

Bioinformatik 12. 7. 2004, M. Sauerhoff

Neighbor-Joining

Gegeben: n Taxa, Distanzen d_{ij} , $1 \leq i, j \leq n$, die Metrik bilden.

Ein *additiver Baum* ist ein Baum mit den Taxa als Blättern und der Eigenschaft, dass für beliebige verschiedene Blätter i, j die Summe der Kantengewichte auf dem Pfad $i \rightsquigarrow j$ gleich d_{ij} ist.

Aufgabe: Konstruiere einen additiven Baum mit folgenden weiteren Eigenschaften:

- Der Baum ist binär und hat keine Wurzel, alle inneren Knoten haben Grad 3.
- Alle Kantengewichte sind positiv.

Nenne solche Bäume *positiv additiv*.

Algorithmus Neighbor-Joining

$A := \{1, 2, \dots, n\}$; (aktive Knoten)

while $|A| \geq 3$ **do**

$$(1) \forall i \in A: r_i := \frac{1}{|A|-2} \sum_{k \neq i} d_{ik};$$

$$\forall i, j \in A, i \neq j: D_{ij} := d_{ij} - (r_i + r_j);$$

(2) wähle $i \neq j$ mit minimalem D_{ij} ;

(3) $A := A - \{i, j\} + \{(i, j)\}$;

$$(4) d_{i,(i,j)} := \frac{1}{2}d_{ij} + \frac{1}{2}(r_i - r_j);$$

$$d_{j,(i,j)} := \frac{1}{2}d_{ij} + \frac{1}{2}(r_j - r_i);$$

$$\forall k \in A, k \neq (i, j): d_{k,(i,j)} := \frac{1}{2}(d_{ik} + d_{jk} - d_{ij});$$

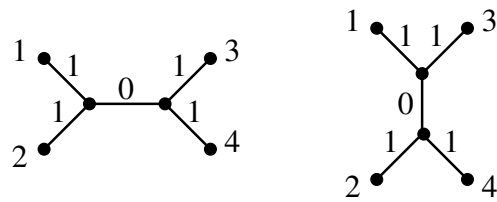
od;

Behandlung der Fälle $|A| \leq 2$.

Theorem 1: Falls zu gegebenen metrischen Distanzen ein positiver additiver Baum existiert, dann rekonstruiert der Algorithmus Neighbor-Joining diesen. Insbesondere gibt es damit also auch nur genau einen positiven additiven Baum.

Bemerkung: Mit Hilfe der Dreiecksungleichung folgt sofort, dass in einem beliebigen additiven Baum alle Kantengewichte zumindest nichtnegativ sind. Der Neighbor-Joining-Algorithmus liefert auch für den Fall additiver Bäume, in denen Kanten mit Gewicht 0 vorkommen, einen additiven Baum zurück. Allerdings kann es in diesem Fall mehrere verschiedene Bäume geben.

Beispiel:



Man kann den allgemeinen Fall auf den positiver Kantengewichte zurückführen, indem man zunächst Kanten mit Gewicht 0 kontrahiert. Der entsprechende Beweis der Korrektheit benutzt im Wesentlichen dieselben Ideen wie hier vorgestellt, wird durch die Mehrdeutigkeiten allerdings unübersichtlicher. Wir verbieten daher der Einfachheit halber Kanten mit Gewicht 0. Für den Beweis von Theorem 1 benutzen wir folgende zwei Lemmata.

Lemma 2: In einem additiven Baum mit drei Blättern und einem inneren Knoten u gilt für verschiedene Blätter i, j, k :

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2.$$

Beweis. Aufgrund der Additivität gilt:

$$d_{iu} + d_{uj} = d_{ij},$$

$$d_{iu} + d_{uk} = d_{ik},$$

$$d_{ju} + d_{uk} = d_{jk}.$$

Addieren der letzten zwei Gleichungen und Subtrahieren der ersten liefert

$$2d_{ku} = d_{ik} + d_{jk} - d_{ij}$$

und damit die Behauptung. □

Lemma 3: Sei ein positiver additiver Baum T gegeben und sei D_{ij} das Minimum in der D -Matrix. Dann sind i, j Geschwister in T .

