

# Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2004/05

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

**Blatt 4 · Ausgabe am 8.11.2004**

**Abgabe am 15.11.2004 vor Beginn der Vorlesung**

**Aufgabe 10 (Parameterwahl für lokale Alignments).** In dieser Aufgabe sollen Sie den in der Vorlesung beschriebenen Phasenübergang beobachten. Wählen Sie zwei Zufallssequenzen der Länge 500 und berechnen Sie ein lokales maximales Alignment mit linearen Gapkosten (Vorschlag für die Wahl der Parameter: *match* 1,3 und *mismatch*  $-1,7$ ). Verwenden Sie möglichst Ihre eigene Implementierung des Smith-Waterman, beziehungsweise Waterman-Eggert.

1. Variieren Sie die Gapkosten und berechnen Sie jeweils ein maximales lokales Alignment und seinen Score. Sie sollten näherungsweise eine Schwelle für die Gapkosten identifizieren unter der auch lokale Alignments 'zerrissen' aussehen und über der die Alignments kompakt sind.
2. Prüfen Sie, ob für diese Schwelle die in der Vorlesung beschriebenen Eigenschaften gelten: Die Scores der lokalen maximalen Alignments wachsen logarithmisch für Gapkosten oberhalb der Schwelle und linear für solche unterhalb.
3. Der erwartete Score des *globalen* Alignments mit Gapkosten oberhalb der Schwelle ist kleiner Null und mit solchen unterhalb der Schwelle grösser Null. In diesem Bereich verliert also die Karlin-Altschul-Statistik ihre Gültigkeit.

**Aufgabe 11 (Multiple Alignments).** 1. Laden Sie aus dem Internet folgende Proteinsequenzen:

- Human alpha Hämoglobin
  - Human beta Hämoglobin
  - Huhn (*Gallus gallus*) alpha Hämoglobin
  - Huhn (*Gallus gallus*) beta Hämoglobin
  - Sperm whale Myoglobin
2. Alignieren Sie diese, benutzen Sie dazu ClustalW (z.B. auf dem EBI Server).
  3. Was passiert, wenn Sie noch eine nicht dazugehörige Sequenz hinzufügen (z.B. Cytochrom c)?
  4. In dem Alignment (ohne eine fremde Sequenz) werden Sie unterschiedlich stark konservierte Regionen finden. Um herauszufinden, welche biologische Bedeutung die stark konservierten Sequenzbereiche haben, schneiden Sie z.B. vom Human alpha Hämoglobin diese Bereiche aus und suchen in einer Domain Datenbank. Dafür bietet sich SMART (EMBL) an.

**Aufgabe 12 (Markov-Ketten).** In dieser Aufgabe wollen wir die Essensabfolge in der Mensa modellieren. Zur Vereinfachung unterscheiden wir nur zwischen drei verschiedenen Gerichten von denen pro Tag jeweils genau eines angeboten wird:

- $M$ : Fleisch (Meat)
- $F$ : Fisch
- $V$ : Gemüse (Vegetable)

Nun haben wir schon seit Monaten die Essensausgabe beobachtet und jeden Tag notiert, welches Gericht es gab, also beispielsweise

$MFMVMSMFMMVFMVMVSMMSMSMVVMSVMSSVM\dots$

Die richtige Folge finden Sie auf der Vorlesungsseite zum Herunterladen (ab Montag abend).

1. Modellieren Sie die Essensausgabe als Markovkette 0ter Ordnung. Zeichnen Sie die Markovkette und schätzen Sie die Parameter. Mit was für einer Wahrscheinlichkeit gibt es nach einander  $VVVMF$ ?
2. Nun benutzen wir eine Markovkette 1ter Ordnung zur Modellierung. Was für eine Wahrscheinlichkeit hat dann  $VVVMF$ ? (Und geben Sie bitte auch die geschätzten Parameter an.)
3. Berechnen Sie die Wahrscheinlichkeit der ursprünglichen Sequenz in beiden Modellen. Was stellen Sie fest? Warum?
4. Wenn es an einem Tag Fleisch gab, mit welcher Wahrscheinlichkeit gibt es dann was vier Tage später? *Bonus: Mit welchem kleinen Trick können Sie die Berechnung im Computer beschleunigen?*