Westfälische
Wilhelms-Universität
Münster

Fachbereich 10
Mathematik und
Informatik

Merle Erpenbeck

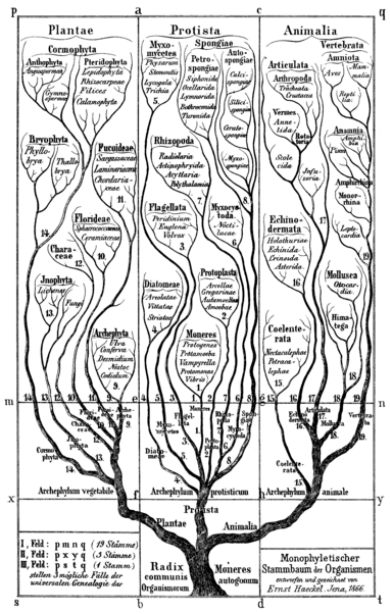
Phylogenetische Rekonstrukion

Sparsamkeits- und Abstandsmethoden

5. Juli 2012

Seminarausarbeitung im Seminar Mathematische Biologie
vorgelegt von Merle Erpenbeck
Matrikelnummer: 358396
Betreuer: Prof. Dr. Matthias Löwe,
Dr. Felipe Torres

Inhaltsverzeichnis



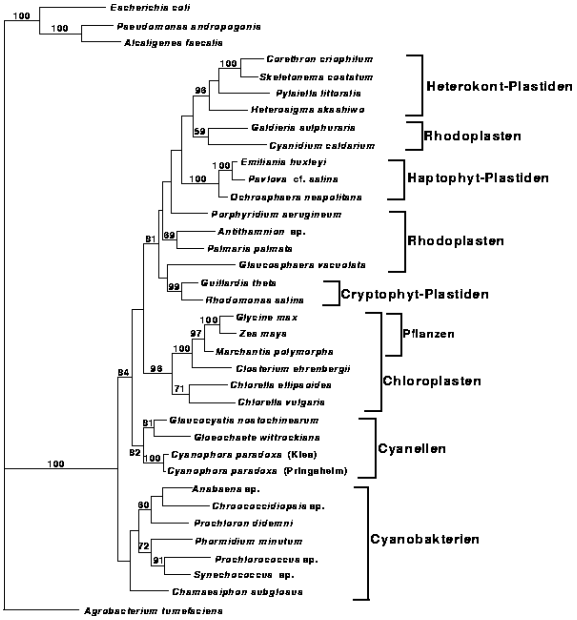


Abbildung 1: aus E.Haeckel: „Generelle Morphologie der Organismen [...]“, 1866

Abbildung 2: Phylogenetischer Stammbaum der Plastidenentwicklung; berechnet aus 16S rRNA-Sequenzen. Aus H. Schmidt: „Parallelisierung phylogenetischer Methoden zur Untersuchung der Crown Group Radiation'",1997

1 Einführung:Phylogenetische Rekonstruktion

1.1 Allgemeines

Ein phylogenetischer Baum soll die evolutionäre Verwandtschaft verschiedener Spezies darstellen. Früher wurden diese Bäume anhand von morphologischen Eigenschaften, also dem Erscheinungsbild der Tiere aufgestellt (siehe Abbildung 1). In den letzten Jahrzehnten ist man jedoch dazu übergegangen, den Bäumen Gene und Proteinsequenzen zugrunde zu legen (siehe Abbildung 2), was auch den Namen „Phylo*genetisch*“ erklärt. Sequenzbasierte Methoden sind sensitiver, da Veränderungen in den Gensequenzen den Veränderungen in der Morphologie vorhergehen. Daher sollte man phylogenetische Bäume eher als Gen-Bäume denn als Spezien-Bäume betrachten.

Es passiert häufig, dass, abhängig davon, welche Sequenzen der einzelnen Spezies man zugrundelegt, verschiedene Bäume herauskommen – auch bei gleichem Datenmaterieal können sich unter Umständen verschiedene Bäume ergeben. Auch wenn es historisch einen einzigen „wahren“ Baum gibt, der die Entwicklung der Spezies beschreibt, ist das Problem, den richtigen Baum zu finden, bis heute ungelöst.

1.2 Phylogenetische Bäume

Wir wollen uns hier nur mit binären Bäumen befassen, also mit Bäumen, bei denen aus jedem Knoten genau zwei Äste entspringen. Dabei wird zwischen gerichteten Bäumen, die eine Wurzel haben und ungerichteten Bäumen, die keine Wurzel haben, unterschieden. Die äußeren Knoten heißen „Blätter“ und werden mit den Namen der ihnen zugeordneten Spezies bezeichnet, den „operational taxonomic units“ (OTUs).

Gerichtete Bäume beschreiben die Evolution von einem gemeinsamen Vorfahren aller OTUs zu den OTUs. Entfernt man die Wurzel des Baumes und fasst die beiden Äste, die dieser entspringen zu einem zusammen, so erhält man einen ungerichteten Baum. Diese Bäume enthalten nur Informationen über die Verwandtschaft verschiedener Spezies, nicht aber über die Richtung der Evolution.

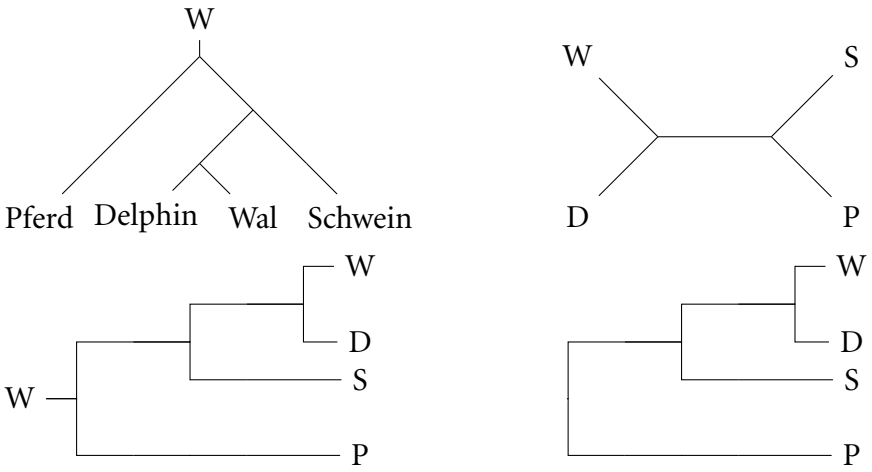


Abbildung 3: Zwei verschiedene Darstellungen von phylogenetischen Bäumen; jeweils einmal gerichtet, einmal ungerichtet.

