Phylogenetische Inferenz

Peter Krupp

22. Juni 1998

1 Einleitung

Bei phylogentischer Inferenz handelt es sich um folgendes Problem: Gegeben eine Menge von heutigen (rezenten) biologischen Arten (Art = phylon). (Allgemeiner: Taxa = taxonomische Einheiten (Art, Gattung, Ordnung, Stamm etc.)) Man möchte wissen, wie diese Arten entstanden sind (Entstehung = Genese) oder genauer: wie ihr Verwandschaftsverhältnis ist, d.h. man möchte den gemeinsamen Stammbaum ermitteln.

Rezente Arten haben sich ja im Lauf der Evolution entwickelt aus Vorläuferarten, die heute ausgestorben sind. Eine neue Art entsteht, indem sich sich eine vorhandene Art aufspaltet. Aus irgendwelchen Gründen (z.B räumliche Trennung) entstehen aus einer Fortpflanzungsgemeinschaften zwei, die sich getrennt weiterentwickeln. Man spricht von einer Bifurkation (Aufspaltung, Gabelung). Der Stammbaum (im folgenden jetzt phylogenetischer Baum genannt) ist also ein binärer Baum, dessen Blätter den rezenten Arten, und dessen innere Knoten den ausgestorbenen Vorläuferarten entsprechen.

Im Allgemeinen weiß man nichts über diese Vorläuferarten. Das einzige was man zur Verfügung hat, sind experimentell gewonnene Daten die die rezenten Taxa beschreiben. Man kennt Merkmale z.B. morphologischer, physiologischer, ethologischer oder molekularer Art, anhand derer die betrachteten Arten charakterisiert werden können. Das kann man sehr wörtlich nehmen: Eine Taxon wird beschrieben duch einen String von Länge n, dessen einzele characters für eine bestimmte Ausprägung des entsprechenden Merkmals stehen. Diese Merkmale müssen *homolog* sein. Das heißt gerade, daß sie gemeinsamen Ursprungs sind. Dann kann man die Arten anhand dieser Strings vergleichen, und versuchen den phylogenetischen Baum zu rekonstruieren. Die Idee dabei ist natürlich, daß zwei ähnliche Arten wahrscheinlich eng verwandt sind, also im phylogenetischen Baum dicht beieinander liegen sollten. Dabei geht man davon aus, das man die Zeichen unabhängig voneinander betrachten kann (oft eine Vereinfachung).

Genau genommen handelt es sich bei dem beschriebenen Problem nicht um ein mathematisch definiertes Problem: Man will den Baum bestimmen, der den tatsächlichen historischen Ablauf widerspiegelt, das läßt sich mathematisch nicht überprüfen (In manchen Fällen läßt es sich biologisch bzw. paläontologisch überprüfen, aber im Allgemeinen überhaupt nicht). Mathematisch definiert wird das Problem erst dadurch, daß man es als Optimierungsproblem formuliert:

1. Definiere ein Maß dafür, wie gut ein phylogenetischer Baum mit den Daten der rezenten Taxa in Übereinstimmung steht (Optimalitätskriterium).

2. Suche den (bzgl. des Optimalitätskriteriums) optimalen phylogenetischen Baum.

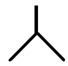
Der vorliegende Text folgt diesem zweistufigen Aufbau des Problems, indem zunächst vier verschiedene Ansätze beschrieben werden, wie man das Optimalitätskriterium definieren kann. Danach wird besprochen wie man Schritt 2 ausführen kann. Es wird sich herausstellen, daß man tatsächlich nicht viel mehr tun kann als "suchen", da das Problem NP-hart ist, wie im letzten Abschnitt gezeigt wird.

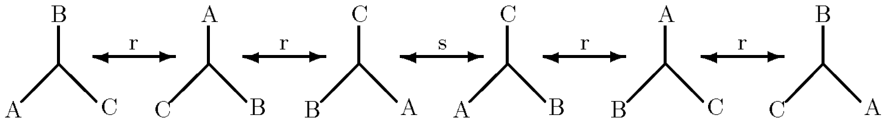
Inhaltsverzeichnis

2 Phylogenetische Bäume

Welche Art von Baum suchen wir? Ein phylogenetischer Baum, auch Kladogramm, Dendrogramm, oder Phylogenie genannt, ist ein gewurzelter oder ungewurzelter binärer Baum, die inneren Knoten haben Grad 3, die rezenten Taxa entsprechen den *Terminalen* (oder Blättern), die *inneren Knoten* den Vorläufertaxa. Jeder phylogenetische mit T Terminalen hat $T-2$ innere Knoten und $2T-3$ Kanten, davon T inzident mit einem Terminal.

Theorem 1. *Es gibt* $B(T) = \prod_{i=3}^{T}(2i-5) = calo(2^T)$ *verschiedene ungewurzelte binäre Bäume mit* T *Terminalen.*

Beweis. Es gibt genau einen phylogenetischen Baum mit drei Terminalen: Es gibt nur eine mögliche Topologie, nämlich . Es gibt $3! = 6$ Permutationen der Terminale, die alle durch Drehung und Spiegelung ineinander überführt werden können:



Es handelt sich also um denselben Baum.

Jeder phylogenetische Baum mit $T+1$ Terminalen kann konstruiert werden durch Einfügung eines neuen inneren Knotens, verbunden mit einem neuen Terminal, in eine Kante eines phylogenetischen Baums mit T Terminalen. Da ein phylogenetischer Baum mit T Terminalen $2T-3$ Kanten hat, gilt nach Induktionsvoraussetzung

$$B(T+1) = \left(\prod_{i=3}^{T}(2i-5)\right) * (2T-3) = \prod_{i=3}^{T+1}(2i-5)$$

□

Fügt man eine Wurzel hinzu, erhöht sich die Anzahl der inneren Knoten und die Anzahl der Kanten jeweils um 1. Da die Wurzel in jede der Kanten eingefügt werden kann, erhöht sich die Anzahl der Bäume um Faktor $2T-3$.

Der Baum den man sucht ist eigentlich ein Wurzelbaum, die Wurzel ist der gemeinsame Vorfahre aller Taxa, es ist aber mit den meisten Methode nicht möglich die Wurzel zu lokalisieren, deshalb beschränkt man sich oft auf ungewurzelte Bäume.

Multifurkierende Bäume haben innere Knoten mit Grad > 3. Solche phylogenetischen Bäume sind nicht *voll aufgelöst*, durch Einfügung weiterer innerer Knoten lassen sie sich auflösen zu streng bifurkierenden Bäumen.

3 Typen von Daten

Die rezenten Taxa werden beschrieben durch experimentell gewonnene Daten, die in zwei grobe Kategorien fallen:

- Character-Daten
  Ein Taxon wird durch N *characters* beschrieben. Wir wollen unterscheiden zwischen einem character als einer Variablen, und dem *charakter-state* den dieses charakter hat. Die Wertebereiche sind meistens diskret, manchmal aber auch kontinuierlich. Im diskreten Fall unterscheidet man zwischen binary und multistate characters. Binäre characters beschreiben typischerweise die Anwesenheit oder Abwesenheit einer Eigenschaft (z.B das Vorhandensein eines bestimmten Enzyms). Multistate characters können geordnet oder ungeordnet sein. Man sagt, die characters sind geordnet wenn eine transitive Relation R (keine Ordnungsrelation) existiert die angibt, welche character-Transformationen erlaubt sind: Eine Transformation eines characters von Zustand a nach Zustand b ist dann möglich wenn $a \mathbf{R} b$. Sequenzdaten sind ein Beispiel für ungeordnete multistate characters. Bei DNA z.B. kann ein Nuleotid jederzeit duch ein beliebiges anderes ausgetauscht werden. Man spricht von Polarität, wenn R nicht symmetrisch ist.

