KONSTRUKTION PHYLOGENETISCHER BÄUME

Seminararbeit zum Vortrag

Seminar:	Bioinformatik
Fachhochschule:	FH Wedel
Seminarleiter:	Prof. Dr. Sebastian Iwanowski
Vortragstermin:	11.12.2012, 11:00 Uhr, HS6
Vortragender:	Timo Jacobs (Winf9122)
Druckdatum:	23.12.2012

INHALTSVERZEICHNIS

ÜBER DIE KONSTRUKTION PHYLOGENETISCHER BÄUME

Phylogenetische Bäume repräsentieren die Evolutionsgeschichte von Spezies als Graphen. Anhand dieser Bäume können Spezies zueinander in eine Beziehung gebracht werden und es kann festgestellt werden, wie weit sie voneinander entfernt sind.

Das Problem bei der Konstruktion besteht darin, dass die exakte Evolution unbekannt ist. Weder konnte sie beobachtet werden, noch weiß man, ob die Ursprungsspezies in dieser Form heute noch existiert. Man kann auch nicht davon ausgehen, dass von jeder älteren Spezies Überreste gefunden werden. Es müssen also Annahmen getroffen werden, anhand derer man dann die Evolutionsgeschichte rekonstruiert.

Weiterhin basiert die Konstruktion von phylogenetischen Bäumen auf Daten aus paarweisen Sequenzabgleichen. Das sind Analysen, die von zwei Spezies jeweils DNA Stränge vergleichen und sie zueinander positionieren. Mit diesen Analysen als Basis können evolutionäre Distanzen gemessen werden und mithilfe von Modellen und statistischen Korrekturen angepasst werden. Diese Analysen werden für jede Paarung von zu analysierenden Daten durchgeführt.

Sind die phylogenetischen Bäume konstruiert, kann man diese weiter analysieren und beurteilen. Hierbei wird zum einen die Konsistenz zu den Daten betrachtet und zum anderen die Güte der Topologie, also der Nachbarschaftsbeziehungen von Spezies.

BASIS: EVOLUTIONÄRE MODELLE UND DISTANZEN

ALLGEMEIN

Wie eingangs erwähnt, werden die Daten aus paarweisen Sequenzabgleichen noch weiter analysiert um den Input für die Konstruktion von phylogenetischen Bäumen zu erhalten. Diese Analysen bestehen aus Distanzmessungen, statistischen Korrekturen und der Anwendung von Formeln, die aus Modellen abgeleitet werden.

Evolutionäre Distanzen beziehen sich auf je zwei Sequenzen. Sie geben ungefähr an, wie viel Mutationen geschehen sind, seitdem sie sich vom gemeinsamen Vorfahren abgespaltet haben. Mit der Anzahl der Mutationen steigt auch die Zeit, die wahrscheinlich vergangen ist. Somit sind die Distanzen auch gewissermaßen eine Zeitmessung.

Evolutionäre Modelle beziehen sich auf den gesamten evolutionären Kontext. Sie beschreiben, wie die Mutationen von Statten gingen und welche Zusammenhänge bei Mutationen bestehen. Aus den Modellen lassen sich Formeln für unterschiedliche Zwecke ableiten. Man kann zwischen Formeln unterscheiden, die man auf Sequenzdaten anwendet, um diese zu modifizieren, damit sie dem Modell entsprechen und dieses besser abbilden und Formeln, die man auf Baumtopologien anwendet, um eine Wahrscheinlichkeit zu erhalten, dass die gegebene Topologie die gesuchten Daten auf diese Weise erzeugt.

P-DISTANCE

Die „p-distance“ ist eine sehr einfache Distanzmessung, die für andere Techniken jedoch eine gute Ausgangslage bietet. Sie wird durch die folgende Formel definiert:

$$p = \frac{unterschiedliche\ Abgleichspositionen}{gesamte\ Anzahl\ Positionen\ im\ Abgleich}$$

Dadurch, dass die „p-distance“ eine sehr einfache Messung ist, treten relativ viele Probleme auf, wenn man ausschließlich die p-distanz als Distanzmessung verwendet. Da die Anzahl an Positionen im Abgleich für gewöhnlich sehr hoch ist, im Vergleich zu den unterschiedlichen Abgleichspositionen, ist die „p-distance“ relativ klein. Gerade bei niedrigen Mutationsraten oder kurzen Zeiten werden sich nicht viele Unterschiede in den Sequenzen ergeben. Dadurch wird die „p-distance“ Messung ungenau. Weiterhin kann die „p-distance“ nur maximal eine Mutation pro Abgleichposition abbilden. Es ist aber durchaus wahrscheinlich, dass manche Positionen sich in dem gegebenen Zeitraum öfters entwickelt haben. Außerdem stehen unterschiedliche Sequenzen zu unterschiedlichen Zeiten unter unterschiedlichen evolutionären Druck. Dies kann ebenfalls nicht durch die „p-distance“ abgebildet werden.

Um diesen Problemen entgegenzuwirken, kann man statistische Korrekturen verwenden. So existieren beispielsweise die „poisson distance correction“, um mehreren Mutationen an derselben Position gerecht zu werden, und die „gamma distance correction“, um unterschiedliche Raten an unterschiedlichen Positionen zu beachten.

JUKES-CANTOR MODELL

Das Jukes-Cantor Modell berücksichtigt, ähnlich der „poisson distance correction“ mehrere Mutationen an derselben Abgleichsposition, jedoch basiert es nicht auf statistischen Messungen, sondern auf chemischen Zusammenhängen.

Es wird angenommen, dass eine konstante Mutationsrate zwischen allen Basenmutationen vorliegt, Mutationen voneinander unabhängig sind, dass die Basenkomposition gleich bleibt und dass die Sequenzlänge identisch bleibt. Die benannte Mutationsrate ist der einzige Parameter dieses Modells. Diese Annahmen können in einer Matrixschreibweise verdeutlicht werden. Die Mutationen treten von der i. Zeile zur j. Spalte auf.

