

Einführung in die angewandte Bioinformatik

Klausurübung am 02.07.2009

Mit Lösungen

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?

Mit `needleman[au] AND Wunsch[au]` findet man: **proteins**

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

8. Dies findet man mit `ukkonen e[au]` und z. B. Benutzung der Limits heraus. `ukkonen e[au] 2000:2005[dp]` funktioniert auch.

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)?

26 129 mal am 3.7.2009, die Zahl wird regelmäßig größer.

Wie wird der Journalname *Nature Biotechnology* offiziell abgekürzt?

Nat Biotechnol (auf der JournalSeek-Webseite herauszufinden)

Aufgabe 10.2

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

Suchbegriff: `homo sapiens[organism] birc5[gene]`. Ergebnis: 3

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

NP_001159.2 (NM_001168.2 ist die Zugriffsnummer für die mRNA)

Aufgabe 10.3

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

26 157 Da

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

Bei den Cross-references auf den Link mmu:11973 hinter "KEGG" klicken, dann die Zeilen mit "PATH:" zählen. Ergebnis: 2

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)?

Yarrowia lipolytica

Aufgabe 10.5

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

P07850

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

304/456 bzw. 66% (dran denken Protein-BLAST zu nehmen)

Aufgabe 10.6 – Strukturvorhersage

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

-53.70 kcal/mol

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?

mean(tabelle\$Groesse)

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

sum(tabelle\$Groesse[tabelle\$Groesse < 1.4])

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

qqnorm(log(tabelle\$Groesse))

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

Die Punkte liegen dann auf einer Geraden

Aufgabe 10.8

Laden Sie sich von der Seite <http://bioinfo.wikidot.com/klausur> die Dateien phdB.fasta und php.fasta herunter. Benutzen Sie EMBOSS mit Standardeinstellungen, um ein globales Alignment der beiden DNA-Sequenzen zu erstellen. Wie groß ist die Identität?

4259/6026 (70.7%) oder 4258/6029 (70.6%). Was man herausbekommt, hängt davon ab, welche Sequenz man zuerst angegeben hat. Man muss aber auf jeden Fall auf DNA-Sequenzen umschalten! Der absolute oder relative Wert allein reicht auch.

Wie lang ist der erste Abschnitt aufeinanderfolgender Mismatches?

2. Dort, wo die beiden Punkte (. .) sind, ca. Position 370.

Theorieaufgaben

Aufgabe 10.9

Im Folgenden sind zwei Alignments zweier Sequenzen dargestellt.

Alignment A:

```
Sequenz1  1  CCCCCC-----TTTTTGGGGG      18
           | | | | | | | | | | | | | | | |
Sequenz2  1  CCCCCCATATATATATATGGGGG      23
```

und Alignment B:

```
Sequenz1  1  CCCCCC-T-T-T-T-TGGGGG      18
           | | | | | | | | | | | | | |
Sequenz2  1  CCCCCCATATATATATATGGGGG      23
```

Welche Parameter wurden zwischen beiden Alignments verändert und wie wurden diese Parameter geändert?

Im zweiten Alignment wurden geringere Lückenöffnungskosten oder höhere Lückenerweiterungskosten oder beides gewählt.

Aufgabe 10.10

Eins der folgenden ist *kein* zulässiges Alignment. Warum? Wie kann man es zu einem zulässigen Alignment machen?

(1) ATCCG-A (2) ATCCG-A
A-C-TGA A-CTG-A

Alignment 2 ist unzulässig, da eine Lücke an eine Lücke aligniert wird. Indem man die vorletzte Spalte weglässt, wird es zulässig.

Aufgabe 10.11

Wofür steht jeweils das E bzw. e in folgenden Ausdrücken? Ist einer der Ausdrücke eine Zahl, geben Sie sie in Dezimalschreibweise an!

- E-value (E-Wert)

Erwartungswert

- $1.2E+03$

$1.2 \cdot 10^3 = 1200$ (das E bedeutet: "... mal zehn hoch ...")

- $3.7e-02$

$3.7 \cdot 10^{-2} = 0.027$, wie eben (so gibt BLAST E-values an)

- $\ln(e^2)$

$\ln(e^2) = 2$. e ist Eulersche Zahl (Basis des natürlichen Logarithmus).

Aufgabe 10.12

Was ist der Unterschied zwischen einer Datenbank und einem Datenbanksystem?

Ein Datenbanksystem enthält eine Datenbank und ein Datenbankmanagementsystem.

Aufgabe 10.13

Wie ist die Zeiteinheit 1 PAM definiert?

Zeit, in der 1% der Aminosäuren mutiert ist.

Wofür wird die PAM250-Scorematrix verwendet?

Vergleich von evolutionär weit entfernten, aber noch eindeutig verwandten Proteinen.

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.

$\mathcal{O}(n^2)$

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

"NP-schwer" ist eine Aussage, die sich auf Probleme bezieht, nicht auf Algorithmen