

# SplitsTree and Phylogenetic Networks

Betreuer:  
Tobias Klöpper

## **Inhaltsverzeichnis**

1.	Einleitung .....	3
2.	Theorie.....	4
2.1	Evolutionäre Verbindungen in Netzwerken .....	4
2.2	Die „Split De-composition“ Theorie.....	5
2.3	Buneman Bäume .....	6
2.4	Split decomposition.....	7
2.5	Von schwach kompatiblen Splits zu Netzwerken .....	9
3.	Anwendung.....	11
3.1	Das SplitsTree Programm .....	11
3.2	Beispiel: mtDNA Datensatz .....	11
3.3	Beispiel 2: HIV-1 Datensatz .....	13
4.	Quellenangaben: .....	15

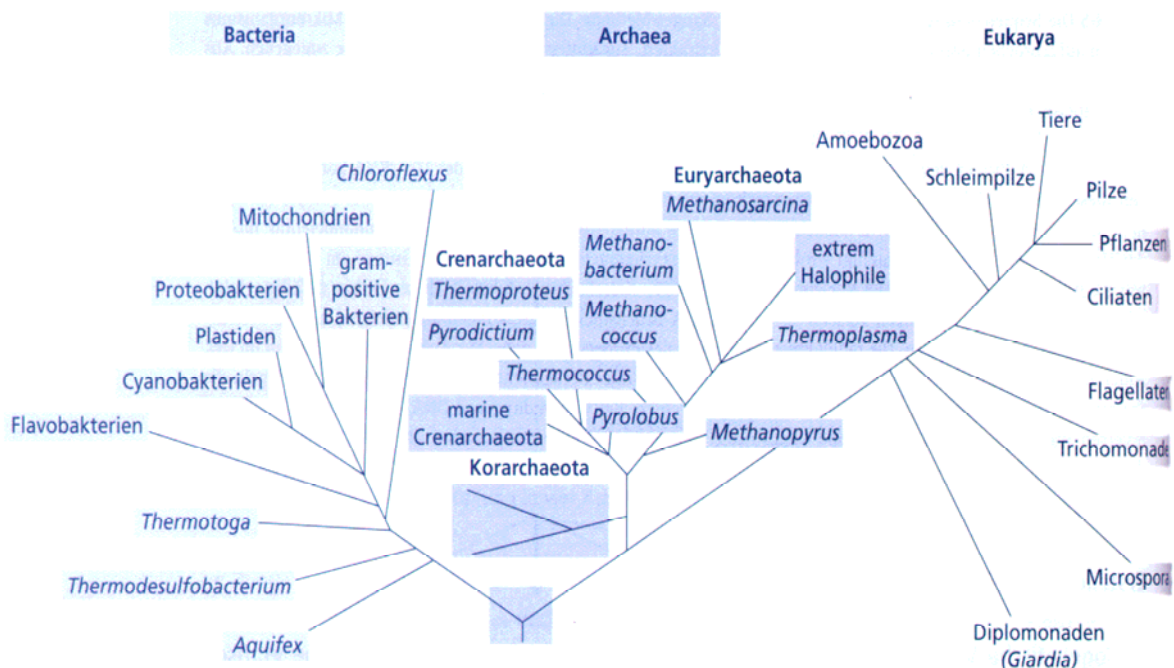
# 1. Einleitung

In den vergangenen Jahrzehnten ist man, nach der Entdeckung der DNA, immer mehr dazu übergegangen Organismen nicht nur anhand ihrer phänotypischen Eigenschaften sondern auch anhand ihres Genotyps zu vergleichen. Mittlerweile gibt es einige gute Verfahren die die Ähnlichkeit und den Verwandtschaftsgrad zweier oder auch mehrerer Organismen bestimmen. So ist die Maus genetisch dem Menschen sehr ähnlich und eignet sich damit auch als Forschungsobjekt.

Um diese komplexen Verwandtschaften nun auch graphisch übersichtlich darzustellen, benötigt man ausgereifte mathematische Verfahren.

Ein Programm, das einige dieser Verfahren, die aus einem gegebenen Datensatz einen graphisch übersichtlichen Zusammenhang liefern, ist SplitsTree (Huson 1998), welches, wie der Name schon sagt, aus einer gegebenen Datenmenge einen Phylogenetischen Baum oder Netzwerk aufbaut. Diese Phylogenetischen Netzwerke können zur visuellen Analyse der erhaltenen Daten genutzt werden.

SplitsTree bietet die Möglichkeit Bäume, ähnlich dem unten abgebildeten Beispiel, oder Netzwerke über eine beliebige Eingabe an Taxa und den damit verbundenen Daten zu erstellen.



## 2. Theorie

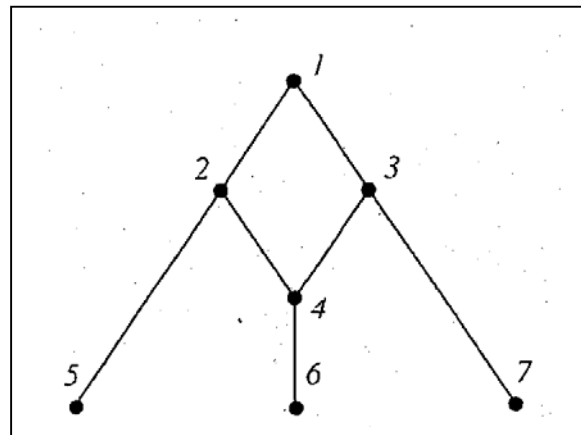
### 2.1 Evolutionäre Verbindungen in Netzwerken

Der klassische Weg evolutionäre Zusammenhänge eines gegebenen Datensatzes an Taxa zu veranschaulichen ist ein binärer Baum, hierbei sind interne Knoten als mögliche Vorfahren dargestellt und die Blätter stellen die aktuell existierenden Taxa dar.

Für den Fall, dass die verwandtschaftlichen Zusammenhänge gar keinen Baum bilden bei dem es immer nur genau 2 Nachfahren gibt, wäre ein Baum mit einer unbestimmten Anzahl Ästen je Knoten ein adäquates Mittel.