Die Länge eines Astes ist (idealerweise) eine positive Zahl, die die Nähe der Verwandtschaft zwischen den OTUs, für die die benachbarten Knoten stehen, angibt. Diese wird häufig durch das Produkt der Länge des Zeitintervalls, dass beide Sequenzen historisch trennt und einem speziellen Wert, der Evolutionsrate berechnet. Die Einbeziehung der Evolutionsraten trägt dem Umstand Rechnung, dass sich manche Gene schneller entwickeln als andere. Die Astlänge wird meist als Zahl an den Ast angefügt (siehe z.B. Abb. 4).

Das Astmuster eines Baumes (ohne die Astlängen) wird als *Baumtopologie* bezeichnet. Ein Baum, dessen Blätter bestimmten OTUs zugeordnet sind, *verbindet* diese OTUs.

In diesem Vortrag werden zwei Methoden vorgestellt, um vorgegebene OTUs optimal zu verbinden. Man hofft, auf diese Weise zumindest einige Rückschlüsse auf die tatsächliche Evolution ziehen zu können.

Hierbei werden folgende Annahmen getroffen:

- Die Genveränderungen treten plötzlich auf, bzw. benötigen im Vergleich zur Länge des Astes sehr wenig Zeit.
- Aus jedem Knoten entspringen genau zwei Äste. Nach dieser Definition wäre der Baum aus Abbildung 1 also kein phylogenetischer Baum, da aus der Wurzel drei Äste entspringen. Um solche Bäume doch zu ermöglichen, werden später aber Äste der Länge 0 zugelassen.

1.3 Phylogenetische Rekonstruktion

Der Vorgang, für eine gegebene Menge von OTUs einen optimalen phylogenetischen Baum zu erstellen, nennt sich phylogenetische Rekonstruktion. Hierzu macht man folgende Schritte:

1. *Auwählen einer geeigneten Familie homologer Sequenzen*

   Homolog heißt in diesem Zusammenhang, dass sich die Sequenzen hinreichend ähnlich sind, um einen evolutiven Zusammenhang zu vermuten. Sind sich die Sequenzen nicht hinreichend ähnlich, so kann man zwar einen Baum berechnen – dieser wird jedoch nicht sehr informativ sein. Geeignete OTUs auszuwählen ist eine Wissenschaft für sich, auf die hier nicht weiter eingegangen werden soll.

2. *Aus den Sequenzen ein reduziertes multiples Alignment bilden*

   Zuerst bildet man (nach vorgegebenen Gütekriterien) ein multiples Alignment (siehe z.B. Vortrag von Alime Karadöl) und bildet daraus ein reduziertes Alignment, indem man die Spalten, die Lücken enthalten, löscht. Es gibt auch Methoden, die aus beliebigen Alignments Bäume bilden können, der Einfachheit halber wird darauf hier jedoch nicht eingegangen.

   1.1 Beispiel Gegeben seien folgende Sequenzen:

   a GCTGCA
   b GCTGA
   c GTCC
   d GCTCCC

   Daraus bilden wir das folgende Alignment:

   a GCTGCA
   b GCTG–A
   c G–TCC–
   d GCTCCC

Das reduzierte Alignment ist dann:

a GTG
b GTG
c GTC
d GTC

3. *Aus dem reduzierten multiplen Alignment einen phylogenetischen Baum konstruieren*
   Die größte Schwierigkeit hierbei ist in der Regel das Aufstellen der Baumtopologie. Im vorangegangenen Beispiel ist schnell zu sehen, dass a und b sowie c und d zusammengefasst werden müssten und somit die Baumtopologie aus Abbildung 3 angemessen ist. Das Bestimmen von Astlängen und dem Ort der Wurzel ist ein Problem, für das noch mehr Informationen benötigt werden.

Es gibt verschiedene Methoden, aus reduzierten multiplen Alignments einen phylogenetischen Baum zu erstellen:

- Methoden der maximalen Sparsamkeit (parsimony methods)
- Abstandsmethoden (distance methods)
- probabilistische Methoden, die auf dem Konzept des Maximum Likelyhood beruhen

Die beiden erstgenannten Methoden sollen nun näher vorgestellt werden.

2 Methoden der maximalen Sparsamkeit

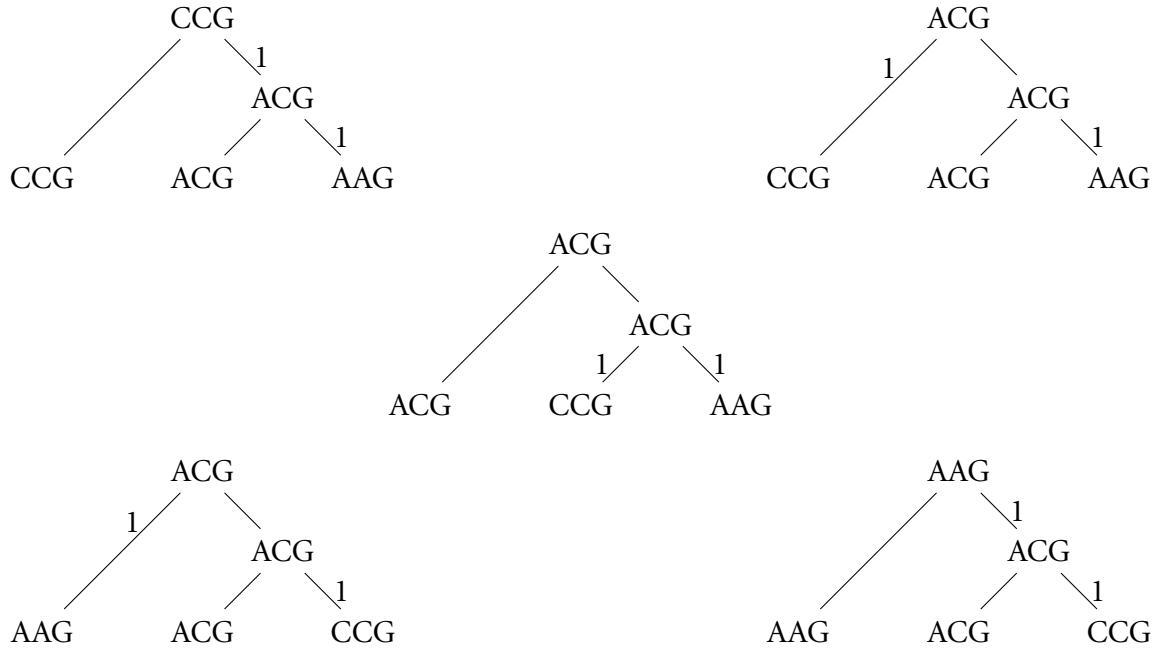
Mit Methoden der maximalen Sparsamkeit findet man Topologien gerichteter Bäume, jedoch keine Astlängen. Auch kann man mit ihnen die Sequenzen an den inneren Knoten des Graphen finden. Bei diesem Ansatz werden die *totalen Kosten* eines Baumes berechnet und die Topologien mit den geringsten Kosten als optimal betrachtet – diese werden die „*sparsamsten*“ Topologien genannt. Dieser Ansatz geht also davon aus, das die Evolution in gewissem Sinne „ökonomisch“ verläuft.

