Phylogenetische Bäume

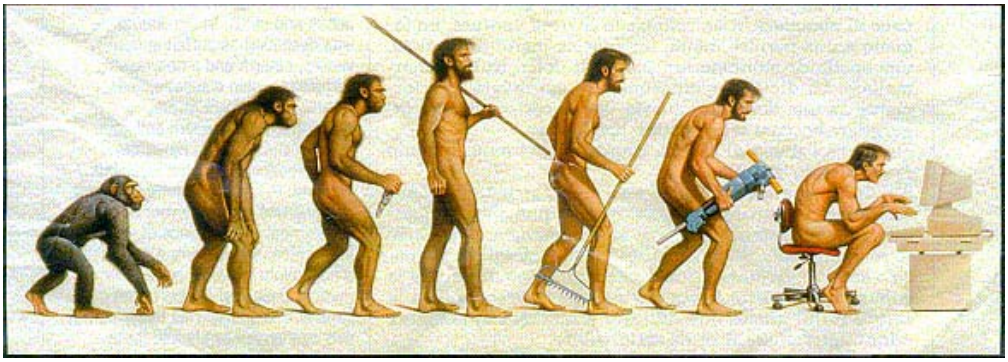
Evolution und Mathematik

Seminararbeit zur Vorlesung Mathematik/Informatik B

2. Semester
Molekulare Biotechnologie
Universität Heidelberg

von

Annegret Kramer · Luisa Schubert · Jonas Schaefer · Lorenz Steinbock · Frank Holtrup



„Nothing in biology makes sense except in the light of evolution.“

Theodosius Dobzhansky

Inhaltverzeichnis

Seite

1. Biologische Grundlagen

1.1 Evolution und Phylogenie

Unter Evolution versteht man Aufbau und Veränderung von Programmen in der Generationenfolge, die zu Anpassung und Vielfalt führen.
Der Begriff Phylogenese (Phylogenie, Stammesgeschichte) beschreibt den Prozess, durch den aus einer gemeinsamen Stammart durch fortlaufende Artaufspaltungen neue Artengruppen entstehen. Evolution ist also Voraussetzung für Phylogenese.
Phylogenetik meint die Wissenschaft von der Rekonstruktion stammesgeschichtlicher Entfaltung. Dazu liefert heutzutage die Analyse des Genotyps (genetische Ausstattung eines Organismus) die Grundlage.

1.2 Darwins Evolutionstheorie

Der Naturforscher Charles R. Darwin begründete bereits im 19. Jahrhundert die Evolutionstheorie, die mit heutigen molekulargenetischen Erkenntnissen modifiziert und ergänzt die moderne Evolutionstheorie bildet.

1835 nahm Darwin an einer Expedition teil, die ihn unter anderem auf die Galapagos-Inseln etwa 1000 km vor der Westküste Südamerikas führte. Dort beobachtete er bei den einzig dort vorkommenden rund 80 verschiedenen Finkenarten unterschiedlichste Schnabelformen und Lebensweisen, die der Nahrung und dem Lebensraum der jeweiligen Finkenart angepasst waren. Darwin fragte sich, welchen Einfluss die Isolierung von Arten auf Inseln auf die Spezialisierung haben könnte.
1859 erschien Darwins maßgebliches Werk „Die Entstehung der Arten durch natürliche Zuchtwahl“, indem er einerseits die Evolution als historisches Ereignis und andererseits die natürliche Selektion als Mechanismus der Evolution manifestierte.
Unter Ersterem verstand er, dass sämtliche Organismen der Erde in einem universellen Stammbaum (Tree of life) mit einem einzigen Ursprung dargestellt werden können. Mit Letzterem ist eine naturbedingte Auslese der bestangepassten Individuen einer Art (Survival of the fittest) gemeint. Dadurch, dass nur (bzw. überwiegend) diese Individuen zur Fortpflanzung kommen, sind es ihre Gene, die den Genbestand (Genpool) folgender Generationen ausmachen.

Darwins Schlussfolgerungen können folgendermaßen zusammengefasst werden: Alle Lebewesen haben die Fähigkeit zu exponentiellem Wachstum, d.h. sie vermehren sich stärker, als es zum Fortbestand der Population nötig ist. Trotz dieser Überproduktion an Nachkommen bleibt die Population stabil, da die Ressourcen (Nahrung, Lebensraum, etc.) begrenzt sind. Unter den Individuen einer Art findet somit ein Wettbewerb um die beste Reproduktionsfähigkeit statt (Struggle for life). Des Weiteren stellte Darwin fest, dass sogar die Nachkommen von ein und demselben Elternpaar sich stets in irgendeiner Weise unterscheiden. Er postulierte, dass jedes Individuum einzigartig ist, da „Varietäten“ auftreten. Durch natürliche Auslese über viele Generationen hinweg überleben nur die Bestangepassten, man spricht von Evolution.

1.2.1 Das Entstehen genetischer Variation: Rekombination und Mutation

Rekombination ist ein wichtiger Vorgang, der zur genetischen Vielfalt führt, und wird deswegen als Evolutionsfaktor eingestuft. Bei den Keimzellen (Eizellen und Spermien) kommt es zur Neuverteilung der homologen (ähnliche, je eins von Mutter und eins von Vater) Chromosomen der diploiden Urkeimzelle (diploid = doppelt vorliegender Chromosomensatz). Dies bezeichnet man als interchromosomale Rekombination. Die Zahl der Kombinationsmöglichkeiten der Chromosomen bzw. die Zahl möglicher unterschiedlicher Keimzellen beträgt bei n Chromosomenpaaren 2^n. Außerdem findet bei der Verschmelzung von Ei- und Samenzelle eine Neuverteilung der Gene statt („Neukombination“).

Mit *Mutationen* bezeichnet man Veränderungen des Erbgutes. Mutationen können einzelne Gene oder ganze Chromosomen betreffen. Genetiker gehen davon aus, dass Mutationen zufällig und ungerichtet auftreten. Des Weiteren werden Mutanten mit vorteilsbringenden Mutationen von der natürlichen Selektion begünstigt, solche mit nachteiligen Mutationen sind dagegen weniger lebensfähig. Die spontane Mutationsrate eines Gens pro Generation ist zwar mit 10^{-6} sehr gering, da aber ein eukaryotischer Organismus etwa 10^4-10^6 Gene besitzt, nimmt man an, dass 10- 40 Prozent aller Keimzellen des Menschen ein mutiertes Gen tragen. Mutationen verändern den Genpool einer Population und sorgen für genetische Variation, so dass man sie ebenfalls als Evolutionsfaktor bezeichnet.

Rekombination ist allerdings wichtiger für die Entstehung neuer Genotypen als Mutation: Entfielen alle Mutationen, entstünden durch Rekombination noch über hunderte von Generationen hinweg ständig neue Genotypen.

1.2.2 Auswahl aus der Vielfalt: Selektion

Mutation und Rekombination schaffen genetische Variation, die *Selektion* gibt dem Evoluitionsprozess eine Richtung. *Selektion* versteht man als statistischen Prozess, bei dem es weniger um das Überleben eines einzelnen Individuums, als vielmehr darum geht, welchen Beitrag es zum Genpool der Folgegeneration leistet. Die Selektion (natürliche Auslese) wird durch Umweltfaktoren, die Individuen mit unterschiedlichen Merkmalen entsprechend zurückdrängen oder begünstigen (negativer bzw. positiver Selektionsdruck), bedingt.

1.2.3 Der natürliche Fortgang: Isolation

Werden mehrere Individuen einer Population durch geographische Isolation (z.B. ein Wirbelsturm, der Tiere auf die Galapagos-Inseln führt) vom Rest der Population isoliert, können sie ihre Gene nicht mehr ungestört austauschen (der Genfluss ist unterbrochen). Es bilden sich zunächst neue Rassen oder Unterarten. Die Ansammlung weiterer Mutationen führt in einem lang andauernden Evolutionsprozess zur Bildung einer neuen Art.

1.3 Der Weg zum Stammbaum

Um phylogenetische Bäume rekonstruieren zu können, benötigt man Datenmaterial, das die stammesgeschichtliche Entwicklung der Lebewesen widerspiegelt. Also müssen die Spuren des Entwicklungsprozesses in diesen Daten erkennbar sein. Früher verwendete man zu diesem Zweck hauptsächlich morphologische Daten, d.h. äußere typische Erscheinungsbilder der Organismen (Phänotypen). Seit etwa 15 Jahren nun ist man zunehmend dazu übergegangen, molekulargenetische Sequenzdaten von DNA und Proteinen als Datenquelle zu nutzen.

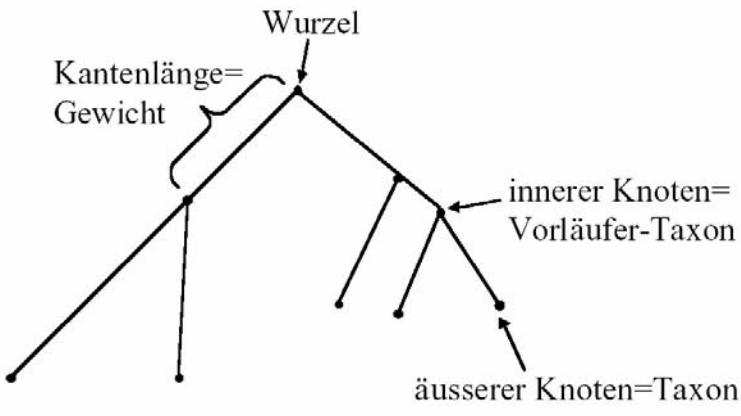
Jedes Gen besteht aus einer einzigartigen Abfolge von Bausteinen, die in der Regel die Anleitung zum Bau eines bestimmten Proteins enthält. Aminosäuren werden durch Basentripletts codiert. Aufgrund des degenerierten genetischen Codes kann

eine Aminosäure durch mehrere verschiedene Basentripletts codiert werden. Da der genetische Code also redundant ist, wandelt eine Mutation nur manchmal das entsprechende Protein ab. Daraus entsteht das Problem, dass auf der Ebene der Aminosäuresequenz Mutationen in der DNA-Sequenz verdeckt bleiben können (stille Mutationen). Auch auf Ebene der DNA können Mutationen durch multiple Substitutionen an einer Position unerkannt bleiben. So ist die tatsächliche Anzahl akkumulierter Gensequenzunterschiede heutzutage noch nicht feststellbar.
Diejenigen Mutationen, die die Funktion des Proteins nicht beeinträchtigen oder gar verbessern, sammeln sich im Laufe der Evolution im Erbgut an. Wenn also zwei Arten einen gemeinsamen Vorfahren besitzen, weichen ihre Gensequenzen, die sie gemeinsam haben, voneinander ab, man spricht von Sequenzdivergenz. Bestimmt man folglich die Sequenzdivergenz von Genen oder Proteinen der zu betrachtenden Organismen, so kann man deren Phylogenese rekonstruieren.
Der Vorteil molekulargenetischer Sequenzdaten gegenüber morphologischen Merkmalen besteht darin, dass jede Base einer Gensequenz bzw. jede Aminosäure eines Proteins einzeln betrachtet und somit ein größerer Merkmalsumfang bzw. eine feinere Auflösung der Untersuchungsergebnisse erzielt werden kann.

