

# Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2002/03

Dipl.-Math. Sven Rahmann · Prof. Dr. Knut Reinert

**Blatt 9 · Ausgabe am 19.12.2002**

**Abgabe am 9.1.2003 vor Beginn der Vorlesung**

**Aufgabe 34 (Entdecken von nichteindeutigen Unitigs).** In der Vorlesung wurde die Teststatistik

$$T := c - K \ln 2$$

verwendet, um unique Unitigs von Unitigs aus kollabierten Repeats zu unterscheiden. Dabei ist  $c := R\rho/G$  die erwartete Anzahl der Read-Starts in einem Sequenzfenster der Länge  $\rho$ ,  $G$  ist die Länge der Zielsequenz,  $R$  die Anzahl der Reads insgesamt, und  $K$  die beobachtete Zahl der Read-Starts im betrachteten Fenster.

Zur Klarstellung der Definitionen: Wir gehen davon aus, dass ein Unitig aus mindestens zwei Reads besteht. Die zugehörige Fensterlänge  $\rho$  ist die Differenz zwischen den Startpunkten des ersten und letzten Reads, und  $K \geq 0$  ist die Anzahl der *internen* read-Starts (also ohne erstes und letztes Read).

1. Motivieren Sie die angegebene Form von  $T$ . Gehen Sie dabei auch auf unausgesprochene Annahmen und Vereinfachungen ein.
2. Wie ist ein Wert von  $T = -10$ ,  $T = 0$ ,  $T = +10$  jeweils zu interpretieren?
3. Gegeben ist eine Sequenz der Länge  $G = 500$  Kb, sowie  $R = 7000$  reads. Kann man ein Unitig der Länge  $\rho = 5000$ , das aus  $K = 250$  reads besteht, als zuverlässig nicht-repetitiv (unique) ansehen?

**Aufgabe 35 (Distanzschätzung zwischen Unitigs).** Das Konzept der Mate-pairs erlaubt, die wahre Distanz  $d$  zwischen zwei Unitigs zu schätzen. Wir haben  $n$  Mate-Pairs, welche Schätzungen  $X_i$  ( $i = 1, \dots, n$ ) mit  $\mathbb{E}[X_i] = d$  und bekannten Varianzen  $\text{Var}[X_i] = \sigma_i^2$  liefern. Daraus soll nun eine Gesamtschätzung  $\hat{d}$  für  $d$  gebildet werden.

1. Ein erster Vorschlag ist der Mittelwert  $\hat{d}_m = \sum_{i=1}^n X_i/n$ . Berechnen Sie die Varianz dieses Schätzers.
2. Intuitiv ist es geschickter, die Einzelschätzungen  $X_i$  mit hoher Varianz weniger stark zu gewichten. Das führt auf den gewichteten Ansatz

$$\hat{d}_w = \sum_{i=1}^n w_i X_i$$

mit den Nebenbedingungen  $\sum_i w_i = 1$  und  $w_i \geq 0$ . Zeigen Sie, dass für beliebige Gewichte  $w_i$ , die diesen Nebenbedingungen gehorchen,  $\mathbb{E}[\hat{d}_w] = d$  gilt, und berechnen Sie  $\text{Var}[\hat{d}_w]$ .

3. Wählen Sie die Gewichte  $w_i$  unter den angegebenen Nebenbedingungen so, dass  $\text{Var}[\hat{d}_w]$  minimal wird. Überprüfen Sie, dass Ihre Lösung der Intuition entspricht, dass  $w_i$  klein ist, wenn  $\sigma_i^2$  groß ist.

**Aufgabe 36 (Intersection Graph eines Eulerschen Graphen).** Sei  $G = (V, E)$  ein Eulerscher Graph, also ein Graph, in dem mindestens ein Eulerkreis  $C$  existiert. Um zu testen, ob es *genau* einen Eulerkreis gibt, verwenden wir den “Intersection Graph”  $I = (W, F)$ , der wie folgt definiert ist: Man zerlegt  $C$  in eine Familie  $(C_1, \dots, C_t)$  von kantendisjunkten *einfachen* Kreisen (in einem einfachen Kreis wird jeder Knoten nur einmal durchlaufen). Die einfachen Kreise bilden die Knoten des Intersection Graphs:  $W = \{C_1, \dots, C_t\}$ . Die Kantenmenge  $F$  ist wie folgt definiert: Die Knoten  $C_i$  und  $C_j$  sind mit genau  $k$  Kanten verbunden, wenn in  $G$  die Kreise  $C_i$  und  $C_j$  genau  $k$  Knoten gemeinsam haben (es handelt sich also um eine Multimenge; zwei Knoten können mit mehr als einer Kante verbunden werden).

Beweisen Sie:  $G$  besitzt genau dann nur einen Euler-Kreis, wenn  $I$  ein Baum ist.

**Aufgabe 37 (Wie packt man alle  $q$ -grams in einen String?).** Sei  $\Sigma$  ein endliches Alphabet der Größe  $n$ . Sei  $q$  eine ganze Zahl. Dann gibt es über  $\Sigma$  also  $n^q$  verschiedene  $q$ -grams. Gesucht wird ein möglichst kurzer Superstring (nicht Supersequenz)  $s$ , der alle  $n^q$  verschiedenen  $q$ -grams als Substrings enthält.

1. Beweisen Sie die trivialen Schranken  $n^q + q - 1 \leq |s| \leq qn^q$ .
2. Geben Sie einen optimalen Superstring für  $n = q = 3$  an ( $\Sigma = \{a, b, c\}$ ).
3. Beweisen Sie, dass stets ein Superstring  $s$  mit minimal möglicher Länge (also mit  $|s| = n^q - q + 1$ ) existiert (Hinweis: Euler).

*Hinweise zur Klausur: In der Klausur könnten unter anderem Aufgaben vorkommen, die den Quizfragen ähneln (aber komplexer sind). Des weiteren bieten sich insbesondere in vereinfachter Form die Aufgaben 1, 6, 9, 10, 11, 12, 14, 16, 17, 18, 20, 22, 25, 27, 29, 31, 32, 33, 34 an. Natürlich ist diese Liste nicht vollständig.*