Technische Universität München
Institut für Informatik
Lehrstuhl für Effiziente Algorithmen
Volker Heun

SS 2000
Lösungsblatt 5
29. Juni 2000

Algorithmen der Bioinformatik

Aufgabe 1

Geben Sie eine auf lokalem Sequenzen Alignment basierende Methode an, um eine optimale Wiederholung, d.h. ein Paar $(s_i \cdots s_{i'}, s_j \cdots s_{j'})$ von Teilstrings mit maximalem Score, innerhalb eines gegebenen Strings $s \in \Sigma^*$ zu finden. Das gefundene Alignment darf in der Dynamic Programming Tabelle die Diagonale nicht berühren.

Lösungsvorschlag:
Wir verwenden einfach die Methode der dynamischen Programmierung. Da sich Wiederholung zwar überlappen, aber sich nicht selbst „überholen dürfen“, darf der zu einem optimalen Score gehörige Weg die Hauptdiagonale nicht kreuzen. Wir bleiben daher ohne Beschränkung der Allgemeinheit unterhalb der Hauptdiagonalen und erhalten:

$$D[i,j] := \begin{cases} -\infty & \text{falls } i = j \\ 0 & \text{falls } j = 0 \\ \max \left\{ \begin{array}{l} 0, \\ D[i-1,j] + s(x_i, -), \\ D[i-1,j-1] + s(x_i, x_j), \\ D[i,j-1] + s(-, x_j) \end{array} \right\} & \text{falls } i < j \end{cases} .$$

In der Maximumsbildung taucht noch die Null auf, da wir bei einem lokalen Alignment ja an jeder Position mit dem Alignment beginnen dürfen. Aus diesem Grund wird auch die erste Spalte auf Null gesetzt ($D[i, 0] = 0$).

Aufgabe 2

Beweisen oder widerlegen Sie: Ein optimales mehrfaches Sequenzen Alignment induziert mindestens ein optimales paarweises Sequenzen Alignment.

Lösungsvorschlag:
Wir betrachten zunächst die folgenden drei Zeichenreihen $s_1 = AC$, $s_2 = AG$ und

$s_3 = CG$. Wir zeigen zuerst, dass das folgende Alignment optimal ist:

$$\begin{array}{cc} A & C \\ A & G \\ C & G \end{array}$$

Offensichtlich sind die Kosten gemäß dem SP-Maß genau 4 Substitutionen.

Nur ein MSA mit Länge größer als 2 könnte besser sein. Betrachten wir das folgende Alignment der Länge 3:

$$\begin{array}{ccc} A & C & - \\ A & - & G \\ - & C & G \end{array}$$

Hier betragen die Kosten 6 Indel-Operationen. Wir nehmen im Folgenden an, dass eine Indel-Operation mehr als 2/3 einer Substitution kostet (wir wissen sowieso, dass sie immer mehr als die Hälfte der Kosten einer Substitution kosten). Dann ist das obige MSA der Länge 3 teurer. Man sieht leicht, dass das obige MSA das billigste der Länge 3 ist. Ebenso leicht überzeugt man sich, dass ein MSA der Länge größer als 4 auch höchstens teurer werden kann.

Im ersten MSA für s_1, s_2 und s_3 gilt, dass die induzierten paarweisen Alignments von s_1 mit s_2 und s_2 mit s_3 optimal sind. Im Gegensatz dazu ist das induzierte paarweise Alignment von s_1 mit s_3 nicht optimal, da

$$\begin{array}{ccc} A & C & - \\ - & C & G \end{array}$$

mit zwei Indel-Operationen billiger ist als das induzierte mit 2 Substitutionen. Wir nehmen hier an, dass eine Indel-Operation billiger als eine Substitution ist.

Wir konstruieren nun drei neue Zeichenreihen wie folgt:

$$\begin{array}{rcl} t_1 & := & s_1 \, TTT \, s_2 \, TTT \, s_3 \\ t_2 & := & s_2 \, TTT \, s_3 \, TTT \, s_1 \\ t_3 & := & s_3 \, TTT \, s_1 \, TTT \, s_2 \end{array}$$

Dafür erhalten wir das folgende optimale MSA:

$$\begin{array}{cccccccccccc} A & C & T & T & T & A & G & T & T & T & C & G \\ A & G & T & T & T & C & G & T & T & T & A & C \\ C & G & T & T & T & A & C & T & T & T & A & G \end{array}$$

Gäbe es ein besseres MSA für t_1, t_2 und t_3, dann könnte man daraus ein besseres MSA für s_1, s_2 und s_3 konstruieren, da die Teilzeichenreihen TTT immer untereinander stehen müssen.