2. Mathematische Algorithmen zum Erstellen phylogenetischer Bäume

Zunächst sollen einige Termini zum Verständnis erklärt werden.

2.1 Terminologie



Die Abbildung illustriert die Terminologie, die bei der Beschreibung phylogenetischer Bäume verwendet wird.
Die phylogenetische Analyse beschränkt sich im Allgemeinen auf binäre Bäume. Unter binären Bäumen versteht man kreislose Graphen, bei der jeder innerer Knoten drei Kanten besitzt. Diese inneren

Knoten repräsentieren ausgestorbene Vertreter, die häufig als HTU – hypothetical taxonomic unit – bezeichnet werden. Blätter dagegen, auch OTU – operational taxonomic unit – genannt, repräsentieren die Sequenzen bzw. die Organismen, für die entsprechende Daten vorhanden sind. Der Vorläufer aller Sequenzen, die den Baum bilden, ist die Wurzel des Baumes. Oft ist jedoch dieser Vorfahre bzw. dessen DNA-Sequenz nicht bekannt, so dass diese Bäume wurzellos sind. In diesem Fall repräsentiert der aufgestellte Baum die Verwandtschaftsverhältnisse aller Taxa untereinander, nicht aber die evolutionären Wege. Phylogenetische Bäume repräsentieren immer Hypothesen, so dass die Wurzel des Baumes als hypothetischer Vorgänger zu sehen ist.

Ein gewurzelter Baum mit n Blättern hat genau 2n – 1 Knoten und 2n – 2 Kanten. Existiert keine Wurzel so sind 2n – 2 Knoten und entsprechend 2n – 3 Kanten zu zählen. Ein Problem bei der Rekonstruktion von phylogenetischen Bäumen ist die große Anzahl der möglichen Bäume. Für einen wurzellosen Baum mit n Blättern gibt es $\prod_{k=3}^{n}(2k-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$ mögliche Kombinationen. Somit sind für 7 Spezies sogar bereits 945, für 10 Spezies schon mehr als 2.000.000 mögliche Bäume erstellbar.

Prinzipiell lässt sich die phylogenetische Analyse in vier verschiedene Schritte einteilen. Zuerst wird eine Menge an Objekten (z.B. Organismen, DNA-Codes ..) betrachtet und ein Multiples Alignment erstellt. Nach der Wahl eines Subtsitutionsmodells wird ein Maß zur Beurteilung eingeführt und mit dessen Hilfe eine Matrix erstellt. Daraus kann letztendlich ein phylogenetischer Baum generiert und erstellt werden. Die verwendeten Verfahren lassen sich in zwei große Gruppen unterteilen.

2.2 Distanzbasierte Verfahren

Distanzbasierte Methoden beruhen auf der Idee, dass sich bei bekannten Distanzen zwischen allen terminalen Taxa eines Datensatzes sich leicht die evolutionäre Geschichte dieser Sequenzen rekonstruieren lässt. Ausgegangen wird hierbei von der *„Molecular Clock Theory“*, die Emile Zuckerkandl und Linus Pauling Anfang der 60er Jahre aufstellten. Nach dieser Theorie ist die Zahl der zulässigen Mutationen in den Genen pro Zeiteinheit ungefähr konstant. Daher ist es möglich, dass die Distanz-

messung einen numerischen Wert mit einem Paar Sequenzen assoziiert. Hierbei indizieren niedrige Werte eine hohe Ähnlichkeit.

Die einfachste Form der Distanzmessung ist die sogenannte *Hamming-Distanz* d_H. Sie gibt für zwei Sequenzen mit gleicher Länge die Anzahl der unterschiedlichen Merkmalsausprägungen der analysierten Sequenzen an. So haben die Sequenzen A: AGCACGAT und B: ATCACACT eine *Hamming-Distanz* von $d_H = 3$, da sie sich in drei Basen unterscheiden. Diese Distanz geht in eine Distanzmatrix ein, aus der dann über verschiedene Verfahren die gesuchten Bäume konstruiert werden können. Auf diese Verfahren soll nun im Folgenden näher eingegangen werden.

2.2.1 UPGMA – unweighted pair group method with arithmetic mean

Der UPGMA-Algorithmus ist ein einfaches Clusterverfahren, das ursprünglich von Sokal und Mitchener (1958) entwickelt wurde. Wie bei allen distanzbasierten Algorith-men wird davon ausgegangen, dass sich die Sequenzen gemäß der molekularen Uhr evolviert haben.

2.2.1.1 UPGMA-Algorithmus

1. Initialisierung:
Jede Sequenz i wird seinem eigenen Cluster C_i zugeordnet. Weiterhin wird jeder Sequenz ein Blatt mit der Höhe 0 zugewiesen.

2. Iteration:
In jedem Iterationsschritt werden die beiden Cluster mit dem kleinsten Abstand zueinander gesucht (d(i,j) ist also minimal). Es wird ein neues Cluster $C_k = C_i \cup C_j$ erstellt und die mittlere Distanz zwischen den zwei Clustern bestimmt, indem über alle paarweise Distanzen gemittelt wird. Hierbei gilt: $$d(k,l) = \frac{1}{|C_k||C_l|} \sum_{p \in C_k, q \in C_l} d(p,q)$$

Weiterhin wird ein Knoten k mit den Tochterknoten i und j definiert und mit der Höhe d(i,j)/2 platziert. Der Abstand des neuen Clusters C_k und den übrigen Clustern wird folgendermaßen berechnet und in einer neuen Distanzmatrix angegeben:

$$d(k,l) = \frac{d(i,l)|C_i| + d(j,l)|C_j|}{|C_i| + |C_j|}$$

3. Ende:

Die Iteration bricht ab, wenn nur noch zwei Cluster C_i und C_j übrigbleiben. Der Baum wird mit einer Wurzel auf der Höhe d(i,j)/2 vervollständigt.

2.2.1.2 Beispiel:

Gegeben seien die fünf DNA-Sequenzen:

A: A**T**CG**A**ATACAG**A**TTCGGT B: A**A**CG**A**ATACAG**A**TTCGGT

C: A**C**CG**T**ATGCAG**C**TTCGGT D: A**G**TG**C**ATCCAG**T**TTCAGT

E: A**G**AG**C**ATCCAG**T**TTCCGT

Hieraus lässt sich folgende Distanzmatrix M bestimmen:

	A	B	C	D	E
A	-	**1**	4	6	6
B		-	4	6	6
C			-	6	6
D				-	2
E					-

Bestimmen des Minimums: d(A,B) = 1
$\Rightarrow$ C = {A,B}

Berechnung der Kantenlänge bis zum Knoten:
$\Rightarrow$ d(A,B)/2 = 0,5

Berechnung einer neuen Distanzmatrix:

	{A,B}	C	D	E
{A,B}	-	$\frac{4+4}{2}=4$	6	6
C		-	6	6
D			-	**2**
E				-

Bestimmen des Minimums: d(D,E) = 2
$\Rightarrow$ C = {D,E}

Berechnung der Kantenlänge bis zum Knoten:
$\Rightarrow$ d(D,E)/2 = 1

Berechnung einer neuen Distanzmatrix:

	{A,B}	C	{D,E}
{A,B}	-	**4**	$\frac{1}{4}(6+6+6+6)=6$
C		-	$\frac{1}{2}(6+6)=6$
{D,E}			-

Bestimmen des Minimums: d({A,B},C) = 4
$\Rightarrow$ C = {A,B,C}

Berechnung der Kantenlänge bis zum Knoten: $\Rightarrow$ d({A,B},C)/ 2 = 2

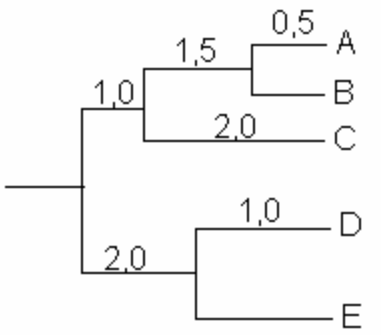
Berechnung einer neuen Distanzmatrix:

	{A,B,C}	{D,E}
{A,B,C}	-	$\frac{1}{6}(6+6+6+6+6+6)=6$
{D,E}		-

Da nun nur noch zwei Cluster vorhanden sind, bleibt lediglich die Länge der Kanten zum Knoten zu berechnen:

$\Rightarrow d(\{A,B,C\},\{D,E\})\ /2 = 3.$

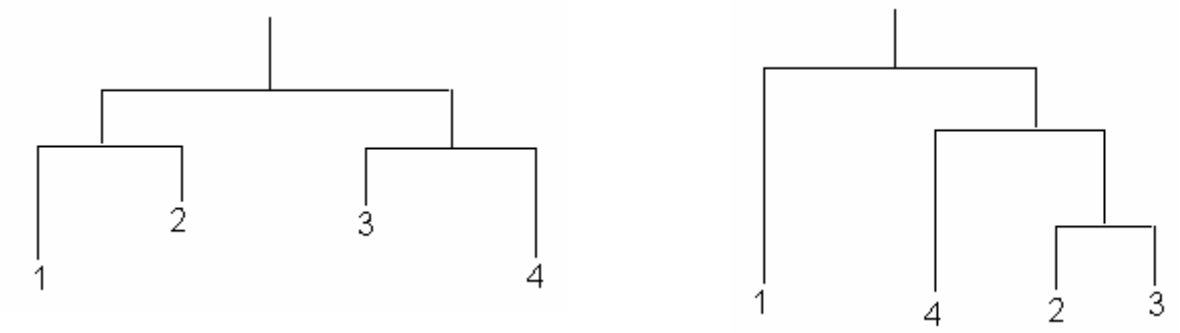
Daraus ergibt sich folgender Baum:



2.2.1.3 Ultrametrik-Eigenschaft

Die Blätter eines durch UPGMA konstruierten Baumes befinden sich alle auf einer Höhe, haben also alle den gleichen evolutionären Abstand von der Wurzel. Umgekehrt wird ein zugrunde liegender Baum, der diese Eigenschaft besitzt, auch von UPGMA korrekt konstruiert. Ein solcher Baum erfüllt die sogenannte Ultrametrik-Eigenschaft: Für alle Sequenzen x_i , x_j und x_k gilt, dass sie den gleichen Abstand haben ($d_{ij}= d_{ik}= d_{jk}$) oder zwei der Abstände gleich und größer als der dritte sind ($d_{ij}= d_{ik} \geq d_{jk}$). Somit ist das Maximum der drei Abstände nicht eindeutig und auch nicht singulär.