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

Aus diesen Annahmen lassen sich Formeln aufstellen. Durch einige zusätzliche Annahmen, wie der, dass α so klein gewählt wird, dass α^2 als 0 angenommen werden kann, erhält man u.a. eine Formel, die man auf die „p-distance“ anwenden kann:

$$d_{JC} = -\frac{3}{4} * \ln(1 - \frac{4}{3}p)$$

Durch die Basis der Substitutionsmatrix, lässt sich dieses Modell gut erweitern, um so falsche Annahmen zu berichtigen. Eine solche Erweiterung ist das Kimura 2 Parameters-Modell.

KIMURA 2 PARAMETERS

Das Kimura 2 Parameters Modell erweitert das Jukes-Cantor Modells auf zwei Parameter. Diese zwei Parameter sind die Transitions- und Transversionsrate. Eine Transition ist hierbei eine Mutation von einer A-Basis auf eine G-Basis, bzw. eine Mutation von einer C-Basis auf eine T-Basis, oder jeweils

umgekehrt. Diese Parameter werden α und β benannt, wobei α diesmal eine andere Bedeutung halt, als noch im Jukes-Cantor-Modell. α ist eine für alle Positionen identische Transitionsrate und β ist eine für alle Positionen identische Transversionsate. Dieser Sachverhalt lässt sich erneut in einer Substitutionsmatrix darstellen.

	A	C	G	T
A	$-2\beta-\alpha$	β	α	β
C	β	$-2\beta-\alpha$	β	α
G	α	β	$-2\beta-\alpha$	β
T	β	α	β	$-2\beta-\alpha$

Aufgrund der unterschiedlichen Ausgangslage, wird eine etwas unterschiedliche Formel aus dem Modell abgeleitet. P wird hierbei den Transitionsanteil aus dem Sequenzabgleich benennen und Q bezeichnet den Transversionsanteil. Addiert erhält man somit den Anteil an unterschiedlichen Positionen, die „p-distance“.

$$d_{K2P} = -\frac{1}{2} * \ln(1-2P-Q) - \frac{1}{4}\ln(1-2Q)$$

Sowohl beim Jukes-Cantor als auch beim Kimura 2 Parameters Modell wird neben der Annahme, dass die Sequenzlänge identisch bleibt (die Spalten addieren sich auf 0) angenommen, dass die Basenkomposition identisch bleibt (die Zeilen addieren sich auf 0). Da dies nicht immer der Fall ist, gibt es weitere Erweiterungen dieser Modelle, beispielsweise das HKY85-Modell.

OPTIMALITÄTSKRITERIUM FÜR EINE TOPOLOGIE

ALLGEMEIN

Wenn man einen phylogenetischen Baum vorliegen hat, oder einen konstruieren möchte, ist es gut, eine quantitative Messung der Optimalität zu haben. Bei der Konstruktion kann man anhand dieser Messung möglichst gute Resultate erzielen; wenn man einen Baum vorliegen hat, kann man ausgehend von diesem versuchen einen besseren zu finden, denn leider existiert keine Methode, die ohne alle möglichen Topologien zu durchsuchen, den global optimalen Baum finden kann.

Grundsätzlich kann man die Messung der Optimalität mit einer Funktion tätigen. Der Funktionswert wird im Rahmen dieser Ausarbeitung als S betitelt. Es ist jedoch unbestimmt, ob es optimaler ist, einen höheren Funktionswert zu erzielen oder einen niedrigeren. Dies bleibt abhängig on der gewählten Funktion. Es gibt zwei grundsätzliche Prinzipien, nach denen sich die Funktionen einteilen lassen: Das Prinzip der maximalen Sparsamkeit („maximum parimony“) und das Prinzip der maximalen Wahrscheinlichkeit („maximum likelihood“). Ersteres Prinzip besagt, dass eine Topologie gut ist, wenn sie mit möglichst wenigen Mutationen auskommt. Letztere meint, dass eine Topologie gut ist, wenn sie möglichst wahrscheinlich die Sequenzen an den Blättern bilden kann.

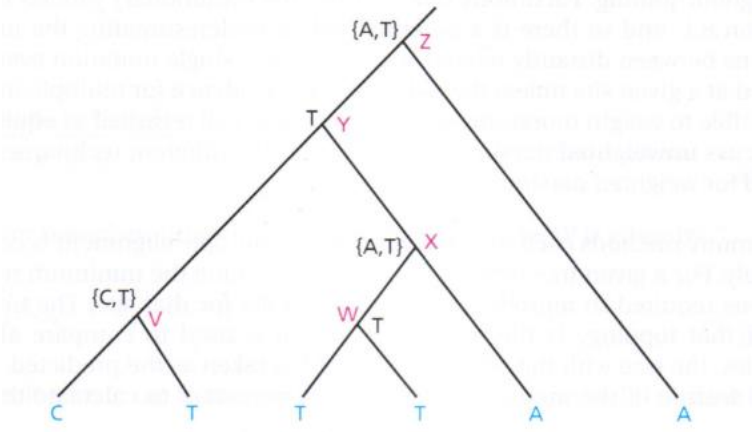
MAXIMUM PARSIMONY

Wie bereits erwähnt, besagt das Prinzip der maximalen Sparsamkeit, dass ein Baum dann optimal ist, wenn es keinen anderen Baum gibt, der weniger Mutationen erlaubt. Hierzu gibt es mehrere Formeln für den Funktionswert S. Eine davon ist die „unweighted parsimony“-Funktion, die nachfolgend erläutert wird.

Die „unweighted parsimony" Funktion berechnet für eine gegebene Topologie, wie viele Mutationen mindestens benötigt werden. Dies wird separat für jede Abgleichsposition getätigt.

Als erster von insgesamt zwei Schritten werden die Positionen eines mehrfachen Sequenzabgeiches untersucht. Sind alle Basen einer Position identisch, handelt es sich um eine „invariable site". Diese ist für eine Bewertung der Topologie irrelevant und uniformativ, da in keiner Topologie eine Mutation notwendig wäre. Es existiert darüber hinaus eine weitere Art der uninformativen Position: die sogenannten „singleton sites". Hierbei handelt es sich um Positionen, in denen nur eine Basenart öfters als einmal auftaucht und die anderen entsprechend maximal ein Mal. Auch hier wird unabhängig von der Topologie immer die gleiche Anzahl an Mutationen benötigt. Schließt man diese beiden Sorten von uninformativen Positionen von der Berechnung aus, kann man die Berechnungszeit insofern verbessern, als dass der nachfolgende Algorithmus nichtmehr für jede Position des Abgleichs durchgeführt werden muss. Der zweite Schritt besteht aus der sogenannten „post order" Traversion.