Aber selbst dieser Fall ist in der Biologie noch nicht allgemein genug. Als Beispiel sei hier die Interaktion von Bakterien genannt bei denen es innerhalb einer Generation zu Hybridisierungen und Rekombinationen kommen kann. Ein Baum eignet sich hierbei nur bedingt um die vollständigen Beziehungen korrekt darzustellen, da ein Baum unter der Bedingung aufgebaut wird, dass einmal getrennte Äste später nicht mehr zusammen geführt werden oder interagieren.

Dieser Fall kann, wie in Abb. 2.1, vereinfacht dargestellt werden. Hierbei werden die Knoten 1, 2, 3, 4 als Vorfahren und die Blätter 5, 6 und 7 als real existierende Taxa betrachtet. Wie bei einem Baum mit einer Wurzel geht man hierbei davon aus,



(Abb. 2.1)

dass 1 den Ursprungsknoten darstellt. Der Unterschied zwischen diesem Netzwerk und einem normalen Baum ist, dass es hier zu einem Ringschluss der Knoten 1-4 kommt. Derartige Netzwerke eignen sich nicht nur für spezielle Arten von Evolution, wie der im obigen Beispiel genannten Rekombination von Bakterien, sondern können in all jenen Fällen verwendet werden wo es unangebracht ist Daten in eine Baumstruktur zu zwingen. Es gibt zwar auch bei anderen Programmen als SplitsTree die Möglichkeit sich Daten in verschiedenen Arten von Bäumen anzeigen zu lassen aber dennoch kann es vorkommen, dass keiner dieser Bäume die Zusammenhänge korrekt wiedergibt. Es mag sogar soweit kommen, dass erst in einem Netzwerk in dem Ringschlüsse erlaubt sind die eigentliche Struktur der Evolution anschaulich und begreifbar wird. Ein Beispiel hierfür wäre der Gebrauch von Netzwerken zur „Phylogenetischen Analyse“ der *Canterbury Tales* (Barrbook et. Al. 1998.)

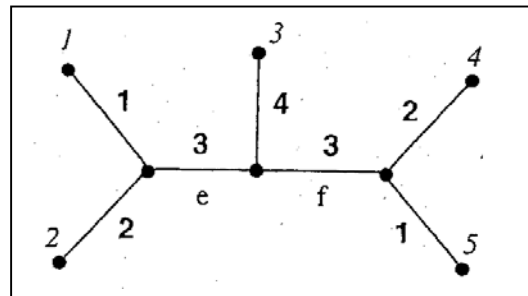
Die Frage die sich nun stellt ist, welche Netzwerke es gibt und für welche Arten von Daten sie geeignet sind. So werden zum Beispiel für die Darstellung der Evolution von mtDNA häufig median Netzwerke benutzt. Wir konzentrieren uns hier jedoch auf eine spezielle Art des Zugangs zur Phylogenetischen Analyse,

dem SplitsTree Programm (Huson 1998). Die hierbei erzeugten SplitGraphen basieren hauptsächlich auf Distanzen die mit der Split-decomposition Theorie errechnet wurden (Bandelt, Dress 1992/1993). Dieser Theorie widmen wir uns nun im folgenden Kapitel.

Weitere Beschreibungen hierzu findet man auch in Dress, Huson, Multon (1996), Page, Holmes (1998), und Swafford et. Al. (1996).

## 2.2 Die „Split De-composition“ Theorie

Der wichtigste Punkt der Split de-composition Theorie ist, dass ein Netzwerk in sogenannte Splits zerlegt werden kann. Würde man z.B in dem in Abb. 2.2 dargestellten Baum ( $T_X$ ) eine beliebige Kante entfernen, so erhielte man 2 disjunkte Teilbäume A und B. Entfernte man beispielsweise die Kante f so erhielte man die Bipartitionen  $A=\{1,2,3\}$  und  $B=\{4,5\}$ . Wie man sieht induziert hierbei jede Kante genau einen Split. Die Menge aller durch die Kanten erzeugten Splits wird  $\Sigma(X)$  genannt, hierbei ist  $|\Sigma|$  genau die Anzahl der Kanten des Baumes.



(Abb. 2.2)

**Zwei Splits  $U=\{A,B\}$  und  $V=\{K,L\}$  heißen kompatibel falls gilt:**

$$\exists ! \emptyset \in \{A \cup K, A \cup L, B \cup K, B \cup L\}$$

Es muss also genau eine der Schnittmengen aus  $U \cup V$  leer sein. Andernfalls heißen die beiden Splits „nicht kompatibel“.

Ein Split bei dem min. eine der beiden Partitionen genau 1 Element enthält bezeichnet man als trivialen Split.

Einen maßgeblichen Beitrag leistete 1971 Bunman indem er bewies, dass die Vereinigung aller Splits genau dann mit der Vereinigung aller Kanten eines Phylogenetischen Baumes übereinstimmt wenn alle Splits paarweise kompatibel sind.

**Die Vereinigung aller paarweiser kompatibler Splits stimmt genau mit der Vereinigung aller Kanten eines Phylogenetischen Baumes überein.**

Man kann um einen Baum, der die evolutionäre Entwicklung eines gegebenen Datensatzes an Taxa darstellt, zu erstellen nach kompatiblen Splits dieser Taxa suchen.

Zu beachten sei hier, dass es für z.B. 5 Taxa 15 mögliche Splits und für n Taxa  $2^{(n-1)} - 1$  mögliche Splits gibt. Um einen vollständigen binären Baum aufzubauen, muss man hierzu nach  $2n-3$  kompatiblen aus den oben genannten

$2^{(n-1)} - 1$  möglichen Splits herausuchen. So gibt es zu 15 Taxa 27 kompatible Splits von 16.383 möglichen. Man muss also nun eine Möglichkeit finden möglichst einfach zu einem optimalen Ergebnis zu kommen. Am effizientesten ist es hierbei nach auffälligen Splits zu suchen, und, obwohl es auch hierzu mehrere Wege gibt wird im Folgenden nur auf den von Buneman (1971) weiter eingegangen, da dieser auch gleichzeitig eine gute Basis liefert um die „Split-decomposition“ Theorie zu verstehen.

