Institut für Informatik
Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer

Wintersemester 2018/19

Übungen zum Bioinformatik-Tutorium

Blatt 2

Termin: Dienstag, 30.10.2018, 11 Uhr

1. Wildcards und Brace Expansion

(a) Finde heraus wie man mit `mkdir` einen Pfad mit seinen Sub-Ordnern erstellen kann, und erstelle dann mit Hilfe von Brace Expansion in einem Befehl das Verzeichnis `wildcards/ordnerXX` wobei `XX` für die Zahlen von 01 bis 10 steht.

(b) Erstelle nun in `ordner01` 1000 Dateien mit den Dateinamen `0001, 0002, ..., 1000`.

(c) Versuche dir nun mit `echo` und Wildcards nur folgende Dateinamen ausgeben zu lassen:

(i) Nur solche, die irgendwo im Namen die Zeichenfolge 20 enthalten.

(ii) Nur solche die an den ersten 3 Stellen eine 0 haben

(iii) Alle, die *nicht* die 1 an erster Stelle haben

2. Pipes und Umleitungen

(a) Finde heraus, wie man mit `tail` eine Datei auf Änderungen beobachtet.

(b) Die Datei `~/tutorium/material/yeastract.csv` enthält das genregulatorische Netzwerk von Hefe. In der ersten Spalte steht ein Transkriptionsfaktor und in der zweiten Spalte ein Gen dessen Expression durch den Transkriptionsfaktor reguliert wird. Suche nun mit `grep` nach jedem Eintrag, in dem das Gen mit dem Identifier **YDR213W** erwähnt wird und zähle Diese.

(c) Finde nun alle Gene die durch den Transkriptionsfaktor **YGL073W** reguliert werden. Speichere die Identifier der regulierten Gene (ohne Duplikate) in eine neue Datei in deinem Tutoriumsordner. Schaffst du das mit einem Befehl? (Zur Kontrolle: Es sollten 571 einzigartige Gene sein.)

(d) Die Datei `~/tutorium/material/cathscop.inpairs` enthält in den ersten beiden Spalten Paare von Protein IDs. In den nächsten beiden Spalten stehen die jeweils zugehörigen CATH-Nummern. Wie viele einzigartige CATH-Nummern enthält diese Datei?

(e) Die Datei `~/tutorium/material/C_elegans.pep.all.fasta` enthält das Referenzproteom von C. elegans im FASTA Format. Speichere alle FASTA header in eine neue Datei. Wie viele Proteine sind in dieser Datei? Sortiere die Namen der Proteine alphabetisch und gib die ersten 20 aus.

3. Packen und Entpacken

(a) Entpacke und öffne die Datei `/usr/share/doc/bash/README.commands.gz`.

(b) Packe das Verzeichnis `/usr/share/doc/bash/` als `.tar.gz`.

4. Arbeiten mit der Shell

Informiere dich auf `http://www.ensembl.org/info/website/upload/gff.html` über das `gtf`-Format und Betrachte im Folgenden die Datei
`~/tutorium/material/saccharomyces_cerevisiae/Saccharomyces_cerevisiae.R64-1-1.75.gtf`
Benutze shell-Kommandos um folgende Fragen zu beantworten. Hinweis: Das Anlegen eines symlinks von der `gtf`-Datei irgendwo in deinen tutoriums-Ordner hinein erspart einiges an Tipparbeit.

(a) Speichere die Zeilen, die eine Coding Sequence (CDS) definieren in eine Datei. Kontrolliere, dass es 7055 sind und korrigiere gegebenenfalls Fehler.

(b) Finde heraus welche verschiedenen Features in der Datei gespeichert sind.

(c) Speichere alle Pseudogene auf dem "+" Strang in eine Datei. Wie viele Pseudogene gibt es auf dem "–" Strang?