Es gibt verschiedene Kostenfunktionen, die man verwenden kann. Als Beispiel wird hier die einfachste Funktion verwendet, nämlich die, die jeder Substitution eine Kosteneinheit zuordnet. Für eine gegebene gerichtete Baumtopologie ordnen wir der Wurzel und allen inneren Knoten Sequenzen zu, die die selbe Länge wie das reduzierte multiple Alignment haben. Die Kosten dieser Zuordnungen definiert man als die Summe der Kosten ihrer Äste, wobei die Kosten eines Astes, der zwei Knoten verbindet, die minimale Anzahl von Substitutionen ist, die man braucht, um von der Sequenz an dem einen Knoten zu der Sequenz des anderen Knotens zu kommen. Die Kosten der Topologie sind die minimalen Kosten der verschiedenen Belegungen.

2.1 Beispiel Gegeben sei das Alphabet Q = {A,C,G,T} und das folgende reduzierte multiple Alignment:

x_1 AAG
x_2 ACG
x_3 CCG

Es gibt drei mögliche gerichtete Topologien für drei OTUs, die alle gleich sparsam sind. Die Kosten jeder dieser Topologien ist 2.



Es gibt einen schnellen Algorithmus (Algorithmus von Fitch), um die Kosten für eine gegebene Topologie zu berechnen. Dieser ist schnell genug, um bei moderater Anzahl von OTUs die Kosten jeder Topologie berechnen zu können und die sparsamste bestimmen zu können. Bei großem N ist es jedoch sehr aufwendig, die möglichen Topologien aufzustellen.

3 Abstandsmethoden

Abstandsmethoden rekonstruieren (gerichtete oder ungerichtete) Bäume aus einer Menge von Abständen zwischen je zwei Sequenzen eines gegebenen reduzierten multiplen Alignment.

3.1 Definition Sei M eine Menge. Eine Funktion $d : \mathscr{M} \times \mathscr{M} \to \mathbb{R}$ heißt *Abstandsfunktion* auf $\mathscr{M}$, wenn

- $d(u,v) \geq 0$ für alle $u, v \in \mathscr{M}$
- $d(u,v) = d(v,u)$ für alle $u, v \in \mathscr{M}$
- $d(u,v) \leq d(u,w) + d(w,v)$ für alle $u, v, w \in \mathscr{M}$

Dies ist keine Metrik nach üblicher Definition, da wir nicht fordern, dass der Abstand zwischen verschiedenen Punkten echt positiv ist. Wir nennen $d(u,v)$ Abstand zwischen u und v.

In unserem Zusammenhang interessieren wir uns für Abstandsfunktionen auf einer endlichen Menge von Sequenzen (OTUs), aus denen wir einen phylogenetischen Baum konstruieren wollen. Dabei gehen wir davon aus, dass die Abstandsfunkton *biologisch relevant* ist, dass sie also Informationen über die biologische Verwandtschaft der Sequenzen trägt; das heißt, dass aus $d(x_i, x_j) > d(x_i, x_k)$ folgt, dass sich x_i und x_j stärker von einem gemeinsamen Vorfahren abweichen als x_i und x_k es tun. Der Einfachheit halber schreiben wir d_{ij} statt $d(x_i, x_j)$. Wir können d durch eine symmetrische Abstandsmatrix $\mathscr{M}_d = (d_{ij})$ darstellen.

Fixieren wir einen ungerichteten Baum $\mathscr{T}$ mit angegebenen Astlängen, der gegebene OTUs verbindet, so können wir die vom Baum erzeugte Abstandsfunktion $d^{\mathscr{T}}$ auf $\mathscr{M}$ betrachten, indem wir $d^{\mathscr{T}}(x_i, x_j) =: d^{\mathscr{T}}_{ij}$ als die Länge des kürzesten Weges zwischen x_i und x_j in $\mathscr{T}$ setzen. Dies ist eine Abstandsfunktion (Nachrechnen!).

Die Aufgabe der Abstandsmethoden zur Rekonstruktion eines phylogenetischen Baumes ist nun, zu einer gegebenen Abstandsfunktion d auf einer Menge von OTUs einen Baum zu finden, so dass $d^{\mathscr{T}}$ möglichst gut mit d übereinstimmt. Daher liefern Distanzmethoden in der Regel ungerichtete Bäume mit Astlängen, es wird aber auch eine Methode vorgestellt, die gerichtete Bäume mit Astlängen liefert.

Die Frage ist nun, ob oder unter welchen Bedingungen es einen Baum $\mathscr{T}$ gibt, der eine gegebene Abstandsfunktion d erzeugt, also dass $d^{\mathscr{T}} = d$ gilt. Abstandsfunktionen, die von einem Baum erzeugt werden, heißen additiv.

Für drei verschiedene OTUs und eine gegebene Abstandsfunktion auf diesen lässt sich immer ein Baum finden, der die Abstandsfunktion erzeugt. Dazu muss man drei positive Zahlen finden, für die gilt:

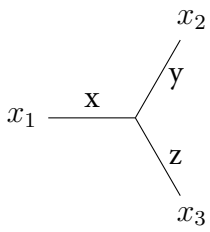
$$x + y = d_{12}$$
$$x + z = d_{13}$$
$$y + z = d_{23}$$

Die Lösung des Gleichungssystems ist

$$\begin{aligned} x &= 0,5(d_{12} + d_{13} - d_{23}) \\ y &= 0,5(d_{12} + d_{23} - d_{13}) \\ z &= 0,5(d_{13} + d_{23} - d_{12}) \end{aligned} \tag{1}$$

Wegen der Dreiecksungleichung sind x, y und z nichtnegativ, sie müssen allerdings nicht echt positiv sein. (Deswegen hatten wir auch Äste der Länge 0 zugelassen.) Biologisch können wir Äste der Länge 0 als „sehr kurze“ Entwicklungszeit deuten.

Also gibt es für beliebige Abstände zwischen drei OTUs genau einen Baum der diese Abstandsfunktion erzeugt. Der Baum, der die drei OTUs verbindet, sieht also folgendermaßen aus:



3.1 neighbor-joining-Algorithmus

3.2 Definition (4-Punkt-Bedingung) Sei d eine Abstandsfunktion auf einer Menge $\mathscr{M}$ und $N \geq 4$. Dann erfüllte d die 4-Punkt-Bedingung, wenn folgendes gilt: Für jede Menge von vier verschiedenen Zahlen $1 \leq i, j, k, l \leq N$ sind zwei der Summen $d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk}$ identisch und nicht kleiner als die dritte Summe.

