Übungen Algorithmen der Bioinformatik II
Wintersemester 2004/05

Stefan Posch, Birgit Möller, Denis Williams

Institut für Informatik
Universität Halle

Blatt 2

Aufgabe 2.1 Wir betrachten eine DNA-Sequenz der Länge N. Unter der Null-Hypothese seinen alle Positionen statistisch unabhängig und jede Position gleichverteilt.

(a) Wir betrachten wie in der Vorlesung die (überlappend) gezählte Anzahl k des Auftreten des Muster *gcgg*. Bestimmen Sie Erwartungswert und Varianz.

(b) Benutzen Sie nun als Approximation der Verteilung von k eine Normalverteilung. Beschreiben Sie das Vorgehen eines statistischen Test zum Signifikanzniveau $\alpha = 0.01$, um obige Null-Hypothese zu verwerfen oder zu akzeptieren.

(c) Wir nehmen nun (fälschlicherweise) an, daß die Anzahl k der Auftreten des Musters binomial verteilt ist.

  (i) Warum ist diese Annahme falsch?
  (ii) Geben Sie den resultierenden Erwartungswert und Varianz an.

(d) Erweitern Sie die Überlegungen aus Teilaufgabe (a) für ein beliebiges Muster der Länge 5.

Aufgabe 2.2 Wir betrachten die idealisierte shot-gun Fragmentierung einer Sequenz der Länge G mit insgesamt N Fragmenten jeweils der Länge L. Das linke Ende jedes Fragments liege (wiederum vergröbert) gleichverteilt auf einer der Positionen im Interval $[1, G]$.

(a) Für welche Anzahl N der Fragmente wird die erwartete Anzahl $C = e^{-a}N$ maximal? Wir halten dabei G und L fest, $a = \frac{LN}{G}$ sei die coverage.
Vergleichen Sie dieses Ergebnis mit den Beispielzahlen der Vorlesung.

(b) Für $|x|$ klein gilt folgende Appoximation: $e^x \approx 1 + x$.
Nuzten Sie diese Appoximation für die coverage a in der Abschätzung über die erwartete contig-Länge $L\frac{e^a-1}{a}$. Interpretieren Sie das Ergebnis.

(c) Simulieren Sie diese shotgun-Fragmentierung für verschiedene Werte von G, L und N und vergleichen Sie die Resultate mit den Vorhersagen.
Ändern die nun die Simulation, indem die Länge L der Fragmente gleichverteilt aus dem Intervall $[400, 600]$ gezogen wird und vergleichen Sie.

Abgabe: 28.10.04