Bei der „posit order" Traversion wird berechnet, wie viele Mutationen die entsprechende Topologie benötigt, um die Basen an der informativen Position zu generieren. Hierzu haben alle Konten eine Menge von Basen zugewiesen. An den Blättern stehen einelementige Mengen, in denen nur die Basis selber steht. Die internen Basenmengen werden während der Traversion aufgebaut. Haben zwei Nachbarn eine gemeinsame Schnittmenge, die nicht leer ist, wird diese Schnittmenge dem Vorfahrenknoten zugewiesen. Der globale Mutationsanzahlzähler wird nicht erhöht, da keine Mutation notwendig ist, auf die Nachfahrenpositionen zu gelangen. Ist die Schnittmenge leer, wird der globale Mutationszähler um eins erhöht und dem Vorfahrenknoten wird die Vereinigungsmenge zugewiesen. Bei dem nachfolgenden Beispiel werden mindestens 3 Mutationen benötigt.



Die berechnete minimale Mutationszahl bildet den Funktionswert S und sollte minimiert werden. Diese Methode der Funktionswertberechnung hat eine Besonderheit gegenüber den meisten anderen: es werden keine Annahmen über Astlängen getroffen und diese werden auch nicht berechnet. Weiterhin werden alle Mutationsformen gleich schwer gewichtet – mit 1. Es gibt Erweiterungen dieser Methode, die diese Eigenarten entfernen.

MAXIMUM LIKELYHOOD

Das Prinzip der maximalen Wahrscheinlichkeit besagt, dass eine Topologie dann gut ist, wenn sie sehr Wahrscheinlich eintritt. Die Funktion der maximalen Wahrscheinlichkeit ist abhängig von einem

evolutionären Modell, welches eine Funktion ergibt, die die Wahrscheinlichkeit berechnet, dass eine Basis i zu einer Basis j in einer Zeit t mutiert. Schreiben kann man diese Funktion als $P(j|i,t)$. Der Funktionswert wird nun für alle möglichen Basenbelegungen der internen Knoten berechnet. So erhält man eine Gesamtwahrscheinlichkeit der Topologie, welche als Funktionswert S angesehen werden kann, den es zu maximieren gilt.

KONSTRUKTIONSMETHODEN INITIALBAUMTOPOLOGIE

UPGMA

Die UPGMA-Methode („unweighted pair-group method using arithmetic averages“) ist, 1985 entwickelt, eine der ältesten Methoden zur Konstruktion von phylogenetischen Bäumen. Als eine Clustermethode hat sie den Vorteil sehr schnell zu sein und somit auch auf große Datenmengen anwendbar zu sein. Ein großer Nachteil ist allerdings, dass die Daten passen müssen, denn diese Methode nimmt an, dass die Entwicklungen unter den Bedingungen einer konstanten molekularen Uhr stattfanden. Das bedeutet, dass Mutationen immer zu bestimmten Zeitpunkten für alle Spezies stattfanden und somit, dass alle Blätter des erzeugten Baumes dieselbe Entfernung zur Wurzel haben. Weiterhin ist der erzeugte Baum ultrametrisch, das bedeutet, dass bei drei Sequenzen mindestens zwei Distanzen zwischen ihnen gleich sind und die dritte, sofern sie nicht auch gleich ist, kürzer sein muss. Als weitere Eigenschaft des resultierenden Baumes gilt, dass dieser eine Wurzel hat, was bei den noch folgenden Konstruktionsmethoden nicht der Fall ist.

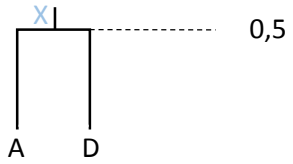
Die Methode hat eine weitere große Schwachstelle: sie arbeitet nicht mit dem Optimalitätskriterium, welches im vorherigen Abschnitt vorgestellt wurde.

Zunächst wird bei dieser Methode festgestellt, welche Distanz in den Eingangsdaten die kürzeste ist. Damit das Kriterium der Ultrametrik eingehalten wird, identifiziert diese Distanz die ersten Nachbarn der entstehenden Topologie. Die identifizierten Nachbarn bilden zusammen einen neuen Cluster, dessen Distanzen zu den übrigen Sequenzen und Clustern anhand von arithmetischen Mitteln berechnet wird. Diese Schritte wiederholen sich, bis die komplette Topologie identifiziert ist. Die Astlängen zwischen den einzelnen Knoten berechnen sich einfach anhand der Distanzen. Da die Distanz zu anderen Sequenzen bei den beiden Sequenzen aufgrund der Ultrametrik identisch ist, muss auch die Distanz zum internen Vorfahrenknoten identisch sein. Deshalb wird die Distanz, die durch die Eingabedaten festgelegt ist, halbiert und als Distanz zum internen Vorfahrenknoten festgelegt.

Ein Beispiel veranschaulicht dieses Vorgehen. Die Distanzen zwischen den Sequenzen werden in folgender Matrix veranschaulicht:

d_{ij}	**A**	**B**	**C**	**D**	**E**	**F**
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

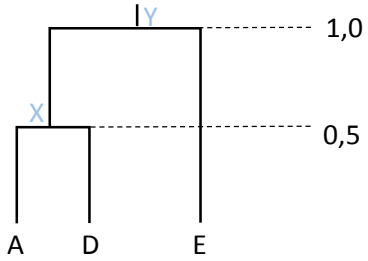
Die kleinste Distanz ist d_{AD}, weshalb A und D das erste, als Nachbarn identifizierte Paar sind. Sie werden zu dem neuen Cluster X zusammengefasst. Die Distanz zum internen Knoten beträgt 0,5, da die Distanz untereinander 1 ist. Somit sieht der Baum bislang wie folgt aus:



Die Distanzen vom neuen Cluster X zu den restlichen Sequenzen werden durch das Arithmetische Mittel der Distanzen von A und D zum Rest berechnet. Es folgt eine neue Matrix mit Distanzen:

d_{ij}	X	B	C	E	F
X	-	6	8	2	6
B		-	8	6	4
C			-	8	8
E				-	6

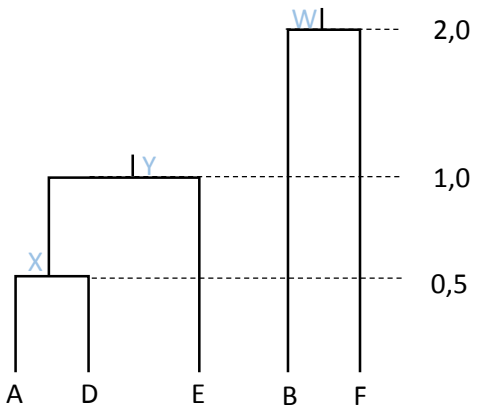
Erneut wird das nächste Nachbarpaar durch die geringste Distanz gewählt, also X und E. Die Evolutionäre Höhe, auf der der neue Cluster Y liegt, ist 1 (die Hälfte von 2). Aktuell ergibt sich der folgende Baum:



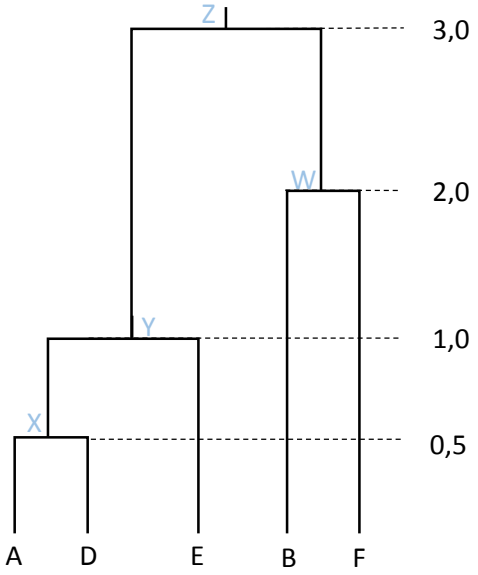
Wie im vorigen Schritt werden nun die Distanzen vom internen Knoten des Clusters Y zu den restlichen Sequenzen durch das arithmetische Mittel berechnet. Nachfolgend werden die weiteren

Distanzmatrizen und Teilbäume bis zum kompletten phylogenetischen Baum aufgezeigt. Auf weitere Erläuterungen, die demselben Schema folgen würden, wird verzichtet.

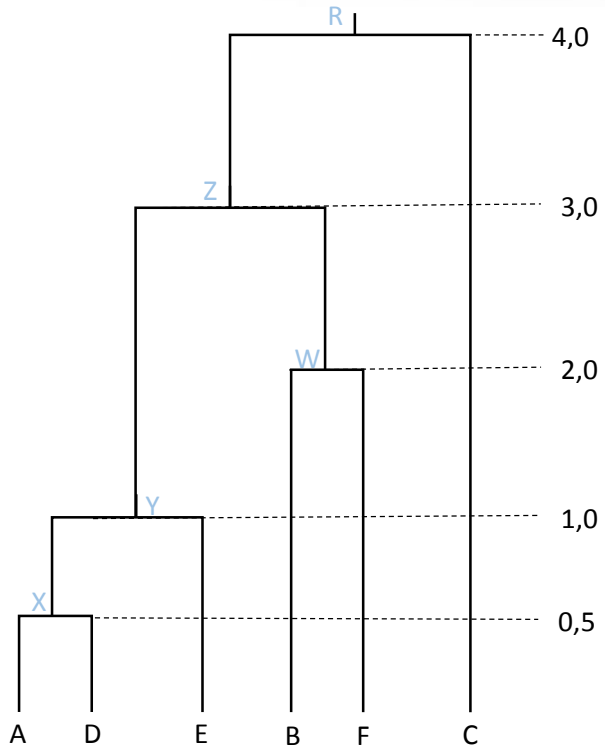
d_{ij}	Y	B	C	F
Y	-	6	8	6
B		-	8	4
C			-	8



d_{ij}	Y	C	W
Y	-	8	6
W		-	8



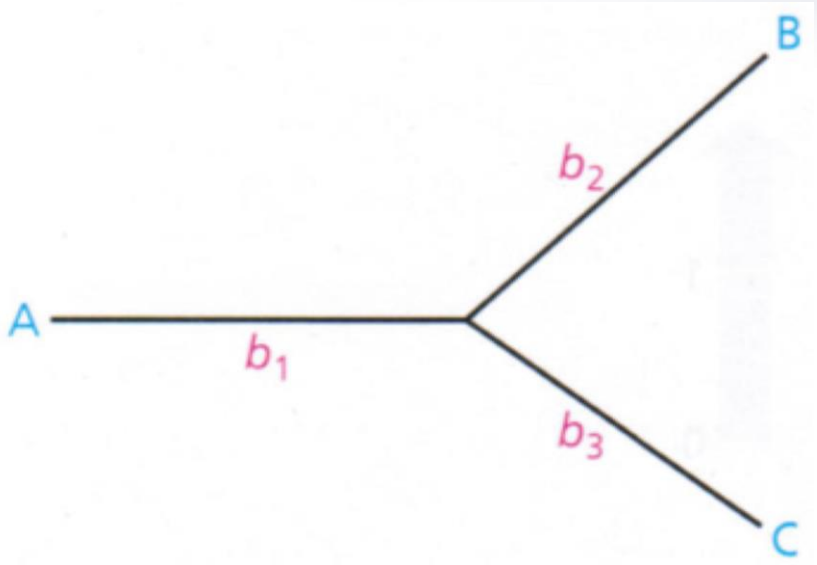
d_{ij}	Z	C
Z	-	8



FITCH-MARGOLIASH

Die Fitch-Margoliash Methode zur Baumkonstruktion unterscheidet sich insofern von der UPGMA-Methode, als dass die keinen ultrametrischen Baum erzeugt, sondern nur einen additiven. Es wird keine konstante molekulare Uhr vorausgesetzt. Außerdem hat der erzeugte Baum keine Wurzel. Die Topologie hingegen bleibt die gleiche, weil die Art und Weise, wie das nächste Nachbarpaar ausgewählt wird die gleiche ist. Es ändern sich bloß die Astlängen. Die Fitch-Margoliash Methode kann auch zur Berechnung von Astlängen für gegebene Topologien genutzt werden.