Wir geben nun einen Algorithmus, um aus gegebenen Sequenzen und Abständen, die die 4-Punkt-Bedingung erfüllen, einen Baum zu konstruieren, der d erzeugt. Dieser Algorithmus heißt *neighbor-joining-Algorithmus*. Er ist iterativ und ersetzt in jedem Schritt ein Paar von OTUs durch eine neue OTU und iteriert auf den verbleibenden OTUs weiter, bis es für drei verbleibende OTUs nur noch eine mögliche Topologie gibt. Dann wird der Baum aufgebaut, indem die zusammengefassten OTUs wieder durch die Paare von OTUs ersetzt werden.

Nun im Detail:

Für jedes $i = 1, \ldots, N$ sei

$$r_i := \frac{1}{N-2} \sum_{k=1}^{N} d_{ik}. \tag{2}$$

Sei weiterhin für alle $i, j = 1, \ldots, N, i < j$

$$D_{ij} = d_{ij} - (r_i + r_j). \tag{3}$$

Wir können die D_{ij} in eine obere Dreiecksmatrix $D = (D_{ij})$ schreiben. Nun nehmen wir ein Paar $1 \leq i, j \leq N$, sodass D_{ij} minimal ist (dies muss nicht eindeutig sein): Wir ersetzen x_i und x_j durch ein neues Element x_{N+1}. Diese neue OTU repräsentiert einen inneren Knoten der phylogenentischen Baumes, der x_i und x_j verbindet und von diesen die folgenden Abstände hat:

$$d_{N+1\,i} = 0,5(d_{ij} + r_i - r_j) \tag{4}$$

$$d_{N+1\,j} = 0,5(d_{ij} + r_j - r_i) \tag{5}$$

Die Abstände zwischen x_{N+1} und den übrigen x_m werden nun wir folgt definiert:

$$d_{N+1\,m} = 0,5(d_{im} + d_{jm} - d_{ij}) \tag{6}$$

Nun haben wir eine neue Familie von $N-1$ OTUs $\mathscr{M}' = \{x_m, x_{N+1} | m \neq i, j\}$ und können das Verfahren wiederholen. Dieses wird iteriert, bis nur noch drei OTUs übrig bleiben. Für diese drei OTUs gibt es eine eindeutige Baumtopologie und eindeutig bestimmt Astlängen (Formel 1). Verfolgt man den Algorithmus nun rückwärts und ersetzt die neu gebildeten OTUs wieder durch die entsprechenden Paare, erhält man den gesuchten Baum.

3.3 Beispiel Sei $N = 6$ und die Abstandsmatrix wie folgt:

M_d	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	8	3	14	10	12
x_2	8	0	9	10	6	8
x_3	3	9	0	15	11	13
x_4	14	10	15	0	10	8
x_5	10	6	11	10	0	8
x_6	12	8	13	8	8	0

Man rechnet schnell nach, dass d eine Abstandsfunktion ist und die 4-Punkt-Bedingung erfüllt. Nun konstruieren wir den Baum $\mathscr{T}$:

Es gilt nach Formel 2:

$$r_1 = \frac{1}{6-2} \cdot (0 + 8 + 3 + 14 + 10 + 12) = \frac{47}{4},$$

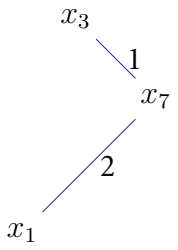
$$r_2 = \frac{41}{4}, r_3 = \frac{51}{4}, r_4 = \frac{57}{4}, r_5 = \frac{54}{4}, r_6 = \frac{49}{4}$$

Das ergibt die folgende Matrix D mit den Einträgen D_{ij} (Formel 3):

D	x_1	x_2	x_3	x_4	x_5	x_6
x_1		-14	$-21,5$	-12	-13	-12
x_2			-14	$-14,5$	$-15,5$	$-14,5$
x_3				-12	-13	-12
x_4					$-15,5$	$-18,5$
x_5						$-15,5$

Der kleinste Eintrag der Matrix ist $D_{13} = -21,5$. Wir führen also eine neue OTU x_7 ein, die das Paar x_1, x_3 ersetzt. Die Abstände von x_7 zu x_1 und x_3 setzen wir wie folgt (Formeln 4 und 5):

$$d_{71} = 0,5\,(d_{31} + r_1 - r_3) = 1$$
$$d_{73} = 0,5\,(d_{31} + r_3 - r_1) = 2$$



Nun berechnen wir die Abstände zwischen x_7 und den anderen OTUs (Formel 6):

$$d_{72} = (d_{12} + d_{32} - d_{13}) = 7$$
$$d_{74} = (d_{14} + d_{34} - d_{13}) = 13$$
$$d_{75} = (d_{15} + d_{35} - d_{13}) = 9$$
$$d_{76} = (d_{16} + d_{36} - d_{13}) = 11$$

Das ergibt folgende Matrix:

M_d	x_2	x_4	x_5	x_6	x_7
x_2	0	10	6	8	7
x_4	10	0	10	8	13
x_5	6	10	0	8	9
x_6	8	8	8	0	11
x_7	7	13	9	11	0

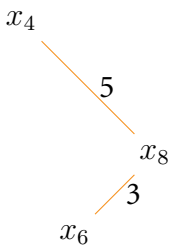
Nun wiederholen wir den Prozess für die neue Abstandsmatrix und erhalten:

$$r_2 = \frac{31}{3}, r_4 = \frac{41}{3}, r_5 = 11, r_6 = \frac{35}{3}, r_7 = \frac{40}{3}$$

und damit:

D	x_2	x_4	x_5	x_6	x_7
x_2		-14	$-\frac{46}{3}$	-14	$-\frac{50}{3}$
x_4			$-\frac{44}{3}$	$-\frac{52}{3}$	-14
x_5				$-\frac{44}{3}$	$-\frac{46}{3}$
x_6					-14

Der kleinste Eintrag ist $D_{46} = -\frac{52}{3}$. Wir führen also eine neue OTU x_8 ein mit $d_{84} = 5$ und $d_{86} = 3$.



