Hierarchische Clusteranalyse

Marvin Meusel, Bertram Vogel

12.01.2015

Eingabe und Zielstellung

Eingabe

- eine Menge von Objekten
- es gibt Cluster
- keine Kenntnis über Anzahl oder Struktur der Cluster

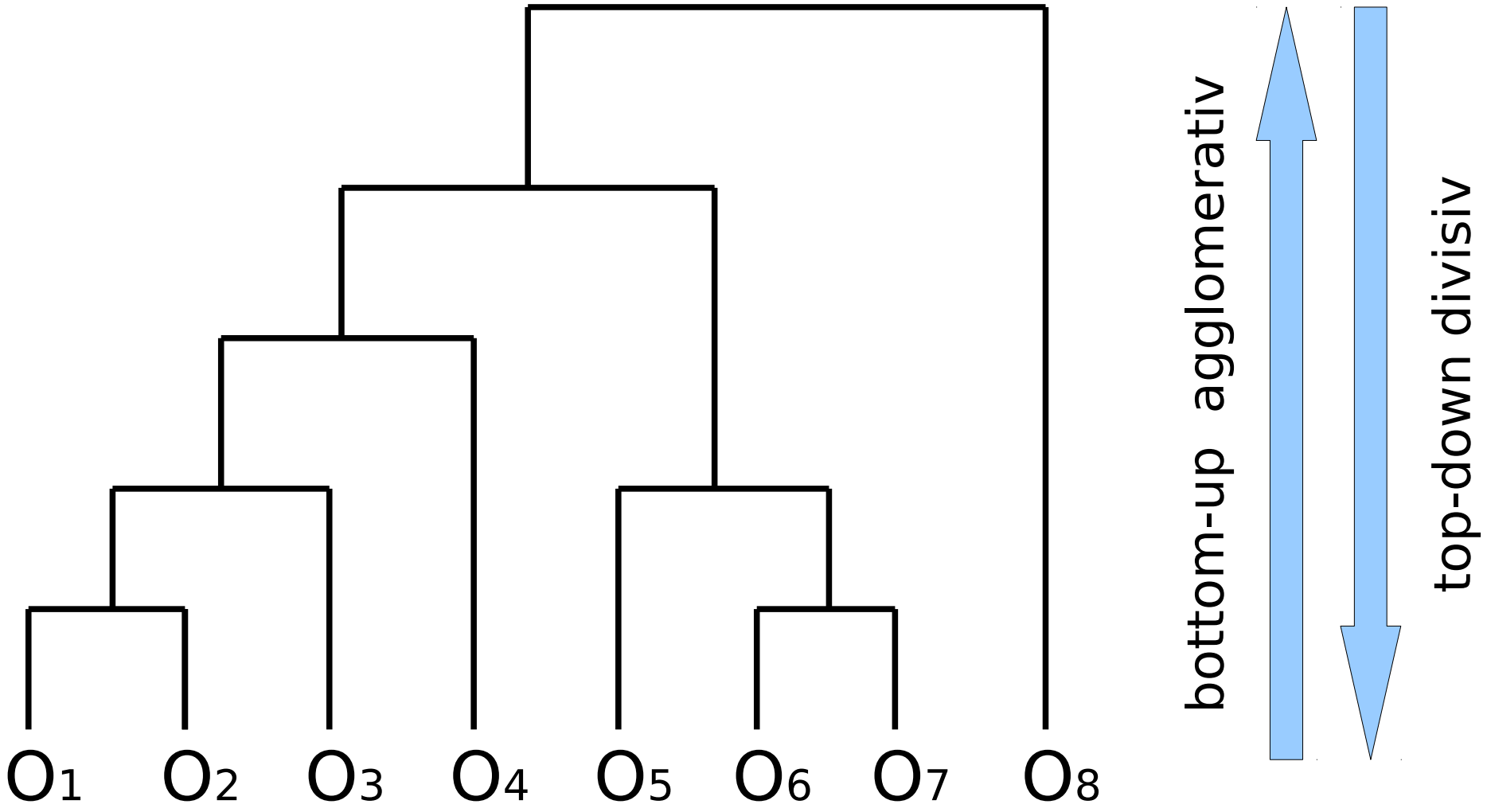
Zielstellung

- die Cluster finden

Cluster?

- Gruppe von Objekten mit ähnlichen Eigenschaften

Hierarchisches Clustern



Die Distanzmatrix D

D[i][j] = Distanz zwischen Objekt i und j

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	
0	2	1	5	5	8	9	5	O_1
	0	5	9	1	3	5	7	O_2
		0	8	2	4	4	9	O_3
			0	6	3	2	8	O_4
				0	9	7	4	O_5
					0	2	1	O_6
						0	5	O_7
							0	O_8

Metriken

Metrik

- D[i][j] = 0 gdw. i = j
- D[i][j] = D[j][i]
- $D[i][j] \leq D[i][k] + D[j][k]$

Ultrametrik (3-Punkte Bedingung)

- für alle i,j,k gilt $D[i][j] \leq \max(D[i][k] , D[j][k])$

Additive Baummetrik (4-Punkte Bedingung)