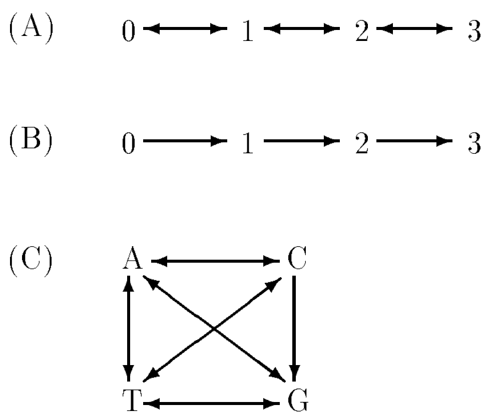


Abbildung 1: (A) Geordnete multistate characters ohne Polarität, (B) geordnete multistate charkters mit Polarität, (C) ungeordnete multistate characters

In diesem Text wollen wir uns auf molekulare Sequenzdaten beschränken, also nach der obigen Terminologie ungeordnete multistate charakters ohne Polarität. Sequnzdaten werden z.B gewonnen durch *multiples Alignment* von homologen DNA-Sequenzen. Homolog ist eine Menge von Sequenzen, wenn alle Sequenzen auf einen gemeinsamen Vorfahren zurückgehen. Wenn das nicht der Fall ist, gibt es auch keinen Baum den man aus diesen Daten vernünftigerweise ableiten könnte.

E s müßen aber nicht nur die Sequenzen homolog sein, sondern sogar jede einzelne Position aus diesen Sequenzen. Bereiche mit insertions oder deletions aus einem multiplen Alignment sind also nicht zu gebrauchen. Es werden nur die Positionen ausgewählt, die nur matches bzw. mismatches enthalten.

```
C...GGACAC---GTTTA...C
A...CGACAC---GT--G...C
C...GGACACAACGTTTA...C
G...A--AAC---GTTTA...A
C...GGACAC---GTTTC...C
```

Abbildung 2: Auswahl von homologen Positionen aus einem multiplen Alignment.

- Paarweise Distanzen bzw. Ähnlichkeiten
  DNA-DNA Hybridisierung liefert als einziges experimentelles Verfahren direkt paarweise Distanzen. Dieses Verfahren spielt aber in der Praxis keine Rolle. Meistens werden paarweise Distanzendaten durch paarweises Alignment von Sequenzen oder Ermittlung von Hamming-Abstand gewonnen.

4 Optimalitätskriterien

Wie schon gesagt, muß zunächst ein Kriterium definiert werden, nach dem man die Güte eines phylogentischen Baums beurteilen kann. In diesem Abschnitt werden verschiedene Kriterien vorgestellt.

4.1 Maximum Compatibility

Die maximum compatibility Methode arbeitet auf binären charakters mit Polarität. Eine Substitution ist nur von Zustand 0 zu Zustand 1 möglich. Ein charakter ist dabei typischerweise als Anwesenheit oder Abwesenheit einer Eigenschaft zu verstehen, z.B. das Vorhandensein bzw. Nicht-Vorhandensein einer bestimmten Restriktionstelle in einer Sequenz (Ansatzstelle für ein Restriktionsenzym, das eine Sequenz an dieser Stelle zerschneidet).

Gegeben sei ein gewurzelter binärer Baum T mit n terminalen Taxa jeweils beschrieben durch m *binäre* charakters. Die Wurzel sei mit (0,0,...,0) gelabelt. Dem i-ten character ist eine Kante e_i zugeordnet. Das i-te Zeichen eines Knotens v ist 1 g.d.w v in dem Teilbaum liegt, der über e_i erreicht wird. Hier werden implizit folgende Annahmen gemacht:

- Keine Konvergenz: Jedes Merkmal ist genau einmal entstanden ("uniquely derived").

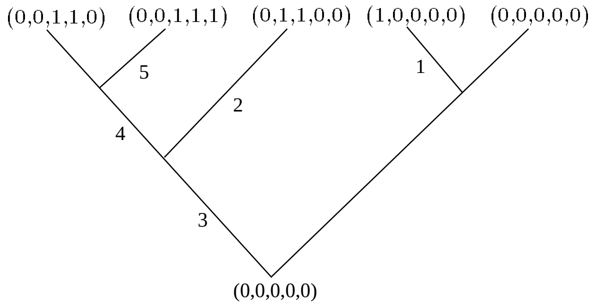


Abbildung 3: Hennig-Baum. Jedes Merkmal ist genau einmal entstanden, und bildet sich nicht wieder zurück

- Keine Reversionen: Es kommt nicht vor, daß ein einmal entstandenes Merkmal wieder verschwindet.

Das Beispiel in Abb. 3 ist so gewählt, daß der Baum mit allen characters in Übereinstimmung steht. Was ist wenn es einen solchen Baum nicht gibt? Die maximum compatibility Methode wählt die größte Teilmenge der characters aus, für die ein Hennig-Baum existiert. Das Maß für die Güte des Baums ist die Kardinalität dieser Teilmenge. Ihre Bestimung ist allerdings NP-hart (MAX CLIQUE).

4.2 Maximum Parsimony

Parsimony ist ein in der Naturwissenschaft weit verbreiterter Begriff der in etwa besagt:

- Einfachere Hypothesen sind komlizierteren vorzuziehen.
- *Ad hoc* Hypothesen sind zu vermeiden.

Im Zusammenhang mit phylogenetischer Inferenz bedeutet Parsimony: Die einfachste Erklärung für gemeinsame Merkmale von Taxa ist, daß diese von einem gemeinsamen Vorfahren geerbt wurden. Homoplasien (Konvergenzen, Paralleltät, Reversionen) sind die komplizierteren Hypothesen.

Den inneren Knoten des Baums wird explizit ein character state zugeordnet, und zwar so, daß die Anzahl der character-Übergänge minimal ist. Der so gelabelte Baum wird *most parsimonious reconstruction (MPR)* genannt (ist nicht eindeutig). Die Länge des Baums ist gleich der Anzahl der Übergänge, d.h. eine Kante hat die Länge 1 wenn ein Übergang stattfindet, sonst 0 (möglich ist auch ein Gewichtung nach character und Art des Übergangs (s.u.)) In mathematischer Ausdrucksweise kann man die Parsimony-Methode beschreiben wie folgt (Verallgemeinertes Parsimony). Aus der Menge aller möglichen phylogentischen Bäume finde alle Bäume τ

mit

$$L(\tau) = \sum_{k=1}^{E} \sum_{j=1}^{C} w_j * diff(x_{k',j}, x_{k'',j})$$

ist minimal. Es wird summiert über alle Kanten des Baums und alle characters. Dabei ist $L(\tau)$ die Länge des Baums, E die Anzahl der Kanten, C die Anzahl der characters, w_j sind Gewichte für die verschiedenen characters, k' und k'' sind die Knoten die inzident sind mit der Kante k, diff(y,z) ist eine Matrix mit den Kosten einer Transition von Zustand y zu Zustand z. In den folgenden Beispielen wird bei der Berechnung der Länge des Baums nur ein character zu Grunde gelegt. Aufgrund der Annahme daß die characters unabhängig sind, kann die Gesamtlänge durch Summierung über alle characters berechnet werden. Die Parsimony-Variante die jetzt vorgstellt wird ist bekannt als *Fitch Parsimony*. Sie ist anwendbar auf Daten mit ungeordneten multistate characters (z.B. Squenzdaten).

4.2.1 Algorithmus von Fitch

Der Algorithmus von Fitch findet in $O(n)$ Zeit eine *most-parsimonous reconstruction* (MPR) eines gegebenen Baums, d.h. der Algorithmus labelt die inneren Knoten so, daß daß die Anzahl der character-Übergänge L minimal ist. Es gibt es keine Gewichtung der Transitionen (Kosten für jede Transition = 1). Zunächst wird der Baum an einem beliebigen Terminal gewurzelt. Der Zustand S der Terminale entspricht ihrem character state, der Zustand der inneren Knoten ist zu Anfang undefiniert. t sei der innere Knoten der unmittelbarer Nachfolger der Wurzel r ist.