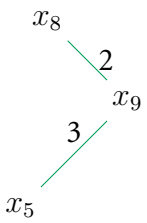
Im nächsten Schritt entstehen folgende Matrizen:

M_d	x_2	x_5	x_7	x_8
x_2	0	6	7	5
x_5	6	0	9	5
x_7	7	9	0	8
x_8	5	5	8	0

und

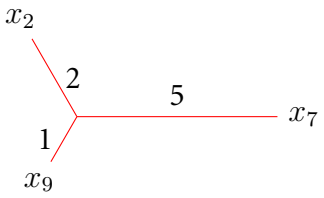
D	x_2	x_5	x_7	x_8
x_2		-13	-14	-13
x_5			-13	-14
x_7				-13

In diesem Fall sind sowohl D_{27} als auch D_{58} minimal; der konstruierte Baum hängt nicht davon ab, welche beiden OTUs wir zusammenfassen. Wir fassen nun x_5 und x_8 zu x_9 zusammen. Es gilt $d_{59} = 3$ und $d_{98} = 2$.Wir erhalten:

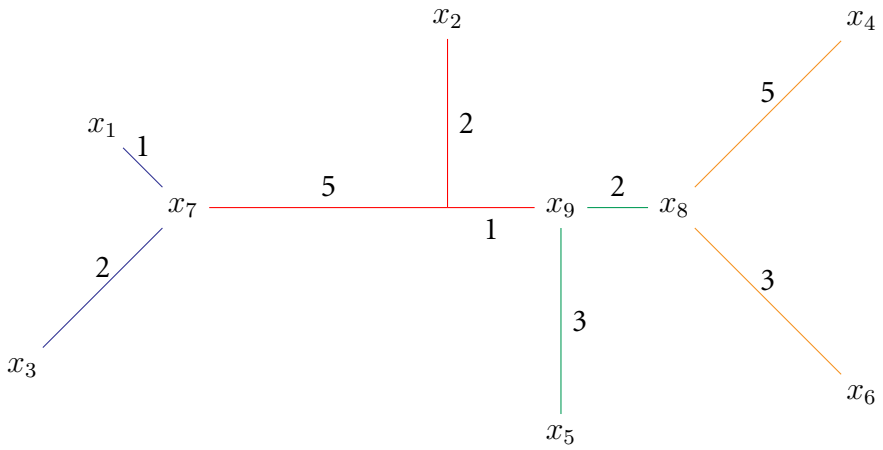


M_d	x_2	x_7	x_9
x_2	0	7	3
x_7	7	0	6
x_9	3	6	0

Aus diesen Abständen lässt sich nun ein eindeutiger Baum erstellen:



Nun kann man alle Teilbäume zusammensetzen und erhält:



Vergleicht man diesen Baum mit der anfänglichen Abstandsmatrix, so sieht man, dass d und $d^{\mathscr{T}}$ übereinstimmen.

Den neighbor-joining-Algorithmus kann man auch anwenden, wenn die Abstandsfunktion nicht die 4-Punkt-Bedingung erfüllt. Dies ist sehr praktisch, da die meisten Abstandsfunktionen, die man aus gegebenen Sequenzen erhält, diese Bedingung nicht erfüllen. Auch die Dreiecksungleichung ist nicht immer erfüllt, so dass man meistens mit *Semimetriken* arbeitet. Semimetriken sind Abstandsfunktionen, die die Dreiecksungleichung nicht unbedingt erfüllen. Wir werden Semimetrikmatrizen wie die Abstandsmatrizen mit $M_d = (d_{ij})$ bezeichnen.
Wendet man den neighbor-joining-Algorithmus auf Semimetriken ohne 4-Punkt-Bedingung an, so können verschiedene Anormalitäten auftreten. Es kann sein, dass aus dem selben Datensatz verschiedene Bäume konstruiert werden können, es können negative Astlängen auftreten (Biologen können dies in einigen Fällen interpretieren) und die Funktionen $d^{\mathscr{T}}$ müssen nicht mit dem gegebenen d übereinstimmen.

3.4 Beispiel Gegeben sei $N = 4$ und folgende Semimetrikmatrix:

M_d	x_1	x_2	x_3	x_4
x_1	0	5	2	7
x_2	5	0	1	8
x_3	2	1	0	3
x_4	7	8	3	0

Die Dreiecksbedingung ist nicht erfüllt, da $d_{14} > d_{13}+d_{34}$. Wegen $d_{13}+d_{24} = 10, d_{12}+d_{34} = 8, d_{14} + d_{23} = 8$ ist auch die 4-Punkt-Bedingung nicht erfüllt.
Wenden wir den neighbor-joining-Algorithmus an, erhalten wir nach Formel 2:

$$r_1 = 7, r_2 = 7, r_3 = 3, r_4 = 9$$

und nach Formel 3:

D	x_1	x_2	x_3	x_4
x_1		-9	-8	-9
x_2			-9	-8
x_3				-9

D_{12}, D_{14} und D_{34} sind minimal. Wir werden nun sehen, dass verschiedene Paarungen von OTUs zu verschiedenen Bäumen führen.
Fassen wir x_1 und x_2 zu x_5 zusammen, ergibt sich folgende neue Matrix:

M_d	x_3	x_4	x_5
x_3	0	3	-1
x_4	3	0	5
x_5	-1	5	0

Diese Funktion dieser Matrix ist keine Semimetrik, da die Matrix negative Einträge enthält. Trotzdem können wir nach Formel 1 einen Baum (mit einem Ast negativer Länge) daraus bilden:

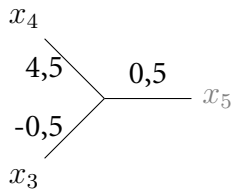


Abbildung 4

Auch wenn dieser Baum einige Äste negativer Länge hat, lässt sich der Baum zu $\mathscr{T}_1$ vervollständigen (Abbildung 5):

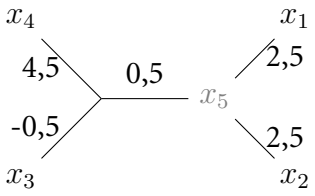


Abbildung 5: $\mathscr{T}_1$

Die von $\mathscr{T}_1$ erzeugte Metrik lässt sich an dem Baum ablesen:

$M_{d\mathscr{T}}$	x_1	x_2	x_3	x_4
x_1	0	5	1,5	7,5
x_2	5	0	1,5	7,5
x_3	1,5	1,5	0	3
x_4	7,5	7,5	3	0

Diese Semimetrik stimmt offensichtlich nicht (wie gewünscht) mit M_d überein.