Für jedes induzierte paarweise Alignment von t_i mit t_j mit $i < j \in [1:3]$ gilt nun aber, dass dort s_1 und s_3 als Teilzeichenreihen übereinanderstehen. Für diese Zeichenreihen können wir jedoch ein besseres MSA mit Indel-Operationen angeben. Damit gibt es MSA, so dass jedes induzierte paarweise Sequenzen Alignment nicht optimal ist.

Aufgabe 3

Zeigen Sie, dass es einen String $s \in \{s_1, \ldots, s_k\}$ gibt, der als Zentrum bei der Center Star Methode eine nicht-konstante Approximationsgüte liefert.

Lösungsvorschlag:
Wir betrachten die folgenden $2k+1$ Zeichenreihen der Länge $O(n)$, dabei sein n gerade:

$$\begin{aligned}
s_0 &:= T^n \\
s_1 &:= T^{n/2}(AG)^k T^{n/2} \\
&\vdots \\
s_k &:= T^{n/2}(AG)^k T^{n/2} \\
s_{k+1} &:= T^{n/2}(GA)^k T^{n/2} \\
&\vdots \\
s_{2+1} &:= T^{n/2}(GA)^k T^{n/2}
\end{aligned}$$

Beachte, dass $s - 1 = \cdots = s_k$ und $s_{k+1} = \cdots = s_{2k}$ gilt.

Wie man sich leicht überlegt sieht das optimale MSA gemäß des SP-Maßes wie folgt aus:

$$\begin{array}{cccccccccccccc}
T & T & \cdots & T & - & - & - & - & \cdots & - & - & - & T & \cdots & T \\
T & T & \cdots & T & A & G & A & G & \cdots & A & G & - & T & \cdots & T \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
T & T & \cdots & T & A & G & A & G & \cdots & A & G & - & T & \cdots & T \\
T & T & \cdots & T & - & G & A & G & \cdots & A & G & A & T & \cdots & T \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
T & T & \cdots & T & - & G & A & G & \cdots & A & G & A & T & \cdots & T
\end{array}$$

In der ersten und letzten Spalte ohne T's haben wir jeweils $k(k+1)$ Indel-Operationen. In den anderen Spalten ohne T's sind es jeweils $2k$ Indel-Operationen. Damit hat dieses MSA $2k(k+1) + (2k-1)2k = 6k^2$ Indel-Operationen.

Wählen wir für die Stern-Methode die Zeichenreihe s_0 als Zentrum, so erhalten wir das folgende MSA:

$$\begin{array}{ccccccccccccccc}
T & T & \cdots & T & - & - & - & - & \cdots & - & - & - & T & \cdots & T \\
T & T & \cdots & T & A & G & A & G & \cdots & G & A & G & T & \cdots & T \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
T & T & \cdots & T & A & G & A & G & \cdots & G & A & G & T & \cdots & T \\
T & T & \cdots & T & G & A & G & A & \cdots & A & G & A & T & \cdots & T \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
T & T & \cdots & T & G & A & G & A & \cdots & A & G & A & T & \cdots & T
\end{array}$$

Damit enthält jede Spalte ohne T's genau k^2 Substitutionen und $2k$ Indel-Operationen. Damit werden also insgesamt $2k^3$ Substitutionen und $4k^2$ Indel-Operationen benötigt.

Da die Kosten einer Substitution teurer als die Kosten einer Indel-Operation ist, ist die Lösung um mindestens den Faktor

$$\frac{2k^3 + 4k^2}{6k^2} \geq \frac{k}{3}$$

schlechter als die optimale Lösung. Somit können wir bei einer beliebigen Wahl des Zentrums bei der Stern-Methode keine konstante Approximationsgüte garantieren.

Anmerkung: Man kann jedoch zeigen, dass man bei freier Wahl des Zentrums mit Wahrscheinlichkeit mindestens $\frac{1}{2}$ eine 3-Approximation erhält.