Besitzt der zugrunde liegende Baum die obige Ultrametrik-Eigenschaft nicht, dann kann UPGMA zu falschen Ergebnissen führen. Ein Beispiel ist im folgenden dargestellt: Der linke Baum erfüllt die Eigenschaft nicht, hier besitzen die Sequenzen x_2 und x_3 zwar den gleichen Abstand, nicht aber den gleichen Vorgänger. Der rechte Baum, der die Ultrametrik-Eigenschaft erfüllt, führt zu einem sinnvollen Resultat.



Allgemein lässt sich sagen: Die Distanzen in evolutionären Bäumen, die der molekularen Uhr gehorchen, genügen der Ultrametrik-Bedingung.

2.2.1.4 Additivitäts-Eigenschaft

UPGMA verwendet implizit eine weitere Eigenschaft der Abstände zwischen Sequenzen: die Additivität. Die Kantenlängen eines Baumes sind additiv, wenn der Abstand zwischen je zwei Sequenzen der Länge des Pfades (Summe der Längen der

Kanten) zwischen diesen Sequenzen entspricht. Zum Nachweis der Additivität ist die Vier-Punkt-Bedingung zu erfüllen: $d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$. Hieraus wird ersichtlich, dass jede ultrametrische Distanz auch die Vier-Punkt-Bedingung erfüllt. Unter der Annahme der Additivität existiert für alle Sequenzen i, j , l ein Knoten k mit $d_{kl} = ½ (d_{il} + d_{jl} – d_{ij})$. Ist k nämlich der Knoten, an dem sich die zu den angegebenen Sequenzen führenden Äste treffen, gilt: $d_{il} = d_{ik} + d_{kl}$, $d_{jl} = d_{jk} + d_{kl}$ und $d_{ij} = d_{ik} + d_{kj}$. Hieraus ergibt sich die genannte Beziehung, aus der direkt folgt: Sind i und j Nachbarn mit dem gemeinsamen Vorgänger k und definiert man den Abstand zwischen k und jedem Knoten l gemäß dieser Beziehung, dann ist d_{kl} exakt der Abstand im zugrunde liegenden Baum.

2.2.2 Neighbour-Joining-Methode

Dieser Algorithmus, der zuerst von Saitou und Nei 1987 veröffentlicht wurde, konstru-iert ebenfalls Bäume nach der distanzbasierten Methode. Im Gegensatz zum UPGMA-Verfahren, bei dem die zwei Taxa als benachbart angesehen werden, deren Abstand minimal ist, werden bei diesem Verfahren Distanzen gebildet, die die mittlere Distanz zu allen anderen Taxa abziehen. Hier werden also die Cluster jeweils miteinander verbunden, die sowohl voneinander gering als auch weit von den anderen entfernt sind. Bei diesem Verfahren wird ein ungewurzelter, additiver Baum konstruiert.

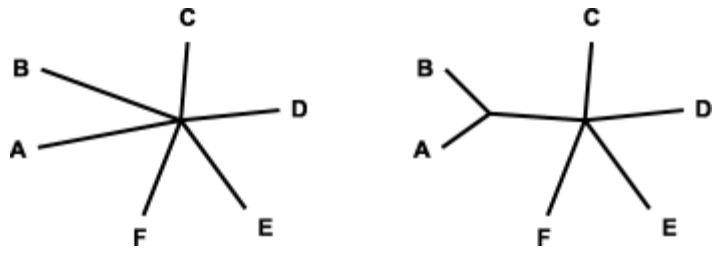
2.2.2.1 Neighbour-Joining-Algorithmus

1. Initialisierung:
Ausgehend von einem Busch (Sterngraph) wird die Menge der Blätter L durch die Menge der Sequenzen initialisiert.

2. Iteration:
In jedem Iterationsschritt werden zwei Knoten zu Nachbarn verschmolzen, also durch einen gemeinsamen Vorfahren ersetzt. Hierbei werden diejenigen Knoten ausgewählt, für die der Ausdruck $S(i,j) = d(i,j) – (r_i + r_j)$ minimal ist, wobei gilt:

$$r_i = \frac{1}{|L|-2} \sum_{k \in L} d(i,k)$$

Ausgehend von der Additivität wird ein Knoten k definiert und die Längen der zugehörigen Kanten auf $d_{ik} = ½ (d_{ij} + r_i - r_j)$, $d_{ij} = d_{ij} - d_{ik}$ gesetzt. Weiterhin werden die Abstände zwischen k und allen Knoten $l \in L$ definiert durch $d_{kl} = ½ (d_{il} + d_{jl} – d_{ij})$.



In jedem Iterationsschritt wird also die Anzahl der Knoten um eins erhöht, da der gemeinsame Vorfahre eingefügt wird.

3. Ende:

Die Iteration bricht ab, sobald $|L| = 2$. Der Baum wird anschließend durch eine Kante zwischen den beiden verbleibenden Knoten vervollständigt. Die Länge dieser Kante entspricht dem Abstand der beiden Knoten.

2.2.2.2 Beispiel:

Gegeben seien 4 Sequenzen, die sich in der folgenden Distanzenmatrix darstellen lassen. ($|L|=4$)

	A	B	C	D
A	-	8	7	12
B		-	9	14
C			-	11
D				-

Berechnung von S^0:

Hierfür werden zuerst aus den Spaltensummen die r_i berechnet:

$r_A = ½ (8+7+12) = 13.5$, $r_B = ½ (8+9+14) = 15.5$,

$r_C = ½ (7+9+11) = 13.5$, $r_D = ½ (12+14+11) = 18.5$.

$S^0 =$

	A	B	C	D
A	-	8	7	12
B	-21	-	9	14
C	-20	-20	-	11
D	-20	-20	-21	-

Das Minimum hierbei ist S(A,B)=-21. Es wird ein neuer Knoten k definiert, der A und B verbindet und die Kantenlänge von k zu A bzw. B berechnet:

$d(A,k) = ½ (d(A,B) + r_a - r_b) = ½ (8 + 13.5 – 15.5) = 3$

$d(B,k) = d(A,B) – d(A,k) = 8 – 3 = 5$

Es wird eine neue Distanzmatrix erstellt :

	K	C	D
k	-	d(k,C)	d(k,D)
C		-	11
D			-

Hierbei ist:

$d(k,C) = ½ (d(A,C) + d(B,C) – d(A,B)) = 4$

$d(k,D) = ½ (d(A,D) + d(B,D) – d(A,B)) = 9$

Berechnung von S^1:

Wiederum werden die r_i aus den Spaltensummen berechnet:

$r_k = ½ (4+9)=6.5$, $r_C = ½ (4+11)=7.5$,

$r_D = ½ (9+11)=10$.

$S^1 =$

	k	C	D
k	-	4	9
C	-10	-	11
D	-7.5	-6.5	-

Das Minimum hierbei ist S(k,C)=-10. Es wird ein neuer Knoten l definiert, der k und C verbindet, und die Kantenlänge von l zu k bzw. C berechnet:

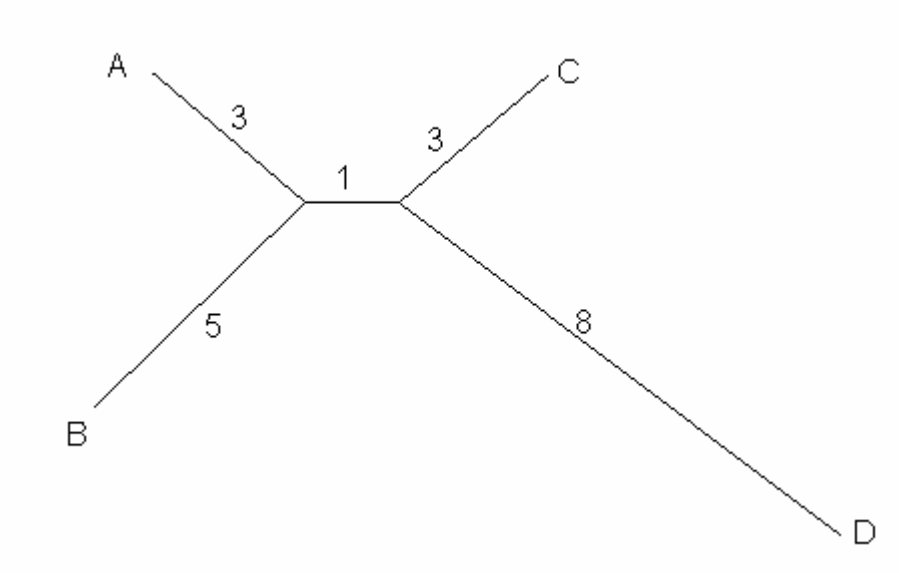
$d(C,l) = ½ (d(k,C) + r_c – r_k) = 3$

$d(k,l) = d(k,C) – d(C,l) = 4 – 3 = 1$

Da nun nur Knoten D und l übrig sind, kann das Verfahren gestoppt werden. Es bleibt nur noch die Länge der Kante d(l,D) zu berechnen:

$d(l,D) = d(C,D) – d(l,C) = 11 – 3 = 8$

Mit den so gewonnenen Erkenntnissen und Ergebnissen lässt sich ein ungewurzelter Baum konstruieren, der die vorgegebenen 6 Distanzen korrekt wiedergibt.



Der mit Hilfe der Neighbour-Joining-Methode konstruierte Baum

2.3 Charakterbasierte Methoden

Zur Rekonstruktion phylogenetischer Bäume aus Sequenzdaten biologischer Makromoleküle haben sich unter den merkmalbasierten Methoden zwei Ansätze etabliert. Diese lassen sich allgemein in zwei Gruppen unterteilen, in den Maximum-Parsimony-Ansatz und den Maximum-Likelihood-Ansatz (Heuristischer Ansatz).