$L \leftarrow 0$
while Es gibt einen inneren Knoten der keinen Zustand hat **do**
  Wähle einen inneren Knoten k, mit $S_k = undef$ der zwei Sohne i und j hat mit $S_i \neq undef$, $S_j \neq undef$
  if $S_i \cap S_j \neq \emptyset$ **then**
    $S_k \leftarrow S_i \cap S_j$
  else
    $S_k \leftarrow S_i \cup S_j$
    $L \leftarrow L + 1$
  end if
end while
if $S_r \not\subseteq S_t$ **then**
  $L \leftarrow L + 1$
end if

An diesem Punkt hat der Algorithmus die Länge des Baums korrekt bestimmt, aber es ist noch kein MPR bestimmt worden. Dazu sind folgende weiteren Schritte erforderlich:

while Es gibt einen inneren Knoten, dem noch kein character state zugeordnet wurde **do**
  Wähle einen inneren Knoten k, der noch keinen character state x_k hat, aber dessen Vaterknoten m ein character state x_m hat
  if $x_m \in S_k$ **then**
    $x_k \leftarrow x_m$
  else
    Wähle x_k zufällig aus S_k
  end if
end while

Es ist klar, daß der Algorithmus Laufzeit $O(n)$ hat, da jeder innere Knoten in den beiden while-Schleifen genau einmal besucht wird (*post-order* Traversierung) und der Rumpf der Schleifen in konstanter Zeit durchlaufen wird.

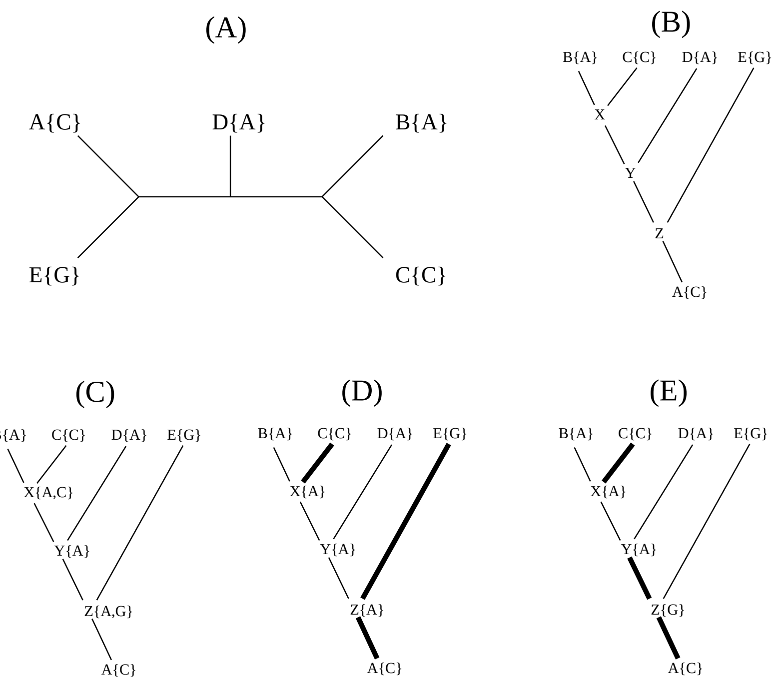


Abbildung 4: Beispiel-Durchlauf des Fitch-Algorithmus. (A) Der ungewurzelte Baum mit character states. (B) Baum mit Wurzel A. (C) Baum nach Phase 1 des Algorithmus. Den inneren Knoten wurden Mengen von character state zugeordnet. (D) Ein MPR nach Phase 2. (E) Alternatives MPR. Fett gezeichnete Kanten entsprechen character-Übergängen.

4.2.2 Parsimony und Inkonsistenz

Die Fähigkeit einer Schätzmethode mit zunehmender Datenmenge gegen den wahren Wert zu konvergieren, wird Konsistenz genannt. Man sagt dann auch, das die Schätzung frei von systematischen Fehlern ist. Es ist erhellend zu fragen, unter welchen Annahmen Parsimony ein *inkonsitenter* Schätzer ist, also einen falschen Baum favorisiert. Dazu kann es kommen, wenn man von Szenarios ausgeht, in denen bestimmte evolutionäre Distanzen vorausgesetzt werden. Ein Beispiel ist in Abb. 5 gegeben.

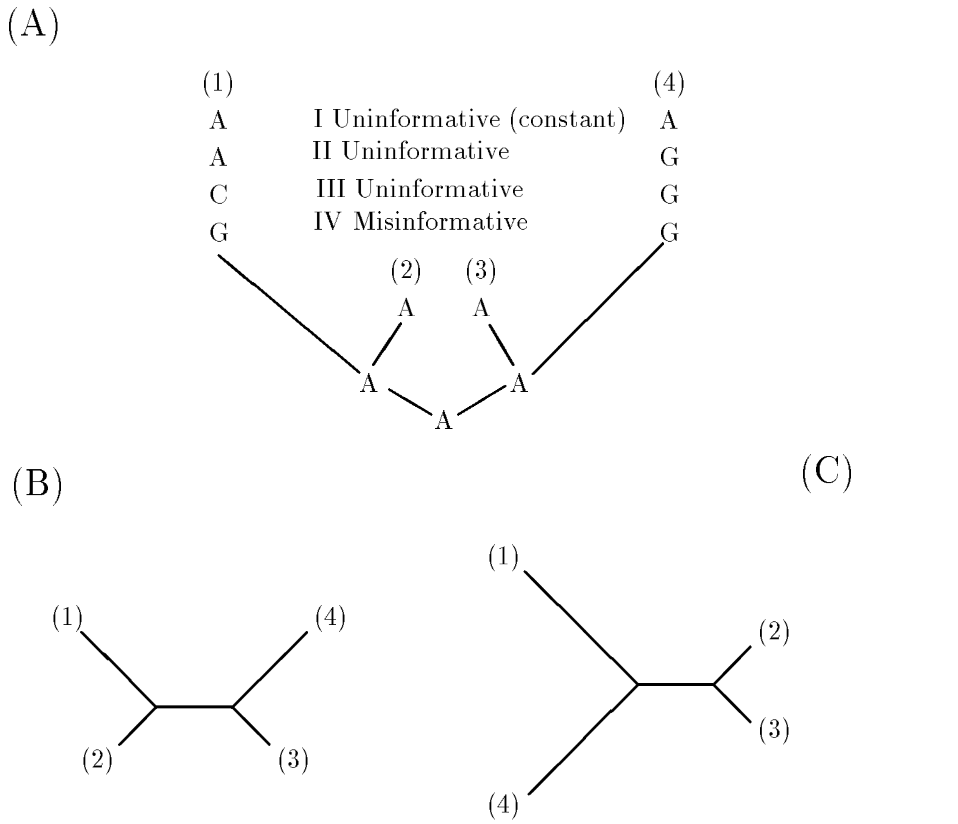


Abbildung 5: Beispiel für die potentielle Inkonsistenz von Parsimony. (A) Baum für vier Taxa mit zwei langen peripheren Kanten, alle anderen Kanten sind sehr kurz. (B) Entsprechender ungewurzelter Baum. (C) Falscher Baum der von Parsimony favorisiert wird.

(A) zeigt einen hypothetischen phylogentischen Baum für die Taxa (1), (2), (3), (4). Es sei vorausgesetzt, daß die peripheren Kanten die inzident mit (1) bzw. (2) sind, lang sind, daß also die evolutionäre Distanz groß ist (die evolutionäre Distanz ist das Produkt von Zeit und Evolutionsrate), alle anderen Kanten seien sehr kurz. Die charakters I bis III favorisieren keinen der beiden Bäume (B) und (C), sie sind *uninformativ.* Charakter IV favorisiert den falschen Baum (C). Parsimony würde also Baum (C) besser bewerten als den wahren Baum (B). Patterns die den wahren Baum unterstützen, sind solche mit einer Substitution entlang der kurzen, zentralen Kante. Diese sind aber unwahrscheinlich (wegen der geringen evolutionäre Distanz). Mißinformative patterns sind dagegen ziemlich wahrscheinlich. Aus diesen Grund wird die Wahrscheinlichkeit, daß Parsimony den falschen Baum auswählt mit zunehmender Datenmenge immer größer. Das ist mit Inkonsistenz gemeint. Man beachte daß, das Problem dain liegt, daß außer den charakters der Taxa noch bestimmte Kantenlängen=evolutionäre Distanzen angenommen werden, wir aber keine Möglichkeit haben, diese in unserem Modell zu repräsentieren. Es gibt keine Möglichkeit evolutionäre Distanzen in Parsimony als Input hineinzustecken. Evolutionäre Distanz ist für Parsimony gleich der Anzahl der character-Substitutionen entlang einer Kante. Je größer aber die evolutionäre Distanz ist, desto mehr Substitutionen werden von nachfolgenden Substituitionen üüberdeckt (gerade bei Sequenzdaten), und Parsimony wird die evolutinäre Distanz desto mehr unterschätzen.

