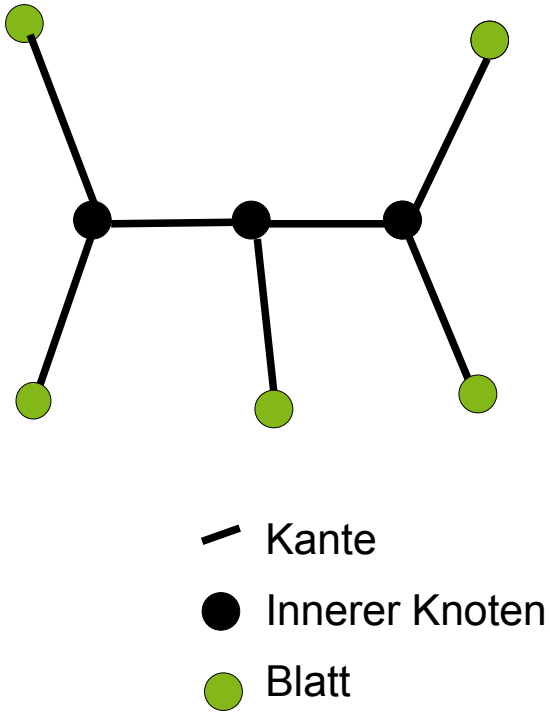
Der **Grad** eines Knoten
ist die Anzahl seiner Nachbarn.

Grundlagen: ungewurzelte Bäume

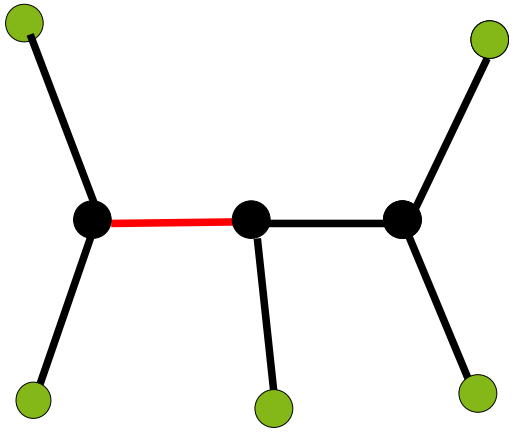
Ungewurzelter Baum (unrooted tree) $T = (V,E)$:
Graph ist
- zusammenhängend
  (:= jeder Knoten wird von jedem anderen erreicht)
- kreisfrei
  (:= es gibt nur einen Weg zwischen je zwei Knoten)

In einem Baum unterscheidet man zwischen
- inneren Knoten (inner nodes N), und
- Blättern = äußeren Knoten (leaves L).

Baum heißt Binärbaum,
wenn innere Knoten Grad 3 und Blätter Grad 1 haben.
Ungewurzelte binäre Bäume erfüllen
$|E| = 2|L| - 3$ und $|N| = |L| - 2$.

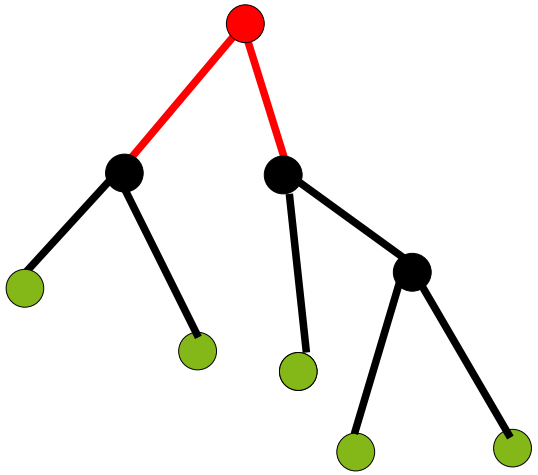


Grundlagen: Gewurzelte Bäume



Ungewurzelter Baum kann gewurzelt werden,
indem man eine Kante auswählt,
in deren Mitte einen Knoten (Wurzel) einfügt,
und den Baum daran aufhängt.

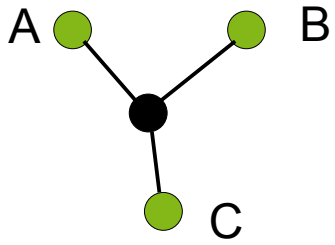
Wurzel hat Grad 2, stets ganz oben.
Kanten werden von der Wurzel weg gerichtet
(Kanten verlaufen von oben nach unten.)
Blätter unten.



Anzahl Binärbäume

Ungewurzelte Binärbäume:

- 3 Blätter: 1 Baum mit 3 Kanten
- 4 Blätter: 3 verschiedene Bäume, je 5 Kanten
- 5 Blätter: 3*5 = 15 Bäume, je 7 Kanten
- 6 Blätter: 15*7 = 105 Bäume ...

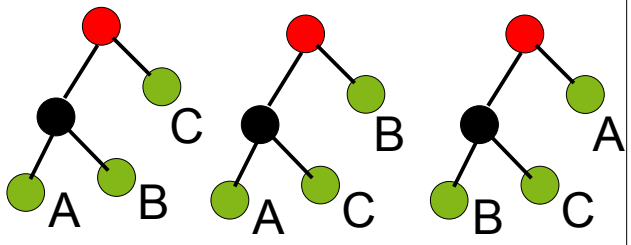


3 Blätter: 1 ungewurzelter Baum

Gewurzelte Binärbäume:

Man kann die Wurzel wie ein weiteres Blatt bei ungewurzelten Bäumen behandeln:

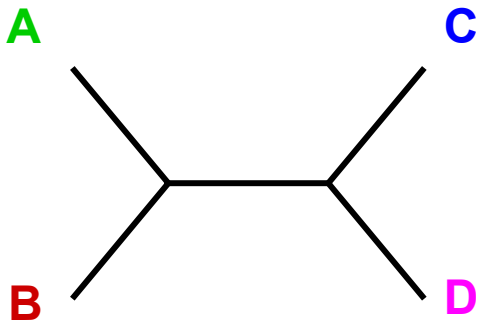
- 3 Blätter: 3 Bäume
- 4 Blätter: 3*5 = 15 Bäume
- 5 Blätter: 15*7 = 105 Bäume ...