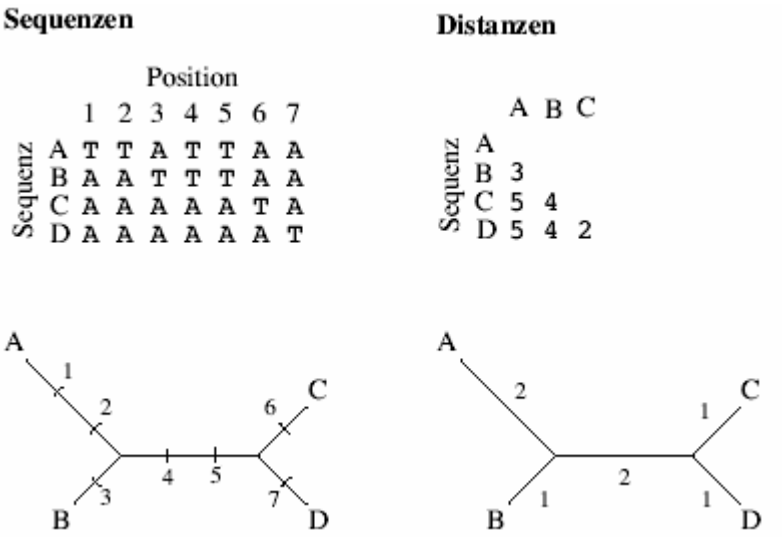
Maximum-Parsimony (MP) sucht den Baum aus, der die kleinstmögliche Anzahl an Veränderungen (Mutationen) benötigt, um die genetischen Unterschiede der verschiedenen Spezies zu erklären.
Maximum-Likelihood (ML) dagegen errechnet die Wahrscheinlichkeit, dass ein Modell (Baum) die beobachtete Sequenzvariation verursacht.

2.3.1 Maximum-Parsimony-Methode

Parsimony kommt aus dem Englischen und bedeutet Sparsamkeit bzw. Geiz. Aus den verschiedenen Phylogenien (Bäumen) werden derjenige herausgesucht, der im Allgemeinen der „most parsimonious tree“ ist, also am sparsamsten mit Mutationen umgeht.
Das Maximum-Parsimony-Verfahren konstruiert für alle internen Knoten eines vorgegebenen Stammbaumes Sequenzen, die die von diesen Knoten repräsentierten Organismen gehabt haben könnten. Diese Sequenzen werden so

konstruiert, dass die Sequenzen entlang des Baumes während der vom Stammbaum vorgegebenen evolutionären Entwicklung möglichst wenigen Mutationen unterworfen sind. Die Gesamtsumme aller im Baum nötigen Mutationen ist dann das Maß für die Qualität des Baumes. Mit der Maximum-Parsimony-Methode wird versucht, aus allen möglichen Stammbäumen denjenigen zu finden, für den die geringste Anzahl an Mutationen nötig ist.
Nukleotidpositionen, die in allen Sequenzen gleich oder nur bei einer Sequenz unterschiedlich sind, sind phylogenetisch nicht informativ, da sie nicht zwischen verschieden alternativen Bäumen unterscheiden. Man unterteilt die variablen Positionen daher in Parsimonie-informativ und in Parsimonie-nicht-informativ (Autapomorphie).

Das Maximum-Parsimony-Ansatzes bietet einige Vorteile. Diese Methode wurde ursprünglich für morphologische Daten entworfen und hat sich bewährt, wenn sich die beobachteten Merkmale nur selten ändern. Das gilt im allgemeinen für morphologische Daten. Diese Methode scheitert jedoch, wenn die beobachteten Merkmale hochvariabel sind oder sehr lange Kanten im gesuchten Baum vorkommen. Außerdem können Parsimony-Methoden als Näherungen von Maximum-Likelihood-Verfahren angesehen werden, da sie schneller als diese sind und damit weniger Rechnerressourcen benötigen. Ein weiterer Vorteil gegenüber distanzbasierten Methoden ist die Tatsache, dass es die Sequenzinformationen nicht auf eine Zahl reduzieren (siehe Beispiel: der linke Graph gründet auf einem merkmalbasierten, der rechte auf einem distanzbasierten Verfahren).



Nachteil des Maximum-Parsimony-Verfahrens ist eine längere Rechenzeit gegenüber Distanzmethoden und die Benutzung von relativ kleinen Sequenzinformationen.

Zum besseren Verständnis des Parsimony-Methoden werden sowohl das Hennig- als auch das Wagner-Verfahren näher betrachtet.

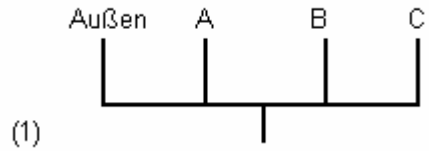
2.3.1.1 Hennig-Verfahren

Die in der Datenmatrix gegebenen Informationen werden bei diesem Verfahren Merkmal für Merkmal abgearbeitet werden und dabei der Stammbaum jeweils dem Kenntnisstand angepasst wird.

Als Beispiel soll eine Matrix mit 4 Gruppen (Außengruppe, A , B und C) dienen, die fünf verschiedene Merkmal besitzen (1) oder nicht (0).

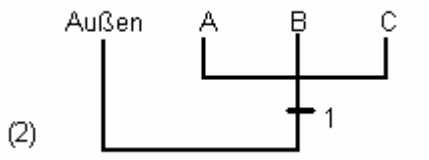
Merkmal	1	2	3	4	5
Außengruppe	0	0	0	0	0
A	1	0	0	0	0
B	1	1	0	1	0
C	1	0	1	1	1

1. Ausgangspunkt ist ein sog. "Busch", d.h. es werden alle Taxa (Spezies) in einer Polytomie (phylogenetischen Baum) zusammengefasst (1).



(1)

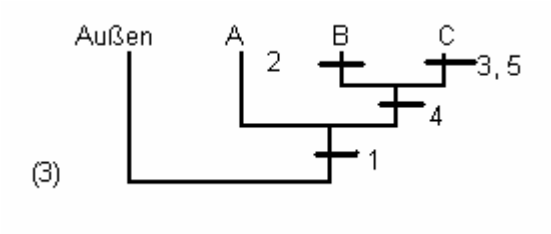
2. Nimmt man das Merkmal Nr. 1 hinzu, so ist es möglich, die Taxa A, B, C zusammen zu gruppieren und von der Außengruppe zu trennen (2).



(2)

3. Merkmal 2 ist eine Autapomorphie (Merkmal, das erst in der Stammlinie des betrachteten Art entstanden sind) von Spezies B und liefert somit keine Informationen über Verwandtschaftsverhältnisse (3).
- Merkmal 3 ist eine Autapomorphie von Taxon (Spezies) C

- Merkmal 4 tritt im Zustand 1 nur bei den Taxa B und C auf und kann daher als Synapomorphie (gemeinsame Merkmale zweier Spezies) dieser beiden Taxa gedeutet werden
- Merkmal 5 ist Autapomorphie von Taxon C



Damit sind alle Merkmale berücksichtigt und der Stammbaum konnte komplett in Dichotomien (Verzweigungen) aufgelöst werden.

2.3.1.2 Wagner-Verfahren

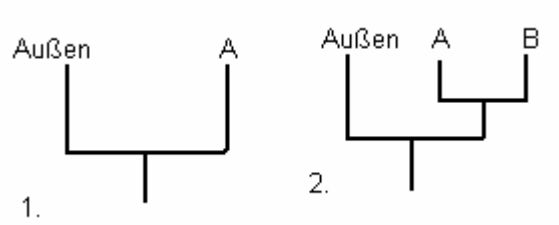
Bei dieser Methode werden die einzelnen Taxa nacheinander zu einem Baum zusammengefügt. Dabei wird die Anzahl der Merkmalsänderungen auf dem resultierenden Kladogramm minimal gehalten.

Als Beispiel dient wieder die gleiche Matrix:

Charakteristik	1	2	3	4	5	#abg
Außengruppe	0	0	0	0	0	0
A	1	0	0	0	0	1
B	1	1	0	1	0	3
C	1	0	1	1	1	4

1. Die Reihenfolge, in der die Spezies zum Kladogramm hinzugefügt werden, ermittelt man zunächst anhand der Anzahl der abgeleiteten Merkmale (#abg) für jedes Taxon (Spezies). Taxon A unterscheidet sich in einem Merkmal zu der Außengruppe und hat daher #abg 1.#

2. Danach wird das Taxon mit der geringsten Zahl an abgeleiteten Merkmalen (A) mit der Außengruppe verbunden.



Es folgt das Taxon mit der nächsthöheren Zahl abgeleiteter Merkmale, also B. Es wird mit A verbunden und am Knotenpunkt von A und B wird die am weitesten abgeleitete Merkmalsausprägung notiert, die diese beiden Taxa gemeinsam haben.

Im Beispiel also 1|0|0|0|0 (AxB). Dies kann so interpretiert werden, dass, wenn A und B Schwesternspezies sind, die Stammart von A & B die Merkmalsausprägung 1|0|0|0|0 besessen hat.

AxB	1	0	0	0	0
AxC	1	0	0	0	0
BxC	1	0	0	1	0
(AB)xC	1	0	0	0	0

3. Als nächstes wird nun wieder die Spezies mit der nächsthöheren Zahl abgeleiteter Merkmale (C) in den Baum eingefügt. Dies kann an drei verschiedenen Positionen geschehen:

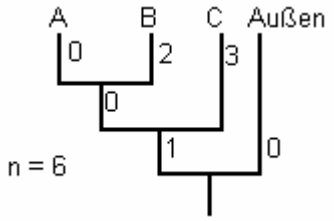
a) als Schwesternspezies zur Gruppe A+B,

b) als Schwesternspezies zu B

c) als Schwesternspezies zu A.

Spezies C muss so eingefügt werden, dass nur eine minimale Zahl von Merkmalsänderungen erforderlich ist. Die entsprechende Stelle im Kladogramm wird durch Differenzbildung zwischen den Merkmalen des einzufügenden und des benachbarten Taxon ermittelt:

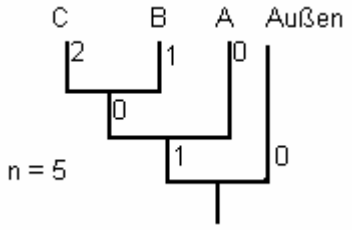
C	10111
AxB	10000
Differenz	00111



d.h. würde C als Schwestergruppe zu A in das Kladogramm eingefügt, so würden 3 weitere Merkmalsänderungen zur "Länge" des Baumes hinzukommen.