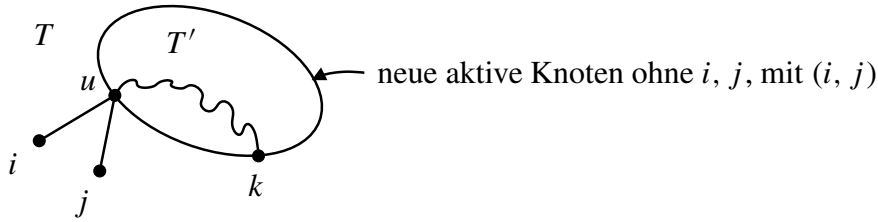
Bemerkung: Zwei Knoten sind *Geschwister*, wenn sie über einen Pfad der Länge 2 verbunden sind. Der Begriff *Nachbarn* ist in der Graphentheorie für Knoten reserviert, die direkt über eine Kante verbunden sind.

Wir verschieben den Beweis von Lemma 3 auf später und beweisen zunächst das Theorem.

Beweis von Theorem 1. Wir zeigen dies per vollständiger Induktion über n . Für $n = 2$ ist die Aussage trivial. Sei also $n \geq 3$ und sei gezeigt, dass der Algorithmus einen vorgegebenen positiven additiven Baum auf $n - 1$ Knoten rekonstruiert.

Wir betrachten n Knoten mit einem additiven Baum T . Sei D_{ij} das Minimum in der D -Matrix, d. h. i, j werden vom Algorithmus zusammengefasst. Um die Induktionsvoraussetzung anwenden zu können, zeigen wir zunächst, dass auf der neuen Menge von $n - 1$ aktiven Knoten auch wieder ein positiver additiver Baum existiert.

Nach Lemma 3 sind in T die Knoten i, j Geschwister. Sei u der Knoten, der i und j verbindet und sei T' der positive additive Teilbaum, der aus T durch Löschen von i und j entsteht.



Wir zeigen, dass T' zu den Distanzen nach dem Update im Algorithmus passt, wenn wir u mit dem Knoten (i, j) identifizieren. Die Distanz zwischen Blättern $k, \ell \neq u$ in T' ist $d_{k\ell}$, wie im Algorithmus nach dem Update. Betrachte nun die Distanz zwischen einem Blatt $k \neq u$ und dem Knoten u in T' . Aufgrund von Lemma 2 ist

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2.$$

Dies ist aber gerade die vom Algorithmus berechnete neue Distanz $d_{k,(i,j)}$. Also ist T' ein positiver additiver Baum auf den neuen aktiven Knoten und neuen Distanzen nach dem Update im Algorithmus und wir können die Induktionsvoraussetzung anwenden, dass der Algorithmus diesen Baum rekonstruiert.

Es bleibt zu zeigen, dass die Kantengewichte $d_{i,(i,j)}$ und $d_{j,(i,j)}$ positiv sind. Einsetzen der Definition und Rechnen liefert:

$$\begin{aligned} d_{i,(i,j)} &= \frac{1}{2}d_{ij} + \frac{1}{2}(r_i - r_j) \\ &= \frac{1}{2}d_{ij} + \frac{1}{2(n-2)} \left(\sum_{k \neq i} d_{ik} - \sum_{k \neq j} d_{jk} \right) \\ &= \frac{1}{2}d_{ij} + \frac{1}{2(n-2)} \sum_{k \notin \{i,j\}} (d_{ik} - d_{jk}) \\ &= \frac{1}{2(n-2)} \sum_{k \notin \{i,j\}} (d_{ij} + d_{ik} - d_{jk}). \end{aligned}$$

Wieder mit Hilfe von Lemma 2 folgt:

$$d_{ij} + d_{ik} - d_{jk} = 2d_{iu},$$

wobei u wie oben der Verbindungsknoten von i und j ist. Da alle Kantengewichte in dem per Voraussetzung vorgegebenen Baum positiv sind, ist $d_{iu} > 0$ und damit $d_{i,(i,j)} > 0$. \square

Es bleibt das entscheidende Lemma 3 zu zeigen.

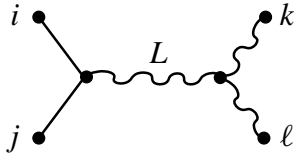
Beweis von Lemma 3. Die Aussage ist trivial für $n \leq 3$, da dann alle Knotenpaare Geschwister sind. Sei also $n \geq 4$. Wir brauchen einige Vorbereitungen und zeigen zunächst zwei Zwischenbehauptungen.