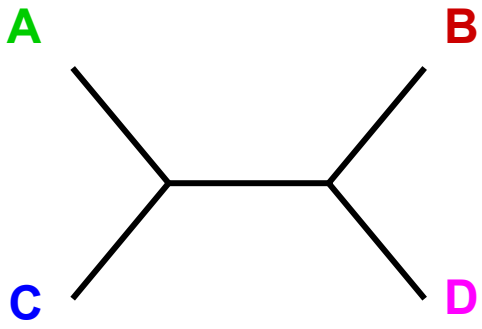
3 Blätter: 3 gewurzelte Bäume mit veschiedenen „Nachbarschaftsverhältnissen“

Anzahl der binären Bäume wächst super-exponentiell in der Zahl der Blätter n (schneller als c^n für jede Konstante c)

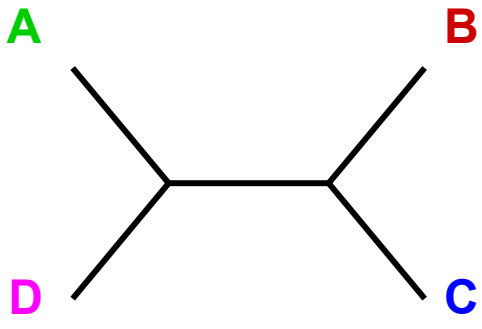
Die drei Quartette (ungewurzelte Bäume mit 4 Knoten)



AB || CD



AC || BD



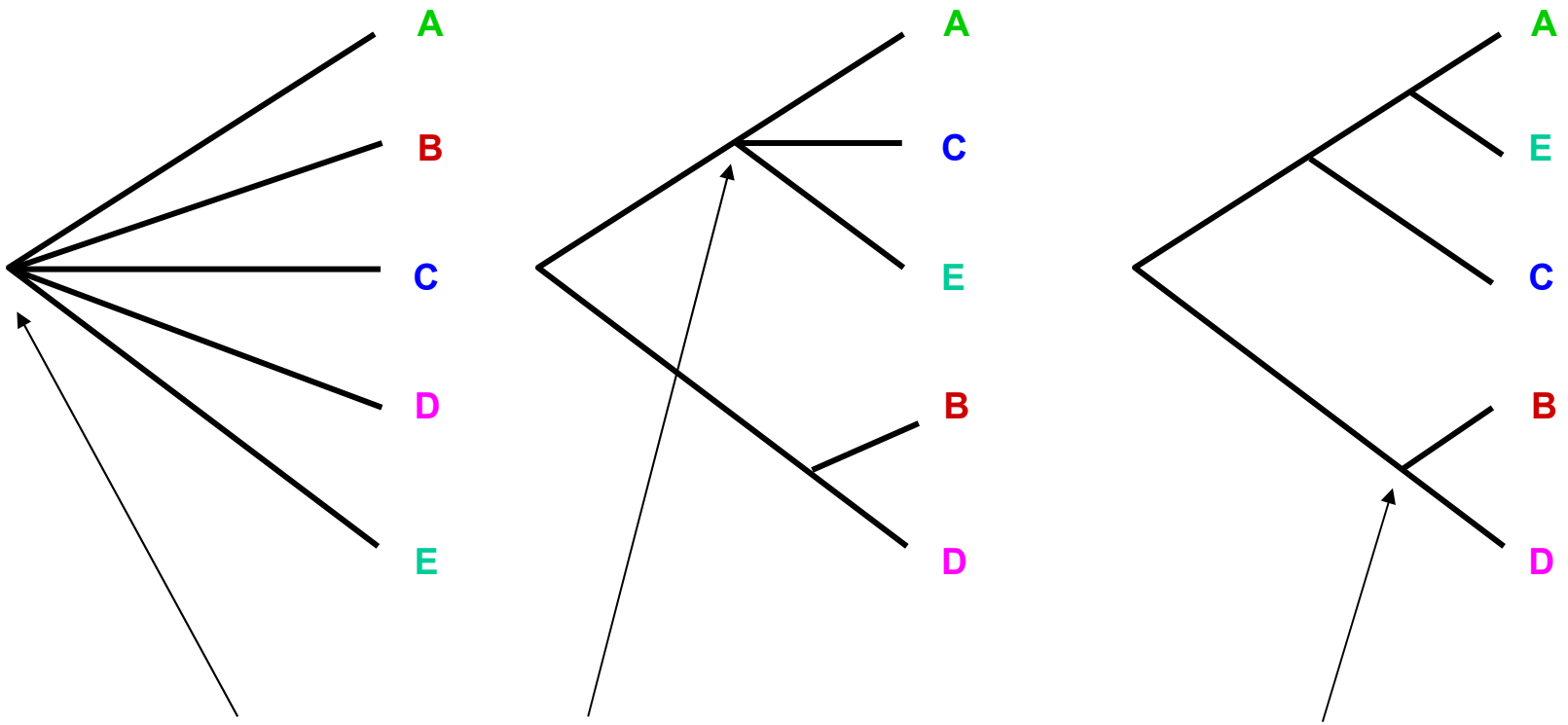
AD || BC

Grad der Auflösung des Verzweigungsmusters

"Stern"-Baum, nicht aufgelöst

Teilweise aufgelöst

Vollständig aufgelöst (Binärbaum)

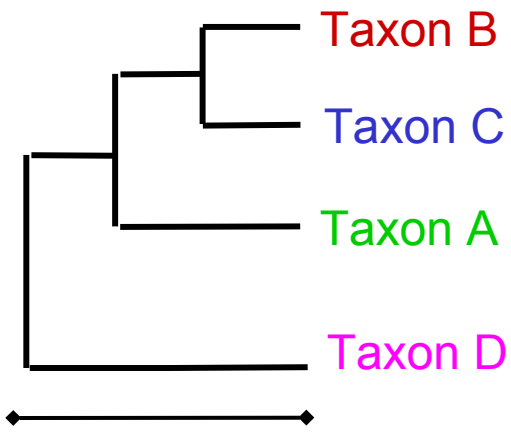


Polytomie, Multifurkation

Bifurkation

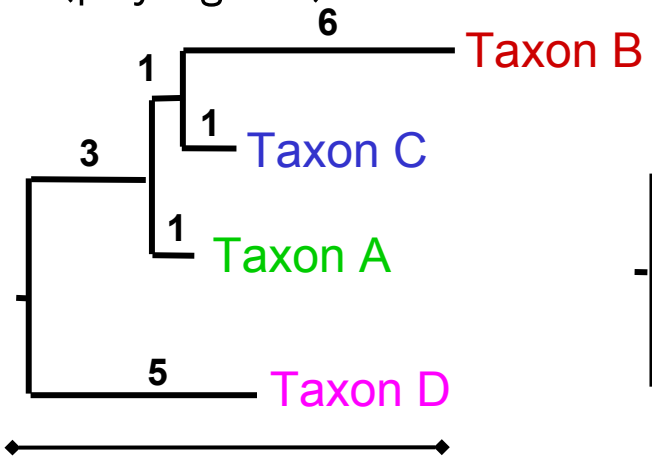
Darstellung phylogenetischer Bäume

Kladogramm
(cladogram)



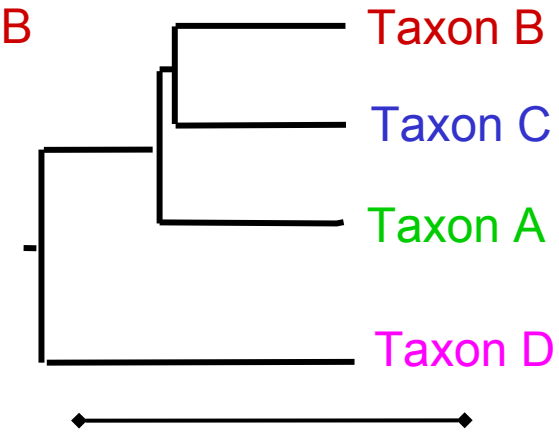
Achse hat
keine Bedeutung

Phylogramm
(phylogram)