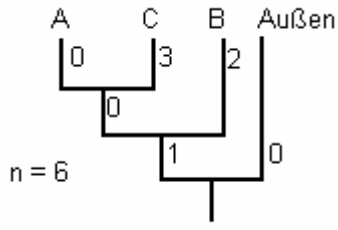
C	10111
B	11011
Differenz	01101

→ 3 Änderungen.



C	10111
A	10000
Differenz	00111

→ 3 Änderungen



4. Es stehen drei Verknüpfungspunkte mit gleicher Wertigkeit zur Verfügung. Man muss also alle drei Varianten ausprobieren und jedes Mal die Gesamtlänge des Baumes, d.h. die Gesamtzahl der Merkmalsänderungen (n) ermitteln. Danach ist der mittlere Baum mit einer Länge von 5 Schritten, d.h. mit 5 Merkmalsänderungen, zu bevorzugen.

2.3.2 Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode wird benutzt, um unbekannte Parameter zu schätzen, von denen eine bekannte Wahrscheinlichkeitsfunktion für einen stochastischen Prozess abhängt. Mit dieser Methode werden dann anhand einer festen Stichprobe, in diesem Fall der Sequenzdaten, die unbekannten Parameter so geschätzt, dass der Wert der Wahrscheinlichkeitsfunktion (bei fester Stichprobe als Likelihood-Funktion bezeichnet) sein Maximum erreicht. Die in der Stammbaumanalyse zu schätzenden unbekannten Parameter sind die Kantenlängen in einem vorgegebenen Baum.
Dabei ist es wichtig zu unterscheiden, zwischen der Wahrscheinlichkeit die beobachteten Daten zu bekommen und der Wahrscheinlichkeit, dass das zugrunde liegende Modell das Richtige ist. Mit einem Modell, welches die Wahrscheinlichkeit, verschiedener Ereignisse zu beobachten, beschreibt, kann man die Wahrscheinlichkeit L für den Erhalt der beobachteten Daten berechnen.
Es gilt: L_D = Whs.(D|H). Dabei ist L_D = Whs.(D|H) die Wahrscheinlichkeit die Daten D zu erhalten unter der Hypothese H zu erhalten. Da der Wert L meist sehr klein ist, wird er als natürlicher Logarithmus dargestellt. $\ln L = \sum_{j=1}^{N} \ln(J)$
Diese Berechnung erlaubt es, verschiedene Modelle auf ihre Wahrscheinlichkeit zu testen.
Das Maximum-Likelihood hat mehrer Vorteile gegenüber dem MP und anderen distanzbasierten Verfahren:

- Die Varianz der Ergebnisse ist sehr klein
- Selbst bei kleinen Sequenzen liefern sie bessere Ergebnisse als alternative Methoden
- Das Verfahren ist statistisch wohl begründet
- Alle Sequenzinformationen werden genutzt

Der Hauptnachteil der ML-Methode ist, dass enorme Rechenzeiten nötig sind, um die große Anzahl der möglichen Stammbäume zu überprüfen, die auch bei heuristischen Methoden meist exponentiell mit der Anzahl der benutzten Spezies wächst.

2.4. Abschätzung des Stichprobensfehlers

Über die vorgestellten Baumrekonstruktionsmethoden erhält man einen oder mehrere Bäume, ohne aber zu wissen, wie sehr den Daten, die in diesen Bäumen stecken, vertraut werden kann.
Um dieses Problem zu lösen, werden zwei Algorithmen genutzt: *Bootstrap* (Felsenstein 1985) und *Jackknife*. Bei beiden Verfahren wird die Datenmatrix, die dem Baum zugrunde liegt, zufällig modifiziert. Ausgehend von diesen modifizierten Matrizen werden neue Bäume (100 bis 1000 Stück) erstellt. Finden sich die Grundzüge des ursprünglich erstellten Baumes besonders häufig in den neuberechneten Bäumen, so wird die Genauigkeit und Aussagekraft von diesem unterstützt.

2.4.1 Bootstrapping

Aus den vorhandenen Daten einer Matrix werden unabhängige Stichproben gewonnen (Pseudomatrizen), indem aus dem originalen Datensatz zufällig Positionen (Matrixspalten) gezogen und zurückgelegt werden. Das heißt, eine zufällige Position der Datenmatrix wird kopiert und nimmt die Position Eins der ersten Pseudomatrix ein. Eine weitere zufällige Position wird kopiert (es kann theoretisch dieselbe Position sein) und bildet Position Zwei dieser Pseudomatrix. Dies wird solange wiederholt, bis die Pseudomatrix die gleiche Größe wie die Originalmatrix aufweist. Durch dieses Verfahren können manche Positionen mehrmals in der Pseudostichprobe vorhanden sein, andere gar nicht. Folglich erhält die neugewonnene Stichprobe nur Positionen wie im Originaldatensatz, aber mit veränderter Frequenz. Auf Grundlage dieser Pseudostichprobe wird über gleiche Baumrekonstruktionsmethode ein phylogenetischer Baum (Bootstrap-Baum) konstruiert. Dieser Prozess der Generierung von Pseudostichproben wird 100 – 1000 mal wiederholt und man erhält so diese Anzahl von Bootstrap-Bäumen. Die Häufigkeit des ursprünglich erstellten Baumes kann nun gemessen werden, um dessen Relevanz angeben zu können.

2.4.2 Jackknifing

Das Bootstrap-Verfahren ist bei umfangreichem Datensatz kaum durchzuführen, da hierfür herkömmliche Computer- Rechenleistungen nicht ausreichen. Für diesen Fall eignet sich das Jackknife-Verfahren (auch Eliminierungsmethode genannt):

Auch hierbei werden Matrizen durch die zufällige Auswahl von Merkmalen, d.h. Positionen aus der Originalmatrix aufgebaut, wobei die entstehende Pseudomatrix allerdings nur aus einem gewissen Prozentsatz der Merkmale besteht. In der Regel wird die Matrixgröße um 50% reduziert, d.h. bei jedem der 100 – 1000 Durchgänge wird immer nur die Hälfte der Merkmale berücksichtigt. Der weitere Vorgang ist dem des Bootstrapping identisch. Die Wahrscheinlichkeit für die Eliminierung eines Merkmals sollte nicht zu hoch liegen, da ansonsten keine für die Auswertung sinnvollen Ergebnisse zu erwarten sind.

2.5 Substitutionsmodelle

Um phylogenetische Bäume rekonstruieren zu können, ist man auf Datenmaterial angewiesen, das die Entwicklungsgeschichte der Organismen widerspiegelt, d.h. in diesen Daten müssen die Spuren des Entwicklungsprozesses erkennbar sein. Da der Evolutionsprozess jedoch ein hochkomplexer, noch nicht in seiner Vollständigkeit verstandener Prozess ist, müssen die Betrachtungen, die in die Baumrekonstruktion miteinfließen, immer eine Vereinfachung der Wirklichkeit sein. Man bedient sich sog. Evolutions- bzw. Substitutionsmodelle, welche Evolution als einen zufälligen Prozess ansehen, bei dem mit einer gewissen Wahrscheinlichkeit Nucleotide der DNA durch andere ausgetauscht werden (Substitutionen).

Bei Substitutionen muss zwischen *Transversionen* (Purinbase wird durch Pyrimidinbase ersetzt und umgekehrt ({A,G}⇔{C,T})) und *Transitionen* (Purinbase durch Purinbase (A⇔G) oder Pyrimidinbase durch Pyrimidinbase (C⇔T)) unterschieden werden.

Unter der Voraussetzung, dass die Nucleotidsubstitutionenanzahl über die Zeit konstant bleibt, kann man in einer einfachen Matrix die Substitutionswahrscheinlichkeit als

$$p_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

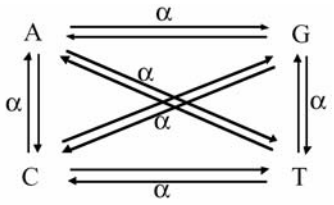
darstellen.

Hierbei gibt p_{AC} die Wahrscheinlichkeit wieder, dass im Zeitintervall t an einer gegebenen Position der Basensequenz ein Austausch von A nach C stattfindet.
Die diagonalen Einträge beschreiben die Wahrscheinlichkeit, dass (anscheinend) keine Substitution stattgefunden hat; eine bestimmte Position trägt z.B. zum Zeitpunkt 0 das Merkmal A und zum Zeitpunkt t ebenfalls. Es ist nicht möglich zu entscheiden, ob an dieser Stelle keine Substitution stattgefunden hat oder ob die Beobachtung das Ergebnis multipler Substitutionen ist. Die Wahrscheinlichkeit dafür ist 1 minus die Wahrscheinlichkeit für die Substitution des A durch ein C, G oder T.

Mathematisch lässt sich diese Wahrscheinlichkeit ausdrücken als $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$.

2.5.1 Jukes-Cantor-Modell

Das Substitutionsmodell von Jukes und Cantor (1969) ist hierfür ein sehr einfaches Modell, denn es nimmt an, dass Substitutionen zufällig zwischen allen 4 Nucleotiden vorkommen. Die Wahrscheinlichkeiten für Transitionen und Transversionen werden gleichgesetzt. In diesem Modell, das auch 1-Parameter-Modell genannt wird, ist die Substitutionsrate in alle Richtungen gleich α.



Die Matrix der Substitutionswahrscheinlichkeit α und der dazugehörige Basenvektor (jedes der vier Merkmale (A, C, G und T) ist mit einer Frequenz von 25% vertreten) lassen sich wie folgt darstellen:

$$p_t = \begin{bmatrix} \bullet & \alpha & \alpha & \alpha \\ \alpha & \bullet & \alpha & \alpha \\ \alpha & \alpha & \bullet & \alpha \\ \alpha & \alpha & \alpha & \bullet \end{bmatrix} \qquad f = [¼ , ¼, ¼, ¼]$$

Die Zahl der Substitutionen seit der Divergenz zwischen 2 Sequenzen, kann über den natürlichen Logarithmus ln in folgender Formel berechnet werden

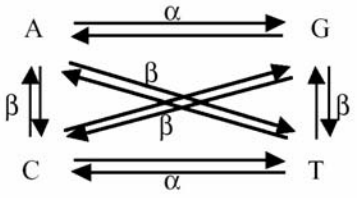
$$K = - ¾ \ln (1 - 4/3\ p)$$

wobei p die Anzahl der Merkmale (Nucleotide) ist, die in beiden Sequenzen unterschiedlich ist (p = Anzahl der Substitutionen / Länge L der Sequenz)

Im Jukes-Cantor-Modell wird aber der Einfluss von Selektion auf einzelne Genabschnitte nicht berücksichtigt und die Sequenzevolution wird als mechanischer Zufallsprozess behandelt. Auch die Annahme, dass der Austausch aller Nucleotide gleich zufällig ist, erweist sich in den meisten biologischen Sequenzen als nicht realistisch. Sequenzvergleiche haben gezeigt, dass Transitionen häufiger vorkommen und somit wahrscheinlicher sind als Transversionen. Diese Tatsachen hat Kimura 1980 in seinem 2-Parameter-Modell aufgegriffen.