Es gibt Verfahren, die die Daten entsprechend der a priori angenommenen Distanzen korrrigieren, bevor sie in Parsimony hereingesteckt werden. Das ist eine Möglichkeit diesem Problem zu begegnen. Eine andere ist, Methoden anzuwenden, die auf einem expliziten Modell des evolutionären Wandels beruhen, und die mit evolutionären Distanzen besser umgehen können. Eine solche Methode wird im nächsten Abschnitt vorgestellt.

4.3 Maximum Likelihood

Maximum Likelihood (ML) ist eine Methode die bei der Berwertung eines phylogentischen Baums ein explizites (stochastisches) Modell der Evolution zugrunde legt. Im Fall von DNA-Daten z.B. basiert das Modell auf einer Matrix von Austauschratenraten zwischen den verschiedenen Basen. ML versucht die Wahrscheinlichkeit der beobachteten Daten (die characters der Terminale) zu berechnen, unter der Hypothese daß eine bestimmte Baumtopologie die Evolution zutreffend beschreibt. Dabei wird den inneren Knoten nicht, wie bei Parsimony, ein bestimmter Zustand zugeordnet. Vielmehr werden alle Szenarios, d. h. möglichen Labelungen der inneren Knoten berücksichtigt. Natürlich sind manche Szenarios viel plausibler als andere, aber jedes Szenario besitzt eine Wahrscheinlichkeit > 0, die zur Likelihood des Baums beiträgt. Informell funktioniert ML so:

Gegeben: Ein (stochastisches) Modell der Evolution.

Gegeben: Die beobachteten Daten (characters der Terminale).

Hypothese: Eine mögliche Baumtopologie und *Längen der Kanten.*

Berechne: Die Wahrscheinlichkeit der beobachteten Daten unter der Hypothese.

Finde: Die Hypothese die zur höchsten Wahrscheinlichkeit führt.

Maximum-Likeklihood Schätzer Maximum-Likelihood Schätzung ist ein Standardverfahren aus der Statistik, daß eingesetzt werden kann um phylogentische Bäume auf der Basis von evolutionären Modellen und evolutionären Dsitanzen zu bewerten.

Ein *Schätzproblem* ist gegeben durch

- eine nichtleere höchtens abzählbare Menge Ξ, der Stichprobenraum.
- eine Familie$\{P_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf Ξ
- eine zu schätzende Funktion $g(\theta)$

In unserem Fall sind die beobachteten Eigenschaften x der terminalen Taxa ein Element aus Ξ. Θ ist die Menge der möglichen Bäume (Topologie und Kantenlängen). Jeder Baum aus Θ stellt einen Parameter für ein Wahrscheinlichkeitsmaß dar, d.h. durch einen bestimmten Baum wird ein Wahrscheinlichkeitsmaß festgelegt, anhand dessen wir die Wahrscheinlichkeit der beobachteten Eigenschaften x berechnen können. Die Funktion g wird in diesem Fall nicht gebraucht (identische Abbildung: $g(\theta) = \theta$. Die Funktion $L_x(\delta) := P_\delta(x)$ nennen wir die *Likelihood-Funktion.* Wenn L_x einen Maximalwert in $\hat{\delta}(x)$ annimmt, dann nennen wir $\hat{\delta}(x)$ einen *Maximum-Likelihood Schätzer* von δ. $\hat{\delta}$ ist also ein Baum, der die Wahrscheinlichkeit unserer Beobachtungen x maximiert.

Modelle für die Evolution von Sequenzen Wie wird nun die Wahrscheinlichkeit eines Szenarios berechnet? Normalerweise werden *Markov-Modelle* benutzt um Sequenzevolution zu modellieren. In einem Markov-Modell ist die Wahrscheinlichkeit einer Zustandsänderung nur abhängig vom momentanen Zustand, nicht aber von den früheren Zuständen. Die Wahrscheinlichkeit, daß an eine Stelle in einer Sequenz ein Austausch von Nukleotid A gege en Nukleotid B erfolgt, ist also nur von A abhängig (und der evolutionären Distanz), aber nicht von füheren Austäuschen. Die Wahrscheinlichkeiten werden berechnet auf der Grundlage einer Matrix von Austauschraten. Im Fall von DNA-Sequenzen ist das eine 4×4-Matrix. Wenn die Matrix symmetrisch ist, dann ist das Modell *Zeit-reversibel*, d.h. es spielt keine Rolle, in welcher Richtung die Substitutionen ablaufen.

Ein spezielles Modell ist das *Jukes-Cantor-Modell.* Es ist gegeben durch die Matrix von Substitutionsraten

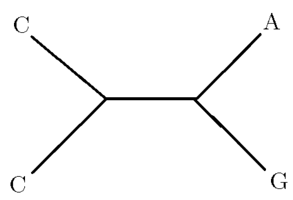
$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

(A)

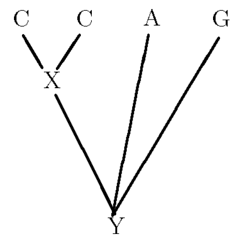
```
      1                 j                N

(1)   C ... G G A C A | C | G T T T A ... C
(2)   C ... A G A C A | C | C T C T A ... C
(3)   C ... G G A T A | A | G T T A A ... C
(4)   C ... G G A T A | G | C C T A G ... C
```

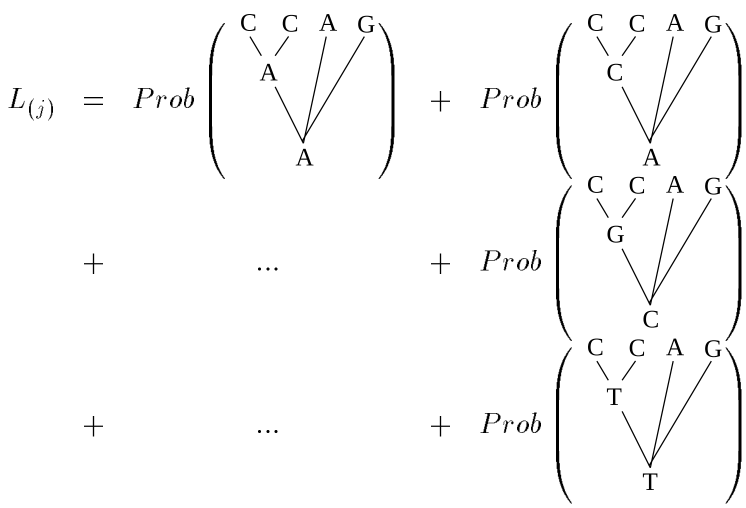
(B)



(C)



(D)



(E) $L = L_{(1)} * L_{(2)} * ... * L_{(N)} = \prod_{j=1}^{N} L_{(j)}$

(F) $\ln L = \ln L_{(1)} + \ln L_{(2)} + ... + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$

Abbildung 6: Überblick über die Maximum Likelihood Methode. (A) characters für 4 Taxa (multiples Alignment). (B) Hypothetischer Baum T. (C) Baum C an einem inneren Knoten gewurzelt. (D) Die Likelihood für ein character ist die Summe der Wahrscheinlichkeiten aller möglichen Labelungen der inneren Knoten. (E) Die Gesamt-Likelihood ist das Produkt der Likelihoods aller characters. (F) Da L sehr klein werden kann, arbeitet man üblicherweise lieber mit der log-Likelihood.