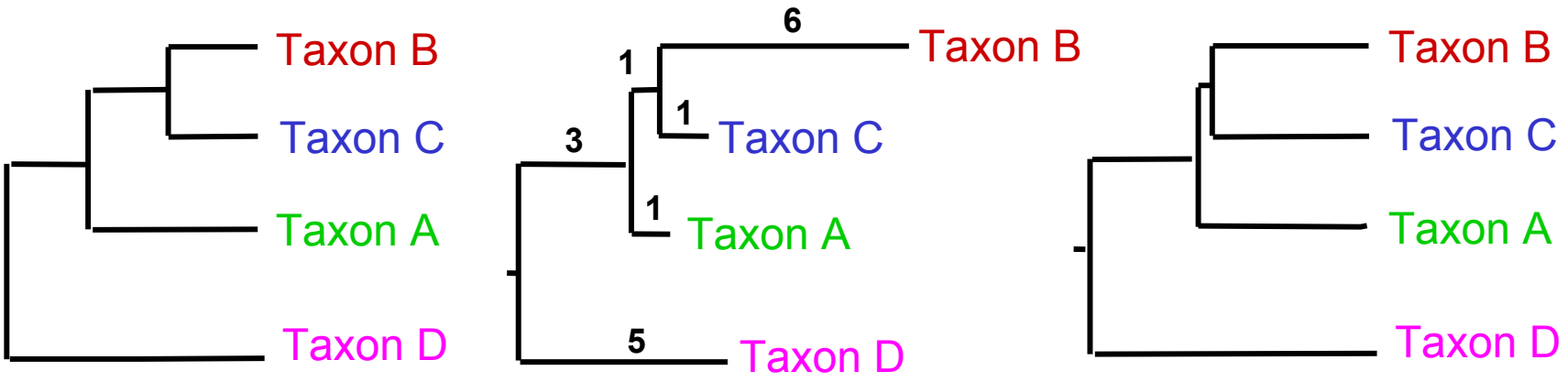
Menge an Evolution
(z.B. PAM-Einheiten)

Ultrametrischer Baum
(ultrametric tree)



Realzeit: gleicher Abstand
aller Blätter zur Wurzel

Darstellung phylogenetischer Bäume



Alle Darstellungsformen (Kladogramm, Phylogramm, ultrametrischer Baum) zeigen die gleiche Baumtopologien (Verzweigungen zwischen den Taxa).
Taxa im selben Unterbaum bilden eine monophyletische Gruppe (clade).

Darstellung durch geschachtelte Klammern:
- Kladogramm: (((B,C),A),D)
- Phylogramm: (((B:6,C:1):1,A:1):3), D:5)
- ultrametrischer Baum: analog

Die molekulare Uhr (molecular clock)

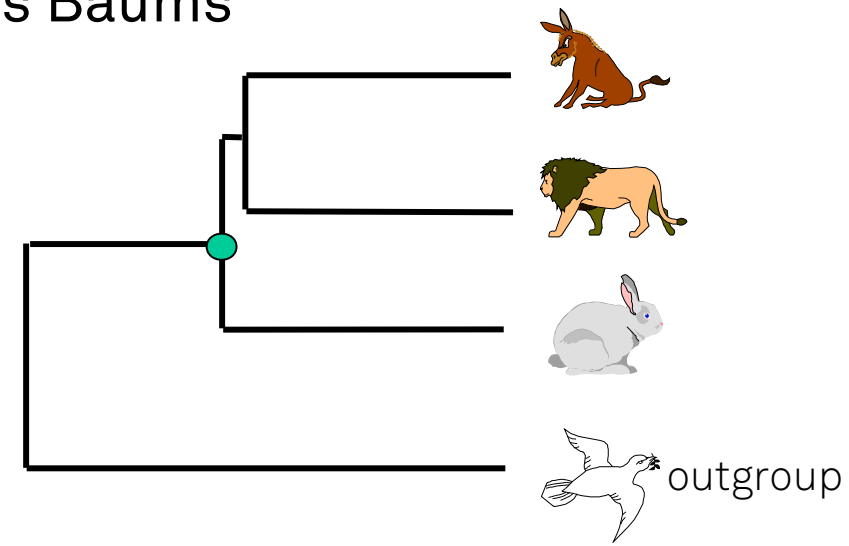
Evolutionäre Abstände zwischen Sequenzen werden gemessen
- nicht in Realzeit,
- sondern als "Evolutionsmenge" [PAM].

Problem: Distanz ist nicht Realzeit.
Evolutionsrate variiert idR zeitlich, genspezifisch, gattungsspezifisch.
Berechnete Bäume repräsentieren keine realen zeitlichen Abstände.
Auch die Wurzel kann nicht zuverlässig positioniert werden.
Man berechnet ungewurzelte Phylogramme.

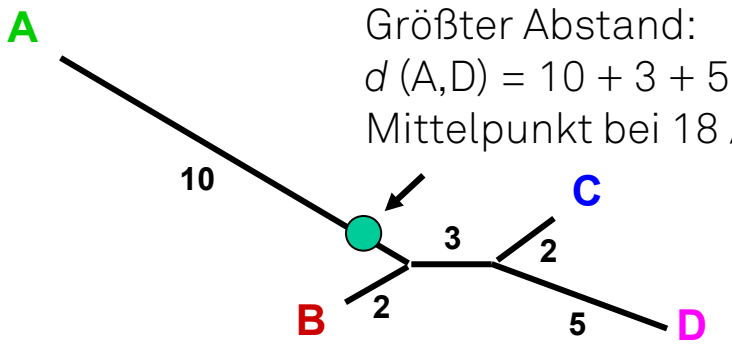
Ausnahme: Evolutionsrate konstant.
Man sagt: Es gilt die Hypothese der „molekularen Uhr“.
Man erhält einen gewurzelten ultrametrischen Baum.

Zwei Methoden zum Wurzeln eines Baums

(1) Mit Hilfe einer „outgroup“:
Spezies, die bekanntermaßen außerhalb der betrachteten Gruppe liegt.
- erfordert Vorwissen
- relativ zuverlässig



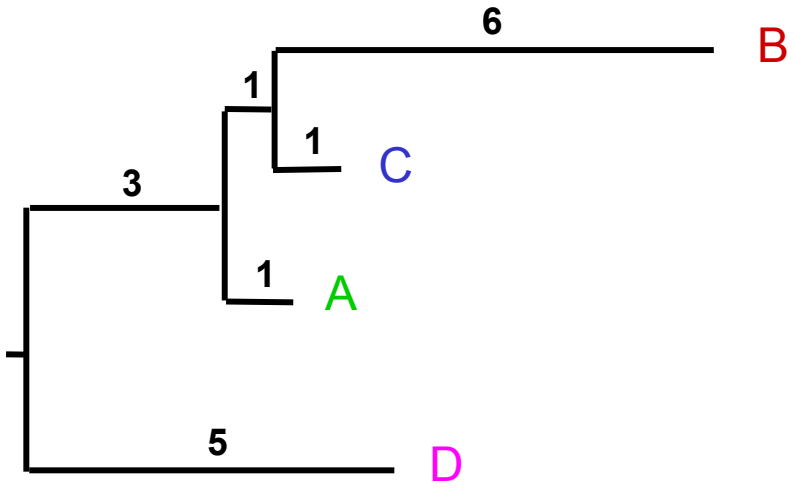
(2) Als Mittelpunkt
der am weitesten voneinander entfernten Taxa
(einfach, relativ unsicher)



Größter Abstand:
d (A,D) = 10 + 3 + 5 = 18
Mittelpunkt bei 18 / 2 = 9

Ähnlichkeit ist nicht Verwandtschaft

Ähnlichkeit, Distanz: beobachtbar, messbar.
Verwandtschaft: historische Tatsache, kann nicht mehr beobachtet werden.

