Übungen zur Vorlesung Sequenzanalyse I

Universität Bielefeld, Wintersemester 2010/2011
Dipl.-Inform Peter Husemann · Dr. Roland Wittler

http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2010winter/SequenzAnalyse

Blatt 12 vom 21.01.2011
Abgabe in einer Woche vor Beginn der Vorlesung.

Aufgabe 1 BLAST (3 Punkte)
Informiere Dich genauer über BLAST.

1. Welche Rolle spielt der `X-Drop Algorithmus` und wie funktioniert er genau?
2. Was sind die Unterschiede zwischen BLAST Version 1.4 und den Versionen ab 2.0?
3. Was bedeutet der Begriff *Sensitivität* im Kontext eines Matches?

Aufgabe 2 Statistik auf Sequenzen (3 Punkte)
Betrachte ein i.i.d.-Modell für DNA-Sequenzen mit den relativen Buchstabenhäufigkeiten $f_{\mathtt{A}} = f_{\mathtt{T}} = 0.3$ und $f_{\mathtt{G}} = f_{\mathtt{C}} = 0.2$.

1. Wie wahrscheinlich ist es, dass eine Zufallssequenz X mit den oben angegebenen Häufigkeiten der Sequenz $x = \mathtt{CCGATCGACGTA}$ entspricht?
2. Wie hoch ist der Erwartungswert für die Anzahl der Treffer von x in einer zufälligen Sequenz X der Länge 1200?

Aufgabe 3 Q-Gramm Statistik (4 Punkte)
Seien X und Y zwei Zufallssequenzen nach dem i.i.d.-Modell. Betrachte nun die zwei Q-Gramme $X[i \dots i+q-1]$ und $Y[j \dots j+q-1]$. Gib die Formeln für die folgenden Wahrscheinlichkeiten an und erläutere jeweils deine Herleitung.

1. Wie hoch ist die Wahrscheinlichkeit, dass die Q-Gramme **genau ein** Mismatch enthalten?
   $\mathbb{P}\big(d_H(X[i \dots i+q-1], Y[j \dots j+q-1]) = 1\big) = \dots$
2. Wie hoch ist die Wahrscheinlichkeit, dass die Q-Gramme **höchstens ein** Mismatch enthalten?
   $\mathbb{P}\big(d_H(X[i \dots i+q-1], Y[j \dots j+q-1]) \leq 1\big) = \dots$
3. Wie hoch ist die Wahrscheinlichkeit, dass die Q-Gramme **höchstens zwei** Mismatches enthalten?
   $\mathbb{P}\big(d_H(X[i \dots i+q-1], Y[j \dots j+q-1]) \leq 2\big) = \dots$

Aufgabe 4 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. Seien $m = n = 3000$, $p = 1/4$ und $q = 6$. Wie muss t gewählt werden, damit der p-Value 0.01 ist? Erläutere ob man bei der Wahl eines ganzzahligen t auf- oder abrunden muss.
2. Eine Suchsequenz x hat die Länge 1000. Bei einer FASTA-Suche mit $q = 6$ gibt es in einer Datenbank zwei Treffer mit einem Score über 15. Die Sequenz y_A, mit einer Länge von 1200, hat einen Score von $C(x, y_A) = 17$; die Sequenz y_B, mit der Länge 950, erzielt den Score $C(x, y_B) = 18$. Welcher Treffer ist signifikanter? Berechne jeweils p-Value und Bit-Score.