Fassen wir hingegen x_1 und x_4 zu einer neuen OTU x_5 zusammen, so erhalten wir $d_{15} = 2,5$ und $d_{54} = 6,5$. Die neue „Semimetrikmatrix“ ist:

M_d	x_2	x_3	x_5
x_2	0	1	3
x_3	1	0	−1
x_5	3	−1	0

Auch diese Matrix ist keine Semimetrikmatrix, da sie negative Einträge enthält. Wir können wiederum trotzdem den Algorithmus anwenden:

Dies ergibt folgenden Baum $\mathscr{T}_2$:

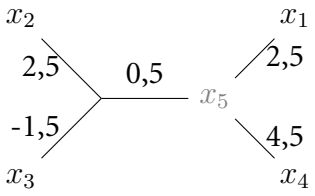


Abbildung 6: $\mathscr{T}_2$

Die Baumtopologie von $\mathscr{T}_2$ unterscheidet sich von der von $\mathscr{T}_1$. Auch $\mathscr{T}_2$ hat wieder negative Äste, wir berechnen aber wieder $d^{\mathscr{T}_2}$:

$M_{d^{\mathscr{T}_2}}$	x_1	x_2	x_3	x_4
x_1	0	$5,5$	$1,5$	7
x_2	$5,5$	0	1	$7,5$
x_3	$1,5$	1	0	$3,5$
x_4	7	$7,5$	$3,5$	0

Diese Matrix unterscheidet sich sowohl von d als auch von $M_{d^{\mathscr{T}_2}}$. Der neighbor-joining-Algorithmus erzeugt hier also zwei verschiedene Bäume, die beide Äste negativer Länge haben und nicht die gegebene Semimetrik erzeugen.

3.2 UPGMA-Algorithmus

UPGMA steht für „Unweighted Pair Group Method Using Arithmetic Averages“. Der UPGMA-Algorithmus liefert gerichtete Bäume. Hierfür benötigen wir zunächst den Begriff der Ultrametrik.

3.5 Definition (Ultrametrik) Eine Abstandsfunktion d auf einer Menge $\mathscr{M}$ von OTUs heißt ultrametrisch, falls zusätzlich gilt:

- $d_{xz} \leq \max\{d_{xy}, d_{yz}\}$ für alle $x, y, z \in \mathscr{M}$

3.6 Bemerkung Ultrametriken erfüllen die 4-Punkt-Bedingung.

Gegeben sei eine ultrametrische Abstandsfunktion d auf einer Menge $\mathscr{M} = \{x_1, \ldots, x_N\}$. Wir setzen die OTUs als Blätter eines gerichteten Baumes und bauen den Baum von unten nach oben hin auf, indem wir neue OTUs einführen, die die inneren Knoten darstellen. Bei UPGMA werden die OTUs zunächst zu Clustern zusammengefasst. Den Abstand zwischen zwei Clustern C_i und C_j aus $\mathscr{M}$ definieren wir wie folgt:

$$d(C_i, C_j) = (\overline{\overline{C_i}} \cdot \overline{\overline{C_j}})^{-1} \sum_{a \in C_i, b \in C_j} d_{ab} \tag{7}$$

wobei $\overline{\overline{C_i}}$ die Anzahl der Elemente in C_i bezeichnet.

Zu Beginn ordne jede OTU x_i einem einelementigen Cluster C_i zu, wir sagen x_i gehöre zu C_i. Wähle nun zwei Cluster, für die $D(C_i, C_j)$ minimal ist. Definiere ein neues Cluster $C_{N+1} = C_i \cup C_j$ und berechne mit Formel 7 die Abstände von C_{N+1} zu den übrigen Clustern.

Setze eine neue OTU x_{N+1} auf die absolute Höhe

$$0,5 \cdot d(C_i, C_j) \tag{8}$$

über x_i und x_j. Die neue OTU gehört zu dem Cluster C_{N+1} und repräsentiert den Knoten,der im fertigen Baum x_i und x_j verbindet. Ersetze die Variablen x_i und x_j

durch x_{N+1} und setze die Abstände zu den anderen OTUs als die Abstände zwischen den zugehörigen Clustern.
Nun haben wir $N - 1$ Cluster, für die wir den Vorgang wiederholen können. Wir wiederholen den Vorgang, bis nur noch zwei Cluster übrigbleiben, sagen wir, C_m und C_l. Über diese setzen wir die Wurzel des Baumes auf der Höhe

$$0,5 \cdot d(C_m, C_l). \tag{9}$$

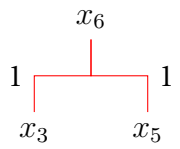
3.7 Beispiel Gegeben sei $N = 5$ und die folgende Abstandsmatrix:

M_d	x_1	x_2	x_3	x_4	x_5
x_1	0	16	6	16	6
x_2	16	0	16	8	16
x_3	6	16	0	16	2
x_4	16	8	16	0	16
x_5	6	16	2	16	0

Da d_{35} minimal ist, bilden wir ein Cluster $C_6 = \{x_3, x_5\}$. Aus Formel 7 bekommen wir

$$d(C_1, C_6) = 0,5(d_{13} + d_{15}) = 6$$
$$d(C_2, C_6) = 0,5(d_{23} + d_{25}) = 16$$
$$d(C_4, C_6) = 0,5(d_{43} + d_{45}) = 16$$

Wir führen nun eine neue OTU x_6 ein und setzen sie auf die Höhe $0,5 \cdot d_{35} = 1$ über x_3 und x_5 (Formel 8).



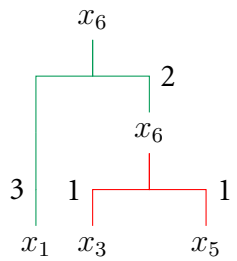
Für die OTUs x_1, x_2, x_4 und x_6 bekommen wir nun folgende Abstandsmatrix:

M_d	x_1	x_2	x_4	x_6
x_1	0	16	16	6
x_2	16	0	8	16
x_4	16	8	0	16
x_6	6	16	16	0

Hier ist d_{16} minimal; deswegen bilden wir ein neues Cluster $C_7 = \{x_1, x_6\} = \{x_1, x_3, x_5\}$. Aus Formel 7 bekommen wir:

$$d(C_2, C_7) = \frac{1}{3} \cdot (d_{21} + d_{23} + d_{25}) = 16$$
$$d(C_4, C_7) = \frac{1}{3} \cdot (d_{41} + d_{43} + d_{45}) = 16$$

Nun definieren wir eine neue OTU x_7 und setzen sie auf die absolute Höhe $0,5 \cdot d_{16} = 3$ über x_1 und x_6 (Formel 8).



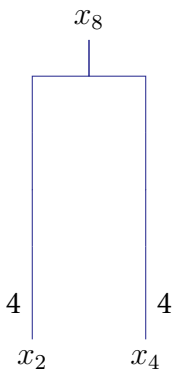
Für die verbleibenden OTU x_2, x_4 und x_7 ergibt sich die folgende Abstandsmatrix:

M_d	x_2	x_4	x_7
x_2	0	8	16
x_4	8	0	16
x_7	16	16	0

Hier ist d_{24} minimal und wir bilden das Cluster $C_8 = \{x_2, x_4\}$. Aus Formel 7 erhalten wir

$$d(C_7, C_8) = \frac{1}{3 \cdot 2}(d_{12} + d_{32} + d_{52} + d_{14} + d_{34} + d_{54}) = 16$$

Dann bilden wir die neue OTU x_8 und setzen sie auf die absolute Höhe $0,5 \cdot d_{24} = 4$ über x_2 und x_4.