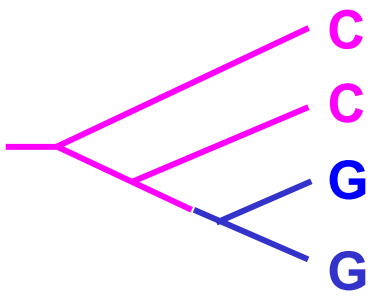


C ist ähnlicher zu A ($d = 3$) als zu B ($d = 7$),
aber C und B sind am nächsten verwandt.

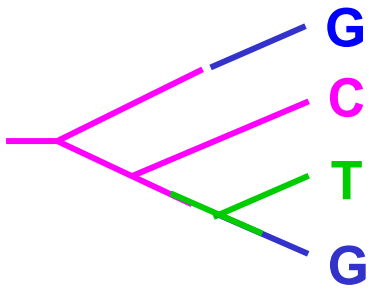
Das heißt, C und B hatten einen
späteren gemeinsamen Vorfahren
als C oder B mit A.

3 Gründe für beobachtete Ähnlichkeit

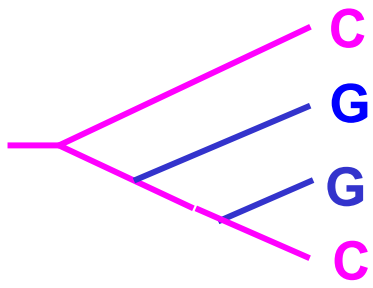
(1) Evolutionäre Verwandtschaft
- gemeinsame Stamm-Merkmale (C)
- gemeinsame abgeleitete Merkmale (G)



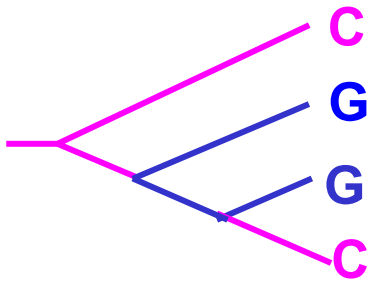
(2) parallele unabhängige Entwicklung von Merkmalen (Homoplasie)



konvergente Mutation zu G



parallele Mutation zu G

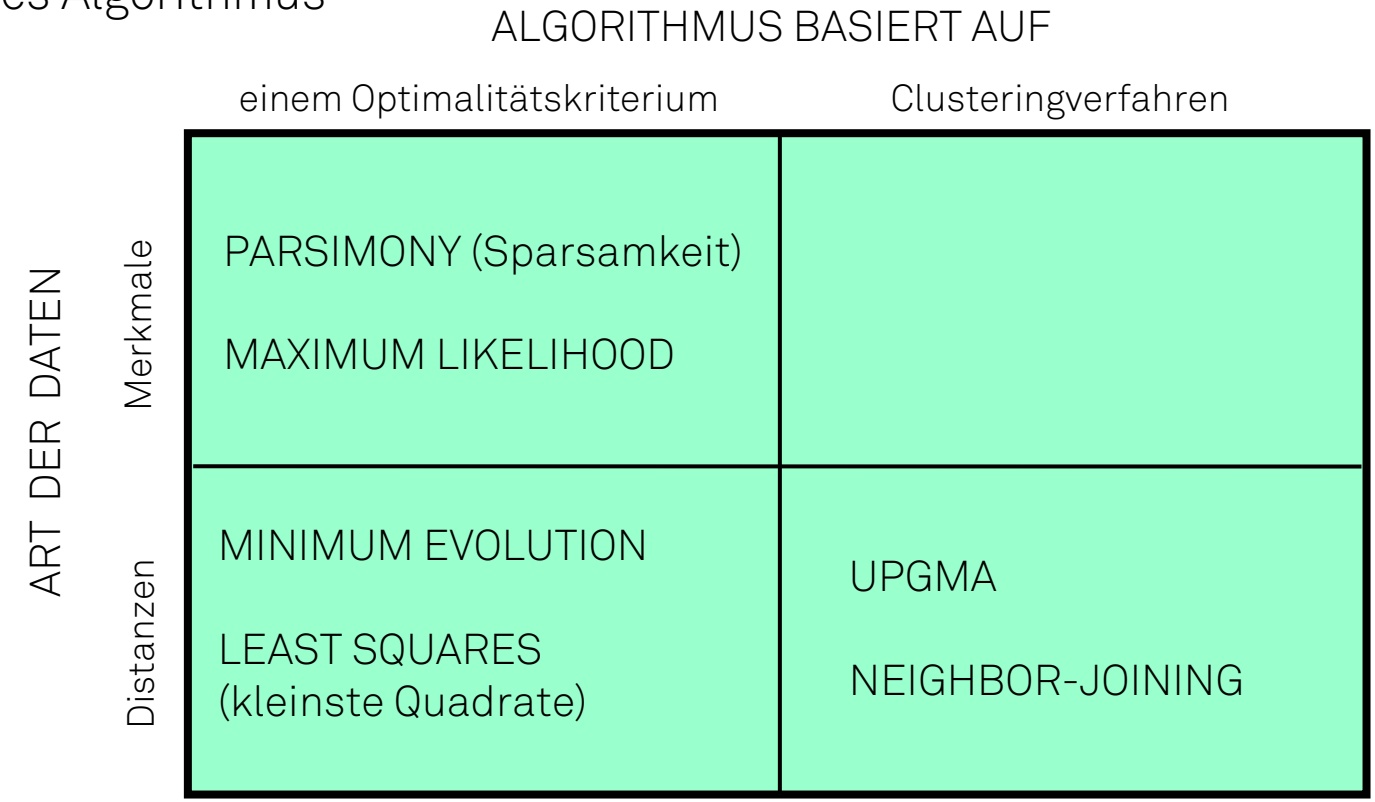


Rückmutation zu C

(3) geringe Evolutionsraten (hoher negativer selektiver Druck)

Methoden der Phylogenetik

Unterscheidung nach:
- Art der Daten (Merkmale oder Distanzen)
- Prinzip des Algorithmus



Merkmals-basierte Methoden

Merkmale:

- Objekte, die einen von mehreren Zuständen annehmen können, z.B. Spalte in einem multiplen Alignment von DNA-Sequenzen.
- In jedem Taxon liegt eine **Merkmalsausprägung** vor.
- Merkmale (Alignments) werden direkt zur Baumkonstruktion verwendet.