2.5.2 Kimuras 2-Parameter-Modell

Wie gerade erklärt, häufen sich mit der Zeit Transitionen wesentlich schneller als Transversionen an. Dieser Beobachtung wird in Kimuras 2-Parameter-Modell Rechnung getragen, indem die Transitionsrate α je Position und die Transversionsrate β je Position zu einer totalen Substitutionsrate von $\lambda = \alpha + 2\beta$ aufaddiert werden. Anders ausgedrückt: für ein Nucleotid gibt es drei Möglichkeiten der Substitution, von denen eine eine Transition und zwei Transversionen sind.



Die Substitutionswahrscheinlichkeit hat dann folgende Form:

$$p_t = \begin{bmatrix} \bullet & \beta & A & \beta \\ \beta & \bullet & B & \alpha \\ \alpha & \beta & \bullet & \beta \\ \beta & \alpha & B & \bullet \end{bmatrix}$$

Für den Anteil der Transitionen P (Anzahl der Transitionen / Sequenzlänge L)
und den Anteil der Transversionen Q (Anzahl der Transversionen / Sequenzlänge L) innerhalb unterschiedlicher Nucleotide in L wird K in diesem Fall über folgende Gleichung berechnet:

$$K = ½ \ln a + ¼ \ln b$$

wobei $a = 1/(1- 2P – Q)$ und $b = 1/(1-2Q)$ ist.

2.5.3 Beispiel:

Gegeben seien 2 Sequenzen bestehend aus 200 Nucleotiden (L= 200), die sich durch 50 Transitionen und 16 Transversionen voneinander unterscheiden.

Unter Zuhilfenahme des Jukes-Cantor-Modells erhält man

$$p = (50 + 16)/ 200 = 0.33$$

$$\rightarrow \quad K = 0.435$$

Das Kimura-Modells ergibt

$$P = 50/ 200 = 0.25$$

$$Q = 16/ 200 = 0.08.$$

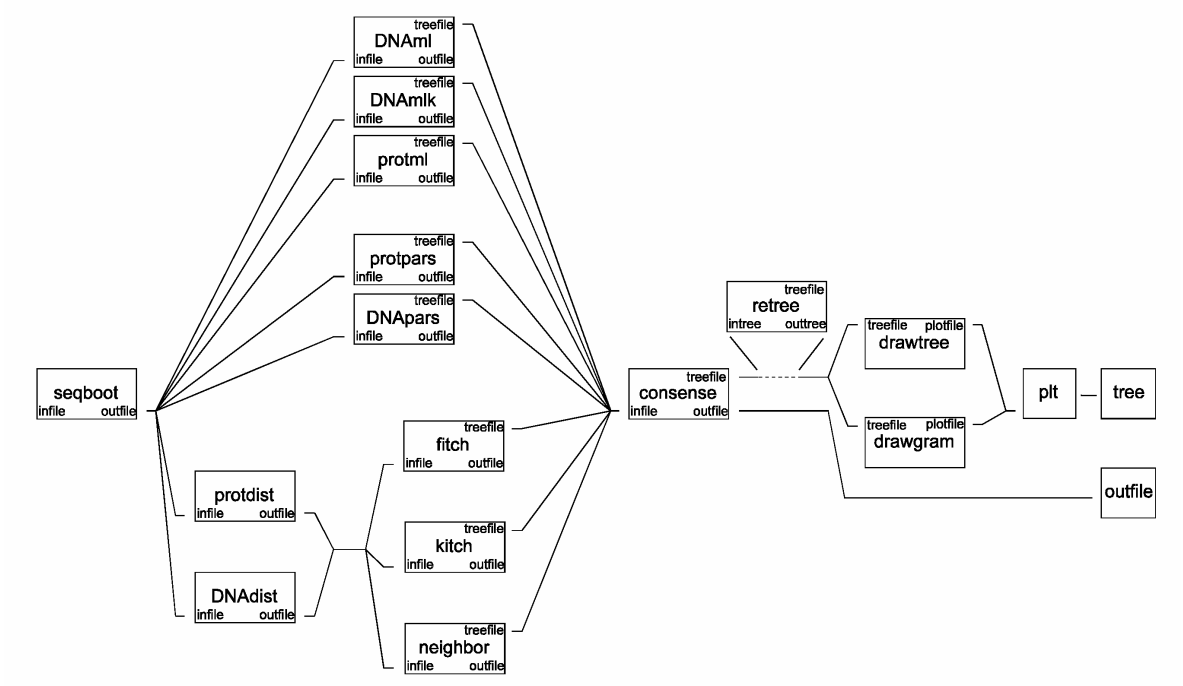
Eingesetzt in die Formel ergibt sich $a = 2.38$ und $b = 1.19$.
Folglich gilt: $K = 0.48$

Wie man sieht, unterscheidet sich die Zahl der Substitutionen ab der Divergenz der Sequenzen je nach Parameterberücksichtigung, wobei das Resultat, welches anhand des Kimura-Modells gewonnen wurde die Realität besser annähert.

3. Computermethoden zur Berechnung phylogenetischer Bäume

Die in den vorhergehenden Kapiteln ausgeführten mathematischen Methoden und Algorithmen zur Berechnung phylogenetischer Bäume lassen sich in Computerprogramme integrieren, so dass die Erstellung von Distanzmatrizen und Bäumen wesentlich erleichtert wird.

Ein besonders umfangreiches und leistungsstarkes Programmpaket stellt PHYLIP (Phylogeny Inference Package) von J. Felsenstein (1986-1996, Universität Washington) dar, anhand dessen in diesem Abschnitt die Computermethoden erläutert werden sollen. Es vereint Programme, die dem Benutzer die Erstellung von Distanzmatrizen aus Sequenzdaten, die Berechung unterschiedlicher phylogenetischer Bäume und die Abschätzung des Stichprobenfehlers (Bootstrapping-Methode) erlauben. Hierbei ist der Benutzer nicht an einen bestimmten Programmablauf gebunden, denn sämtliche Teilprogramme bilden Module, die auch unabhängig voneinander ausgeführt werden können. So ist es möglich, sich von PHYLIP eine Distanzmatrix einer Sequenz erstellen zu lassen oder – wenn diese bereits vorhanden ist – nur den dazugehörigen Baum zu berechnen. Alle Module können aber auch miteinander interagieren, indem die erzeugten Output-Dateien (z.B. Distanzmatrix) als Input-Dateien weiterer Module (z.B. Erstellung von Bäumen) verwendet werden (s. Abb.).



Die PHYLIP-Module können auf unterschiedliche Weise miteinander interagieren und kombiniert werden.

Als Rohdaten dienen entweder DNA- oder Protein-sequenzen. Module, die DNA-Sequenzen bearbeiten, beginnen mit dem Kürzel „DNA“, Programme für Proteinsequenzen mit „prot“. Im Wesentlichen werden aber DNA- wie Proteindaten nach denselben Algorithmen ausgewertet, erkennbar an den Modul-Endungen (z.B. „ml“ für Maximum Likelihood oder „pars“ für Parsimony“).
PHYLIP lässt sich auf den unterschiedlichsten Plattformen wie Windows, DOS, Macintosh, Linux und Unix ausführen. Plattformunabhängig lässt es sich über eine komfortable Webschnittstelle bedienen, die alle In- und Output-Dateien in einem Webbrowser darstellen kann.

Aufgrund der Komplexität des Programmpakets können in dieser Arbeit lediglich die wesentlichen und wichtigsten Schritte/Programme zur Erstellung eines phylogenetischen Baums mit PHYLIP erläutert werden. Alle folgenden Beschreibungen beziehen sich auf die PHYLIP-Webschnittstelle, die unter

http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

zu erreichen ist.

Die wesentlichen PHYLIP-Module, um einen Baum erfolgreich aufzustellen, sind *DNAdist*, *neighbor*/*kitsch* und *drawgram*/*drawtree*. Zwischen ihnen liegen noch viele weitere Module, die entweder nach anderen Algorithmen arbeiten oder die Ergebnisse durch Bootstrapping bzw. Abwandlung der Bäume verändern.

3.1 Beschreibung der wesentlichen Module

DNAdist

DNAdist ist in der Lage, mit Hilfe von vier verschiedenen wählbaren Algorithmen (darunter auch Jukes-Cantor und Kimuras 2-Parameter-Methode) Distanzmatrizen aus zwei oder mehr gegebenen Sequenzen zu berechnen.

neighbor/kitsch

neighbor kreiert einen Baum durch die Neighbour-Joining- bzw. UPGMA-Methode, *kitsch* bedient sich hingegen der ultrametrischen Methode unter der Annahme einer molekularen Uhr.