Wie bereits erwähnt, ist bei der Fitch-Margoliash Methode die Art und Weise der Auswahl des nächsten Nachbarpaares identisch, das nächste Nachbarpaar wird also durch die kürzeste Distanz bestimmt. Ist das Nachbarpaar identifiziert, müssen de Astlängen zum verbindenden internen Knoten berechnet werden. Hierzu wird ein Baum mit drei Blättern als Ausgang für die Ableitung von Formeln angenommen.



Zwei der drei Blätter sind die Sequenzen des Nachbarpaares. Das dritte Blatt ist ein Cluster aus dem Rest der Sequenzen. Wie die folgenden Formeln zeigen werden, werden die Distanzen zwischen allen drei Blättern benötigt. Die Distanz zwischen den Nachbarsequenzen lässt sich aus den Eingabedaten Ablesen. Die Distanz zum Restcluster wird durch ein arithmetisches Mittel der Distanzen zu den Elementen des Clusters gebildet. Die Formeln für die Kanten, welche durch die Additivität des Baumes entstehen, lauten:

$$b_1 = \frac{1}{2}(d_{AB} + d_{AC} - d_{BC})$$

$$b_2 = \frac{1}{2}(d_{AB} + d_{BC} - d_{AC})$$

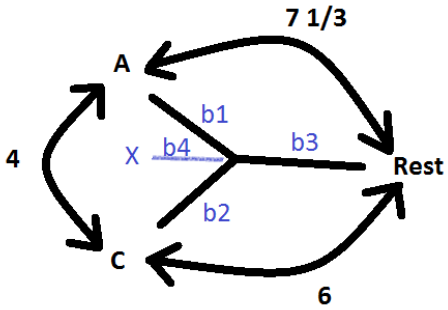
$$b_3 = \frac{1}{2}(d_{AC} + d_{BC} - d_{AB})$$

Weiterhin wird eine vierte Distanz benötigt, um den durch das Nachbarpaar entstehenden Cluster als „Sequenz“ nutzen zu können. Diese Distanz wird im Folgenden b_4 genannt und ist das Mittel aus den beiden Ästen des Nachbarpaares. Wird solch ein Cluster später als ein Nachbar identifiziert, muss bedacht werden, dass die vorher berechnete Astlänge b_4 von der neu berechneten Länge b_1 abgezogen werden muss. Ein Beispiel wird diesen Sachverhalt verdeutlichen.

Als Ausgangdaten für das Beispiel sollen die folgenden gelten:

d_{ij}	A	B	C	D	E
A	-	5	4	9	8
B		-	5	10	9
C			-	7	6
D				-	7

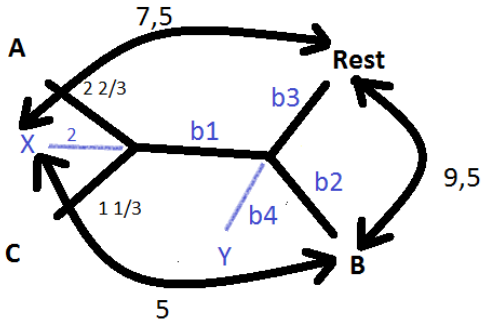
Wie bei der UPGMA-Methode wird die geringste Distanz als erster Nachbar genommen, also A und C. Die Distanz von A zum Rest-Cluster beträgt $\frac{5+9+8}{3} = 7\frac{1}{3}$ und die Distanz von C zum Rest beträgt $\frac{5+7+6}{3} = 6$. Daraus folgt folgende Ausgangslage für die Berechnung:



Mit den oben genannten Formeln lassen sich nun die Astlängen berechnen. Interessanter wird es, wenn der neue Cluster X als Nachbar identifiziert wird. Zunächst müssen jedoch neue Distanzen von dem Cluster X berechnet werden. Auch dies erfolgt durch arithmetische Mittel. Die neuen Distanzen lauten wie folgt:

d_{ij}	X	B	D	E
X	-	5	8	7
B		-	10	9
D			-	7

Als nächstes muss also das Cluster X mit der Sequenz B als Nachbar definiert werden. Erneut wird ein Rest-Cluster gebildet. Die Benennungen b_1 bis b_4 werden im Folgenden neu vergeben.

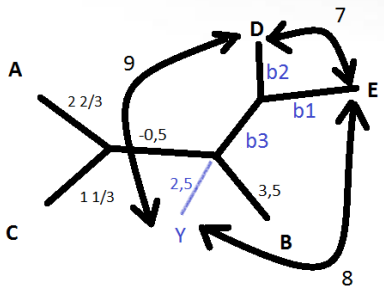


An dieser Stelle muss man beachten, dass b_1 nicht wie vorher die gesamte Stecke bis zum Cluster X bezeichnet, sondern bereits beim internen Knoten endet. Berechnet wird b_1 deshalb wird folgt:

$$b_1 = \frac{1}{2}(7{,}5 + 5 - 9{,}5) - 2 = -0{,}5$$

Man erkennt direkt, dass dieses Ergebnis falsch sein muss, da eine evolutionäre Entwicklung nicht über einen negativen Zeitraum stattfinden kann. Die durch die Fitch-Margoliash Methode erzeugte Topologie kann deshalb falsch sein. Bei den weiteren Berechnungsschritten tritt kein erneuter Fehler auf. Diese werden nachfolgend der Vollständigkeit halber dargestellt.

d_{ij}	Y	D	E
Y	-	9	8
D		-	7



$$b_1 = \frac{1}{2}(7 + 8 - 9) = 3$$

$$b_2 = \frac{1}{2}(7 + 9 - 8) = 4$$

$$b_3 = \frac{1}{2}(8 + 9 - 7) - 2{,}5 = 2{,}5$$

Wie UPGMA hat auch diese Methode bei der Konstruktion kein Optimalitätskriterium, auf das sie hinarbeitet. Deshalb werden auch bei dieser Methode weitere Optimierungen notwendig sein, wenn es notwendig ist, solch einem Optimalitätskriterium gut zu entsprechen. Die folgenden Methoden haben diesen Mangel nicht.