- Merkmalsausprägung der Ahnen kann geschätzt werden.
- Zeitpunkte evolutionärer Ereignisse können geschätzt werden.
- Häufig schwierige (NP-schwere) resultierende Probleme

Taxa	Merkmale
Species A	**ATGGCTATTCTTATAGTACG**
Species B	**ATCGCTAGTCTTATATTACA**
Species C	**TTCACTAGACCTGTGGTCCA**
Species D	**TTGACCAGACCTGTGGTCCG**
Species E	**TTGACCAGTTCTCTAGTTCG**

Distanz-basierte Methoden

Merkmale werden mit mathematischem Modell ("evolutionärer Markovprozess") in paarweise Distanzen zwischen Taxa umgerechnet.
Nur Distanzen werden zur Baumkonstruktion verwendet.
(Merkmale werden "vergessen".)

- vielseitig (auch ausserhalb der Phylogenetik) verwendbare Methoden
- Evolutionäre Ereignisse und Ahnen-Merkmale können nicht geschätzt werden.

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Zwei berechnete Distanzmatrizen
rechts: Anteil Unterschiede, links: PAM-Schätzung / 100

Optimierungs-Methoden

Kriterium (Eigenschaft des zu berechnenden Baumes) wird definiert,
das es zu optimieren gilt.

Beispiele:
- möglichst wenig Mutationsereignisse insgesamt im Baum (Merkmals-basiert)
- möglichst wenig Diskrepanz zwischen gegebenen Distanzen
  und den Distanzen, die die Baumtopologie impliziert (Distanz-basiert)

Aufgabe des Algorithmus:
Finde einen (den) Baum, der das Kriterium optimiert.

- Unterscheide exakte Algorithmen vs. Heuristiken.
- Kann Ausgaben (Güte) verschiedener Algorithmen miteinander vergleichen:
  "Baum A ist (bezüglich des Kriteriums) besser als Baum B".

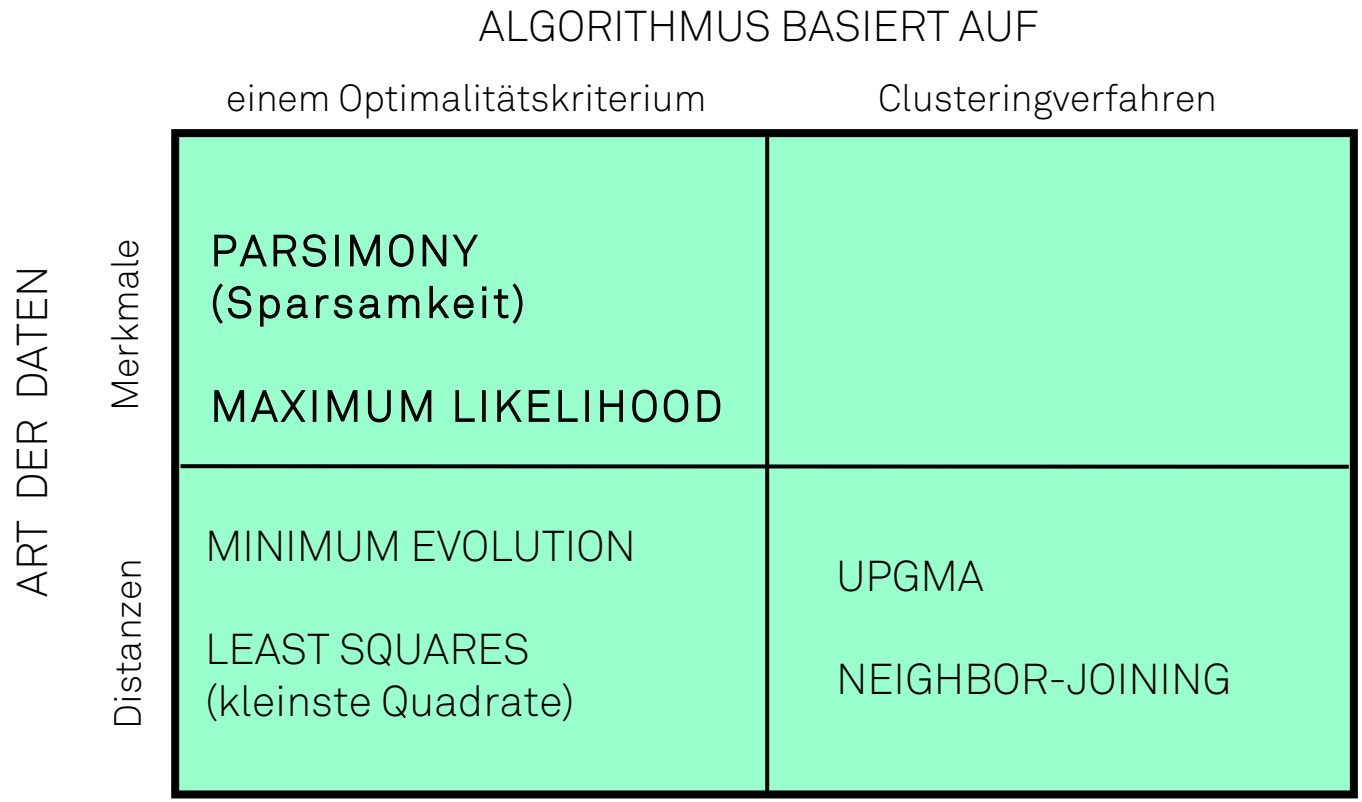
Clustering-Methoden

Definiere Regeln,
nach denen Taxa zu (2er-)Bäumen zusammenzufassen sind,
diese wiederum zu größeren Bäumen, etc.

Liefern immer einen einzigen Baum zurück;
dieser kann nicht gegenüber Alternativen bewertet werden.
(Es gibt kein Optimalitätskriterium.)