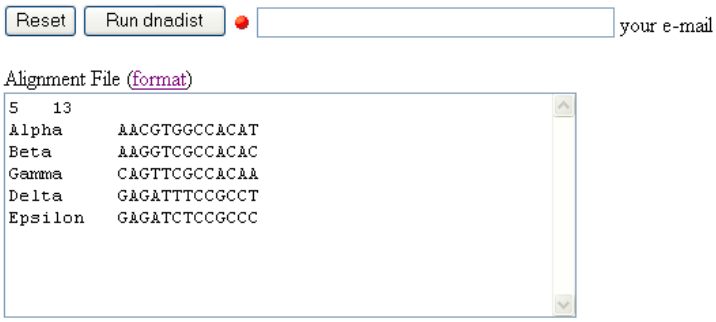
drawgram/drawtree

drawgram gibt aus den Daten von *neighbor*/*kitsch* eine grafische Darstellung des erstellten Baumes mit Wurzel und *drawtree* ohne Wurzel aus

3.2 Vorgehensweise (distanzbasierte Verfahren)

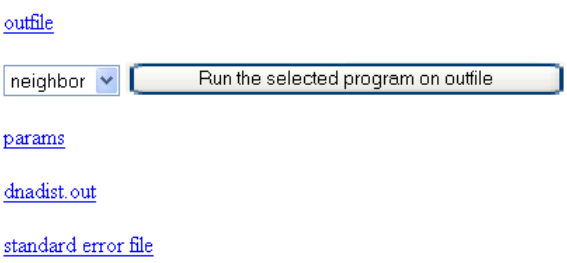
3.2.1 Erstellen der Distanzmatrix mit DNAdist

DNAdist kann die Sequenzdaten aus einer Datenbank (z.B. GenBank, EMBL etc.) durch Angabe der Zugriffsnummern im Textfeld oder durch die direkte Eingabe der Sequenzen beziehen. Dabei muss folgendes Format beachtet werden:
In der ersten Zeile stehen die Anzahl der zu vergleichenden Sequenzen (Spezies) und die Basenanzahl durch ein Leerzeichen voneinander getrennt. In der nächsten Zeile folgen die Sequenzdaten, wobei die ersten zehn Stellen für den Speziesnamen reserviert sind. Eine Eingabe mit den fiktiven Spezies Alpha, Beta, Gamma, Delta und Epsilon könnte also wie folgt aussehen:



Hauptformular von DNAdist mit 5 eingegebenen Beispielsequenzen

Im erweiterten DNAdist-Formular lässt sich die Methode einstellen, nach der die Distanzmatrix erstellt werden soll. Als Beispiel soll in diesem Fall der Jukes-Cantor-Algorithmus dienen. Nach Eingabe der E-Mail-Adresse kann mit „Run DNAdist“ die Distanzmatrix berechnet werden. PHYLIP gibt nach der Rechenzeit eine Website aus, auf der alle Einstellungen sowie die Output-Datei *outfile* eingesehen werden können.



Nach Programmausführung kann outfile eingesehen und mit anderen Modulen weitergenutzt werden.

```
    5
Alpha       0.0000  0.2758  0.5393  0.9492  1.2882
Beta        0.2758  0.0000  0.2758  0.9492  0.5393
Gamma       0.5393  0.2758  0.0000  0.9492  0.7166
Delta       0.9492  0.9492  0.9492  0.0000  0.1722
Epsilon     1.2882  0.5393  0.7166  0.1722  0.0000
```

Im outfile findet sich die Distanzmatrix, auf deren Grundlage alle weiteren Berechnungen durchgeführt werden.

3.2.2 Berechnung des Baumes

Die berechnete Matrix kann nun als Input-Datei (*infile*) für ein weiteres Modul dienen, das durch ein Dropdown-Menü ausgewählt werden kann. Als Beispiel sollen hier der Neighbour-Joining- und UPGMA-Algorithmus dienen, die mit Hilfe des PHYLIP-Programms *neighbor* auf die Distanzmatrix angewendet werden. Durch „Run the selected program on outfile“ wird das Modul ausgeführt. Auch hier können neben der Methodenwahl (Neighbour-Joining oder UPGMA) wieder verschiedene erweiterte Einstellung getroffen werden. So ist es z.B. möglich, nach der Bootstrap-Methode einen Baum aus den gemittelten Werten zu berechnen, indem man die Anzahl der zu analysierenden Daten-Sets angibt. Darauf soll hier aber verzichtet werden. Mit „Run neighbor“ wird der entsprechende Baum berechnet. Natürlich muss das Programm jeweils einmal für dein Neigbour-Joining- und einmal für den UPGMA-Algorithmus ausgeführt werden. Das Ergebnis wird im *outfile* textbasiert ausgegeben:

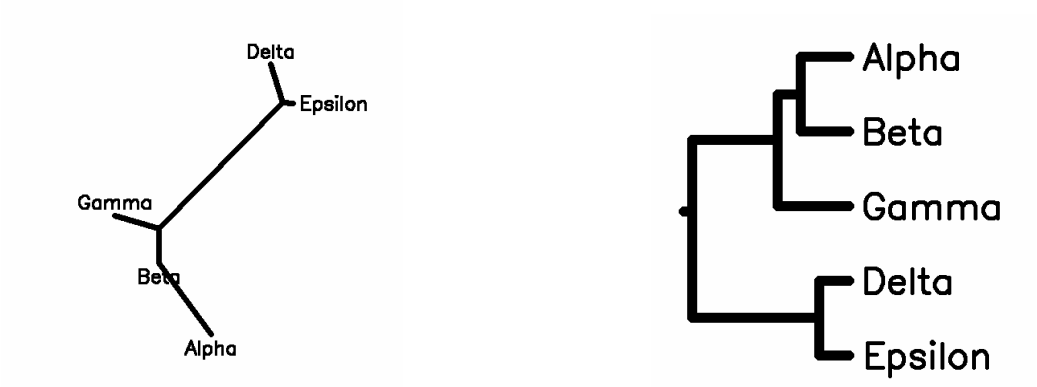
```
   5 Populations

Neighbor-Joining/UPGMA method version 3.6a2.1


 Neighbor-joining method

 Negative branch lengths allowed


  +Beta
  !
  !      +--------Gamma
  2------3
  !      !                                  +--------Delta
  !      +----------------------------------1
  !                                         +--Epsilon
  !
  +-----------------Alpha


remember: this is an unrooted tree!

Between        And            Length
-------        ---            ------
   2          Beta          -0.02159
   2             3           0.11516
   3          Gamma           0.15449
   3             1           0.59231
   1          Delta           0.13668
   1          Epsilon         0.03552
   2          Alpha           0.29739
```

```
   5 Populations

Neighbor-Joining/UPGMA method version 3.6a2.1


 UPGMA method

 Negative branch lengths allowed


                       +-------Alpha
                   +---2
  +----------------3   +-------Beta
  !                !
--4                +-----------Gamma
  !
  !                      +----Delta
  +----------------------1
                         +----Epsilon


From     To            Length          Height
----     --            ------          ------
   4        3          0.24553         0.24553
   3        2          0.06587         0.31141
   2     Alpha         0.13790         0.44931
   2     Beta          0.13790         0.44931
   3     Gamma         0.20377         0.44931
   4        1          0.36321         0.36321
   1     Delta         0.08610         0.44931
   1     Epsilon       0.08610         0.44931
```

Im outfile befinden sich nach der Anwendung von neigbor die Bäume zur Neigbour-Joining- (links) und UPGMA-Methode (rechts) mit den zugehörigen Distanz-angaben.

3.2.3 Grafische Darstellung des Baumes

Mit dem Programm *drawtree* lässt sich der Baum ohne Wurzel aus der Neighbour-Joining-Methode grafisch darstellen. Wichtig hierbei ist das Format des ausgegebenen Plot-Files. Für die Darstellung in einem Webbrowser empfiehlt sich die Wahl von „MS-Windows Bitmap“ unter den „Drawtree options“. Anschließend kann die Berechnung mit „Run drawtree“ durchgeführt werden. Ebenso verfährt man mit dem Programm drawgram, welches den gewurzelten Baum aus der UPGMA-Methode plottet.



Grafische Darstelung des ungewurzelten Baums aus der Neigbour-Joining- Methode (links) und des gewurzelten Baums nach dem UPGMA-Algorithmus (rechts)

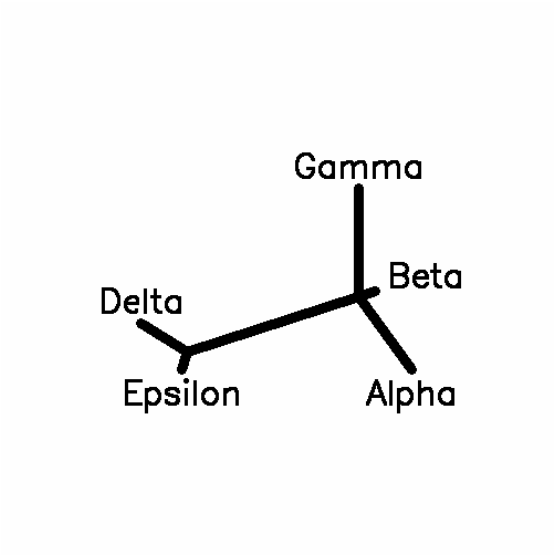
Es ist leicht zu erkennen, dass sich die Abstände der einzelnen Spezies in den beiden Bäumen ähneln. Trotz des unterschiedlichen Aufbaus kann man auf dieselben Verwandtschaftsverhältnisse schließen.

3.3 Parsimony-Methode

Zum Vergleich zwischen distanz- und charakterbasierten Verfahren soll im Folgenden noch ein Baum nach der Parsimony-Methode mit dem Modul *DNApars* erstellt werden.

DNApars errechnet aus den Sequenzdaten direkt einen Baum, ohne dass vorher eine Distanzmatrix erstellt werden muss. Dabei werden die Daten genauso eingegeben wie bei *DNAdist*. Der entstandene Baum kann mit *drawtree* wieder in eine Grafik umgesetzt werden.

```
DNA parsimony algorithm, version 3.6a2.1

One most parsimonious tree found:

                        +-Epsilon
  +---------------------2
  |                     +------Delta
  |
  1-------------Gamma
  |
  +-Beta
  |
  +-----------Alpha


requires a total of     13.000

  between      and       length
  -------      ---       ------
     1           2       0.384615
     2      Epsilon      0.038462
     2      Delta        0.115385
     1      Gamma        0.230769
     1      Beta         0.038462
     1      Alpha        0.192308
```

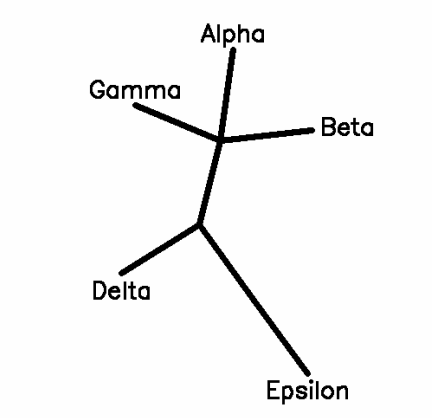


Textbasierte (links) und grafische Ausgabe (rechts) des berechneten Baumes durch die Module DNApars und drawtree

Vergleicht man die beiden wurzellosen Bäume aus der Neigbour-Joining-Methode und dem Parsimony-Algorithmus, so erkennt man einige Unterschiede. Besonders in der Region „Alpha-Beta-Gamma“ weichen die Bäume merkbar voneinander ab. Um den Parsimony-Baum noch etwas zu verbessern, kann zusammen mit *DNApars* ein Bootstrapping durchgeführt werden.