### 2.3 Buneman Bäume

Um überhaupt einen derartigen Baum aufbauen zu können, benötigt man eine vollständige Distanzmatrix die jedem Paar an Taxa einen Wert zuordnet:

$$d : X \times X \rightarrow \mathbb{R}$$

Man definiert  $\beta(uv|xy)$  über den Split  $S = \{A, B\}$  wobei  $u, v \in A$  und  $x, y \in B$  als:

$$\beta(uv|xy) = \min(d(x, u) + d(y, v), d(x, v) + d(y, u)) - (d(x, y) + d(u, v))$$

Der **Buneman Index**  $\beta_S$  des Splits  $S$  ist definiert als:

$$1/2 \min \beta(uv|xy) \text{ über alle } u, v \in A \text{ und } x, y \in B$$

Beispiel:

Betrachtet man den in Abb. 2.2 dargestellten Baum so ist die Distanz zweier Taxa definiert als die Summe der Gewichtungen auf dem Weg zwischen beiden. So ist die Distanz  $d_T(2,5) = 2+3+3+1 = 9$ .

Will man nun  $\beta$  für alle möglichen Paare eines Splits  $S = \{\{1,2\}, \{3,4,5\}\}$  berechnen so ergibt sich

$$\beta(12,34) = 6;$$

$$\beta(12,35) = 6 \text{ und}$$

$$\beta(12,45) = 12.$$

Somit ist der Buneman Index  $\beta_S = 1/2 * 6 = 3$ .

Der wichtigste Fakt aber den Buneman hierbei herausfand ist:

Für einen Satz an Taxa für den die Distanzmatrix bestimmt ist gilt:

**Die Vereinigung aller Splits für die  $\beta_S > 0$  gilt, sind kompatibel und lassen sich somit als Baum repräsentieren.**

Somit ist  $\beta_S$  ein wichtiges Kriterium um zu entscheiden welche Splits wesentlich sind und somit einen Baum konstruieren lassen.

Ein derartiger Baum, dessen Äste jeweils dem Gewicht  $\beta_S$  der durch sie erzeugten Splits entsprechen, wird **Buneman Baum** genannt. Die Entfernungen der gewichteten Äste entsprechen hierbei den errechneten Distanzen der Matrix  $d$ .

Jede Methode die einen Baum aus genetischen Distanzen errechnet, sollte folgenden Kriterien entsprechen:

1. Die Methode angewandt auf die genetischen Distanzen eines gewichteten Baumes  $T$  sollte den Baum  $T$  ausgeben.
2. Die Methode angewandt auf genetische Distanzen sollte von diesen „kontinuierlich“ abhängen. Das heißt kleine Änderungen an  $d$  sollten auch nur kleine Änderungen an  $T$  zur Folge haben und nicht das komplette Erscheinungsbild des Baumes ändern.
3. Es sollte möglich sein die Methode effizient zu implementieren.
4. Der ausgegebene Baum  $T$  sollte unabhängig von der Reihenfolge der Eingabe der Taxa sein.

Dies sind zwar gute Kriterien, jedoch entsprechen selbst einige der gängigsten Methoden zur Rekonstruktion eines Baumes aus gegebenen genetischen Distanzen nicht diesen Bedingungen. UPGMA beispielsweise entspricht nicht immer Kriterium 1 und Neighbour Joining (NJ) entspricht nicht immer den Kriterien 2 und 4. Genauer beschrieben wird dies in Moulton, Steel (1999).

Obwohl der Aufbau eines Buneman Baumes allen diesen Kriterien entspricht sind die erzeugten Bäume nicht immer vollständig aufgelöst, da, wegen der Sortierung nach den Minima der vorkommenden  $\beta$ , oft zu viele Splits verworfen werden, so dass der Baum aufgelöster erscheint als er nach den vorliegenden Daten tatsächlich ist. Das folgende Kapitel befasst sich nun mit genau einer solchen Möglichkeit dieses Problem zu beheben, der Split decomposition.

## 2.4 Split decomposition

Im Gegensatz zu der Methode von Buneman wird bei der Split decomposition nun durch eine Änderung ein neuer Index definiert. Hierbei ist:  $\alpha(uv|xy)$  mit Split  $S=\{A,B\}$  wobei  $u,v \in A$  und  $x,y \in B$  definiert als:

$$\alpha(xy|uv) = \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\} - (d(x,y) + d(u,v))$$

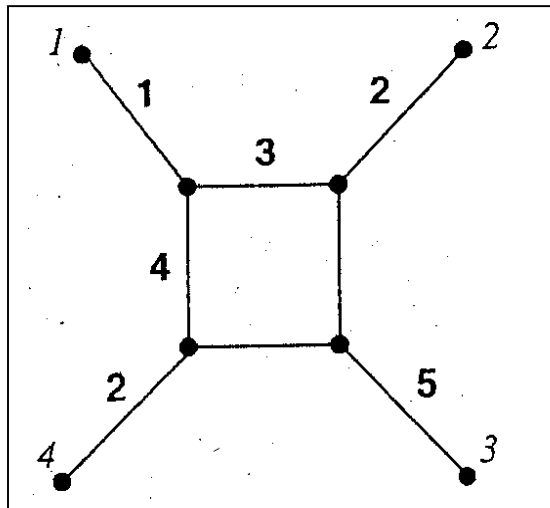
Der **Isolation Index**  $\alpha_s$  ist definiert als:

$$1/2 \min \alpha(uv|xy) \text{ über alle } u,v \in A \text{ und } x,y \in B$$

Beispiel:

Betrachtet man den in Abb. 2.3 dargestellten Netzwerk  $N$  mit den Taxa 1,2,3,4 so ist auch hier die geringste Entfernung zweier Taxa zueinander die geringste Summe der gewichteten Kanten des Netzwerks. Es kann allerdings, im Gegensatz zu Bäumen, wie auch in diesem Beispiel vorkommen, dass zwei unterschiedliche Pfade von Kanten beide die geringste Gewichtung haben.