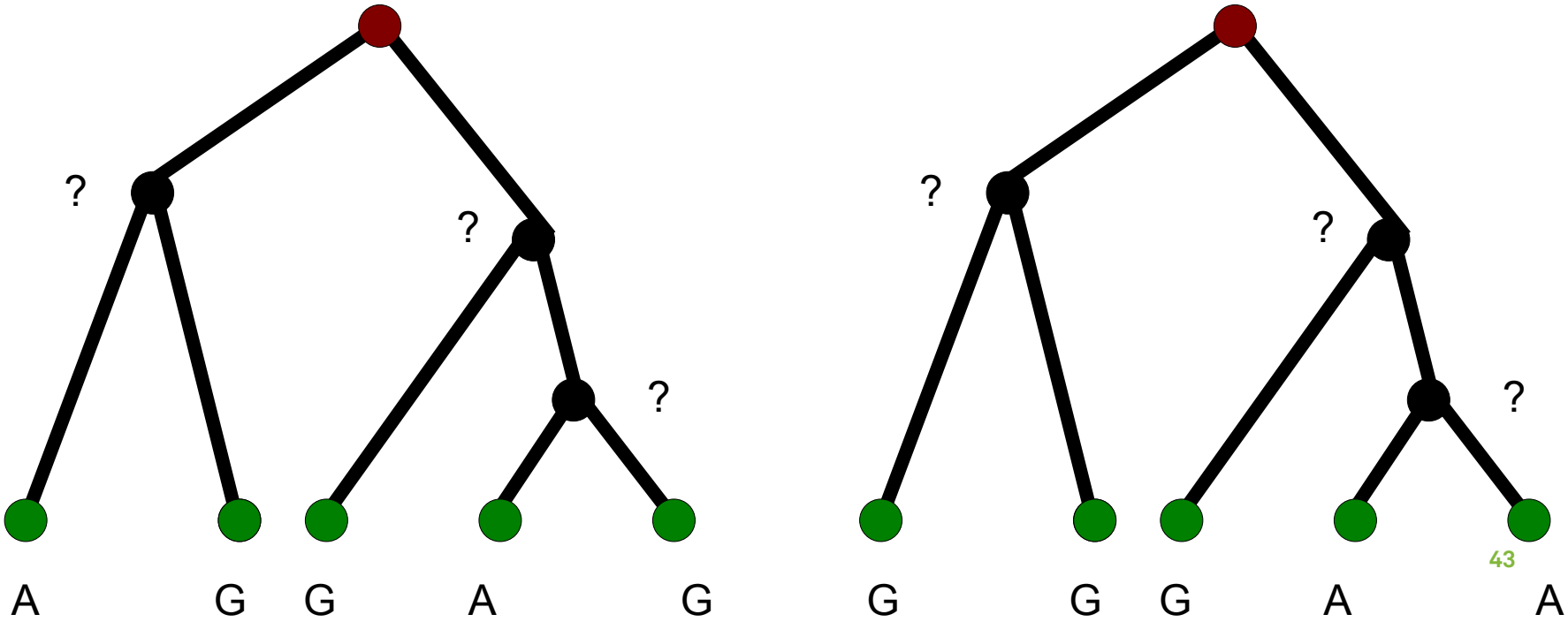
Achtung:
Weder Optimierungs-Methoden noch Clustering-Methoden
garantieren evolutionär korrekten Baum.

Ausgewählte Verfahren in der Übersicht



Einfaches Beispiel zu Maximum Parsimony und ML

Gegeben sind 5 Sequenzen: A, A, G, G, G
Zwei Baumtopologien:
Welche Sequenzen stehen jeweils an den inneren Knoten?
Welcher Baum ist besser?



Maximum Parsimony

Gesucht / Ausgabe:

Baumtopologie, Ahnen-Sequenzen in den inneren Knoten & Wurzel
Optimalitätskriterium: “sparsamster” Baum (“most parsimonious” tree):
über alle Kanten summiert am wenigsten Änderungen.

Vorteile

- Kriterium intuitiv, motiviert durch Occam's Razor:
  „die einfachste, kürzeste Erklärung ist die beste“.

Nachteile:

- Unterschätzt die wahre Anzahl von evolutionären Ereignissen
  bei entfernt verwandten Sequenzen (z.B. wegen Homoplasien)
- liefert falsche Ergebnisse
  bei weit entfernten Sequenzen und stark unterschiedlichen Raten
- NP-schweres Problem

Maximum Likelihood

Gesucht / Ausgabe:

Baumtopologie, Kantenlängen (in PAM-Einheiten o.ä.).
Optimalitätskriterium:
Zu gegebenem statistischem Evolutionsmodell + Baumtopologie + Kantenlängen berechnete Wahrscheinlichkeit, dass beobachtete Sequenzen auftreten.

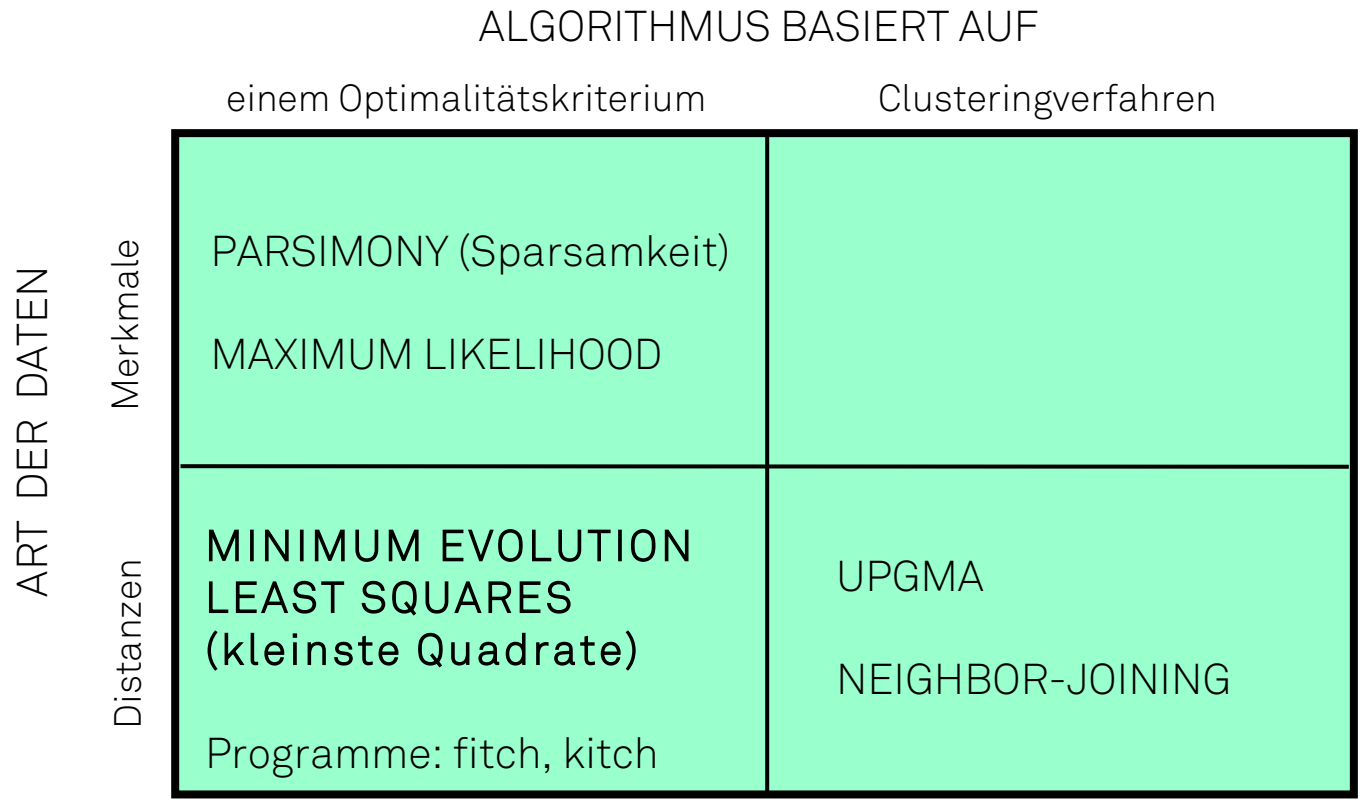
Vorteile

- Modellbasiert: Alle Annahmen werden explizit gemacht
- Mehrfach- und Rücksubstitutionen sind im Modell berücksichtigt
- Konsistente Schätzung evolutionärer Distanzen (Kantenlängen)

Nachteile:

- Ergebnisse Modell-abhängig; falsches Modell führt zu beliebigen Ergebnissen
- schwierig(er) zu verstehen (als MP)
- NP-schweres Problem