STEPWISE ADDITION

Bei der schrittweisen Addition handelt es sich um eine allgemeinere Methode zur Konstruktion von Bäumen. Hierbei wird der Funktionswert S bereits bei der Konstruktion mit einbezogen.

Diese Methode startet mit einem kleinen Initialbaum und fügt in jedem Schritt eine Sequenz hinzu. Das Hinzufügen der Sequenz wird an jeder verfügbaren Position versucht und jeweils wird der Funktionswert S berechnet. Die Stelle, an der er Funktionswert optimal ist, wird übernommen und der Baum wird als Ausgangslage für das Hinzufügen einer neuen Sequenz genommen.

Die Wahl der Reihenfolge des Hinzunehmens von Sequenzen ist entscheidend für das Ergebnis und kann variieren.

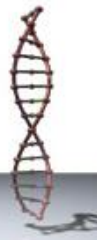
STAR DECOMPOSITION

ALLGEMEIN

Bei der „star decompostion“-Methode handelt es sich um einer der „stepwise addition“ Methode sehr ähnlichen Art und Weise phylogenetische Bäume zu erzeugen. Jedoch wird hier nicht Schrittweise eine Sequenz hinzugenommen. Ausgangslage bildet ein Stern, also ein interner Knoten, mit dem alle Sequenzen verbunden sind. Schrittweise wird nun ein Paar erkannt und abgespaltet. Die Abspaltung wird nun als Cluster erneut an dem Stern verbunden. So wird mit jedem Schritt die Menge an Verbindungen um 1 sinken. Die Wahl des nächsten Paares erfolgt anhand des Optimalitätskriteriums S.

NEIGHBOR-JOINING

Bei der sehr bekannten „Neighbor-Joining“-Methode handelt es sich um eine Art der „star decomposition“. Es ist eine Methode der maximalen Sparsamkeit. Das genutzt Optimalitätskriterium ist allerdings nicht die bereits vorgestellte „unweighted parsimony“-Funktion. Die genutzte Funktion ist einfacher und verzichtet bei der Implementierung auf die Kenntnisname der genauen Abgleichspositionen. Dadurch ist diese Art der Funktionswertberechnung auch effizienter. Der Funktionswert S ist die Summe aller Astlängen des Baumes. Diese Summe gilt es zu minimieren.



Weiterhin konnte gezeigt werden, dass es genügt bei der Auswahl des nächsten Nachbarpaares einen Funktionswert δ_{ij} minimal auszuwählen. Dies optimiert das Laufzeitverhalten dieser Methode weiter.

Der Funktionswert δ_{ij}berechnet sich wie folgt:

$$\delta_{ij} = d_{ij} - \frac{U_i + U_j}{N-2}$$

mit

$$U_i = \sum_{i=1}^{N} d_{1i}$$

Generell werden nachfolgend der interne Knoten des Ausgangssternes X genannt und der neu abgespaltete Knoten wird mit Y gekennzeichnet. Diese Kennzeichnungen sollten nicht mit den Benennungen der Cluster verwechselt werden. Weiterhin werden noch die Formeln zur Bestimmung der Astlängen und zur Berechnung der Distanzen des neuen Clusters(Y) benötigt:

$$b_{iY} = \frac{1}{2}\left(d_{ij} + \frac{U_i - U_j}{N-2}\right)$$

$$b_{jY} = d_{ij} - b_{iY}$$

$$d_{Yk} = \frac{1}{2}\left(d_{ik} + d_{jk} - d_{ij}\right)$$

Auch für die „Neighbor-Joining“ Methode folgt ein Beispiel. Die Ausgangsdaten sind:

d_{ij}	A	B	C	D	E
A	-	5	4	9	8
B		-	5	10	9
C			-	7	6
D				-	7
E					-

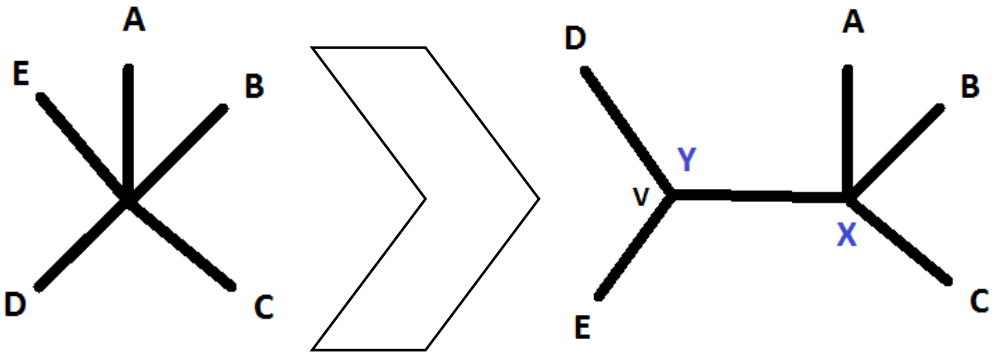
Es müssen nun für jede Sequenz i die Werte U_i berechnet werden und für jedes Sequenzpaar i und j die δ_{ij}. Nachfolgend wird aus Rundungsgründen nicht δ_{ij} sondern $(N-2)\delta_{ij}$ berechnet. Die ersten Berechnungsergebnisse lauten:

d_{ij}	A	B	C	D	E
A	-	5	4	9	8
B		-	5	10	9
C			-	7	6
D				-	7
E					-

U_i
26
29
22
33
30

$3\delta_{ij}$	A	B	C	D	E
A	-	-40	-36	-32	-32
B		-	-36	-32	-32
C			-	-34	-34
D				-	-42
E					-

Hieraus erkennt man, dass das erste abgespaltete Nachbarpaar D und E ist. Aus einem Sternbaum wird demnach ein Paar als Cluster V abgespalten. Folgende Abbildungen sollen diesen Sachverhalt verdeutlichen.