So ist beispielsweise die Entfernung  $d_N(1,3)=1+3+4+5=13$ . Um zu diesem Ergebnis zu gelangen kann man aber 2 verschiedenen Pfaden folgen, nämlich zuerst dem senkrechten und dann dem waagerechten oder umgekehrt. Will man nun für den Split  $S=\{\{1,4\}\{2,3\}\}$  den Isolation Index  $\alpha_S$  von S berechnen so ergibt sich aus  $\alpha(14|23) = 6 \Rightarrow \alpha_S = 3$ . Und für den Split  $T=\{\{1,2\}\{3,4\}\}$  ist  $\alpha_T = 4$ . Hierbei fällt auf, dass die berechneten Indizes genau den Gewichten der parallel verlaufenden Kanten entspricht.



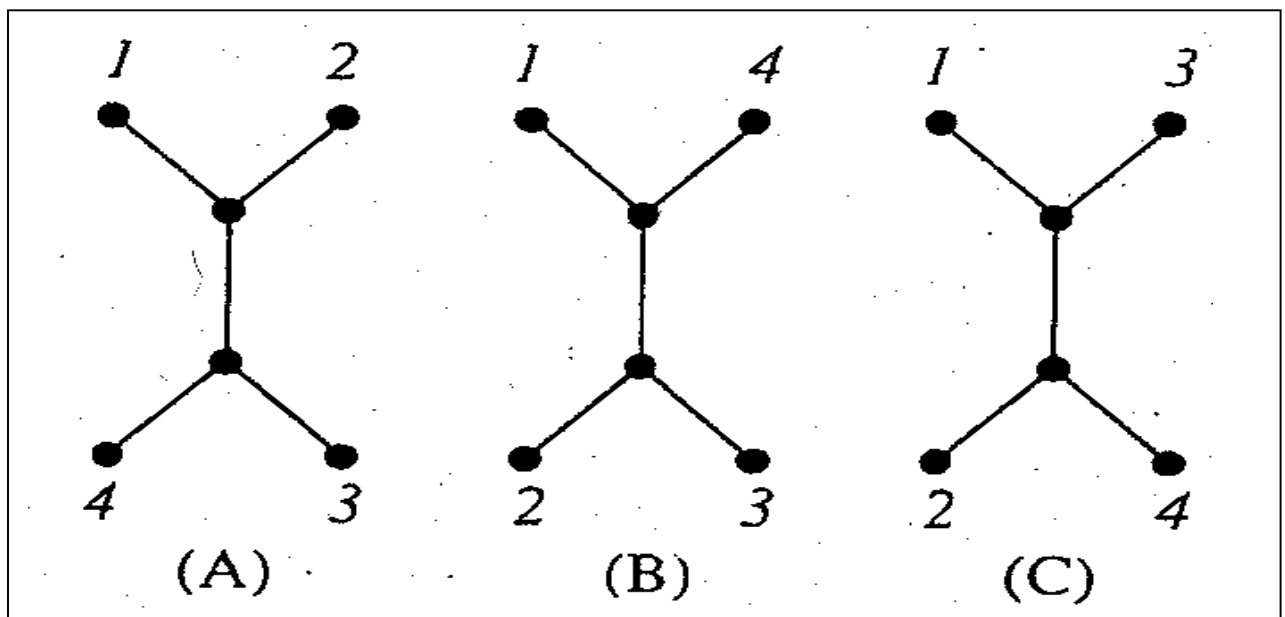
(Abb. 2. 3)

Aus diesem Beispiel lassen sich nun 2 wichtige Dinge erkennen. Erstens, führt die Entfernung parallel verlaufender Kanten zu einer Spaltung des Netzwerks, dessen Isolation Index genau dem Gewicht der jeweils entfernten Kanten entspricht. Und zweitens sieht man, dass die Splits S und T nicht mehr kompatibel sind und somit auch nicht zu einem Baum gehören können. Das bedeutet nun, dass

Splits mit positivem Isolation Index im Gegensatz zu Splits mit positivem Buneman Index nicht mehr unbedingt kompatibel sein müssen. Da kein Vorteil darin liegt mehr Splits als notwendig zu behalten wird nun allen verbleibenden Splits mit Hilfe der **spectral analyse** ein Wert über ihre Wichtigkeit zugewiesen.

Berechnet man hier z.B, wie in Abb. 2.3 zu sehen, den Isolation Index eines Splits  $U=\{\{1,3\}\{2,4\}\}$  so ergibt sich  $\alpha_U=0$ . Da  $\alpha_U$  hiermit kein positiver Index aus der Menge der Taxa  $\{1,2,3,4\}$  ist gehört es auch nicht dazu. Geht man nun weiter und berechnet die Isolation Indizes und die Buneman Indizes der in Abb. 2.4 dargestellten A, B und C so sieht man, dass man mit dem Isolation Index sowohl A als auch B behalten würde und nur C verworfen würde, beim Buneman Index hingegen würden C und auch B verworfen und nur A behalten. Kombiniert man nun A und B miteinander erhält man wieder das in Abb.2.3 dargestellte Netzwerk welches eine Mischung aus A und B darstellt und keinem von beiden eine Priorität einräumt.





(Abb2.4)

Aus dieser Dissonanz zwischen den immer kompatiblen Splits eines positiven Buneman Indexes und den nicht gezwungenermaßen kompatiblen Splits eines positiven Isolation Indexes erklärt sich nun die neue Definition einer schwachen Kompatibilität.

**Drei Splits sind schwach kompatibel, falls mindestens eine Schnittmenge aus der Splits  $S=\{A,B\}$ ,  $T=\{C,D\}$  und  $U=\{E,F\}$  leer ist:**

$$1 \leq |\emptyset| \in \{A \cap C \cap E, A \cap D \cap F, B \cap C \cap F, B \cap D \cap E\}$$

Die wichtigsten Schlüsse die man nun aus dieser schwachen Kompatibilität ziehen kann sind folgende:

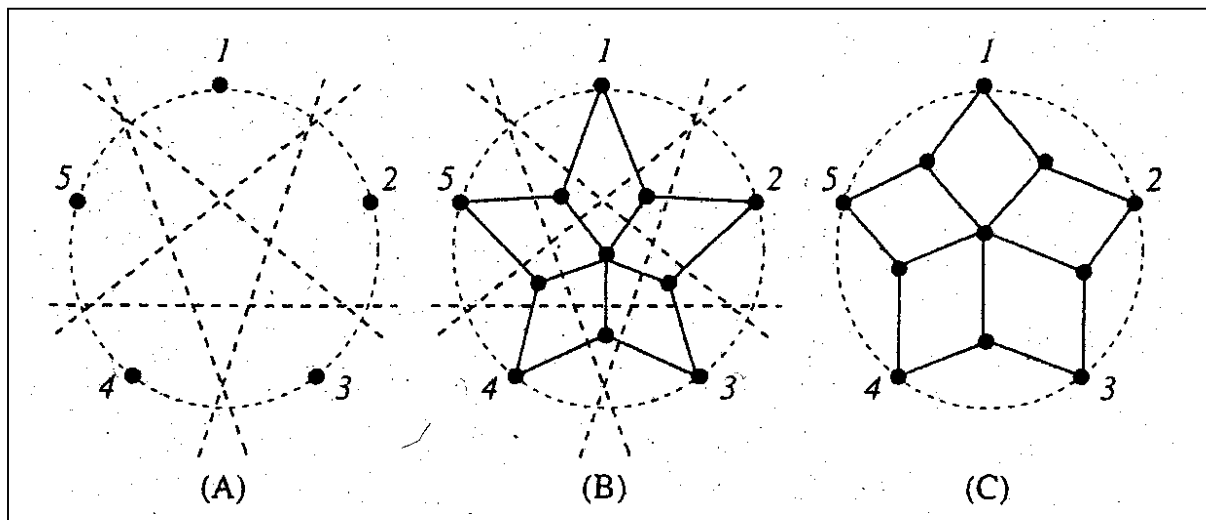
- Hat X n Elemente so ist die Anzahl der Splits mit positivem Isolation Index maximal  $n(n-1)/2$ .
- Diese können effizient berechnet werden.
- Alle 4 der oben geforderten Ansprüche an ein derartiges Verfahren wird genüge getan.

## 2.5 Von schwach kompatiblen Splits zu Netzwerken

Nachdem man nun zu einem solchen Satz an schwach kompatiblen Splits den jeweiligen Isolation Index berechnet hat muss man eine Möglichkeit finden diese in einem gewichteten Netzwerk darzustellen. Im Allgemeinen kann dies immer unter der Verwendung von Median Netzwerken erreicht werden, bei diesen besteht aber das Problem, dass sie nicht immer auch planar sind und somit schwer zu zeichnen. Sofern die berechneten Splits aber zyklisch sind

besteht die Möglichkeit diese in einem sogenannten **äußeren Planaren Netzwerk** darzustellen. Diese Netzwerke sind es auch, die im Allgemeinen von dem Programm *SplitsTree* erzeugt werden.

**Die Menge der Splits eines gegebenen Sets an Taxa ist zyklisch, falls diese auf einem Kreis so angeordnet werden können, dass sich jeder Split durch eine Linie darstellen lässt**



(Abb. 2.5)

Betrachtet man nun das in Abb. 2.5. dargestellte Beispiel so sieht man, dass jede gepunktete Linie einen Split darstellt. Fügt man nun jedem, der in Teil A durch eine gepunktete Linie eingegrenzten Bereiche, einen Knoten hinzu und verbindet diese so kommt man zu Abb. 2.5B. Man sieht, dass Teil C nun schon dem originalen Netzwerk das in Teil C abgebildet ist ähnelt, man erreicht dies indem man die Ecken nun leicht anpasst, so dass diese parallel zueinander verlaufen. Die Methode die hier in diesem Beispiel verwendet wurde basiert auf dem Prinzip der **De Bruijn dualisation**.

Ordnet man nun jeder Kante den Wert des ihres Splits entsprechenden Isolation Indexes zu so lässt sich aus diesem Gewichteten Netzwerk ein repräsentativer Wert der Distanz  $d_N$  errechnen. Ist das mit Hilfe eines positiven Isolation Index der Splits erzeugte Netzwerk zyklisch, so stellt  $d_N$  einen Näherungswert der wirklichen Distanz  $d$  dar. Die verbleibende Differenz zwischen  $d$  und  $d_N$  wird als **split-prime-residue** ( $d - d_N$ ) bezeichnet und ist genau dann 0 falls die erzeugten  $d_N$  der eigentlich errechneten  $d$  entsprechen.

Das Maß für die Genauigkeit diese Näherung der  $d_N$  an  $d$  wird definiert als **Fit Index**:

$$fi = \left| \frac{\sum (d - d_N)(x, y)}{\sum d(x, y)} \right| * 100\% \text{ für alle } x, y \text{ aus } X$$

