Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, WiSe 2009/2010
Prof. Dr. Jens Stoye · Dipl.-Inform. Nils Hoffmann

http://wiki.techfak.uni-bielefeld.de/gi/GILectures/2009winter/SequenzAnalyse

Blatt 12 vom 14.01.2010
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 Index-basierte Datenbanksuche (3 Punkte)
Beschreibe jeweils, wie die folgenden Algorithmen im Vergleich zu BLAST arbeiten.

1. BLAT
2. SWIFT

In welchem Kontext werden die Algorithmen verwendet?

Aufgabe 2 Statistik auf Sequenzen (3 Punkte)
Eine DNA-Sequenz S hat die relativen Buchstabenhäufigkeiten $f_a = f_t = 0.4$ und $f_g = f_c = 0.1$.

1. Wie wahrscheinlich ist dann die Sequenz $x =$ `atccgtattaca` mit den oben angegebenen Häufigkeiten?
2. Wie hoch ist der Erwartungswert für die Anzahl der Treffer von x in S, wenn T die Länge 1200 hat?

Aufgabe 3 FASTA Score-Statistik (4 Punkte)
Die Wahrscheinlichkeit, dass wir einen FASTA-Score $C(X,Y) \geq t$ in zwei zufälligen Sequenzen X und Y erhalten, kann wie folgt approximiert werden:

$$\mathbb{P}\big(C(X,Y) \geq t\big) \approx 1 - e^{-mnp^{t+q-1}}$$

mit $m = |X|$, $n = |Y|$, $p = \sum_{c \in \Sigma} f_c^2$ und q der Länge der verwendeten q-Gramme.

1. Sei $m = n = 1000$, $p = 1/4$ und $q = 8$. Wie muss t gewählt werden, damit der p-Value 0.01 ist?
2. Eine Suchsequenz x hat die Länge 100. Bei einer FASTA-Suche mit $q = 5$ gibt es in einer Datenbank zwei Treffer mit einem Score über 15. Die Sequenz y_A, mit einer Länge von 120, hat einen Score von $C(x, y_A) = 18$; die Sequenz y_B, mit der Länge 90, erzielt den Score $C(x, y_B) = 17$. Welcher Treffer ist signifikanter? Berechne p-value und Bit-Score.