

Übungen zur Vorlesung
**Einführung in die angewandte
Bioinformatik**
Sommersemester 2009

Übungsblatt 4
Bearbeitungszeit:
07.05.2009

Aufgabe 4.1 – PubMed und MeSH

Sie suchen nach Veröffentlichungen über Ausbrüche des Denguefieber. Suchen Sie zunächst in PubMed nach `outbreaks dengue`. Wie viele Artikel werden gefunden? _____

Suchen Sie nun fokussierter mit MeSH (verlinkt von der Vorlesungswebseite). Suchen Sie also in MeSH zunächst nach sinnvollen Kategorien für `outbreaks` und `dengue` und verknüpfen Sie diese mit AND in einer Searchbox. Wie viele Artikel werden mit dieser Searchbox gefunden? _____

Hinweis: Suchen Sie zunächst nach dem ersten Stichwort in MeSH, wählen Sie eine sinnvolle Kategorie aus, indem Sie auf das Kästchen vor der Kategorie klicken, und senden Sie die Kategorie dann über das *Send-to*-Menü in eine Searchbox. Wiederholen Sie dies mit dem zweiten Stichwort und starten Sie die in der Searchbox spezifizierte Suche in PubMed.

Schränken Sie die Suche weiter auf Statistiken ein, indem sie bei der zu `outbreaks` gehörigen Kategorie als Unterkategorie `statistics and numerical data` wählen. Wie viele Veröffentlichung werden nun aufgelistet? _____

Um die folgende Aufgabe zu bearbeiten, müssen Sie über einen Rechner der TU auf die ISI-Webseite gehen.

Aufgabe 4.2 – ISI Web of Knowledge

Gehen Sie auf die Webseite des ISI Web of Knowledge. Suchen Sie ein Paper, von dem Sie wissen, dass der Titel die Wörter *bioinformatics tools* enthält und dass der Autor *Ramana Davuluri* ist. Wann und in welcher Zeitschrift wurde der Artikel veröffentlicht?

Klicken Sie in den Suchergebnissen auf den Titel dieses Artikels, um detailliertere Angaben zu erhalten. Wie oft wurde dieser Artikel zitiert? _____

Wie häufig wird hingegen der Artikel "BLAT - The BLAST-like Alignment Tool" zitiert? _____
Sehen Sie sich zunächst durch einen Klick auf die entsprechende Zahl die Liste der Artikel an, die den BLAT-Artikel zitieren. Betrachten Sie die zitierenden Autoren im *Refine-Results*-Menü. Wie oft hat der Autor in seinen späteren Arbeiten auf diesen Artikel verwiesen? _____

Aufgabe 4.3 – Gensuche in der NCBI-Datenbank

Gehen Sie auf die NCBI-Seite. Wählen Sie die *Genome*-Datenbank aus und suchen Sie das Bakterium *Salmonella typhi CT18*. Schränken Sie die Suche auf Organismen ein, indem Sie die *Limits* oder das Search Field Tag `[Organism]` benutzen. Wie viele Ergebnisse werden gefunden? _____

Wählen Sie von den Ergebnissen das aktuellste aus. Wie lautet die Zugriffsnummer (*Accession Number*) (sie steht hinter „Refseq:“)? _____ Wie ist der GC-Gehalt? _____

Wenn Sie auf die Zugriffsnummer klicken, wird der GenBank-Eintrag angezeigt. In welchem Journal wurde der Artikel veröffentlicht, in dem das *C.-glutamicum*-Genom vorgestellt wurde? _____

Wie viele Basenpaare hat die Sequenz? _____

Suchen Sie nun nach dem Genom des Denguevirus. Nehmen Sie die Variante 1, wenn Ihnen die Ergebnisse angezeigt werden. In welchem Institut wurde diese Variante des Virus sequenziert?

Wählen Sie nun die *Nucleotide*-Datenbank aus und suchen Sie nach dem SRY-Gen. Es sollen Vorkommen in *Homo sapiens* oder in *Mus musculus* gefunden werden. Verwenden Sie die Tags [Gene Name] und [Organism] und verknüpfen Sie die Suche passend mit OR. Wie viele Ergebnisse erhalten Sie? _____
Schauen Sie sich einige der gefundenen Einträge an, z. B. AY601860, NM_003140 und S83252.