## 3. Anwendung

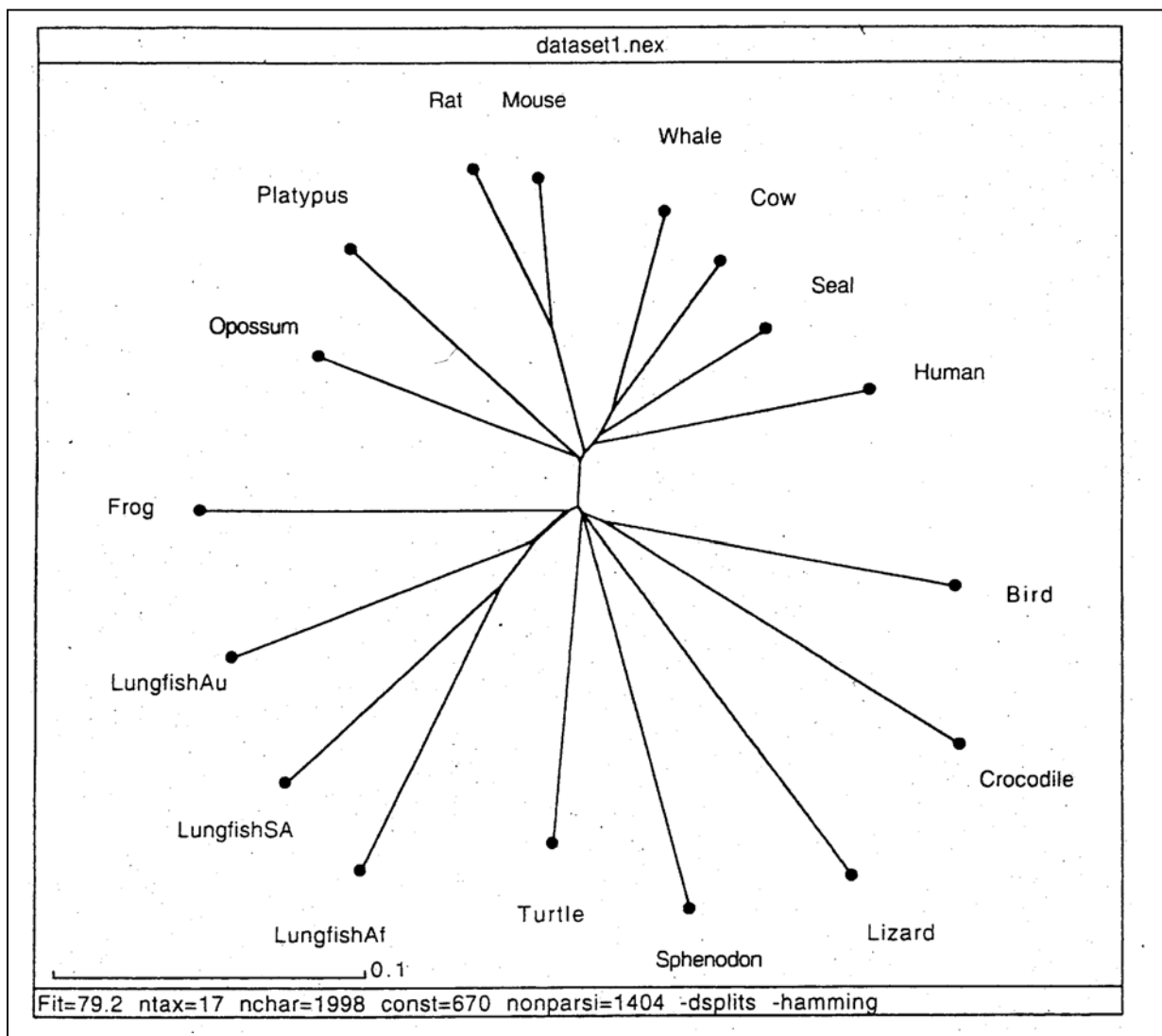
### 3.1 Das SplitsTree Programm

Erhältlich sind mehrere Versionen von SplitsTree, die aktuellste Release Version 3.2 ist verfügbar für Win32 und Unix. Für MacOS ist die Version 2 verfügbar. Eine Java basierte Version 4 Namens Jsplits ist im Betastadium. Alle Versionen sind verfügbar unter:

<http://www-ab.informatik.uni-tuebingen.de/software/splits/>

Für die Version 3.2 für Win32 wird zudem noch die TCL/TK Erweiterung TCL805.exe benötigt. Diese ist zu finden unter <http://www.scriptics.com>. Zudem müssen noch die Dateien TCL80.dll und TK80.dll in den SplitsTree Ordner kopiert werden.

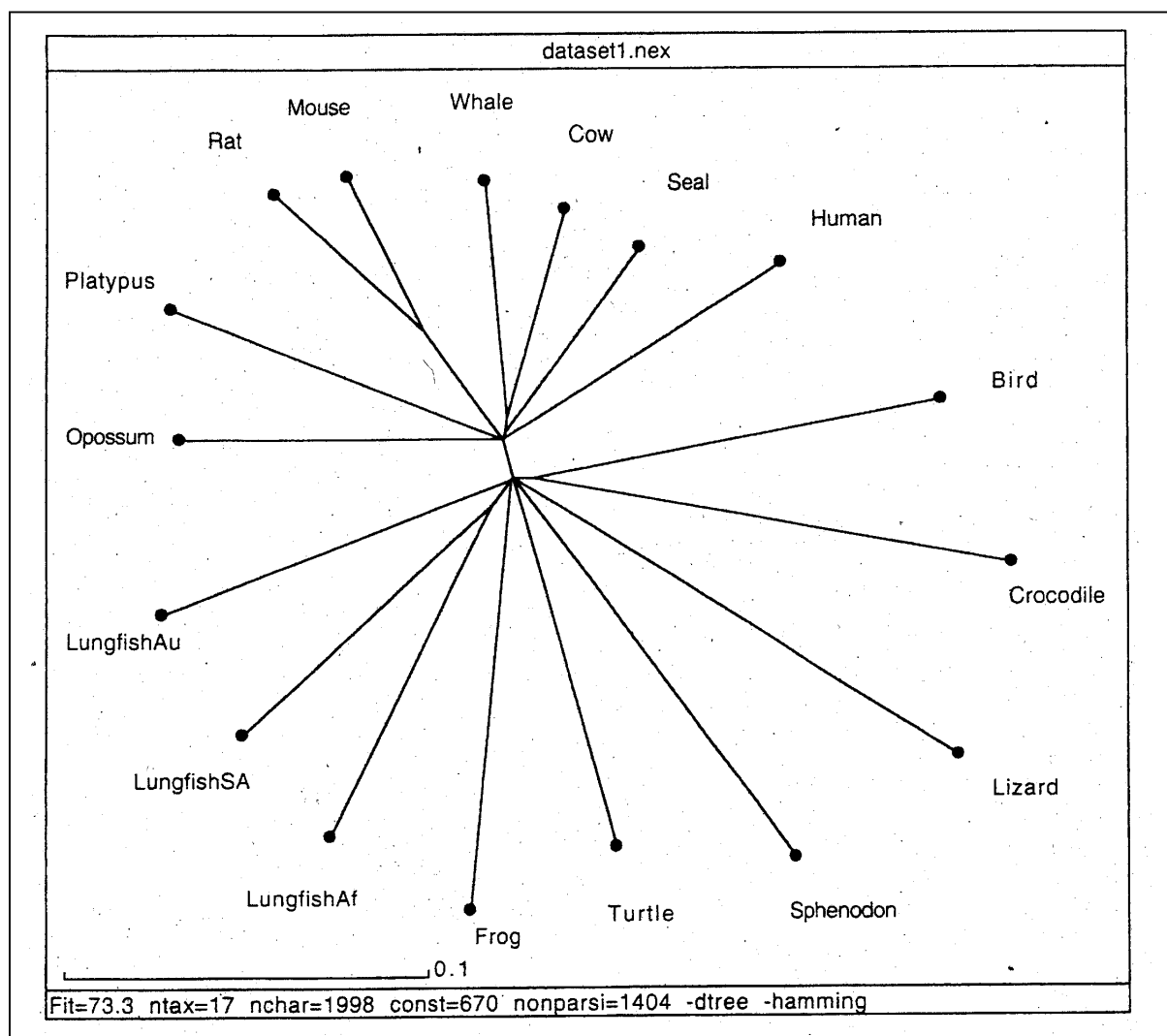
### 3.2 Beispiel: mtDNA Datensatz



(Abb. 3.1)