- für alle i,j,k,l gilt

  $D[i][j] + D[k][l] \leq \max(D[i][k] + D[j][l] , D[i][l] + D[j][k])$

Pair Group Methods with Arithmetic Mean

Eingabe

- n x n Distanzmatrix D

Ausgabe

- Baum mit Distanzen als Kantenlabel

Ablauf

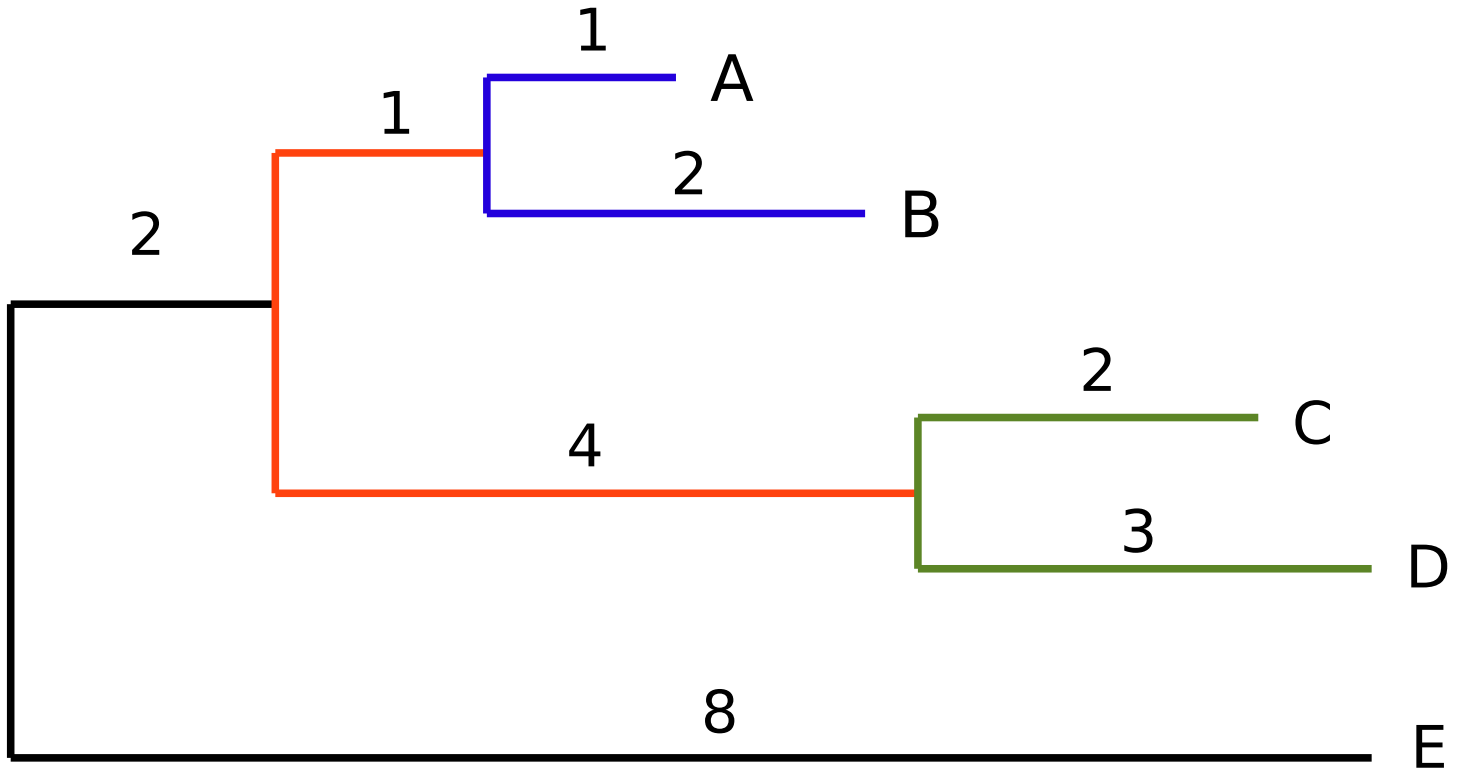
- zu Beginn pro Taxon ein Cluster
- in jedem Schritt
  - suche minimalen Eintrag D[i][j] mit i != j
  - fasse die Cluster i und j zum Cluster k zusammen, lege neuen Knoten im Baum an
  - berechne die Distanzmatrix neu

UPGMA und WPGMA

Neuberechnung der Distanzmatrix

- entferne Cluster i und j, füge neues Cluster k ein
- berechne die Distanzen von k zu den bestehenden Clustern
  - Distanz des neuen Clusters zu einem bereits bestehenden Cluster f
    - WPGMA $D[k][f] = D[f][k] = \frac{D[i][f] + D[j][f]}{2}$
    - UPGMA $D[k][f] = D[f][k] = \frac{|i| \cdot D[i][f] + |j| \cdot D[j][f]}{|i| + |j|}$

Das Newick Format



((A:1,B:2):1,(C:2,D:3):4):2,E:8)

Aufgaben

Anforderungen an die Implementierung

- Einlesen einer Datei mit mehreren DNA-Sequenzen im multiple fasta Format, Berechnung einer Distanzmatrix aus paarweisen Edit-Distanzen (Needleman-Wunsch) und Ausgabe der Matrix.

- Einlesen einer Distanzmatrix. Konstruktion eines Stammbaumes mit UPGMA und WPGMA. Ausgabe im newick Format.

- Bestimmung des Metrik-Typs einer Distanzmatrix. Testet damit jeweils die eingelesene und erzeugte Matrix!

Aufgaben

Aufgaben

- Implementierung der Klassen mit den entsprechenden Algorithmen
- API Dokumentation mit javadoc
- Projektmanagement mit Gradle
- Unit-Tests
- Benutzerinterface als Kommandozeilenprogramm (CLI) + README
- Testläufe und Testdaten mit Protokollieren der Ergebnisse (Testdaten und Aufgabenblatt auf der Kurswebsite)