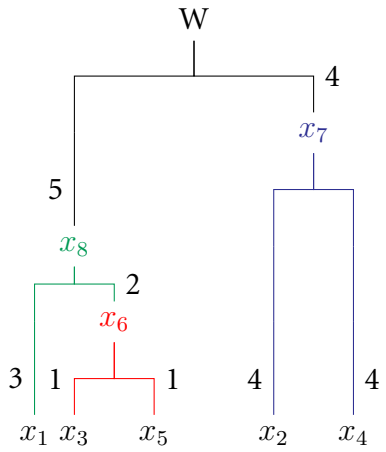


Abbildung 7: $\mathscr{T}$

Zum Schluss setzen wir die Wurzel auf die Höhe $0,5 \cdot d(C_7, C_8) = 8$ über x_7 und x_8 (Formel 9) und haben damit den vollständigen Baum.

3.8 Beispiel Wenden wir nun auf die Abstandsmatrix aus dem obigen Beispiel den neighbor-joining-Algorithmus an:

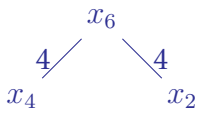
M_d	x_1	x_2	x_3	x_4	x_5
x_1	0	16	6	16	6
x_2	16	0	16	8	16
x_3	6	16	0	16	2
x_4	16	8	16	0	16
x_5	6	16	2	16	0

Wir bekommen

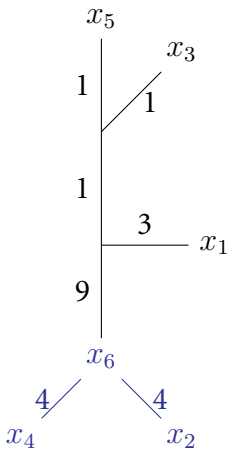
$$r_1 = \frac{44}{3}, r_2 = \frac{56}{3}, r_3 = \frac{40}{3}$$
$$r_4 = \frac{56}{3}, r_5 = \frac{40}{3}$$

D	x_1	x_2	x_3	x_4	x_5
x_1		$-\frac{52}{3}$	-22	$-\frac{52}{3}$	-22
x_2			-16	$-\frac{88}{3}$	-16
x_3				-16	$-\frac{74}{3}$
x_4					-16

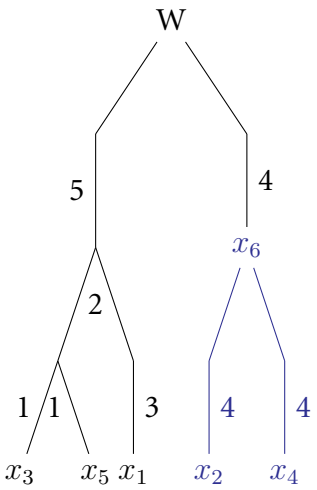
D_{24} ist minimal. Ersetze x_2 und x_4 durch x_6:



Nach zwei weiteren Schritten erhalten wir folgenden Baum:



Wir können eine Wurzel so setzen, dass alle OTU von der Wurzel gleich weit entfernt sind:



Dieser Baum ist (bis auf die Darstellungsform) identisch mit dem von UPGMA erzeugten Baum (Abb.7). Bei ultrametrischen Abstandsfunktionen erzeuget der neighbor-joining-Algorithmus also die ungerichtete Variante des Baumes, den der UPGMA-Algorithmus erzeugt.

Wenn eine Abstandsfunktion die 4-Punkt-Bedingung erfüllt, aber nicht ultrametrisch ist, so kann der UPGMA-Algorithmus den falschen Baum erzeugen. Dazu folgendes Beispiel:

3.9 Beispiel Gegeben sei $N = 4$ und die folgende Abstandsmatrix:

M_d	x_1	x_2	x_3	x_4
x_1	0	3	9	9
x_2	3	0	10	8
x_3	9	10	0	16
x_4	9	8	16	0

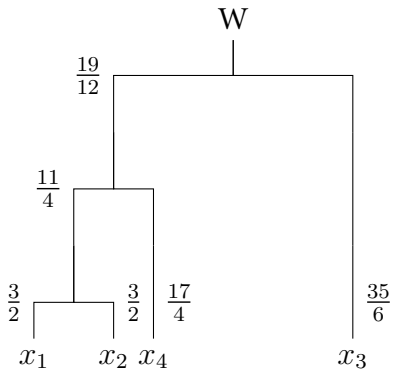
Man kann nachrechnen, dass d eine Abstandsfunktion ist und dass die 4-Punkt-Bedingung erfüllt. Wegen $d_{23} = 10 > 9 = \max\{d_{21}, d_{13}\}$ ist d nicht ultrametrisch.

Wir wenden nun den UPGMA-Algorithmus auf d an. Da d_{12} minimal ist, bilden wir ein neues Cluster $C_5 = \{x_1, x_2\}$ und die neue OTU x_5, die wir auf die absolute Höhe $1,5$ über x_1 und x_2 setzen.

Die Abstandsmatrix für x_3, x_4 und x_5 ist nun:

M_d	x_3	x_4	x_5
x_3	0	16	$9,5$
x_4	16	0	$8,5$
x_5	$9,5$	$8,5$	0

Hier ist d_{45} minimal und wir bilden das Cluster $C_6 = \{x_1, x_2, x_4\}$ und die neue OTU x_6, die wir in der Höhe $4,25$ über x_4 und x_5 setzen. Der Abstand von C_3 zu C_7 ist $\frac{35}{3}$ und wir setzen deswegen die Wurzel über x_3 und x_7 auf die absolute Höhe $\frac{35}{6}$.



Man sieht schnell, dass $d^{\mathcal{T}} \neq d$.

3.3 Methode der kleinsten Quadrate

Wenn die Dreiecksungleichung oder die 4-Punkt-Bedingung nicht erfüllt sind, liefern die beiden vorgestellten Algorithmen in der Regel falsche Ergebnisse. Wir können aber versuchen, für eine gegebende Semimetrik einen optimalen Baum $\mathscr{T}$ zu finden, so dass d und $d^{\mathscr{T}}$ möglichst nah beieinander liegen (wobei natürlich zu definieren ist, was „nah beieinander“ in diesem Zusammenhang heißen soll).