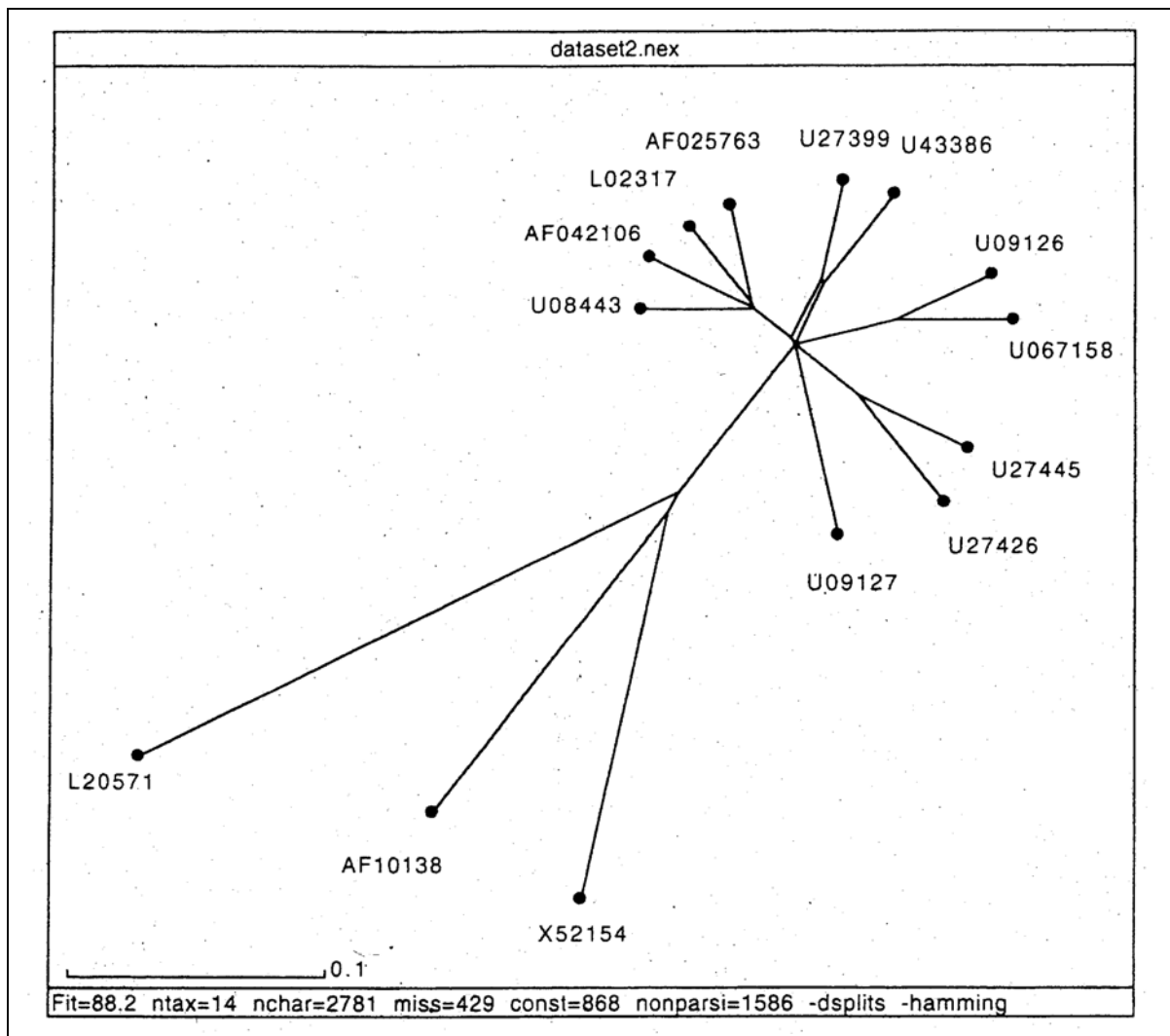
Abb. 3.1 stellt einen mit SplitsTree erstellten Split Graphen dar, dessen Fit Index wie in der Statusleiste angegeben bei 79,2% liegt. D.h. 80% der angegebenen Distanzen sind noch korrekt und 20% der Distanzen weichen von ihrer eigentlich errechneten Distanz ab. Man kann nun leider nicht generell sagen welcher Fit Index für einen SplitGraphen gut ist. Erfahrungsgemäss werden Netzwerke die bei über 80% liegen als akzeptabel betrachtet. Bei Fit Indizes von 70% und weniger kann man davon ausgehen, dass zu viele verworfen wurden um noch ein Netzwerk darstellen zu können, als dass man das Netzwerk noch verwenden könnte.

Man darf davon ausgehen, dass bei einem hohen Fit Index die Ergebnisse anderer Methoden die auf Entfernungen basieren, wie z.B. NJ, sehr ähnlich aussehen würden. Im Folgenden sieht man in Abb3.2 den gleichen Datensatz an Taxa, diesmal allerdings als Buneman Baum aufgebaut.



