

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

Übungsblatt 5
Bearbeitungszeit:
14.05.2009

Aufgabe 5.1 – Ein Protein – drei Datenbanken

Suchen Sie in der Protein-Datenbank des NCBI nach dem Eintrag mit der Zugriffsnummer P00042. Welches Protein wird darin beschrieben? _____

Versuchen Sie nun, mittels der SRS-Seite den entsprechenden Swissprot-Eintrag zu finden. Welche Gesamtmasse hat das Protein? _____

Suchen Sie das Protein nun auch noch auf der UniProt-Webseite.

In welcher Zellkomponente ist das Protein angesiedelt? _____

Wie lauten die letzten drei Aminosäuren der Proteinsequenz? _____

Aufgabe 5.2 – Ein interessantes Protein

Arbeiten Sie in dieser Aufgabe mit der UniProt-Webseite. Suchen Sie nach dem Protein mit der Zugriffsnummer P00533 und nehmen Sie sich ein paar Minuten, um sich die Fülle an Informationen anzusehen.

Wann wurde der Eintrag erstellt? _____

Wann wurde die Sequenz zuletzt korrigiert? _____

Wann wurde die Annotation zuletzt bearbeitet? _____

Wie viele Veröffentlichungen zu diesem Protein werden angegeben? _____ Die Antworten auf diese Fragen legen den Verdacht nahe, dass dieses Protein ziemlich interessant ist. Haben Sie eine Idee warum? _____

Schauen Sie sich die *Sequence annotation* an. Was befindet sich an Position 266? _____

Sehen Sie sich den SWISS-2DPAGE-Eintrag an. Wie viele Spots wurden als das fragliche Protein identifiziert?

Sehen Sie sich den KEGG-Eintrag an. An wie vielen Pathways ist das Protein beteiligt? _____

Zuletzt noch eine Frage zu den Features dieses Proteins: Zwischen welchen Positionen befindet sich die erste α -Helix? _____

Aufgabe 5.3 – Noch ein Protein

Gehen Sie wieder zur UniProt und nutzen Sie die Zugriffsnummer P06968, um sich nun über das Protein des dUTPase-Gen von *E. coli* zu informieren.

Wählen Sie unter den Cross-references "GenBank" als Sequence-Datenbank aus und klicken Sie auf den ersten Locus (X. . .). Sie gelangen zum NCBI. Wie viele Basenpaare lang ist das Gen, welches die dUTPase in *E. coli* kodiert? _____

Wo befinden sich die Promotorsequenzen? _____

Was befindet sich an den Positionen 889 bis 895? _____

Gehen Sie zurück zur UniProt-Seite. Schauen Sie sich den Abstract der fünften Literaturreferenz an. Mit welcher Auflösung wurde die Kristallstrukturanalyse (*X-ray crystallography*) durchgeführt? _____

Aufgabe 5.4 – HMM-Logo

Schauen Sie sich in der UniProt das Protein mit Zugriffsnummer Q15835 an. Klicken Sie unter "Family and domain databases" auf den ersten Pfam-Link (PF00069). Schauen Sie sich das HMM-Logo an. Wo ist (ungefähr) die erste Deletion? _____

Aufgabe 5.5 – R: Verteilungen

Um sich die Normalverteilung in Erinnerung zu rufen, plotten Sie einmal in R die zugehörige Dichtefunktion (dnorm: **d**ensity of **n**ormal distribution).

```
> curve(dnorm(x), from=-4, to=4)
```

Erzeugen Sie sich nun – ähnlich wie auf dem Übungsblatt 3 – einen Vektor mit 1000 normalverteilten Zufallszahlen. Wenden Sie die Funktionen `hist`, `boxplot`, `qqnorm` auf diesen Vektor an und schauen Sie sich die Plots an.

Plotten Sie sich einmal die Dichte der Exponentialverteilung (die Dichtefunktion heißt `dexp`). Wiederholen Sie das Experiment von eben mit 1000 exponentialverteilten Zufallszahlen (erzeugen Sie diese mit `rexp`). Was sagt Ihnen nun der Q-Q-Plot, den Sie mit `qqnorm` erzeugt haben?

Aufgabe 5.6 – R: Normalverteilte Frauen oder Männer

Weiter geht es mit dem Datensatz der Klausurpunktzahlen, den Sie schon auf dem letzten Übungsblatt genutzt haben. Laden Sie erneut die Datei `klausur.dat` in R (wieder mit `read.delim`). Benutzen Sie wieder `attach`.

Vergleichen Sie sowohl die Verteilung der Punktzahlen der Frauen als auch die der Männer mit der Normalverteilung. Benutzen Sie dafür zum Beispiel `qqnorm`. Zur Erinnerung: Die Punktzahlen der Frauen bekommen Sie mit `Punkte[Geschlecht == "w"]`, die der Männer mit `Punkte[Geschlecht == "m"]`.

Wessen Punktezahlen sind einer Normalverteilung am ähnlichsten?