3.4 Bootstrapping

Die Bootstrap-Optionen finden sich in dem erweiterten *DNApars*-Formular. Die Anzahl der Bootstrap-Pseudostichproben wurde auf 99 eingestellt (es muss eine ungerade Zahl sein). Wird DNApars schließlich mit diesen Einstellungen ausgeführt, so werden 99 Bäume berechnet und im *outfile* aufgeführt. Es bietet sich an, aus diesen Bäumen mit dem Modul *consense* einen Mehrheitsregel-Konsensus-Baum (sozusa-gen einen gemittelten Baum) zu berechnen und diesen schließlich mit *drawtree* zu zeichnen.



Konsensus-Baum nach Bootstrapping in DNApars

Es wird deutlich, dass die Parsimony-Methode erheblich unterschiedliche Ergebnisse als der Neighbour-Joining- oder UPGMA-Algorithmus liefert. Allein durch unterschiedliche Einstellungen in den PHYLIP-Modulen lassen sich aus denselben Sequenzen theoretisch unendlich viele verschiedene Bäume berechnen, die zwar gewisse Ähnlichkeiten besitzen, aber nie exakt dieselben Verhältnisse wiedergeben. So ist PHYLIP ein leistungsfähiges Programm, um Phylogenien genauer zu untersuchen, aber auch ein gutes Beispiel für die Tatsache, dass es nie möglich sein wird, einen exakten Baum aus einem Alignment zu berechnen, was im Folgenden noch ausführlicher diskutiert werden soll.

4 Diskussion und Ausblick

4.1 Probleme der „Molecular Clock Theory“

Ein Hauptproblem beim Erstellen phylogenetischer Bäume besteht in der Tatsache, dass distanzbasierte Verfahren auf der „Molecular Clock Theory“ von Zuckerkandl und Pauling basieren. Obwohl bei eng verwandten Organismen anhand molekularer Uhren Zeitabstände geschätzt werden können, ist diese Theorie umstritten, da von verschiedenen Molekülklassen bekannt ist, dass die Zahl der Mutationen pro Zeiteinheit beträchtlich variieren kann.
So haben z.B. Untersuchungen von Vertebraten (Wirbeltieren) ergeben, dass sich die Substitutionsraten zwischen den einzelnen Linien zum Teil stark unterscheiden. Dieses Phänomen wird auf Effekte zurückgeführt, den die unterschiedliche Generationszeiten mehrzelliger Organismen haben können. Für die phylogenetischen Analysen sind lediglich diejenigen Mutationen von Bedeutung, die in der Keimbahn dieser Organismen stattfinden, da nur diese weitervererbt werden können. Haben nun zwei Organismen etwa dieselbe Anzahl an DNA-Replikationen in der Keimbahn, so ist die mögliche Substitutionsrate pro Zeit in dem Organismus höher, bei dem die Generationsdauer kürzer ist. Dies kommt daher, dass in demselben Zeitraum mehr Zellteilungen bzw. DNA-Replikationen in der Keimbahn stattfinden und damit mehr Mutationen entstehen können. Eine zusätzliche Erklärung ist, dass die einzelnen Organismenlinien unterschiedlich gut funktionierende

Reparaturmechanismen besitzen und daher auch verschieden hohe Mutationsraten möglich sind.

Ein weiteres Problem für die Benutzung molekularer Uhren liegt in der Tatsache, dass es Reparaturmechanismen gibt, die bei weitem noch nicht verstanden sind. Dadurch kann man keine Aussage über die zeitliche Verzerrung machen, die durch solche Mechanismen ausgelöst werden. Daher wäre eine Eichung der Zeitabstände sehr ungenau. Ein solches Problem tritt z.B. auf, wenn als Grundlage der phylogenetischen Untersuchungen Gensequenzen verwendet werden, die mit vielen Kopien im Genom vorkommen. Bei diesen ist nicht bekannt, wie der Mechanismus funktioniert, mit dem die Zellen die einzelnen Sequenzen gegen den Mutationsdruck identisch halten.

Aus diesen Gründen ist es nur sehr schwer möglich, genaue Aussagen über die zeitlichen Abstände der zu untersuchenden Sequenzen zu machen. Denn eine Umrechnung der Kantenlänge in Zeiteinheiten ohne genaue Kenntnis der Mutationsraten pro Generation ist nicht möglich.

4.2 Horizontaler Gentransfer

Darwin stellte sich einen einzigen universellen Stammbaum aller Organismen der Welt mit fast überall getrennten, geradlinigen Ästen vor. Zwar werden Gene vertikal, d.h. von Generation zu Generation weitergegeben, doch weiß man heute außerdem von *horizontalem Gentransfer* in der Evolution der Zellen. Hierbei geraten Gene – einzeln oder gebündelt – von einer Art in eine andere zur gleichen Zeit lebende. Dieser Prozess wurde beispielsweise bei der Übertragung von Antibiotikaresistenz von Bakterien auf andere Arten bakterieller Erreger beobachtet.

Vorausgesetzt, es gab während der frühen Entwicklungsgeschichte einen horizontalen Gentransfer, würde dies erklären, wieso Eukaryoten viele stoffwechselwichtige bakterielle Gene besitzen, obwohl sie aus einer Archaeen - Zelle hervorgegangen sind. Zusätzlich würde dies begründen, wie eine Vielzahl von Archaeen bakterielle Gene angesammelt haben. Laut dem etablierten Stammbaum müssten bei den Eukaryoten Erbfaktoren der Mitochondrien- bzw. Chloroplasten-DNA sowie Gene, die durch Cyanobakterien in den Zellkern gelangten, bakteriellen Ursprung haben.

Ferner sollten die übertragenen Gene beim Atmungsstoffwechsel bzw. der Photosynthese mitwirken und nicht bei allgemeinen Prozessen, da diese bereits von Genen, die die Archaeen- Vorfahren lieferten geregelt würden.
Im Widerspruch dazu leiten sich allerdings Kern-Gene der Eukaryoten oft von Bakterien ab, statt ausschließlich von Archaeen. Des Weiteren besagt der Standardstammbaum, dass bakterielle Gene nur in Eukaryoten eingegangen seien. Heute weiß man jedoch von vielen Archaeen, die ebenfalls bakterielle Gene besitzen. Aus diesen Überlegungen folgt, dass im universellen Stammbaum die Weiterentwicklung von den Archaeen zu den Eukaryoten zu vereinfacht bzw. sogar falsch dargestellt ist. Richtig wäre eher, dass die Eukaryoten nicht einer Archaeen-Zelle entstammen, sondern einer Vorläuferzelle teils bakteriellem, teils archealem Ursprungs, welche durch horizontalen Gentransfer entstanden ist.

Nach heutigem Wissensstand behielte der universelle Stammbaum die Verzweigung für vielzellige Tiere, Pflanzen und Pilze an der Spitze. Auch die alten Querverbindungen, als Mitochondrien und Chloroplasten der Eukaryoten aus bakterieller Form entstanden, blieben unverändert. Diese Gentransfers würden als Verschmelzen von größeren Ästen erscheinen, wobei unter- und oberhalb bei den Domänen der Bakterien und Archaeen noch viele zusätzliche Vereinigungen von Ästen zu zeichnen wären.
Im Bereich der Prokaryoten und an der Basis der Eukaryoten könnte man nicht entscheiden, welches der Hauptstamm wäre. Allerdings wäre auch dieses Modell nicht wirklichkeitsgetreu, da die verschmelzenden Äste keine Vereinigung ganzer Genome, sondern nur den Transfer einzelner oder mehrerer Gene repräsentieren.
Vor allem aber hat nach heutigem Erkenntnisstand nie eine einzelne Zelle (Linie) existiert, die der letzte gemeinsame Vorfahre genannt werden könnte.

Zusammenfassend lässt sich daher sagen, dass das attraktive Modell des einzigen universellen Stammbaums experimentell getestet wurde, die Ergebnisse jedoch zeigen, dass das Modell eindeutig zu einfach ist. Folglich sind nun neue Hypothesen zur Beschreibung eines Stammbaums des Lebens gefragt.

5. Literaturangaben

1. Doolittle W. F.: **Stammbaum des Lebens**. *Spektrum der Wissenschaft* **04/2000**, 52

2. Sudhaus W.: **Einführung in die Phylogenetik und Systematik**. *Gustav Fischer Verlag*, Stuttgart **1992**

3. Nieselt-Struwe K.: **Phylogenetische Bäume**. Vorlesungsskript, Algorithmen in der Bioinformatik, Universität Tübingen **2001**
   Quelle: *http://www.zbit.uni-tuebingen.de/pas/archiv_algo1.htm*

4. von Öhsen N.: **Phylogenien und Methoden zu ihrer Rekonstruktion**. Seminar Bioinformatik **2000**
   Quelle: *http://cartan.gmd.de/~ralf/Public/Lehre/WS00_01/Ausarbeitungen/Phylogenie-Seminar-Folien_30_10_00.pdf*

5. Schmidt H.: **Parallelisierung phylogenetischer Methoden zur Untersuchung der Crown Group Radiation**. Diplomarbeit, Universität Köln **1996**
   Quelle: *http://www.dkfz-heidelberg.de/tbi/people/hschmidt/publ/diplom/*

6. Merkl R.: **Informatik in der Biologie**. Universität Göttingen **2002**
   Quelle: *http://www-lehre.img.bio.uni-goettingen.de/edv/*

7. **Evolutionary Trees an Perfect Phylogeny**. Zentrum für Bioinformatik, Universität des Saarlandes
   Quelle: *http://www.zbi.uni-saarland.de/zbi/stud/lehrveranstaltungen/ws01/bioinformatikI/materialien/BioinfI8.ppt*

8. Hüllemeier E.: **Bioinformatik: Methodische Grundlagen**. Vorlesungsskript, Universität Marburg **2002**
   Quelle: *http://www.mathematik.uni-marburg.de/%7Eeyke/BioInformatik.shtml*

9. Kierstein G.: **Phylogenetische Entwicklung asiatischer Wasserbüffel anhand Polymorphismen in der mitochondrialen D-loop Region**. Dissertation, Universität Göttingen **2001**
   Quelle: *http://webdoc.gwdg.de/diss/2001/kierstein/kierstein.pdf*

10. Rashidi H., Bühler L. K.: **Grundriss der Bioinformatik**. *Spektrum Akademischer Verlag*, Heidelberg **2001**

11. **PHYLIP – Phylogeny Analysis Workshop**. University of Virgina **1998**
    Quelle: *http://hsc.virginia.edu/achs/documents/ACHS-311.pdf*

12. Felsenstein J.: **PHYILP – Phylogeny Inference Package**. University of Washington **1993**
    Quelle: *http://www.cmbi.kun.nl/bioinf/PHYLIP/*