Das Jukes-Cantor Modell geht von gleichen Substitiutionsraten für alle Paare von Nukleotiden aus. In anderen Modelle werden unterschiedlich Raten verwendet, zum Beispiel höhere Raten für Transitionen (Purin $\leftrightarrow$ Purin, bzw. Pyrimidin $\leftrightarrow$ Pyrimidin), gegenüber Transversionen (Purin $\leftrightarrow$ Pyrimidin).

Aus den Substititutionsraten der Matrix Q werden folgendermaßen Wahrscheinlichkeiten berechnet: Die Austauschwahrscheinlichkeiten entlang einer Kante der Länge t ist gegeben durch die Matrix

$$P(t) = e^{Qt}$$

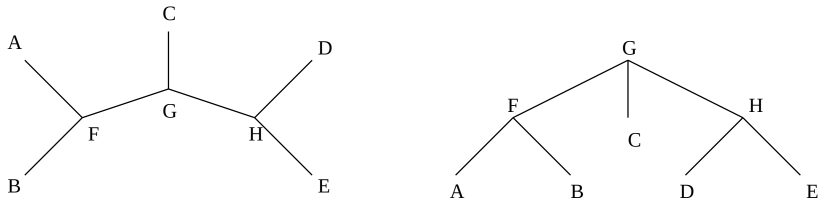
Für das Jukes-Cantor Modell ist $P(t)$ gleich

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

Berechnung der Likelihood eines Baums Bei der Berechnung der Likelihood müssen alle Kombinationen von Zuständen der inneren Knoten des Baums berücksichtigt werden. Zunächst wird der Baum an irgendeinem inneren Knoten gewurzelt. (Die Likelihood ist bei Zeit-reversiblen Modellen unabhängig davon welcher Knoten als Wurzel gewählt wird). Wenn A jetzt eine Sequenz im Baum ist mit unmittelbaren Nachfolgern B und C, dann ist die *bedingte Likelihood* von character state i auf Position j in A gleich

$$L(x_{Aj} = i) = \left(\sum_k P_{ik}(v_{AB}) L(x_{Bj} = k) \right) * \left(\sum_l P_{il}(v_{AC}) L(x_{Cj} = l) \right)$$

dabei ist v_{xy} die Länge der Kante die Sequenz X und Sequenz Y verbindet. $L(x_{aj} = i)$ ist die Likelihood des Baumes unter A, unter der Bedingung daß $x_{Aj} = i$, deshalb *bedingte* Likelihood. Betrachten wir den ersten Term auf der rechten Seite. $P_{ik}(v_{AB})$ ist die Wahrscheinlichkeit, daß entlang der Kante AB aus dem k an Position j in Sequenz B (dem linken Sohn von A), das i in A geworden ist. $L(x_{Bj} = k)$ ist die bedingte Likelihood für charakter k auf Position j in B (wird rekursiv berechnet). Das Produkt $P_{ik}(v_{AB}) * L(x_{Bj} = k)$ wird summiert über alle möglichen charakters k. Der zweite Term berechnet das analoge für den rechten Sohn C von A. Die Likelihood eines gesamten Baums für eine Position j erhält man, indem man die bedingte Likelihood $L(x_{Aj} = i)$ multipliziert mit dem Frequenzparameter π_i und summiert über alle i. Abb. 7 zeigt ein Beispiel. Das Produkt aller Positions-spezifischen Likelihoods ist die Gesamt-Likelihood eines Baums.



$$
\begin{aligned}
L_{(j)} &= \sum_m \pi_m \\
&\times \left[\sum_k P_{m,k}(v_{FG}) P_{k,x_{Aj}}(v_{AF}) P_{k,x_{Bj}}(v_{BF})\right] \\
&\times \left[\sum_l P_{m,l}(v_{GH}) P_{l,x_{Dj}}(v_{DH}) P_{l,x_{Ej}}(v_{EH})\right] \\
&\times P_{m,x_{Cj}}(v_{CG})
\end{aligned}
$$

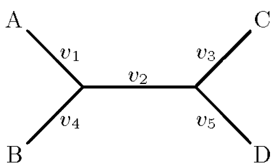
Abbildung 7: (A) Phylogenetischer Baum für 5 bekannte Sequenzen A,B,C,D,E. (B) Knoten G wurde als Wurzel gewählt. (C) Berechnung der Likelihood des Baums für Position j.

4.4 Distanzmethoden

Distanzmethoden arbeiten mit paarweisen Distanzdaten, die meistens durch paarweisen Vergleich von Sequenzen durch Alignment oder Ermittlung von Hamming-Abstand gewonnen werden. Bei der Anwendung einer Distanzmethode für N Taxa geht man also aus von eine $N \times N$ Matrix von paarweisen Distanzen, sowie einem phylogenetischen Baum. Man untersucht, wie gut der Baum in übeinstimmung mit der Distanzmatrix steht.

4.4.1 Additive Distanzen

Wenn die Distanzdaten die wahre evolutionäre Distanz der betrachteten Taxa wiederspiegeln, haben sie die Eigenschaft der *Baum-Additivität.* Das bedeutet, die Distanz jedes Paars (A,B) von Taxa ist gleich der Summe der Distanzen, die auf dem Weg von A nach B im Baum auftreten.



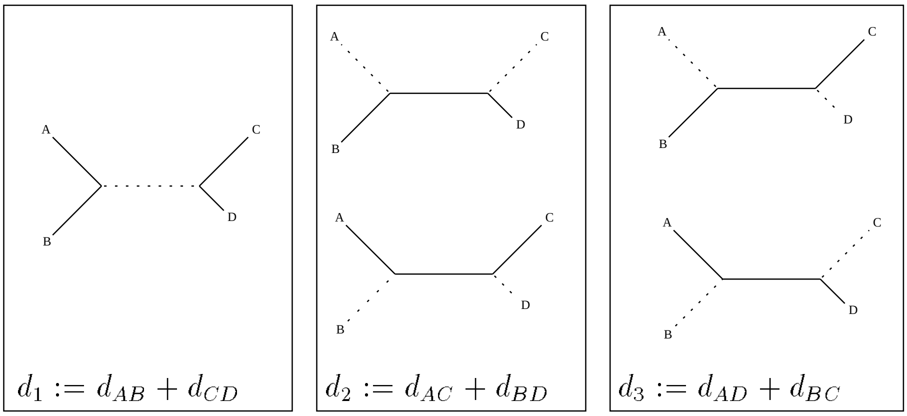
	B	C	D
A	$v_1 + v_4$	$v_1 + v_2 + v_3$	$v_1 + v_2 + v_5$
B		$v_4 + v_2 + v_3$	$v_4 + v_2 + v_5$
C			$v_3 + v_5$

Abbildung 8: Additiver Distanzen. Die Distanzen zweier Taxa im Baum sind gleich den paarweisen Distanzen.

Additive Distanzen genügen der *four-point metric condition*: Für beliebige Taxa A,B,C,D gilt:

$$d_{AB} + d_{CD} \leq max(d_{AC}, +d_{BD}, d_{AD} + d_{BC})$$

Anschaulich bedeutet diese Ungleichung, daß von den drei Summen $d_{ij} + d_{kl}, (i \neq j \neq k \neq l)$ mindestens zwei gleich sind, und die dritte kleiner oder gleich diese beiden ist.



$d_1 := d_{AB} + d_{CD}$ $d_2 := d_{AC} + d_{BD}$ $d_3 := d_{AD} + d_{BC}$

Dies folgt aus den Eigenschaften binärer Bäume. Wie man sieht, gehen in d_2 und d_3 dieselben Kanten ein, und die Kanten die in d_1 eingehen, gehen auch in d_2 und d_3 ein.

Empirische Daten sind aber nie perfekt, sie enthalten Fehler. Schon aus diesem Grund werden die Distanzen nie hundertprozentig additiv sein. Wenn der Baum, den man untersucht, Ähnlichkeit mit der wahren Phylogenie der Taxa hat, sollten die Daten aber annähernd additiv sein.

Methode von Fitch-Margoliash Diese Methode summiert die Abweichungen zwischen paarweisen Distanzen und Distanzen im Baum Über alle Paare von Taxa.

