Übungen Algorithmen der Bioinformatik II
Wintersemester 2006/07

Dr. Ivo Große, Dipl.-Bioinf. Jan Grau

Institut für Informatik
Universität Halle

Blatt 5

Aufgabe 5.1 Beweisen Sie: Wenn $X_1, X_2, \ldots, X_M$ statistisch unabhängig sind, gilt $E(\prod_{m=1}^{M} X_m) = \prod_{m=1}^{M} E(X_m)$. Beweisen Sie weiterhin die folgenden beiden Eigenschaften des Korrelationskoeffizienten $\rho(X,Y)$: $-1 \leq \rho(X,Y) \leq 1$ und $\rho(X,Y) = \pm 1$ gdw. X und Y linear voneinander abhängen. Finden Sie drei Beispiele für statistisch abhängige Zufallsvariable, die unkorreliert sind, ein Beispiel für zwei funktional abhängige Zufallsvariable X und Y mit $-1 < \rho(X,Y) < 1$ und ein Beispiel für zwei funktional abhängige Zufallsvariable X und Y mit $\rho(X,Y) = 0$.

Aufgabe 5.2 In Aufgaben 3.2 und 4.2 haben wir uns mit der Häufigkeitsverteilung überlappender Trimere in Binärsequenzen befasst.

(a) Leiten Sie für jedes Trimer die Varianzen von N_{ijk} und D_{ijk} als Funktionen von N her und vergleichen Sie diese mit denen der Simulation aus Aufgabe 3.2 und den analytisch berechneten Varianzen aus Aufgabe 4.2.

(b) Definieren Sie nun $n_{ijk}^{1/2} = \frac{N_{ijk}^{1/2} - E(N_{ijk}^{1/2})}{\sqrt{Var(N_{ijk}^{1/2})}}$ und $d_{ijk} = n_{ijk}^{1} - n_{ijk}^{2}$ und wiederholen Sie Aufgabe 3.2(a-h), indem Sie überall $N_{ijk}^{1/2}$ durch $n_{ijk}^{1/2}$ und D_{ijk} durch d_{ijk} ersetzen.

(c) Welche Empfehlung würden Sie einem angewandten Bioinformatiker geben, der das *overlapping word paradox* nicht kennt, aber dringend Sequenz- und Expressionsdaten wie in Aufgabe 3 beschrieben analysieren möchte?

Aufgabe 5.3 Wir betrachten ein Genom, das mehrere tausend Replikate eines Sequenzstückes enthält, welches für 5.8 S rRNA kodiert. Diese Sequenzen sind aufgrund evolutionärer Ereignisse unterschiedlich, einige haben auch ihre Funktionalität verloren, sind aber im Genom erhalten geblieben.

Wir interessieren uns für einen Teil dieser Sequenzen und wollen diese klonieren, um sie anschließend sequenzieren zu können. Zunächst nutzen wir PCR, da wir ein Primer-Paar kennen, das alle Sequenzen der 5.8 S rRNA kodierenden Bereiche flankiert. Das PCR-Produkt wird mittels Gelelektrophorese bezüglich der Länge der Sequenzen getrennt, und im weiteren untersuchen wir eine der entstandenen Bande.

Wir nehmen an, daß diese Bande einen gewissen Anteil an funktionalen und nichtfunktionalen Sequenzen enthält, wobei die funktionalen Sequenzen auf Grund Ihrer Funktionalität in zwei Klassen (I und II) eingeteilt werden können. Wir nehmen weiterhin an, dass die Sequenzen in der von uns untersuchten Bande von A_{I} und A_{II}

Replikaten stammen, die funktionale 5.8 S rRNA kodieren, sowie von B Replikaten, die nicht-funktionale 5.8 S rRNA kodieren.

Aus dieser Bande klonieren wir nun N (individuelle) Moleküle und erhalten so eine Klonbibliothek mit N Klonen. Diesen Prozeß modellieren wir als Ziehen einer Stichprobe **mit** Zurücklegen.

(a) Diskutieren Sie die geschilderte Modellierung.

(b) Wieviele Sequenzen für funktionale 5.8 S rRNA aus Klasse I, funktionale 5.8 S rRNA aus Klasse II und nicht-funktionale 5.8 S rRNA haben wir im Mittel in unserer Klonbibliothek?

(c) Wie groß müssen wir N wählen, um mit Wahrscheinlichkeit ϕ mindestens eine funktionale 5.8 S rRNA aus Klasse I, eine funktionale 5.8 S rRNA aus Klasse II und eine nicht-funktionale 5.8 S rRNA in der Klonbibliothek zu haben?

(d) Führen sie für $A_{\mathrm{I}} = 600$, $A_{\mathrm{II}} = 300$ und $B = 100$ Simulationen mit $N = 10, 100, 1000, 10000$ durch und diskutieren Sie die Ergebnisse insbesondere im Vergleich zu den theoretisch bestimmten Resultaten.

Abgabe: 21.11.06