Ausgewählte Verfahren in der Übersicht



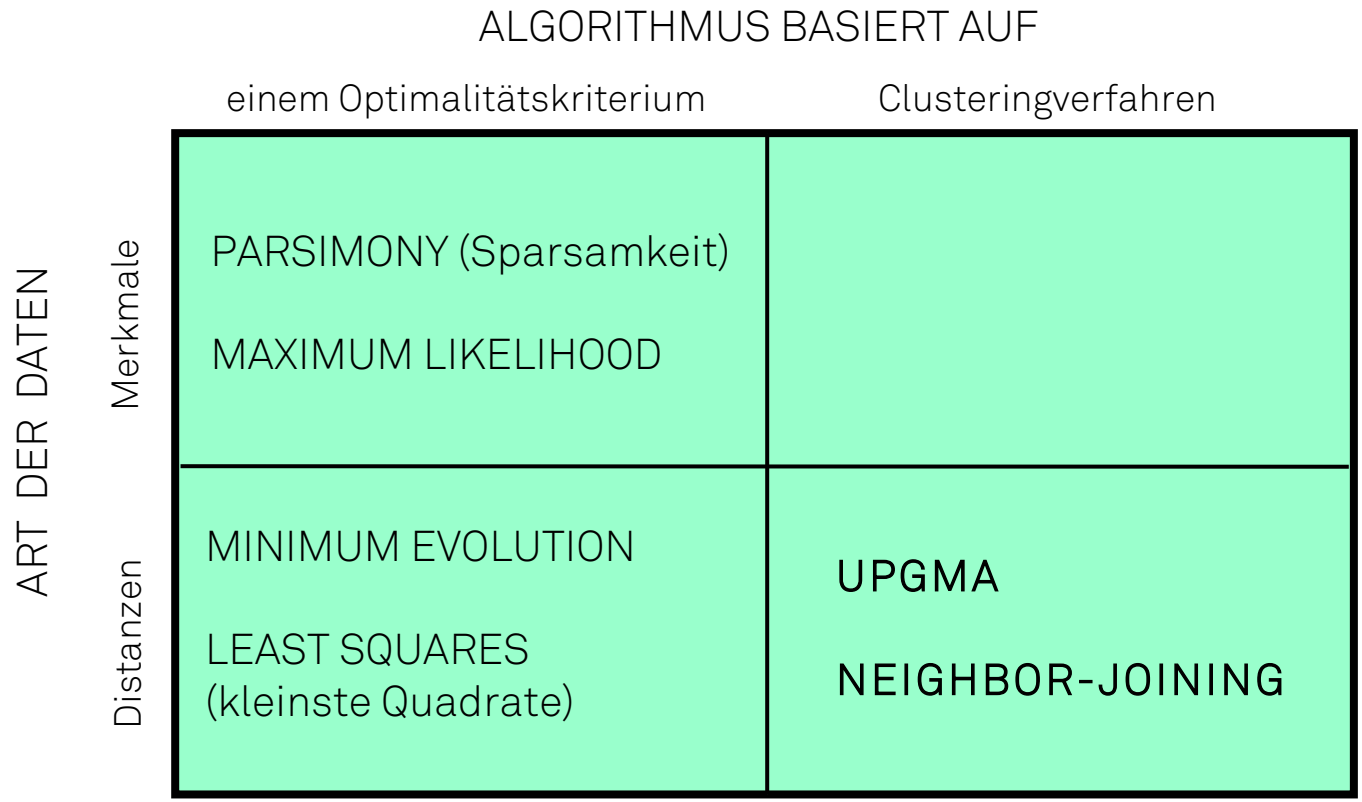
Least Squares und Minimum Evolution

Gesucht / Ausgabe: Baumtopologie, Kantenlängen
Optimalitätskriterium: Zu jeder Baumtopologie werden Kantenlängenbestimmt, so dass die Baum-Distanzen die gegebenen Distanzen optimal approximieren (im Sinne kleinster Quadrate).
Optimal ist die Baumtopologie
- mit kleinstem quadratischen Fehler (bei least squares)
- mit kleinster Gesamtlänge (bei minimum evolution)

Vorteile
- optimalitäts-basiert („Lösungen“ können vergleichend evaluiert werden)
- schneller als Merkmals-basierte Methoden

Nachteile:
- langsamer als Clustering-Methoden NJ und UPGMA

Ausgewählte Verfahren in der Übersicht



UPGMA: Unweighted Pair Group Method Using Averaging

Gesucht / Ausgabe: Baumtopologie, Kantenlängen
Verfahren:
Solange mehr als ein Objekt vorhanden ist:
  Finde das Paar (x,y) von Objekten mit kleinster Distanz
  Ersetze sie durch ein einziges Objekt z={x,y}
  Berechne Distanzen von z={x,y} zu den anderen Objekten a, ...,
    mittels Durchschnittbildung zwischen d(x,a) und d(y,a)
Baum ergibt sich aus der Hierarchie der zusammengefassten Objekte.

Vorteile

- einfach zu verstehen, einfach durchzuführen
- liefert korrektes Resultat bei ultrametrischen Distanzen,
  d.h. wenn es einen ultrametrischen Baum gibt, der die gegebenen Distanzen als Distanzen zwischen den Blättern aufweist, wird dieser (schnell) gefunden.

Nachteile

- Ergebnisse schlecht interpretierbar bei nicht ultrametrischen Eingaben

NJ: Neighbor Joining

Gesucht / Ausgabe: Baumtopologie, Kantenlängen

Verfahren:
Ähnlich wie bei UPGMA, aber die Wahl des zu aggregierenden Objektpaars ist sorgfältiger: Es werden alle Distanzen berücksichtigt, nicht nur die kleinste. Der Baum ergibt sich aus der Hierarchie der zusammengefassten Objekte.

Vorteile
- liefert korrektes Resultat bei additiven Distanzen,
  d.h. wenn es einen (ungewurzelten) Baum gibt, der die gegebenen Distanzen als Distanzen zwischen den Blätten aufweist, wird dieser (schnell) gefunden.
- viele Varianten mit unterschiedlichen Eigenschaften

Nachteile
- Ergebnisse schlecht interpretierbar bei nicht additiven Eingabe-Distanzen