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_{ij} |d_{ij} - p_{ij}|^{\alpha}$$

mit T ist die Anzahl der Taxa, w_{ij} ist ein Gewicht für die Taxa i und j, d_{ij} ist die Distanz von Taxa i und j, p_{ij} ist die Länge des Pfades zwischen i und j im Baum, α ist 1 oder 2. Durch niedrige Gewichte w_{ij} für Paare, deren empirische Distanzen als besonders ungenau oder fehlerhaft angesehen werden, erreicht man das unsichere Daten das Ergebnis nicht zu stark beinflussen. Wenn man die Daten allgemein als ziemlich fehlerhaft einschätzt, ohne ds auf bestimmte Paare eingrenzen zu können, wählt man $\alpha = 1$, damit große fehlerhafte Abweichungen das Ergebnis nicht zu sehr verfälschen, sonst $\alpha = 2$. Folgende Schemata fúr die Gewichtung sind gebräuchlich:

1. $w_{ij} = 1$
2. $w_{ij} = 1/d_{ij}$
3. $w_{ij} = 1/d_{ij}^2$
4. $w_{ij} = 1/\sigma_{ij}^2$ (σ_{ij}^2 erwartetete Varianz von d_{ij})

4.4.2 Ultrametrische Distanzen

Daten die genau die evolutionäre Distanz der Taxa wiedergeben, sollten nicht nur der *four point condition* genügen, sondern auch der stärkeren *three point condition* : Für beliebige Taxa A,B,C gilt:

$$d_{AC} \leq max(d_{AB}, d_{BC})$$

Diese Ungleichung besagt, daß von den Distanzen zwischen drei Taxa immer zwei gleich groß sind, und die dritte kleiner oder gleich diesen ist. Das ist der Fall, wenn die Distanzen additiv sind, *und* der Baum so gewurzelt werden kann, daß alle Taxa den gleichen Abstand zur Wurzel haben. Das entspricht

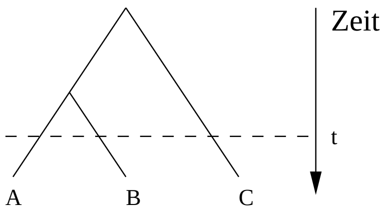


Abbildung 9: Ultrametrischer Baum mit 3 Taxa

dem Konzept der molekularen Uhr, das besagt, daß die Änderungsrate zu jedem Zeitpunkt t in allen Zweigen des Baums gleich ist (das bedeutet nicht unbedingt, daß die Änderungsrate in der Zeit gleich bleibt).

Aufgrund der Ungenauigkeit der Daten, ist es jedoch sehr unwahrscheinlich wirklich ultrametrische Distanzen zu haben.

5 Suche nach dem optimalen Baum

Im vorigen Kapitel wurden Methoden beschrieben, die bewerten, wie gut eine bestimmte Baum-Topologie zu den gegebenen Daten der rezenten Taxa passt. Es wurde nichts darüber gesagt, wie man einen "guten" oder sogar optimalen Baum findet. Darum geht es in diesem Abschnitt. Die Bewertungsmethoden sind glücklicherweise recht billig ($O(n)$ bei Parsimony), doch leider sieht es der Baumsuche nicht so gut aus. Alle Formalisierungen des Problems "Phylogenetsche Inferenz" führen auf NP-harte Probleme, wie hier am Beispiel von Parsimony dargestellt werden soll. Das bekannte "MINIMAL STEINER TREE"-Problem läßt sich auf die Parsimony-Fassung unseres Problems reduzieren. Das bedeutet, phylogenetische Inferenz ist NP-hart. Es ist also zwecklos, nach einem polynomiellen Algorithmus zu suchen. Was uns übrigbleibt ist erschöpfendes Durchsuchen aller Baumtopologien (exakt) sowie die Anwendungen von Heuristiken (nicht exakt). Exakte Suche ist möglich für kleine Eingabgrößen von bis zu 20 Taxa. Manche Optimalitätskriterien erlauben dabei die Eingrenzung des Suchraums durch *branch & bound*. Für größere Eingabegrößen bleibt nur die Anwendung von Heuristiken. Dabei wendet man meistens *greedy* Methoden zur Konstruktion einer anfänglichen Baums, der dann auf verschiedene Weise modifiziert wird. Zunächst soll jetzt das Steiner Baum Problem vorgestellt werden, dann wird erläutert, wieso die Suche nach einem optimalen Baum unter einem Parsimony-Kriterium ein Steiner-Baum Problem ist.

5.1 Das Steiner-Baum Problem

Das Steiner-Baum Problem in Netzwerken ist eine kombinatorische Version des Euklidischen Steiner-Baum Problems. Es kann folgendermaßen formuliert werden:

- **GEGEBEN:** Ein ungerichtetes Netzwerk $G = (V, E, c)$ mit Kantengewichtung $c : V \to \mathcal{R}$ und einer nicht-leeren Menge N, $N \subseteq V$, von Terminalen.
- **FINDE:** Ein Teilnetzwerk $T_G(N)$ von G, so daß
  - es gibt einen Pfad zwischen jedem Paar von Terminalen.
  - die totale Länge $|T_G(N)| = \Sigma_{e_l \in T_G(N)} c(e_l)$ ist minimal.

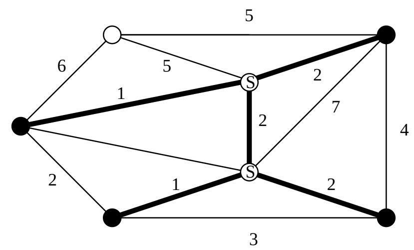


Abbildung 10: Minimaler Steiner Baum. Terminale schwarz, Nicht-Terminale weiß, S Steiner-Punkte, Kanten des Steiner Baums fett.

5.2 Parsimony als Steiner-Baum Problem

Das Problem Pylogenetische Inferenz läßt sich in der Wagner-Parsimony Formulierung in ein Steiner-Baum Problem überführen. Entscheidend ist dabei, daß bei Parsimony den inneren Knoten, also den Vorläuferarten, Zustände zugeordnet werden. Dadurch erhält man einen Koordinatenraum, in dem die möglichen Steinerpunkte liegen. Konkret sieht das so aus: Gegeben ein Menge T von N Taxa, jeweils beschrieben durch m charakters $(c_1, c_2, ..., c_m) \in C_1 \times C_2 \times ... \times C_m$. Die C_i sind die Wertebereiche der einzelnen charakters (Wenn die charakters z.B. homologe Stellen aus einem multiplen Alignment von DNA-Sequenzen sind, dann ist $C_i = \{A, C, T, G\}$ für $1 \leq i \leq m$). Die zugehörige Instanz des Steinerbaum-Problems ist dann:

Der Graph $G = (V, E, c)$ ist der multidimensionale rectilineare Graph $V = C_1 \times C_2 \times ... \times C_m$. Es gibt eine Kante zwischen zwei Knoten g.d.w. sie sie sich nur einem charakter unterscheiden. Die Menge der N Taxa ist die Menge der Terminale. Die Länge der Kanten ist 1. Gesucht ist ein minimaler Steiner-Baum für T mit folgenden Eigenschaften

1. Die Blätter des Baumes sind die Terminale aus T.
2. Der Grad der Knoten ist beschränkt durch 3.

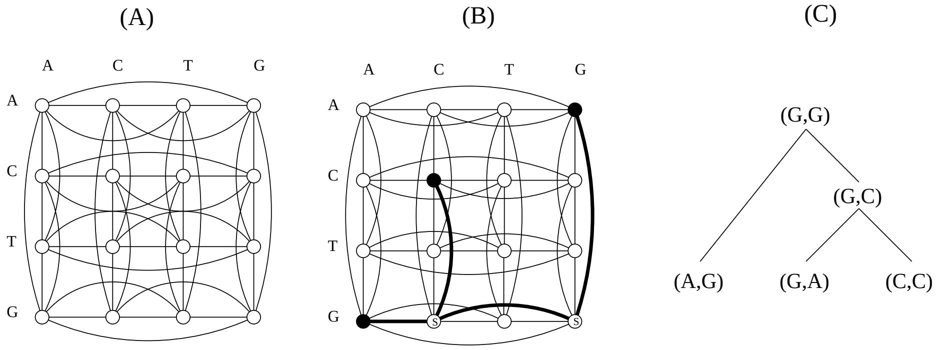


Abbildung 11: Parsimony als Steiner-Baum Problem. (A) 2-dimensionaler rectilinearer Graph für C={A,C,T,G}. Alle "Reihen" und "Spalten" sind Cliquen. (B),(C) Minimaler Steinerbaum für die schwarzen Terminale

5.3 Exakte Suche

Da das Problem, einen optimalen Baum unter einem gegebenen Optimalitätskriterium zu finden NP-hart ist, ist es aussichtslos, nach einem polynomiellen Algorithmus zu suchen. Wollen wir also eine exakte Lösung finden, müssen wir exponentiellen Aufwand in Kauf nehmen. Dann kann eine optimale Lösung mittels erschöpfender Suche gefunden werden. Es müßen also alle möglichen Bäume erzeugt und bewertet werden.