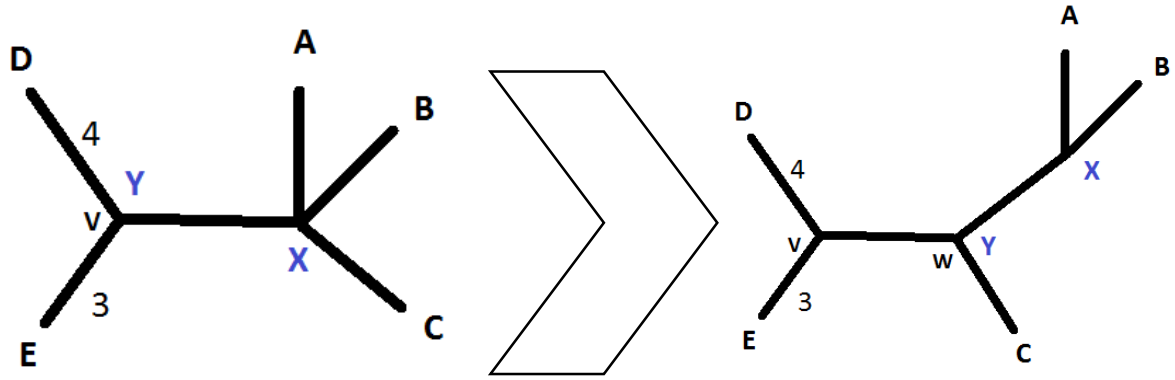
Als nächstes werden, nachdem die Astlängen durch oben beschriebene Formeln berechnet wurden, die Distanzen vom Cluster V zu den Sequenzen A, B und C berechnet und erneut die δ_{ij} ausgewertet. Die Ergebnisse lassen sich in den nachfolgenden Tabellen finden.

d_{ij}	V	A	B	C
V	-	5	6	3
A		-	5	4
B			-	5
C				-

U_i
14
14
16
12

$2\delta_{ij}$	V	A	B	C
V	-	-18	-18	-20
A		-	-20	-18
B			-	-18
C				-

Den niedrigsten Wert in der rechten Tabelle hat das Paar A und B bzw. V und C. Das nächste Paar kann beliebig zwischen ihnen gewählt werden. Im Nachfolgenden wurde das Paar V und C gewählt.

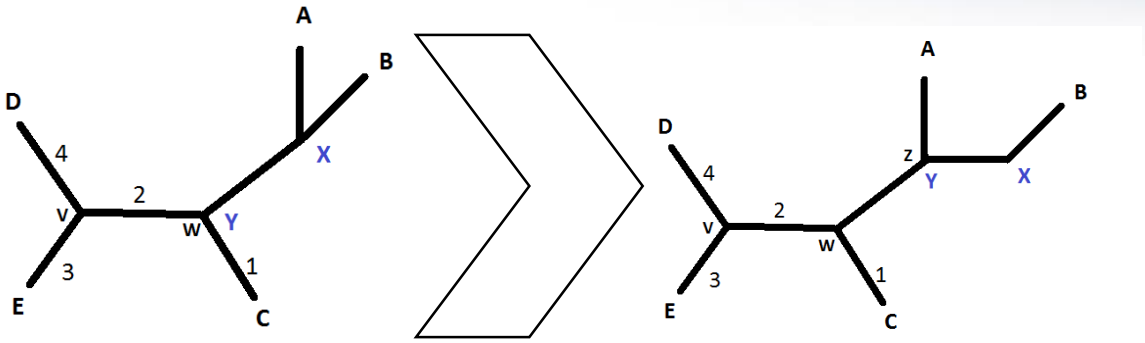


Die nächsten Schritte folgen demselben Prinzip. Nachfolgend werden nur noch die Resultate dargestellt.

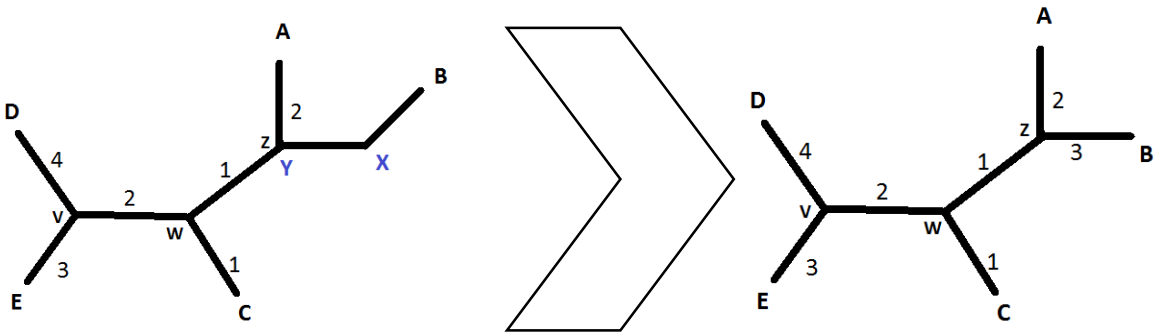
d_{ij}	W	A	B
W	-	3	4
A		-	5
B			-

U_i
7
8
9

δ_{ij}	W	A	B
W	-	-12	-12
A		-	-12
B			-



d_{ij}	Z	B
Z	-	3
B		-



METHODEN ZUR ERKUNDUNG ÄHNLICHER TOPOLOGIEN

ALLGEMEIN

Sämtliche bisher vorgestellten Methoden können nur ein lokales Optimum finden, um ein Globales Optimum zu finden u können, müssen durch Änderungen an der Topologie sich ähnelnde Bäume untersucht werden, ob diese bessere Funktionswerte S ergeben. Branch-Swapping ist in der Lage aus einer Topologie ähnliche Bäume zu erstellen. Dennoch geben auch diese Methoden nicht die Garantie das globale Optimum zu finden, sie erhöhen lediglich die Wahrscheinlichkeit herfür.

BRANCH-SWAPPING

NEAREST-NEIGHBOR INTERCHANGE (NNI)

Ein Interner Ast bildet immer eine Verbindung von zwei Nachbarpaaren, also vier Teilbäumen. Tauscht man diese Teilbäume untereinander, kann man drei verschiedene Topologien bilden. Die Änderungen sind somit minimal, können jedoch an diversen internen Ästen durchgeführt werden. Je öfter das NNI angewendet wird, desto größer sind die möglichen Änderungen und desto größer wird die Wahrscheinlichkeit, ein globales Optimum zu finden.