(Abb. 3.2)

### 3.3 Beispiel 2: HIV-1 Datensatz



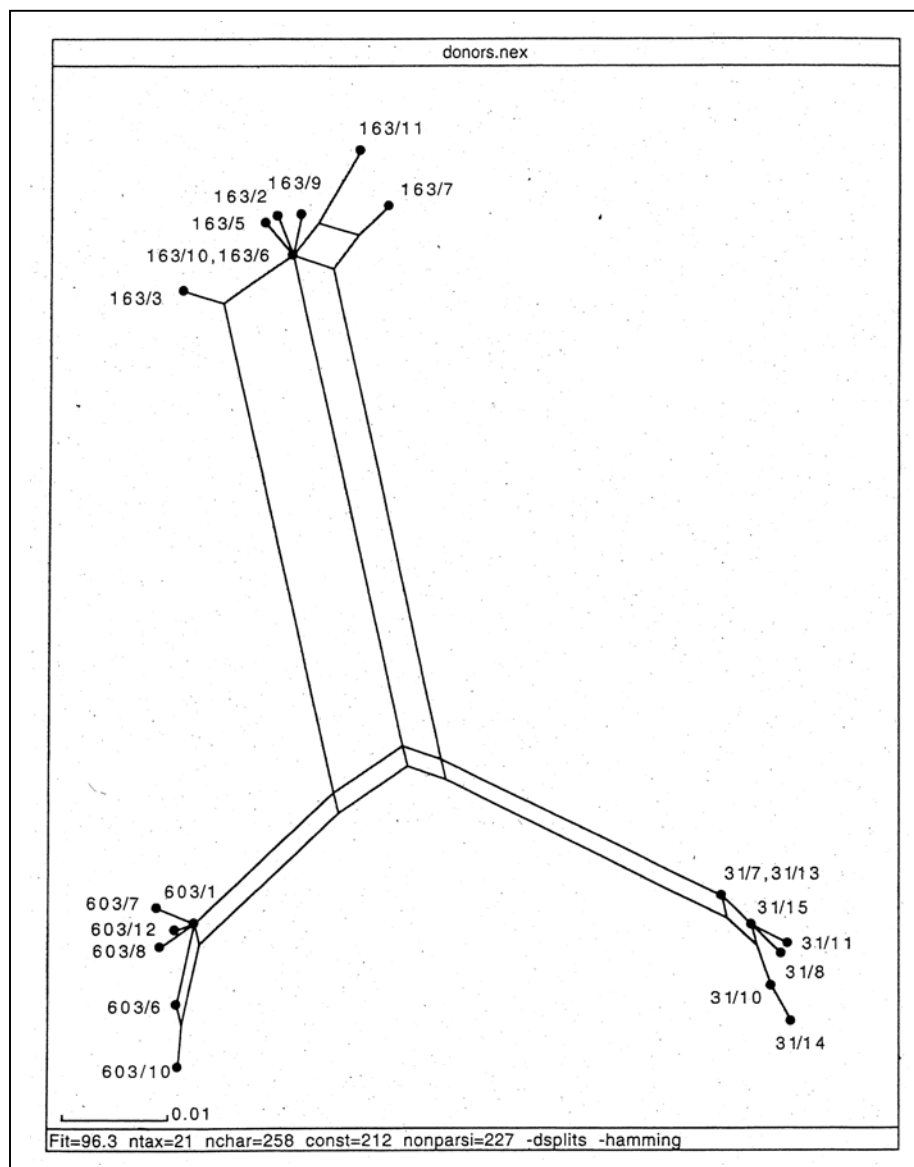
(Abb. 3.3)

In Abb. 3.3 nun dargestellt sieht man das Netzwerk der aus einem HIV Set erstellt wurde. Das Netzwerk ist zwar zum größten Teil baumartig und der Fit Index von 88,2% bestätigt die Korrektheit der Darstellung. Im Gegensatz zum ersten Beispielt tritt hier jedoch eine Ungenauigkeit im Netzwerk vor den Taxa U27399 und U43368 auf. Des Weiteren ist der Zentrale Knoten mit einem Grad von 6 auffällig. Dies lässt auf einen Konflikt der Daten schließen, so dass sich dieser Knoten nicht weiter auflösen lässt.

Bei den bisherigen beiden Beispielen wurde nun die Distanz schlicht mit der Hamming Methode berechnet welche die Anzahl der Unterschiede zwischen zwei Sequenzen als deren Entfernung ausgibt.

Es ist nun aber auch möglich schon im Voraus mit einer Methode berechneten Distanzmatrizen in SplitsTree einzubinden und zu verwenden. Dazu müssen die zu importierenden Daten lediglich im Nexus Dateiformat bereitgestellt werden.

Wie man in Abb. 3.4 leicht erkennt wurde diese nicht aus einem baumartigen Datenset erzeugt sondern aus HCV Daten (Allain et al. 2000) einer Studie über die Immunantwort auf Hepatitis C. Eine baumartige Darstellung dieses Netzwerkes wäre, im Gegensatz zur dieser Abbildung, nur unzureichend. Zumal der Split Index von 96,3% auf eine nahezu korrekte Darstellung der errechneten Distanzen hinweist. Man kann nun das dargestellte Netzwerk grob in drei Einheiten aufteilen. Hierbei wurde die mit 603 gekennzeichneten Taxa aus einem Donor entnommen und die mit 163 und 31 gekennzeichneten aus zwei unterschiedlichen Rezipienten. Des Weiteren beachte man den Knoten der mit 31/7,31/13 gekennzeichnet ist. Dieser ist gleich in zweierlei Weise beachtenswert. Die doppelte Kennzeichnung weist darauf hin, dass kein Splitindex eines Splits gefunden wurde der diese zwei Taxa trennen würde. Die Tatsache, dass dieser Knoten ein interner Knoten und kein Blatt ist deutet darauf hin, dass es sich hierbei um einen Vorfahr der an den Blättern dieses Teilnetzwerks vorhandenen Taxa handelt.-



Weitere Beispiele zur Analyse von Daten findet man in Dopaz et al. (1993) und Plikat, Nielst-Struwe und Meyerhans(1997)

(Abb 3.4)

## Quellenangaben:

Verwendete Abbildungen: *The Phylogenetic Handbook*, M.Salemi,  
A-M. Vandamme, Cambridge University Press, 2003

Verwendete Literatur: *The Phylogenetic Handbook*, M.Salemi,  
A-M. Vandamme, Cambridge University Press, 2003

Studienarbeit zum Vergleich prokaryotischer Genome,  
A. Auch, Uni Tübingen , 2003