Übungen Algorithmen der Bioinformatik II
Wintersemester 2006/07

Dr. Ivo Große, Dipl.-Bioinf. Jan Grau

Institut für Informatik
Universität Halle

Blatt 3

Durch Microarray Studien erhält man oft Datensätze von Genen mit verschiedenen Expressionsprofilen, z. B. differentiell exprimierte Gene auf der einen Seite und eine Kontrollgruppe von nicht differentiell exprimierten Genen auf der anderen Seite. Diese nennen wir im folgenden einfach Kontrollgene. Oft untersucht man nun, ob es Sequenzmuster gibt, die in den Promotorregionen der differentiell exprimierten Gene wesentlich häufiger (oder seltener) auftreten als in den Promotorregionen der Kontrollgene. Diese könnten nämlich auf Bindungsstellen von Transkriptionsfaktoren hindeuten, die für die differentielle Expression der untersuchten Gene verantwortlich sein könnten. Oft wird bei derartigen Analysen ein Phänomen übersehen, welches unter dem Namen *overlapping word paradox* bekannt ist. Diesem Phänomen wollen wir uns durch die folgenden beiden Studien nähern.

Aufgabe 3.1 Studie 1 (mathematisch einfach, biologisch nicht sinnvoll)
Wir betrachten der Einfachheit halber im folgenden Binärsequenzen, die aus den beiden Nukleotiden A und B bestehen. Wir betrachten zwei Datensätze aus je 100 Sequenzen der Länge 300 bp. In jedem Datensatz bestimmen wir die absolute Häufigkeit jedes der 8 möglichen Trimere durch **nichtüberlappendes** Zählen, d. h. in jeder Sequenz werden genau 100 nichtüberlappende Trimere gezählt. Die absolute Häufigkeit von Trimer ijk in Datensatz 1 (2) nennen wir $N_{ijk}^{(1)}$ ($N_{ijk}^{(2)}$), und wir definieren die Differenz $D_{ijk} = N_{ijk}^{(1)} - N_{ijk}^{(2)}$. Generieren Sie 10^4 solcher Paare von Datensätzen 1 und 2, wobei jedes Nukleotid einer jeden Sequenz durch Werfen einen fairen Münze statistisch unabhängig von allen anderen Nukleotiden generiert wird.

(a) Rangordnen Sie für jedes Paar von Datensätzen die 8 D-Werte, und notieren Sie für jedes Trimer dessen Rang $\ell \in \{1, 2, \ldots, 8\}$. Stellen Sie für jedes Trimer die Rangverteilung $P(\ell)$ graphisch dar.

(b) Stellen Sie in **einer** Abbildung graphisch dar, wie häufig jedes der 8 Trimere auf Rang 1 landet.

(c) Stellen Sie für jedes Trimer die empirischen Verteilungen $P_1(N_{ijk}^{(1)})$, $P_2(N_{ijk}^{(1)})$ und $P_3(D_{ijk})$ graphisch dar.

(d) Wie stark unterscheiden sich die empirischen Verteilungen P_1 und P_2?

(e) Wie stark unterscheiden sich die empirischen Verteilungen P_1, P_2, P_3 von Trimer zu Trimer?

Abgabe: 07.11.06

(f) Bestimmen Sie für jedes Trimer den Erwartungswert und die Varianz von $N_{ijk}^{(1)}$, $N_{ijk}^{(2)}$ und D_{ijk} aus den Ergebnissen der Simulation.

(g) Wie stark unterscheiden sich die Erwartungswerte und die Varianzen von $N_{ijk}^{(1)}$ und $N_{ijk}^{(2)}$?

(h) Wie stark unterscheiden sich die Erwartungswerte und die Varianzen von Trimer zu Trimer?

Aufgabe 3.2 Studie 2 (mathematisch schwieriger, biologisch sinnvoller)
Wir betrachten zwei Datensätze aus je 100 Sequenzen der Länge 102 bp. In jedem Datensatz bestimmen wir die absolute Häufigkeit jedes der 8 möglichen Trimere durch **überlappendes** Zählen, d. h. in jeder Sequenz werden genau 100 überlappende Trimere gezählt. Wiederholen Sie nun alle Schritte aus Aufgabe 3.1.

(a) Formulieren Sie eine Vermutung, warum die Verteilungen und Varianzen, aber nicht die Erwartungswerte, von Trimer zu Trimer variieren.

(b) Gibt es eine Korrelation zwischen der Varianz und der Wahrscheinlichkeit, auf Rang 1 zu landen?

(c) Formulieren Sie Ihre Version des *overlapping word paradox*.