Zwischenbehauptung 1. Seien i, j Geschwister in einem positiven additiven Baum mit $n \geq 4$ Blättern. Dann ist D_{ij} striktes Zeilen- und Spaltenminimum in der D -Matrix.

Beweis der Zwischenbehauptung 1. Durch Einsetzen der Definitionen und Rechnung ergibt sich:

$$\begin{aligned}
 D_{ik} - D_{ij} &= d_{ik} - (r_i + r_k) - (d_{ij} - (r_i + r_j)) = d_{ik} - d_{ij} - r_k + r_j \\
 &= d_{ik} - d_{ij} - \frac{1}{n-2} \sum_{\ell \neq k} d_{k\ell} + \frac{1}{n-2} \sum_{\ell \neq j} d_{j\ell} \\
 &= \frac{1}{n-2} \sum_{\ell \notin \{i, j, k\}} (d_{j\ell} - d_{k\ell}) + d_{ik} - d_{ij} + \frac{1}{n-2} (-d_{ki} - d_{kj} + d_{ji} + d_{jk}) \\
 &= \frac{1}{n-2} \sum_{\ell \notin \{i, j, k\}} (d_{ik} + d_{j\ell} - d_{ij} - d_{k\ell}).
 \end{aligned}$$

Betrachte nun die Situation im gegebenen Baum für ein beliebiges Blatt $\ell \notin \{i, j, k\}$ (beachte, dass i und j Geschwister sind):



Es ist dann aufgrund der Additivität $d_{ik} + d_{j\ell} - d_{ij} - d_{k\ell} = 2L$. Da alle Kantengewichte positiv sind, folgt $L > 0$. Wegen $n \geq 4$ ist auch die obige Summe positiv und damit $D_{ik} < D_{ij}$. \square

Für den Rest des Beweises fixiere einen positiven additiven Baum mit $n \geq 4$ Blättern. Sei E dessen Kantenmenge. Für Blätter i, j sei $p(i, j)$ die Menge der Kanten auf dem Pfad $i \leftrightarrow j$ im gewählten Baum. Für $e \in E$ sei $d(e)$ das Gewicht der Kante e . Schließlich bezeichne für ein Blatt i und $e \in E$ mit $b(i, e)$ die Anzahl Blätter, die von i aus über einen Pfad erreichbar sind, der e enthält.

Wir benutzen diese Definitionen, um die r_i -Werte im Algorithmus in den Griff zu bekommen. Es gilt offenbar für beliebige i :

$$r_i = \frac{1}{n-2} \sum_{k \neq i} d_{ik} = \frac{1}{n-2} \sum_{e \in E} d(e) b(i, e).$$

Weiterhin haben wir folgende Abschätzung.

Zwischenbehauptung 2. Seien i, j, k, ℓ verschiedene Knoten, wobei i, j Blätter seien. Es existiere ein Pfad $i \leftrightarrow k \leftrightarrow \ell$ und die Kante $\{k, j\}$. Dann ist

$$r_i - r_j \leq d_{ik} - d_{\ell j} + \frac{1}{n-2} \sum_{e \in p(k, \ell)} d(e) (b(i, e) - b(j, e)).$$

Beweis der Zwischenbehauptung 2. Wir wenden die obige Beobachtung an um r_i und r_j umzuschreiben.

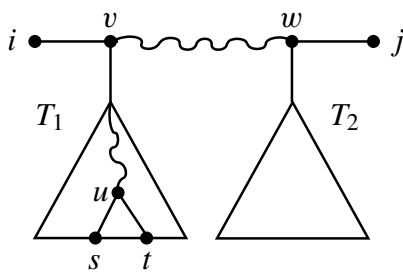
$$\begin{aligned}
 r_i - r_j &= \frac{1}{n-2} \sum_{e \in E} d(e) \underbrace{(b(i, e) - b(j, e))}_{= 0, \text{ falls } e \notin p(i, j)} \\
 &= \frac{1}{n-2} \left(\sum_{e \in p(i, k)} d(e) \underbrace{(b(i, e) - b(j, e))}_{\leq n-1} + \sum_{\substack{e \in p(\ell, j) \\ = \{\ell, j\}}} d(e) \underbrace{(b(i, e) - b(j, e))}_{= 1} \underbrace{}_{= n-1} \right. \\
 &\quad \left. + \sum_{e \in p(k, \ell)} d(e) (b(i, e) - b(j, e)) \right) \\
 &\leq d_{ik} - d_{\ell j} + \frac{1}{n-2} \sum_{e \in p(k, \ell)} d(e) (b(i, e) - b(j, e)).
 \end{aligned}$$

