

Aufgabe 6: Distanzbasierte Phylogenie: Hierarchisches Clustering

Ulf Leser

Wissensmanagement in der
Bioinformatik



Daten

- Wir verwenden neue Daten
- Die müssen sie ausnahmsweise selber suchen
- DNA-Sequenzen des Enzyms ‚Phenylalanine Hydroxylase‘
 - Mutationen im dafür kodierenden Gen führen zu **Phenylketonurie**
 - Häufigste angeborene erbliche Stoffwechselstörung (ca. 1:6000)
 - Verminderter Abbau von Phenylalanin führt zu schweren geistigen Entwicklungsstörungen
 - Einfache Diagnose (Neugeborenen-Screening), gute Behandelbarkeit (Diät)
 - Hauptursache (98%) sind **verschiedene Mutationen im kodierenden Gen** (Chromosom 12)

Aufgabe 1: Daten finden (4 Punkte)

- Suchen Sie die **Sequenz** des Genes beim **Menschen** (*Homo sapiens*) in Genbank (1 Punkt)
 - <http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide>
 - GI-Nummer 4557818, 2680 bp
- Suchen Sie mit **Nucleotide BLAST** die dazu ähnlichsten Referenz-mRNA Sequenzen für die folgenden Spezies (3 Punkte)
 - Geben Sie in den FASTA Dateien die Acc-Nummer der Sequenz an
 1. *Mus musculus*
 2. *Bos taurus*
 3. *Rattus novvegicus*
 4. *Macaca mulatta*
 5. *Xenopus tropicalis*
 6. *Equus caballus*
 7. *Canis familiaris*
 8. *Monodelphis domestica*
 9. *Ornithorhynchus anatinus*
 10. Mit der humanen Sequenz haben wir also **10 Sequenzen**
- Kopieren Sie alle Sequenzen in eine FASTA Datei



Aufgabe 2: Distanzbasierte Phylogenie (8 Punkte)

- Implementieren Sie den UPGMA Algorithmus zum **hierarchischen Clustering von Sequenzen**
- Berechnen Sie damit den Stammbaum der zehn Sequenzen
 - Sequenzen werden durchnummeriert in der Reihenfolge wie auf der vorherigen Folie
 - Sequenzen müssen aus einer FASTA Datei gelesen werden
 - Verwendung derselben Substitutionsmatrix wie in Aufg. 4 und 5
- Ausgabe des Programms
 - Die Distanzmatrix (Spalten / Zeilen in der geg. Reihenfolge)
 - Den Baum als Reihenfolge der **Cluster-Merge-Schritte**
 - Beispiel: Wenn als erstes die Sequenzen 1 und 4 zu 14 gemergt werden, dann 5 und 7 zu 57, dann 14 und 3 zu 143, dann soll die Ausgabe so aussehen: (1,4), (5,7), (14,3)



Aufgabe 3: Ultrametrische Matrizen (4 Punkte)

- Gegeben eine Distanzmatrix D der Größe $n \times n$
- Geben Sie einen Algorithmus an, der in $O(n^2)$ bestimmt, ob D ultrametrisch ist

Aufgabe 4: Ähnlichkeit von Bäumen (4 Punkte)

- Berechnen Sie den Abstammungsbaum auch mit **ClustalW** (im Web)
- Extrahieren Sie den tatsächlichen Abstammungsbaum vom NCBI Taxonomy Projekt
 - <http://www.ncbi.nlm.nih.gov/Taxonomy/>
- Beschreiben Sie, in wie weit sich die drei Bäume unterscheiden (1 Punkt)
- Überlegen Sie sich einen Algorithmus zum Messen der **Ähnlichkeit von Bäumen** (3 Punkte)
 - Hier ist Phantasie gefragt, alle Vorschläge sind willkommen
 - Welche Eigenschaften hat Ihr Vorschlag? Was würde er für die drei obigen Vergleiche ausrechnen?

Aufgabe 5: Wettbewerb (0 Punkte)

- Schreiben Sie ein Programm, das in möglichst kurzer Zeit mittels UPGMA den phylogenetischen Baum für eine beliebige Menge von Sequenzen berechnet und ausgibt
 - Also: Aufgabe 2 möglichst schnell löst
- Wir werden mit mehr Sequenzen ungefähr der gleichen Länge testen

Programmaufruf

- Das Programm (Aufgabe 2) muss wie folgt aufrufbar sein
 - `java -jar Assignment6.jar sequenzfile subfile`
 - Sequenzfile: FASTA Datei mit allen Sequenzen
 - Subfile: Name der Datei mit der Substitutionsmatrix
 - Ausgabe auf STDOUT
 - Distanzmatrix und Merge-Schritte
- Das Programm für den Wettbewerb muss wie folgt aufrufbar sein
 - `java -jar Competition4.jar sequenzfile subfile`
 - Sequenzfile: FASTA Datei mit allen Sequenzen
 - Subfile: Name der Datei mit der Substitutionsmatrix
 - Ausgabe auf STDOUT
 - Distanzmatrix und Merge-Schritte

Abgabe

- Bis **Montag, 11.2.2008**, 23.59 Uhr
- Nur per Mail als TXT Datei
 1. FASTA Datei mit zehn Sequenzen
 2. Die Ausgabe des Programms aus Aufgabe 2
 3. Lösung von Aufgabe 3
 4. Lösung von Aufgabe 4
- Der Code muss im Quelltext und als ausführbare JAR Datei eingeschickt werden