Alle möglichen phylogenetischen Bäume für einen Menge von Taxa können folgendermaßen erzeugt werden:

- Wähle 3 Taxa und konstruiere den einzigen möglichen Baum mit 3 Taxa.

- Wähle ein weiteres Taxon aus, und erzeuge drei Bäume mit vier Taxa, in dem dieses in jede Kante eingefügt wird.

- Fahre fort Taxa auf diese Weise in die Bäume einzufügen, bis alle Taxa enthalten sind.

Die Anzahl der so erzeugten Bäume ist für N Taxa gleich $\prod_{i=3}^{N} 2i - 5$ also exponentiell in N. Dies ist etwa bis N=11 "tractable" (34.459.425 Bäume).

Wenn ein Optimalitätskriterium benutzt wird, das nicht-abnehmend ist (Die Größe die minimiert werden soll (= Länge des Baums) wird nicht kleiner, wenn neue Taxa in den Baum eingefügt werden), kann man die Größe des Suchbaums durch branch & bound reduzieren. Das ist bei Parsimony, Maximum Likelihood und Fitch-Margoliash der Fall. Der Suchbaum wird dabei depth-first durchlaufen. Zunächt wird einen obere Schranke für die Länge des optimalen Baums durch Bewertung eines zufällig gewählten Baums berechnet. Da das Optimalitätskriteriun nicht-abnehmend ist, ist die Länge des Baums an jedem Knoten V im Suchbaum eine untere Schranke S_v für alle Knoten im Teilbaum unter V. Wenn wir an Knoten V also eine obere Schranke S_o kennen mit $S_o \leq S_v$ braucht der Teilbaum unter V nicht mehr durchsucht zu werden. Man kann also zum Vaterknoten von V "backtracken" und dort mit der nächsten Alternative fortfahren. Wird ein Blatt des Suchbaums erreicht (ein phylogenetscher Baum T für alle N Taxa), dessen Länge kürzer ist als die momentane obere Schranke, so nehen wir diese als neue obere Schranke. T ist dann der Kandidat für den optimalen Baum. Wenn der Baum komplett durchlaufen ist, ist der Kandidat gleich dem optimalen Baum.

Auf diese Weise kann der Suchbaum in günstigen Fällen stark beschnitten werden, so daß der optimale Baum für 20 oder mehr Taxa gefunden werden kann. Im worst-case allerdings bringt branch & bound keine Verbesserung gegenüber erschöpfender Suche.

5.4 Nicht-Exakte Suche

In vielen Fällen ist die Anwendung exakter Algorithmen nicht mehr tractable. Dann hilft nur noch die Anwendung von Optimierungsmethoden wie *hill climbing* oder *simulated annealing*. Prinzipiell geht man dabei so vor:

1. Konstruiere *greedy* einen Baum von dem am annehmen kann, daß er nahe am Optimum legt (s.u.).

2. Versuche den Baum lokal zu verändern (s.u.), so daß der score besser wird.

3. Führe Schritt 2 so lange durch, bis keinen Verbesserung mehr gefunden werden kann.

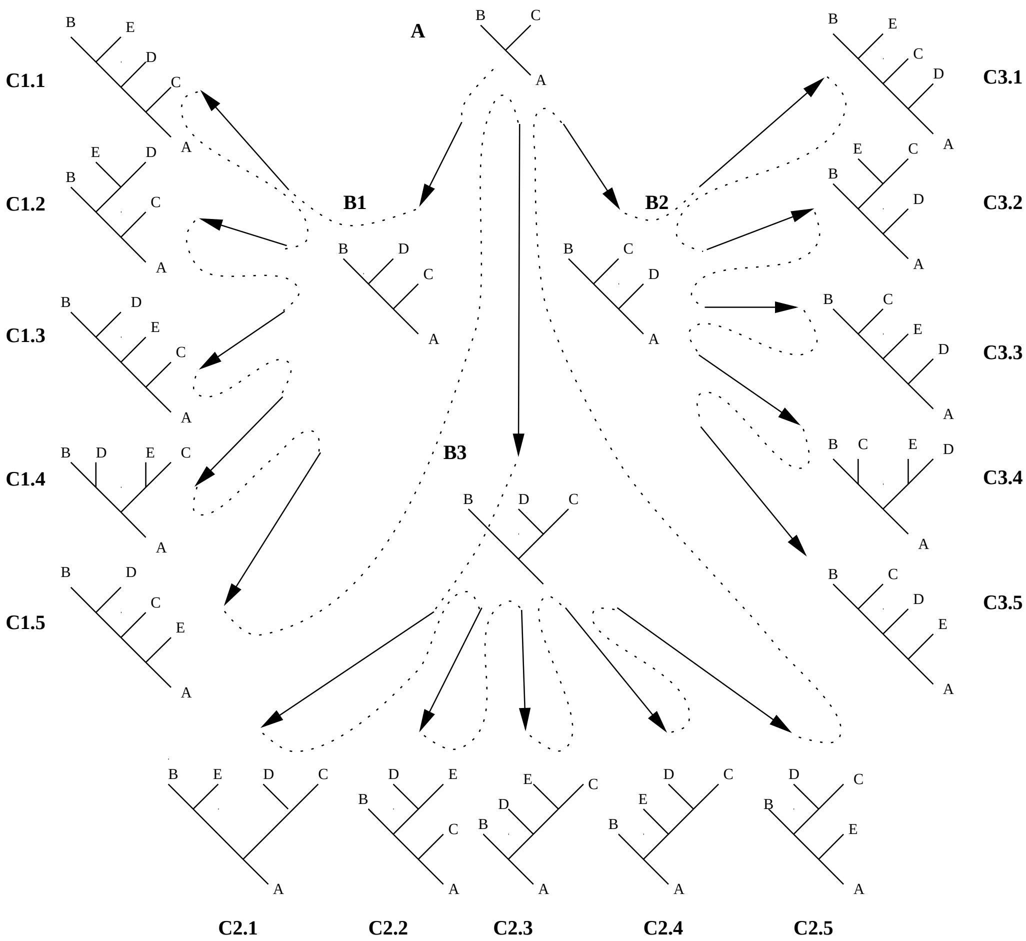


Abbildung 12: Depth-first Durchlauf eines Suchbaums für fünf Taxa. Bei Einsatz von branch&bound wird jeder Knoten bewertet, auch die Inneren. Ist bei Erreichen eines Knoten T eine obere Schranke S für die Länge des optimalen Baums bekannt mit $S \leq L(T)$, braucht der Teilbaum unter T nicht mehr durchsucht zu werden.

Das Problem ist, daß man nie weiß, ob man ein *globales* Optimum oder lediglich ein *lokales* gefunden hat. Man kann sich die score-Funktion wie ein Gebirge vorstellt, dessen höchsten Gipfel man finden will. Man beginnt an irgendeinem Punkt im diesem Gebirge. Wendet man einfaches *hill climbing* an, bewegt man sich von diesem Punkt aus immer bergauf, bis man einen Gipfel erreicht hat. Doch möglicherweise gibt es im Gebirge höhere Gipfel. Simulated Annealing läßt dagegen auch Schritte die zu einer momentanen Verschlechterung des scores führen, mit einer gewissen Wahrscheinlichkeit zu. Dadurch läuft man nicht zwangläufig auf das nächste lokale Optimum zu.