Die in der zweiten Zeile angegebenen Abschätzungen bzw. Werte für $b(i, e), b(j, e)$ ergeben sich direkt aus der Definition. Über eine Kante e auf dem Teilpfad $i \leftrightarrow k$ sind von i aus höchstens $n - 1$ Blätter erreichbar (alle anderen außer i) und von j aus mindestens eins (nämlich i). Da der Teilpfad $p(\ell, j)$ nur aus einer Kante besteht, erhalten wir hier auf analoge Weise sogar die exakten Werte. \square

Wir beweisen nun das Lemma. Wir nehmen an, dass i, j keine Geschwister im betrachteten Baum sind und leiten daraus einen Widerspruch her. Wir betrachten zwei Fälle.

1. Fall: i oder j (oder beide) haben ein Blattgeschwister. O. B. d. A. habe i ein Blattgeschwister k . Dann folgt aus der Zwischenbehauptung 1, dass $D_{ik} < D_{ij}$ (striktes Zeilenminimum). Andererseits ist aber $D_{ij} \leq D_{ik}$. Widerspruch.

2. Fall: Weder i noch j haben ein Blattgeschwister. Dann ergibt sich die Situation in folgender Abbildung.



Die Blätter i, j hängen über je eine Kante an Knoten v bzw. w . Da i, j keine Geschwister sind, sind v, w verschieden und durch einen Pfad mit mindestens einer Kante verbunden. Die Knoten v, w sind innere Knoten vom Grad 3, an denen wie im Bild gezeigt Teilbäume T_1, T_2 hängen. Jeder der beiden Bäume T_1, T_2 hat mindestens zwei Blätter, da ansonsten i bzw. j ein Blattgeschwister hätte. Damit muss es auch jeweils ein Paar von Blättern in diesen Bäumen geben, die Geschwister sind.

Bezeichne mit $|T_i|$ die Anzahl Blätter in T_i und sei o. B. d. A. $|T_1| \leq |T_2|$. Seien s, t Blätter in T_1 , die Geschwister sind, und sei u deren Verbindungsknoten. Wir zeigen nun, dass $D_{st} < D_{ij}$, im Widerspruch zur Minimalität von D_{ij} .

Es ist

$$D_{ij} - D_{st} = d_{ij} - d_{st} - (r_i + r_j - r_s - r_t). \quad (1)$$

Wir wenden die Zwischenbehauptung 2 an, um $r_i - r_s$ abzuschätzen. Dazu betrachten wir den Pfad $i - v \rightsquigarrow u - s$.

$$r_i - r_s \leq d_{iv} - d_{us} + \frac{1}{n-2} \sum_{e \in p(v,u)} d(e)(b(i, e) - b(s, e)).$$

Ein Pfad, der von i aus über eine Kante $e \in p(v, u)$ läuft, erreicht nur Blätter in T_1 , also ist dann $b(i, e) \leq |T_1|$. Ein Pfad, der von s startet und über eine solche Kante läuft, erreicht auf jeden Fall die Blätter i, j und alle Blätter in T_2 (evtl. noch mehr), also ist $b(s, e) \geq |T_2| + 2$. Wir haben damit (unter Ausnutzung von $|T_1| \leq |T_2|$):

$$r_i - r_s \leq d_{iv} - d_{us} + \frac{1}{n-2} d_{vu} (|T_1| - |T_2| - 2) \leq d_{iv} - d_{us} - \frac{2}{n-2} d_{vu}. \quad (2)$$

Auf analoge Weise erhalten wir für den Pfad $j \rightsquigarrow v \rightsquigarrow u - t$:

$$r_j - r_t \leq d_{jv} - d_{ut} - \frac{2}{n-2} d_{vu}. \quad (3)$$

Durch Einsetzen von (2) und (3) in (1) folgt:

$$D_{ij} - D_{st} \geq d_{ij} - d_{st} - \left(\underbrace{d_{iv} + d_{jv}}_{=d_{ij}} - \underbrace{d_{us} - d_{ut}}_{=-d_{st}} - \frac{4}{n-2} d_{vu} \right) = \frac{4}{n-2} d_{vu} > 0,$$

also der gewünschte Widerspruch. □