Bioinformatik I, WS00/01
Übungsblatt 7

Besprechung am 11.01.2001

1 Pair Group Method

1. Implementieren Sie den *unweighted pair group method Algorithmus (UPGMA)* zur Stammbaumkonstruktion (siehe auch Anhang). Das Programm soll eine Distanzmatrix einlesen und einen Baum mit Distanzen (z.B. in geklammerter Darstellung) ausgeben.

2. Implementieren Sie (ausgehend von Ihrem UPGMA-Programm) den *weighted pair group method Algorithmus (WPGMA)* (siehe auch Anhang). WPGMA erzeugt wie UPGMA einen Stammbaum aus einer Distanzmatrix. Im Unterschied zu UPGMA werden die Distanzen zu einem neuen Cluster in folgender Weise berechnet. Bei der Verschmelzung der Cluster c und d sei die Distanz zum neuen Cluster $e = c \cup d$ eines Clusters $f \neq e$ gegeben als

$$\mathrm{dist}(e, f) = \mathrm{dist}(f, e) = \frac{|c| \,\mathrm{dist}(f, c) + |f| \,\mathrm{dist}(f, d)}{|c| + |d|},$$

wobei $|c|$ die Grösse eines Clusters c bezeichne, d.h. die Anzahl der Taxa in diesem Cluster. (zum Vergleich: bei UPGMA wird $\mathrm{dist}(c \cup d, f) = \mathrm{dist}(f, c \cup d)$ als $\frac{\mathrm{dist}(f,c)+\mathrm{dist}(f,d)}{2}$ berechnet.) Offensichtlich erfordert die Gewichtung in WPGMA die Kenntnis der Grössen der zu verschmelzenden Cluster. Deshalb kann/sollte man die Grössen aller Cluster in Variablen mitführen und bei jeder Verschmelzung entsprechend aktualisieren.

Anhang: Algorithmen

Die zu implementierenden Algorithmen werden zur Verdeutlichung im Folgenden nochmal wie im Buch Clote, Backofen aufgeführt.

Algorithm UPGMA

INPUT: $n \times n$ distance matrix D

1. Initialize set $\mathcal{C}$ to consist of n initial singleton clusters $\{1\}, \ldots, \{n\}$

2. Initialize function $dist(c, d)$ on $\mathcal{C}$ by defining for all $\{i\}$ and $\{j\}$ in $\mathcal{C}$

$$dist(\{i\}, \{j\}) = D(i, j)$$

3. Repeat $n-1$ times
   (a) determine pair c, d of clusters in D such that $dist(c, d)$ is minimal; define
   $$d_{\min} = dist(c, d)$$
   (b) define new cluster $e = c \cup d$; define
   $$\mathcal{C} = \mathcal{C} - \{c, d\} \cup \{e\}$$
   (c) define a node with label e and daughters c, d, where the e has distance $\frac{d_{\min}}{2}$ to its leaves (i.e. the leaves of the subtree rooted at e).
   (d) define for all $f \in \mathcal{C}$ with $f \neq e$
   $$dist(e, f) = dist(f, e) = \frac{dist(c, f) + dist(d, f)}{2}$$

Algorithm WPGMA

INPUT: $n \times n$ distance matrix D

1. Initialize set $\mathcal{C}$ to consist of n initial singleton clusters $\{1\}, \ldots, \{n\}$
2. Initialize function $dist(c, d)$ on $\mathcal{C}$ by defining for all $\{i\}$ and $\{j\}$ in $\mathcal{C}$
   $$dist(\{i\}, \{j\}) = D(i, j)$$
3. Repeat $n-1$ times
   (a) determine pair c, d of clusters in D such that $dist(c, d)$ is minimal; define
   $$d_{\min} = dist(c, d)$$
   (b) define new cluster $e = c \cup d$; define
   $$\mathcal{C} = \mathcal{C} - \{c, d\} \cup \{e\}$$
   (c) define a node with label e and daughters c, d, where the e has distance $\frac{d_{\min}}{2}$ to its leaves.
   (d) define for all $f \in \mathcal{C}$ with $f \neq e$
   $$dist(e, f) = dist(f, e) = \frac{|c| dist(c, f) + |d| dist(d, f)}{|c| + |d|}$$