5.4.1 Greedy Methoden zur Konstruktion von Bäumen

Diese Methoden dienen in der Regel dazu, einen guten Startpunkt für die nachfolgende Optimierung zu finden.

Stepwise Addition Bei dieser Methode wird Taxon für Taxon greedy in den wachsenden Baum eingefügt. Man geginnt mit 3 Taxa und dem einzigen möglichen Baum. Dann wird ein vierte Taxon ausgewäht. Von den drei Möglichkeiten, dieses in den Baum einzufügen (in eine der Kanten) , wird diejenige mit dem besten score ausgewählt. So fährt man fort, bis alle Taxa eingefügt sind.

Star Decomposition Möglich für alle Kriterien, die anwendbar sind auf multifurkierende Bäume. Ausgangspunkt ist ein *star tree*, in dem alle terminalen Taxa mit einem inneren Knoten verbunden sind. Das Optimalitätskriterium wird angewendet auf alle Bäume, die enstehen durch Zusammenfassung von zwei Terminalen mittels Hinzufügung eines neuen inneren Knotens. Von diesen Bäumen wird der mit dem besten score ausgewählt. Dieses Vorgehen wird fortgesetzt, bis der Baum vollständig aufgelöst ist, d.h. binär ist.

Cluster Analysis Cluster Analysis ist eine Heuristik die ausgehend von paarweisen Distanzdaten einen Baum konstuiert. Wenn die Daten *ultrametrisch* sind, dann ist der konstuierte Baum sogar optimal. Die Methode ist simpel:

Gegegen sei eine Matrix $d_{i,j}$ mit paarweise Distanzen von N Taxa. Zu Anfang sei eine Menge M von N *clustern* entspechend den N Taxa gegeben

1. Finde das Paar von clustern (i,j) mit der minimalen Distanz $d_{i,j}$
2. Fasse i und j zu einem neuen *cluster* u zusammen.
3. Definiere die Distanz von u zu jedem anderen cluster v als den Durchschnitt $d_{u,v} := (d_{i,v} + d_{j,v})/2$

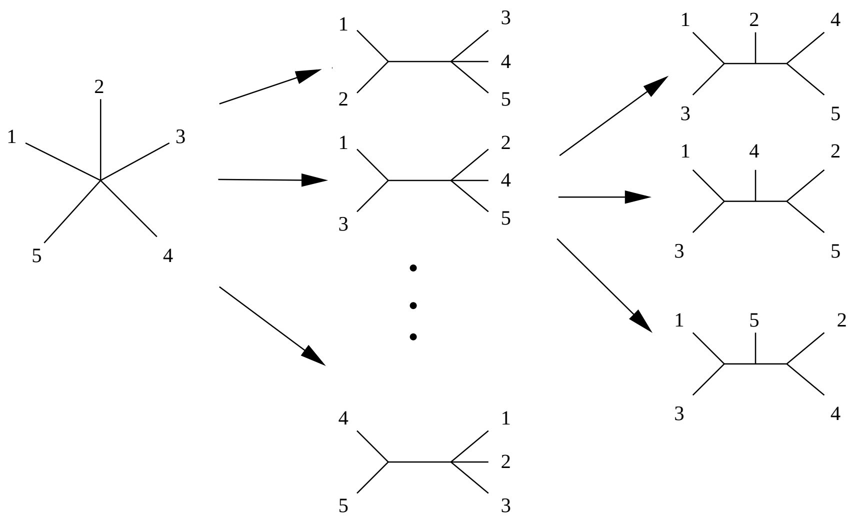


Abbildung 13: Star Decomposition Heuristik. In jedem Schritt wird das Optimalitätskriterium angewendet auf jeden Baum der entseht durch Zusammenfassung von zwei Kanten, die vom zentralen Knoten wegführen, durch Einfügung eines neuen Knotens. Der beste Baum ist Ausgangspunkt fúr den nächsten Schritt.

4. Entferne i und j aus M und füge u hinzu.

5. Falls M noch mehr als ein cluster enthält fahre fort bei Schritt 1.

Das Ergebnis ist ein ultrametrischer Baum T, die Cluster und Subcluster entsprechen auf naheliegende Weise Teilbäumen vo T.

5.4.2 Branch Swapping

Unter der Bezeichnung branch swapping werden verschiedene Methoden zusammengefaßt, die lokale Veränderungen an Bäumen vornehmen. Diese Veränderungen entsprechen den Schritten im "Gebirge" der score-Funktion während der Optimierung. In jedem Schritt wird der Baum durch Anwendung einer Methode auf jede mögliche Weise lokal modifiziert. Die Modifikationen werden bewertet, und falls Verbesserungen des scores auftreten, wird die beste Modifikation als Ausgangspunkt für den nächsten Optimierungsschritt genommen.

Nearest neighbor interchange ist ein Beispiel für eine branch swapping Methode.

6 Zusammenfassung

Ausgehend von heutigen biologischen Taxa möchte man den Auschnitt der Evolution rekonstuieren der zu diesen Taxa führt, man möchte also den phylogenetischen Baum bestimmen, dessen Blätter die heutigen Taxa, und

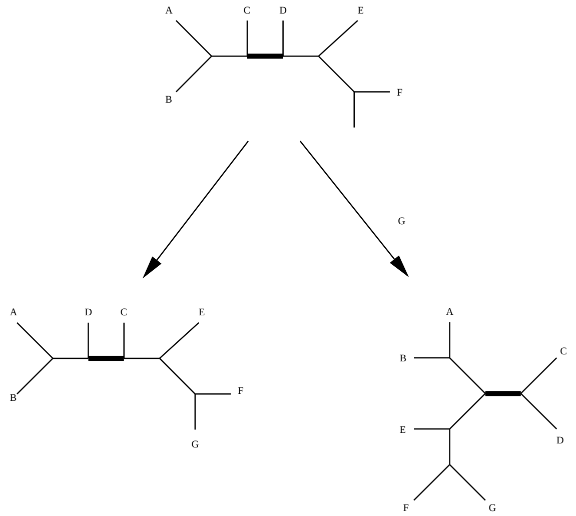


Abbildung 14: Branch Swapping durch *nearest neighbor interchange*. Jede innere Kante eines Baums definiert vier Teilbäume, die mit der Kante verbunden sind. Es gibt zwei Möglichkeiten diese Teilbäume umzugruppieren.

dessen innere Knoten die heute ausgestorbenen Vorläufertaxa sind. Das Problem gliedert sich in zwei Stufen: (1) Lege ein Kriterium fest, nach dem beurteilt wird, wie gut ein bestimmter Baum den Daten entspricht. (2) Finde den optimalen Baum, oder zumindest einen "guten" Baum.

Wir haben verschiedene Optimalitätskriterien kennengelernt: Maximum Compatibility, Maximum Parsimony, Maximum Likelihood, Additive-Baum Methoden. Alle diese Kriterien können einen gegebenen Baum effizient bewerten, Parsimony z.B in $O(n)$ Zeit, ML ist aufwendiger. Die entscheidende Frage ist hier die nach den Annahmen die die verschiedenen Kriterien implizit oder explizit über die biologische Evolution machen und wie gut diese Annahmen in jedem speziellen Fall gerechtfertigt sind.

Die Suche nach einem optimalen Baum ist NP-hart, wie mit Hilfe des Steiner-Baum Problems gezeigt wurde. Deshalb ist es nur für relativ kleine Instanzen des Problems (ca. 20 Taxa) möglich, eine exakte Lösung zu finden. Bei größeren Instanzen ist man auf heuristische Optimierungsmethoden angewiesen.

Literatur

[1] Hwang, Richards, Winter. The Steiner Tree Problem. Annals of Discrete Mathematics 53

[2] Swofford, Olsen, Waddel, Hillis.

[3] Krengel. Einführung in die Wahrscheinlichkeitstheorie und Statistik. Vieweg 1991