Wir werden hier die Methode der kleinsten Quadrate verwenden: Für zwei Semimetriken d und d auf derselben N-elementigen Menge $\mathscr{M}$ sei die Summe der Quadrate definiert als

$$\rho(d, d') = \sum_{1 \le i,j \le N} (d_{ij}, -d'_{ij})^2$$

Wir betrachten nun den Spezialfall $d' = d^{\mathscr{T}}$, wobei $\mathscr{T}$ ein ungerichteter Baum ist, der die gegebenen OTUs verbindet. wir setzen

$$ss_d(\mathscr{T}) = \rho(d, d^{\mathscr{T}})$$

Die Methode der kleinsten Quadrate wählt nun unter allen ungerichteten Bäumen $\mathscr{T}$ den Baum aus, für den $ss_d(\mathscr{T})$ minimal ist (sofern im Raum der ungerichteten Bäume minimale Elemente bezüglich ss_d existieren). Jeder dieser Bäume ist optimal im Sinne der Methode der kleinsten Quadrate.

Im Idealfall minimiert die Methode der kleinsten Quadrate ss_d über *alle* ungerichteten Bäume; ab einer gewissen Anzahl von OTUs ist es natürlich nur noch möglich, einige Baumtopologien und einige Astlängen zu betrachten, was die Sensitivität dieser Methode verringert. Als Astlängen können entweder alle reellen Zahlen oder nur nichtnegative Zahlen zugelassen sein.

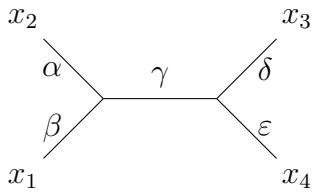
Im folgenden ein Beispiel, in dem eine analytische Lösung möglich ist.

3.10 Beispiel Gegeben sei wie in Beispiel 3.4 $N = 4$ und die Abstandsmatrix

M_d	x_1	x_2	x_3	x_4
x_1	0	5	2	7
x_2	5	0	1	8
x_3	2	1	0	3
x_4	7	8	3	0

Es gibt nun drei mögliche Topologien von Bäumen, die die vier OTUs verbinden.

1. $\mathscr{T}_1(\alpha, \beta, \gamma, \delta, \varepsilon)$:



Sei nun
$\varphi_1(\alpha,\beta,\gamma,\delta,\varepsilon) := ss_d((\mathscr{T}_1))$
$= (\alpha+\beta-5)^2+(\beta+\gamma+\delta-2)^2+(\beta+\gamma+\varepsilon-7)^2+(\alpha+\gamma+\delta-1)^2+(\alpha+\gamma+\varepsilon-8)^2+(\delta+\varepsilon-3)^2$
Um nun das Minimum von φ_1 zu finden, also die Werte $(\alpha,\beta,\gamma,\delta,\varepsilon)$, für die der Abstand zwischen d und $d^{\mathscr{T}}$ minimal ist, leiten wir φ_1 partiell nach den 5 Variablen ab und bestimmen die Nullstellen der Ableitung:
Die Ableitung nach α ist

$$\partial_\alpha(\varphi_1)(\alpha,\beta,\gamma,\delta,\varepsilon) = 2(\alpha+\beta-5)+2(\alpha+\gamma+\delta-1)+2(\alpha+\gamma+\varepsilon-8)$$
$$= 6\alpha+2\beta+4\gamma+2\delta+2\varepsilon-28$$

Um die Nullstellen der Ableitung zu finden, müssen wir also folgende Gleichung lösen:

$$3\alpha+\beta+2\gamma+\delta+\varepsilon-14=0$$

Zusammen mit den anderen partiellen Ableitungen erhalten wir das folgende Gleichungssystem:

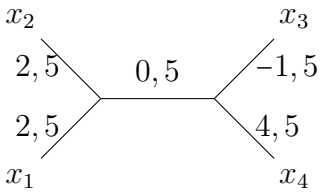
$$\begin{aligned}
3\alpha+\beta+2\gamma+\delta+\varepsilon-14&=0\\
\alpha+3\beta+2\gamma+\delta+\varepsilon-14&=0\\
2\alpha+2\beta+4\gamma+2\delta+2\varepsilon-18&=0\\
\alpha+\beta+2\gamma+3\delta+\varepsilon-6&=0\\
\alpha+\beta+2\gamma+\delta+3\varepsilon-18&=0
\end{aligned}$$

Die Lösung dieses Systems ist

$$\begin{aligned}
\alpha&=2,5\\
\beta&=2,5\\
\gamma&=0,5\\
\delta&=-1,5\\
\varepsilon&=4,5
\end{aligned}$$

An dieser Stelle hat φ_1 sein globales Minimum und es gilt

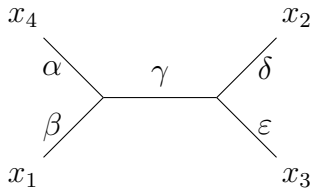
$$\varphi_1(\alpha,\beta,\gamma,\delta,\varepsilon) = (2,5+2,5-5)^2+(2,5+0,5-1,5-2)^2+(2,5+0,5+4,5-7)^2+(2,5+0,5-1,5-1)^2$$



$$+(2,5+0,5+4,5-8)^2+(-1,5+4,5-3)^2=1$$

Der zugehörige Baum ist: Der hier entstehende Baum ist der selbe wie auf Abbildung 5 in Beispiel 3.4.

2. $\mathscr{T}_2(\alpha,\beta,\gamma,\delta,\varepsilon)$:



Analog zeigen wir, dass $\varphi_2(\alpha,\beta,\gamma,\delta,\varepsilon) := ss_d((\mathscr{T}_2))$ von folgendem Baum minimiert wird:

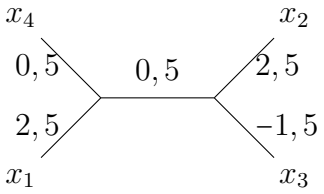
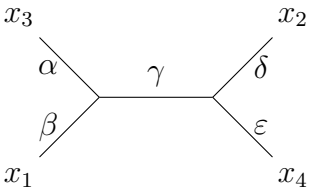


Abbildung 8: $\mathscr{T}_2$

Es gilt $\min(\varphi_2)=1$.

Dieser Baum ist identisch mit dem auf Abbildung 6 aus Beispiel 3.4.

3. $\mathscr{T}_3(\alpha,\beta,\gamma,\delta,\varepsilon)$:



Analog erhalten wir für $\varphi_3(\alpha,\beta,\gamma,\delta,\varepsilon) := ss_d((\mathscr{T}_3))$ folgenden Baum:

Hier gilt $\min(\varphi_3)=0$, der Baum erzeugt also die gewünschte Abstandsfunktion.

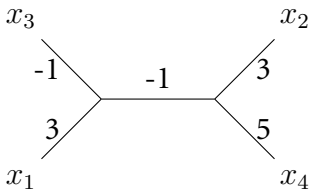


Abbildung 9: $\mathscr{T}_2$