Empfehlung

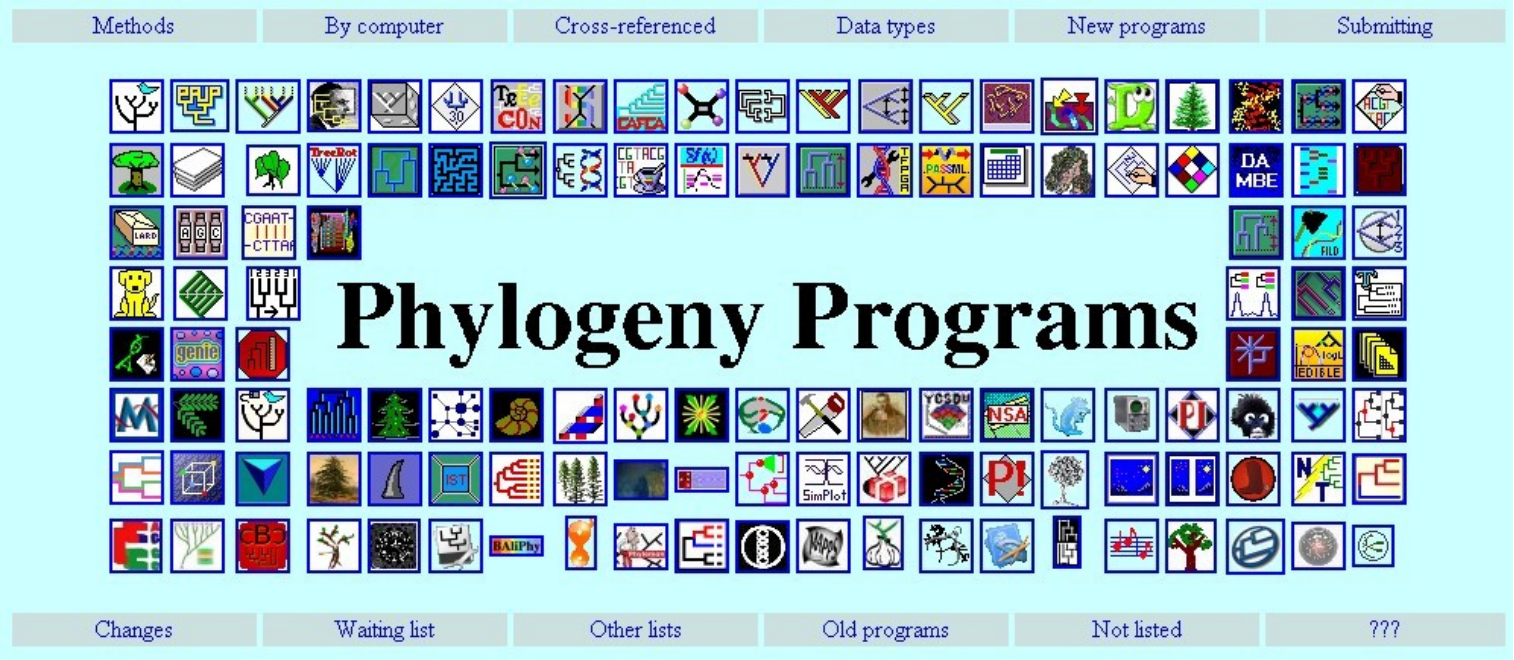
Distanzbasierte Clustering-Methoden
- schnell und gut, wenn das richtige Distanz-Modell verwendet wurde.
- Es empfiehlt sich, eine NJ-Variante zu benutzen, nicht UPGMA.

Probabilistische Methoden (ML, Bayes'sche Methoden)
- sind langsam (und NP-schwer),
- können bei guter Modellwahl bessere Ergebnisse liefern als NJ.
- können leicht Alternativen und relative Konfidenzwerte finden

Parsimony-Methoden
- unterschätzen den evolutionären Abstand (ok bei nah verwandten Sequenzen)
- keine Kantenlängen / Zeitschätzungen
- langsam und NP-schwer

Software

Umfangreiche Sammlung phylogenetischer Software:
http://evolution.genetics.washington.edu/phylip/software.html



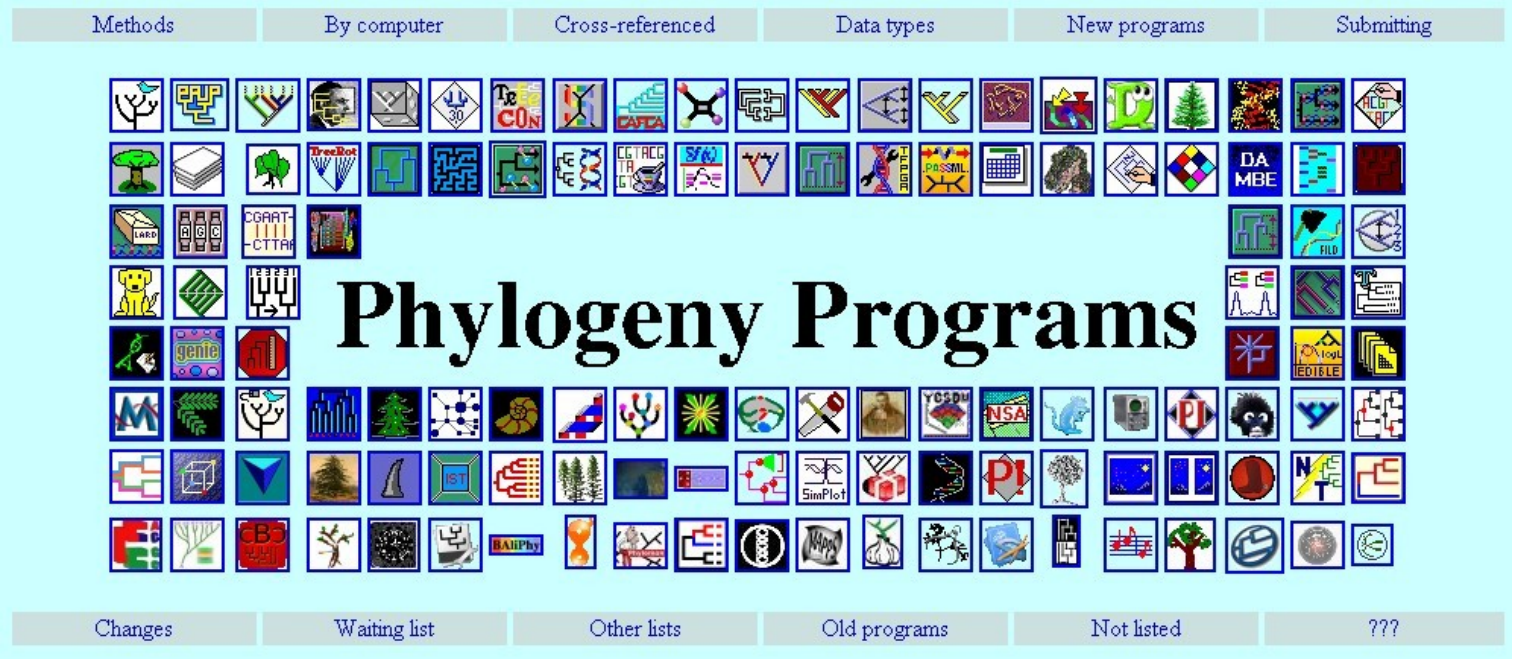
Software: Phylip

- häufig benutztes Paket zum Erstellen von Phylogenien
- enthält zahlreiche Programme für verschiedene Aufgaben

- Projekt-Homepage: http://evolution.genetics.washington.edu/phylip.html
- Web-Interface zu einer Teilmenge von Phylip + mehr: http://mobyle.pasteur.fr

Software

Umfangreiche Sammlung phylogenetischer Software:
http://evolution.genetics.washington.edu/phylip/software.html



Phylip:
das am häufigsten benutzte Paket zum Erstellen von Phylogenien,
enthält zahlreiche Einzelprogramme für verschiedene Aufgaben