Ändern Sie die Datenbank zurück auf *Gene* und wiederholen Sie die Suche. Wie viele Ergebnisse erhalten Sie? _____
Schauen Sie sich die Ergebnisse an. Was ist der Unterschied zu den Einträgen, die Sie in der *Nucleotide*-Datenbank fanden?

Aufgabe 4.4 – Alternatives Spleißen

Benutzen Sie weiterhin die NCBI-Datenbank und finden Sie das Gen *egl-15*. Verwenden Sie wieder ein passendes Tag, um die Suche einzuschränken.

Aus welchem Organismus kommt dieses Gen? _____

Wie viele Proteinvarianten werden von diesem Gen kodiert? _____

Wie lautet die Zugriffsnummer der ersten angegebenen mRNA-Sequenz? _____

Aufgabe 4.5 – Sequence Retrieval System

Benutzen Sie für die nächste Suche das Sequence Retrieval System (SRS@EBI). Suchen Sie in der EMBL-Datenbank nach dem Protein *sonic hedgehog* im Organismus *Rasbora elegans*: Wählen Sie dazu auf der *Library Page* (Link steht oben auf der Seite) die EMBL-Datenbank aus. Gehen Sie zum Reiter *Query Form* und suchen Sie *rasbora elegans* im Feld „Organism Name“ und *sonic hedgehog* im Feld „Description“. Wie viele Ergebnisse erhalten Sie und wie lang sind die jeweiligen Nukleotidsequenzen? _____

Was unterscheidet die beiden Ergebnisse? _____

Suchen Sie nun mit SRS nach dem Locus H06H21.10. Benutzen Sie nur die Datenbank „EMBL (Coding Sequences)“. Wählen Sie auf der Seite, auf der die Ergebnisse angezeigt werden, *Display Options* → *Complete entries*. Vergleichen Sie die Einträge. Auf welchem Chromosom liegt das Gen? _____

Aus wie vielen Exons sind die gefundenen Proteine aufgebaut?

Aufgabe 4.6 – Klausurergebnisse analysieren mit R

Laden Sie sich von der Übungswebseite die Datei *klausur.dat* herunter und speichern Sie sie ab. Sehen Sie sich die Datei an: Sie enthält die Punktzahlen, die bei einer alten Klausur erzielt wurden. Laden Sie diesen Datensatz in eine Variable namens *klausur*.

```
> klausur = read.delim("klausur.dat") # evtl. Pfad angeben, also
# z.B. "Desktop/klausur.dat"!
> dim(klausur)      # Dimensionen herausfinden
> colnames(klausur) # Überschriften (column names) anschauen
> klausur          # alle Daten ansehen
> klausur[1:10,]   # die ersten zehn Einträge anschauen
> klausur$Punkte    # nur die Werte der "Punkte"-Spalte anschauen
> klausur$Geschlecht # nur die "Geschlecht"-Spalte anschauen
```

So zeigen Sie nur die Punkte an, die die männlichen Teilnehmer erreicht haben:

```
> klausur$Punkte[klausur$Geschlecht == "m"]
```

Denken Sie daran, dass „==“ auf Gleichheit überprüft, während „=" für Zuweisungen ist!

Wie können Sie die durchschnittlich erzielten Punkte aller Teilnehmer berechnen?

Wie können Sie die maximal erzielten Punkte der weiblichen Teilnehmer berechnen?

Bevor es gleich weitergeht, möchten Sie sich noch etwas Tipparbeit sparen und benutzen dazu die attach-Anweisung:

```
> attach(klausur) # ab jetzt kann das "klausur$" weggelassen werden
> Punkte          # statt klausur$Punkte
```

Lassen Sie sich nun ein Histogramm und einen Boxplot über alle Punktzahlen anzeigen (probieren Sie auch mal, die Farbe des Plots zu ändern).

```
> hist(Punkte, seq(5, 30, 5))
> boxplot(Punkte)
```