SUBTREE PRUNING AND REGRAFTING (SPR)

Beim SPR handelt es sich um eine Methode, die einen internen Ast entfernt, sodass zwei Teilbäume entstehen, die jeweils eine Verzweigung haben, in der eine Verbindung fehlt. Bei einem Teilbaum wird nun diese Verbindung entfernt. Die verbleibende Verbindung des zweiten Teilbaumes wird nun an alle möglichen Stellen des ersten Teilbaumes getestet. Diese Methode kann größere Veränderungen als das NNI realisieren.

Tree Bisection and Reconnection (TBR)

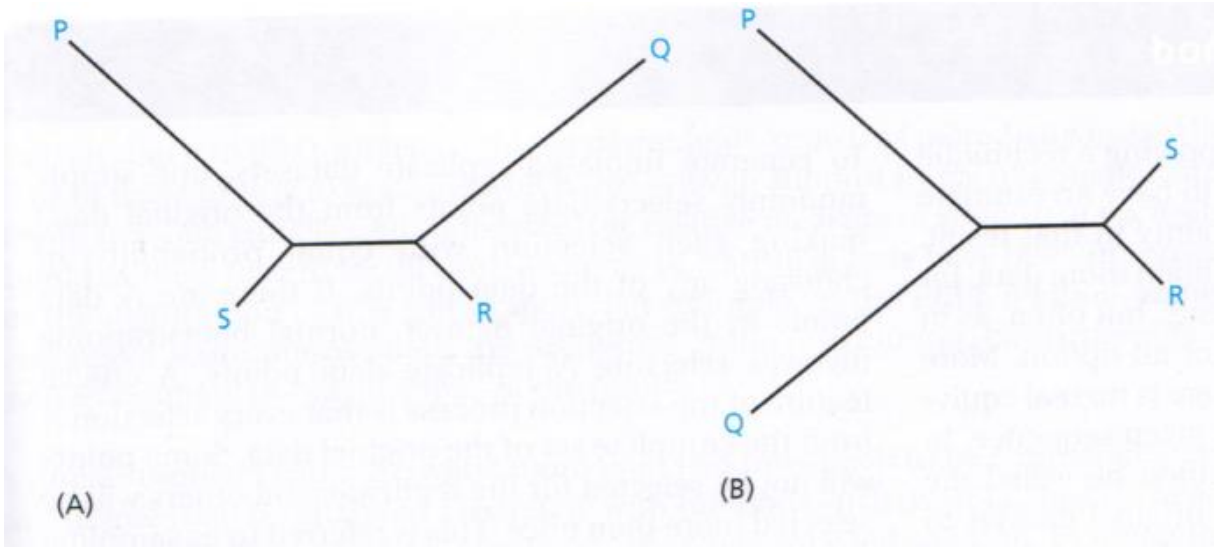
Die TBR-Methode ähnelt der SPR Methode stark. Der Unterschied ist, dass keine Verzweigung ohne Verbindung beibehalten wird und alle möglichen neuen Verbindungen probiert werden. Diese Methode lässt die größten Änderungen an der Topologie zu.

DIE WAHRSCHEINLICHKEIT EINES KORREKTEN ERGEBNISSES

ALLGEMEIN

Auch bei der korrekten Anwendung der vorgestellten Methoden ist es immer noch gut möglich, dass das Ergebnis falsch ist. Um die Ergebnisse zu prüfen fehlt es allerdings aufgrund des hohen Aufwandes der Stichprobenentnahme an alternativen Eingabedaten. Diesem Problem wird die Bootsrapping Analyse Herr.

Ein Problem, das in den Bäumen auftreten könnte, ist die sogenannte „long-branch attraction“. Hierbei handelt es sich um ein Phänomen, das auf unterschiedlichen Wegen ausgelöst werden kann. Insbesondere das Prinzip der maximalen Sparsamkeit verleitet dazu, eine Art „Außenseitergruppe“ zu bilden, also Gruppen, in denen sich sehr lange Astlängenhäufen. Eine Sinnvolle Unterbringung in den bisherigen Daten fehlt. Die folgende Abbildung veranschaulicht dies noch einmal.



A ist hierbei der „richtige“ phylogenetische Baum. B wurde durch das Prinzip der maximalen Sparsamkeit erzeugt.

BOOTSTRAPPING ANALYSE

Die Bootstrapping Analyse ist ein statistisches Werkzeug, durch welches das Problem umgangen wird, dass erneut eine Stichprobe aus der Grundgesamtheit zu ziehen ist. Ziel ist es, eine zweite Stichprobe zu erhalten, die die selbe Verteilung hat, wie die Grundgesamtheit. Ausganglage ist, dass man bereits eine Stichprobe vorliegen hat, die analysiert wurde. Diese Analysen möchte man gerne bestätigen. Deshalb wird eine zweite Analyse auf der gleichen Verteilung benötigt. Dies jedoch ist mit einem hohen Aufwand verbunden. Um dieses Problem zu lösen, wird eine Stichprobe aus der Stichprobe gezogen. Wenn man dies jedoch „ohne“ zurücklegen tätigt, hat man entweder eine Teilmenge mit einer u.U. falschen Verteilung oder dasselbe Ergebnis. Deshalb wird es ausdrücklich gewünscht, dass Duplikate in der neuen Stichprobe möglich sind. So kann man die Verteilung der ursprünglichen

Stichprobe imitieren und kann durch Durchführen derselben Analysen gute Vergleichsergebnisse erzielen.

FAZIT

Im Zuge dieses Seminarvortrages wurden viele verschiedene Möglichkeiten vorgestellt, phylogenetische Bäume zu erzeugen, diese zu beurteilen und zu optimieren. Allein die Menge an unterschiedlichen Methoden zeigt, dass es noch keine wirklich zufriedenstellende gibt, phylogenetische Bäume zu rekonstruieren. Weitere Forschungen im Bereich der Konstruktion von phylogenetischen Bäumen haben demnach noch viel Raum und können noch viele spannende Ergebnisse erwarten lassen.

QUELLEN- UND LITERATURVERZEICHNIS

Die einzige Quelle und Literatur für diesen Seminarvortrag und die Ausarbeitung bildet das folgende Buch:

Autoren	Marketa J. Zvelebil Jeremy O. Baum
Titel	understanding bioinformatics
Erscheinungsjahr	2007
Verlag	Taylor & Francis Ltd.
ISBN-13	978-0815340249
ISBN-10	0815340249