

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2004/05

Utz J. Pape · Ben Rich · Dr. Stefan Röpcke · Prof. Dr. Martin Vingron

Blatt 3 · Ausgabe am 01.11.2004

Abgabe am 08.11.2004 vor Beginn der Vorlesung

Aufgabe 7 (FastA / Blast). In der Vorlesung haben Sie die beiden Datenbanksuch-
Algorithmen FastA und Blast kennengelernt. Die Query-Sequenz ist gegeben mit

ACTTACGACGAACTACGACC

1. Bei FastA wird in der Vorverarbeitung für die Query eine Index-Tabelle mit k -Tupels angelegt. Erstellen Sie diese Tabelle für obige Sequenz mit $k = 3$.
2. Der Blast Algorithmus durchsucht die Datenbank mit einem Automaten. Der Automat erkennt nicht nur die k -Tupel der Sequenz, sondern zusätzlich auch die Neighborhood zu jedem k -Tupel. Geben Sie die Neighborhoods der ersten vier k -Tupels mit $k = 3$ von obiger Sequenz an.
3. Neben Algorithmen, die die Query vorverarbeiten, gibt es auch Algorithmen, welche die Datenbank vorverarbeiten. Zum Beispiel kann man die Datenbank als Suffixbaum speichern.
 - (a) Generieren Sie den Suffixbaum zu obiger Sequenz.
 - (b) *Bonus:* Der Suffixbaum hat den Vorteil, dass Suchanfragen (exaktes Pattern-Matching) sehr schnell beantwortet werden können. Wie schnell?
 - (c) *Bonus:* Welchen Nachteil haben Suffixbäume? Wie löst man das Problem?

Aufgabe 8 (Verteilungen). In der heutigen Vorlesungsstunde haben Sie einige Verteilungen kennengelernt, sowie deren wichtige Funktion für die Sequenzanalyse. Auch die Verteilungen, die Sie in früheren Vorlesungen kennengelernt haben, sind sehr wichtig. Gehen Sie zur Wiederholung auf folgende Punkte für die Binomialverteilung, die Poisson-Verteilung und die Exponentialverteilung ein:

- Dichte- und Verteilungsfunktion als Formel und als Graph
- Parameter und deren Bedeutung (wie verändert sich der Verlauf des Graphen bei Änderung der Parameter)
- das erste und das zweite Moment
- ein Beispiel aus der Natur/dem realen Leben

Aufgabe 9 (Scorestatistik bei lokalen Alignments). In der Vorlesung haben Sie einiges über Score-Statistiken von Alignments gehört. Hier wollen wir nun selbst eine Score-Statistik erstellen. Dabei definieren wir zunächst ein Null-Modell und simulieren mit diesem Zufallssequenzen. Dadurch erhalten wir eine Score-Verteilung für das Nullmodell. Im zweiten Teil der Aufgabe berechnen Sie dann den Score eines Alignments zweier verwandter Sequenzen und können durch das Nullmodell aus dem ersten Teil einen p-value für diesen Alignment-Score bestimmen.

1. Nullmodell

- (a) Erstellen Sie ein Programm zur Erzeugung von Zufallssequenzen der Länge 500 wobei die Basen unabhängig gleichverteilt sind.
- (b) Führen Sie 1000 Smith-Waterman Alignments mit zufällig generierten Sequenzen durch. Bewerten Sie dabei einen Match mit 2, 3 und einen Mismatch mit $-3, 1$. Die Gapkosten sind gegeben durch die Funktion $g(k) = 13,1 \cdot k$ wobei k die Länge des Gaps ist. Stellen Sie die erhaltenen Scores (des jeweils optimalen Alignments) als Histogramm dar.
- (c) In der Praxis ist es oft zeitaufwendig, Simulationen durchzuführen. Daher versucht man bekannte parametrische Verteilungen anstelle der simulierten Verteilung zu benutzen. Die Wahl der Verteilung ist dabei entscheidend. Fitten Sie daher folgende Verteilungen an die Verteilung der simulierten Scores und stellen jeweils die empirische und die gefittete Verteilungsfunktionen zusammen dar:
 - Normalverteilung: Schätzen Sie den Mittelwert und die Varianz aus ihrer Simulation, um die Parameter der Normalverteilung zu erhalten.
 - (Gumbel-) Extremwertverteilung: $P(X \leq x) = \exp(-e^{-(x-\xi)/\theta})$. Die Parameter ξ, θ können Sie durch Schätzung der Momente mit Hilfe folgender Gleichungen approximieren:

$$\begin{aligned}\tilde{\theta} &= \frac{\sqrt{6}}{\pi} S \\ \tilde{\xi} &= \bar{X} - \gamma \tilde{\theta}\end{aligned}$$

Dabei ist \bar{X} der empirische Mittelwert, S die empirische Standardabweichung und γ die Euler-Mascheroni Konstante mit dem Wert $\gamma = 0,577\dots$

2. Nun wollen wir den Score und den entsprechenden p-value des Scores zweier verwandter Sequenzen berechnen. Wir wollen also herausfinden, ob deren lokales optimales Alignment signifikant ist.

- (a) *Bonus*: Generieren Sie zwei verwandte Sequenzen. Gehen Sie dabei wie folgt vor: Erzeugen Sie die erste Sequenz (der Länge 500) zufällig (siehe oben). Die zweite Sequenz erhalten Sie, indem Sie für jede Position zufällig bestimmen, ob an dieser Position eine Mutation (wir verstehen unter einer Mutation auch eine silent Mutation bei der z.B. ein A zu einem C und dieses später wieder zu einem A mutiert) erfolgt. Die Mutationswahrscheinlichkeit soll 65% betragen. Erfolgte eine Mutation, dann ziehen Sie erneut eine Base aus den vier möglichen Basen, wobei diese gleichverteilt sind. Wir betrachten also nur Substitutionen und ignorieren an dieser Stelle der Einfachheit halber Deletionen und Insertionen.
- (b) Nehmen Sie entweder an, dass dieses Alignment einen Score von $s = 30,5$ hat oder *Bonus*: berechnen Sie den Score s des lokalen optimalen Alignments der beiden verwandten Sequenzen von oben.
- (c) Bestimmen Sie den p-value des Scores s mit Hilfe der verschiedenen Nullmodelle aus obiger Aufgabe:

- Simulation: Schauen Sie, wieviele Scores der 1000 durchgeführten Alignments schlechter sind als s . Dividieren Sie diese Zahl durch die Anzahl der Simulationen (also 1000). Schon haben Sie eine Approximation für den p-value p . Der p-value besagt, dass unser optimales Alignment mit Score s mit einer Wahrscheinlichkeit von p zufällig auftritt. Dabei wird durch das Nullmodell festgelegt, was unter zufällig zu verstehen ist.
- Normalverteilung: Standardisieren ($\mu = 0, \sigma^2 = 1$) Sie die geschätzte Normalverteilung aus dem ersten Aufgabenteil durch Transformation der Zufallsvariable $X' = \frac{X - \hat{\mu}}{\hat{\sigma}}$ und ermitteln Sie dann die Wahrscheinlichkeit $P(X \geq s)$ aus einer Tabelle. (Alternativ können Sie auch die Wahrscheinlichkeit durch ein Programm wie R berechnen. Dann benötigen Sie auch die Transformation nicht.) So erhalten Sie eine weitere Schätzung für den p-value.
- Im ersten Aufgabenteil ist die Extremwert-Verteilungsfunktion gegeben. Benutzen Sie diese mit den ermittelten Parametern, um $P(X < s)$ auszurechnen. Der p-value ist $P(X \geq s) = 1 - P(X < s)$.
- Vergleichen Sie die erhaltenen p-values und erklären Sie die Unterschiede.

Wieviel Zeit haben Sie zur Lösung dieses Aufgabenblatts verwendet? Hätten Sie mehr Zeit gebraucht? Wieviel?