Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2018/19
Dr. Roland Wittler · M. Sc. Tizian Schulz
http://gi.cebitec.uni-bielefeld.de/teaching/2018winter/sequaprak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 11 vom 15./16.01.2019
Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Aufgabe 1 (Rekonstruktion eines phylogenetischen Baums)

Wir betrachten erneut die sechs Proteinsequenzen des Haemoglobin-Alpha-Gens aus Aufgaben 1 und 2 des vorigen Übungsblatts. Aus dem mittels *Clustal Omega* (`http://www.ebi.ac.uk/Tools/msa/clustalo/`) erstellten multiplen Alignments soll nun ein phylogenetischer Baum rekonstruiert werden.
Hierfür verwenden wir das Programmpaket *Phylip*, eine sehr umfangreiche Sammlung von klassischen Methoden zur phylogenetischen Analyse. Die einzelnen Applikationen des Pakets sind im CeBiTec-System unter `/vol/biotools/` installiert.

1. Wiederhole die Berechnung des multiplen Alignments in *Clustal Omega* mit dem Ausgabeformat "PHYLIP". Verwende dann das Phylip-Programm `protdist` (mit Standardseinstellungen) um aus dem multiplen Alignment paarweise Distanzen zwischen den sechs Sequenzen zu berechnen. Die Ausgabe des Programms wird in der Datei `outfile` abgelegt, welche sinnvoll umbenannt werden sollte.

2. Aus der erhaltenen Distanzmatrix soll nun mittels hierarchischem Clustering (UPGMA) ein phylogenetischer Baum erstellt werden. Eine entsprechende Methode ist in dem Phylip-Programm `neighbor` implementiert. Hier muss nach Einlesen der Distanzmatrix die Methode `UPGMA` ausgewählt werden. Es werden zwei Ausgaben erzeugt. Die Datei `outfile` enthält eine einfache Darstellung des rekonstruierten Baums mit Angaben zu Kantenlängen in einer Tabelle, und die Datei `outtree` enthält den Baum im Newick-Format, einem klassischen Datenformat für (phylogenetische) Bäume. Auch diese Dateien sollten sinnvoll umbenannt werden.

Aufgabe 2 (Bootstrapping – Implementierung)

Zur Überprüfung der Verlässlichkeit der in Aufgabe 1 rekonstruierten Phylogenie soll nun zunächst eine Methode zur Erzeugung von Bootstrap-Replikaten des multiplen Alignments implementiert werden, welche in Aufgabe 3 weiter analysiert werden.

1. Wähle in *Clustal Omega* für das multiple Alignment ein Ausgabeformat, welches sich einfach einlesen lässt (z.B. "PHYLIP" oder "FASTA"). Impementiere eine Methode, welches ein gegebenes multiples Alignment (das aus Aufgabe 1) in eine einfache Datenstruktur einliest (z.B. eine Liste der Sequenzbezeichner und eine Liste der Alignment-Zeilen).

2. Kern deines Programms ist eine Methode zur Erzeugung von Bootstrap-Replikaten. Ein Replikat ist ein multiples Alignment, welches aus dem gegebenen Alignment erstellt wird, indem zufällig gewählte (gleichverteilt mit Zurücklegen) Alignmentspalten des ursprünglichen Alignments zu einem neuen Alignment gleicher Länge zusammengefügt werden. Ein so erhaltenes Alignment unterscheidet sich zwar vom gegebenen (da einige Spalten mehrfach gewählt werden, andere dafür gar nicht), ist ihm aber sehr ähnlich (da keine neuen Spalten erzeugt werden).

3. Implementiere eine Ausgabefunktion, die ein Replikat im Phylip-Format ausgibt. Erzeuge 100 Bootstrap-Replikate und gib alle 100 Replikate in *einer* Datei aus, d.h. die Ausgaben der einzelnen Replikate werden einfach hintereinander in eine Datei geschrieben. Tipp: Dein Programm kann einfach alle Replikate nacheinander auf dem Terminal ausgeben und du lässt diese Ausgabe mittels `>` in eine Datei schreiben. Eine Beispielausgabe ist in `/prj/seqan/Praktikum/` zu finden.

Aufgabe 3 (Bootstrapping – Anwendung)

Kanten im phylogenetischen Baum, welche erhalten bleiben, auch wenn man die zugrundeliegenden Daten ein wenig verändert, sind vertrauenswürdiger als solche, die sich bei kleinen Änderungen im Alignment nicht mehr rekonstruieren lassen. In Aufgabe 2 wurden leicht veränderte Daten erzeugt, für welche nun überprüft werden soll, inwiefern sie die Kanten im phylogenetischen Baum aus Aufgabe 1 unterstützen. Hierzu solle analog zu Aufgabe 1 für jedes Bootstrap-Replikat eine Phylogenie erstellt werden, welche dann mit dem Baum aus Aufgabe 1 verglichen werden.

1. Verwende `protdist`, um für alle Bootstrap-Replikate Distanzmatrizen zu erzeugen. Als Eingabe können die in Aufgabe 2 in *einer* Datei abgelegten 100 Alignments eingelesen werden, indem eine entsprechende Einstellung vorgenommen wird. Die Ausgabe enthält entsprechend 100 Distanzmatrizen untereinander. (Umbenennen der Ausgabedatei `outfile` nicht vergessen.)

2. Verwende `neighbor`, um auf alle soeben erzeugte Distanzmatrizen jeweils UPGMA anzuwenden. Auch hier können nach Anpassung der Parameter im Menü alle Distanzmatrizen aus *einer* Datei eingelesen werden und die Ergebnisse werden in den Ausgabedateien konkateniert.

3. Abschließend soll jeder Kante des Baums aus Aufgabe 1 ein sog. *Bootstrap-Wert* zugewiesen werden, welcher deren Verlässlichkeit widerspiegeln soll. Dazu wird für jede Kante gezählt, in wie vielen der 100 replizierten Bäume diese Kante ebenfalls vorkommt. Verwende hierzu die Platform *Galaxy* (`https://galaxy.pasteur.fr`). Lade die Daten (Baum aus Aufgabe 1, Bäume aus Aufgabenteil 3.2) hoch und wähle dabei als Datentyp "nhx". Berechne die Bootstrap-Werte mit dem Tool "Booster" und visualisiere das Ergebnis (*Tree with FBP supports*) mit dem Tool "Newick Display". Füge die Abbildung dem Protokoll hinzu.

4. Was bedeutet ein Bootstrap-Wert von 1? Diskutiere das Ergebnis – auch im Vergleich zum Guidetree von Clustal Omega von Übungsblatt 10. Wie stellt sich der Vergleich dar, wenn man beide Bäume "ungewurzelt" betrachtet, also beide von der Wurzel ausgehenden Kanten zu einer verschmelzen würde?