

Einführung in die angewandte Bioinformatik

Klausurübung am 02.07.2009

Hinweise

Möchten Sie dieses Übungsblatt unter realistischen Bedingungen zur Vorbereitung auf die Klausur durcharbeiten, so beachten Sie folgende Punkte.

- Sie haben insgesamt 45 Minuten Zeit zur Bearbeitung des Blatts. Es gibt einen Praxis- und einen Theorieteil. Der Praxisteil ist so ausgelegt, dass er ca. 30 Minuten beansprucht.
- Tragen Sie oben *auf jedem Blatt* Ihren Namen und Ihre Matrikelnummer ein.
- Benutzen Sie zur Sicherheit als Browser nur den Firefox. Manche Webseiten werden sonst unvollständig angezeigt.
- Sie dürfen nur die Webseiten nutzen, die von <http://bioinfo.wikidot.com/klausur> verlinkt sind. Auch die auf diesen Seiten zur Verfügung gestellten Hilfeseiten, wie z. B. die Hilfe zu den *search field tags* auf der NCBI-Seite, dürfen Sie benutzen. Insbesondere **nicht erlaubt** ist die Benutzung von Wikipedia, Google und den auf der Vorlesungswebseite herunterladbaren Vorlesungsfolien.

Webseite für dieses Übungsblatt: <http://bioinfo.wikidot.com/klausur>

Praxisaufgaben

Aufgabe 10.1 – Literatur

Suchen Sie in der PubMed-Datenbank nach dem Paper, das von Needleman und Wunsch geschrieben wurde. Wie lautet das letzte Wort im Titel des Papers?

Wie viele Veröffentlichungen von E. Ukkonen, die in der Zeit von 2000–2005 erschienen, sind in der PubMed-Datenbank gespeichert?

Wie häufig wurde das Paper "Basic local alignment search tool" dem ISI Web of Knowledge zufolge zitiert (diese Aufgabe können Sie nur in der Uni lösen)?

Wie wird der Journalname *Nature Biotechnology* offiziell abgekürzt?

Aufgabe 10.2

Benutzen Sie die NCBI-Webseite. Finden Sie das BIRC5-Gen im Menschen. Wie viele Proteinvarianten werden von diesem Gen kodiert?

Wie lautet die RefSeq-Zugriffsnummer des Proteins der Isoform 1?

Aufgabe 10.3

Benutzen Sie nun die Uniprot-Webseite. Suchen Sie nach dem Protein mit der Zugriffsnummer P50518. Wie groß ist seine Masse?

Wie viele Pathways sind in der KEGG-Datenbank für dieses Protein bekannt?

Aufgabe 10.4

Laden Sie sich von der Klausurwebseite die Datei unbekannt4.fasta herunter. Wie lautet der Organismus, aus dem dieses Stück DNA-Sequenz stammt (benutzen Sie BLAST mit der "NCBI Genomes"-Datenbank)?

Aufgabe 10.5

Wie lautet die Uniprot-Zugriffsnummer des "sulfite oxidase"-Enzyms aus dem Organismus Gallus gallus?

Benutzen Sie diese Zugriffsnummer, um eine BLAST-Suche im Organismus Mus musculus durchzuführen. Wie hoch ist die Sequenzidentität des ersten Treffers?

Aufgabe 10.6 – Strukturvorhersage

Sagen Sie mit RNAfold die RNA-Sekundärstruktur für die auf der Klausurwebseite angegebenen Sequenz vorher. Wie hoch ist die minimale freie Energie? _____

Aufgabe 10.7

In R wurden folgende Anweisungen ausgeführt:

```
> tabelle = read.delim("groessen.txt")
> tabelle
  Kennung Groesse
1      A    1.43
2      B    1.20
3      C    1.70
4      D    1.29
...

```

(Die Punkte deuten an, dass die Tabelle noch mehr Zeilen enthält als hier gezeigt werden.) Wie lauten die Befehle,

- a) um das arithmetische Mittel aller Werte in der Spalte *Groesse* auszurechnen?

-
- b) um die Summe aller Werte in der Spalte *Groesse*, die kleiner als 1,4 sind, auszurechnen?
-

- c) um mit Hilfe eines Q-Q-Plots zu überprüfen, ob die Logarithmen der Werte in der Spalte *Groesse* normalverteilt sind?
-

Wie erkennen Sie im Q-Q-Plot, ob die Werte normalverteilt sind?

- $1.2E+03$

- $3.7e-02$

- $\ln(e^2)$

Aufgabe 10.12

Was ist der Unterschied zwischen einer Datenbank und einem Datenbanksystem?

Aufgabe 10.13

Wie ist die Zeiteinheit 1 PAM definiert?

Wofür wird die PAM250-Scorematrix verwendet?

Aufgabe 10.14 – Laufzeit und Komplexität

Ein Algorithmus benötigt zum Erledigen eines Problems der Größe n nicht mehr als $14 \cdot \log\left(\frac{n}{8}\right) + 3n^2$ Rechenschritte. Geben Sie die Laufzeit in \mathcal{O} -Notation an.

Bei der Beschreibung eines bestimmten Algorithmus sagt Ihnen jemand: "Der Algorithmus ist NP-schwer."
Warum ist diese Aussage unsinnig?