

Übungen zu Theoretische Informatik III

Abgabe bis zum 17.05.05

Aufgabe 3.1 Dynamische Programmierung 12 Punkte

Eine Aufgabenstellung der Bioinformatik ist es, den „Verwandtschaftsgrad“ (Ähnlichkeit) zweier Gene, gegeben als DNA-Sequenzen, zu bestimmen. Für uns sind DNA-Sequenzen einfach endliche Wörter über dem Alphabet $\Sigma = \{A, G, C, T\}$ (abkürzend für die vier Basen Adenin, Guanin, Cytosin und Thymin, aus welchen die DNA aufgebaut ist). Wir gehen von zwei Sequenzen $x = x_1x_2 \dots x_m$ und $y = y_1y_2 \dots y_n$ aus.

Da nicht jede Base einer DNA-Sequenz Information über ein Gen tragen muss bzw. die genaue Position eines Gens in einer DNA-Sequenz u.U. nicht bekannt ist, werden für ein gegebenes Sequenzenpaar (x, y) sogenannte Alignments (\vec{x}, \vec{y}) betrachtet, wobei das Paar (\vec{x}, \vec{y}) aus (x, y) entsteht, indem x bzw. y durch Einfügen so genannter Lücken '-' auf gleiche Länge gebracht werden ($|\vec{x}| = |\vec{y}|$).

Zum Beispiel ist für $x = ACCCGAT$ und $y = CCTATA$ ein mögliches Alignment

$$\begin{array}{cccccccc} \vec{x} & = & A & C & C & C & G & A & T & - \\ \vec{y} & = & - & C & C & - & T & A & T & A \end{array} .$$

Um die Qualität eines Alignments zu messen, bewertet man jede Stelle des Alignments einzeln und summiert dann die einzelnen Bewertungen auf. Das heißt, man geht von einer vorgegebenen Bewertungsfunktion

$$s : \{A, G, C, T, -\}^2 \rightarrow \mathbb{R}$$

für zwei Zeichen aus und dehnt diese auf das gesamte Alignment mittels

$$s^*(\vec{x}, \vec{y}) = \sum_{i=1}^{|\vec{x}|} s(\vec{x}_i, \vec{y}_i)$$

aus. Als Beispiel sei s so gewählt, dass $s(A, A) = s(G, G) = s(C, C) = s(T, T) = 2$, sonst $s(X, Y) = -1$ gilt. Dann hat das obige Alignment den Wert 4.

Von besonderem Interesse sind nun natürlich *optimale* Alignments für ein gegebenes Sequenzenpaar x, y , das heißt, Alignments (\vec{x}, \vec{y}) welche die Bewertung maximieren. Um durch unnötiges Auffüllen von Lücken keinen "Gewinn" zu erzielen und um somit unendliche Alignments zu verhindern, wird immer $s(-, -) < 0$ gesetzt.

- (a) Für ein Wort $w = w_1 \dots w_n$ sei hierfür $w_{[i,k]}$ das Teilwort $w_i w_{i+1} \dots w_k$ (mit $1 \leq i \leq k \leq n$).

Überlegen Sie sich, wie die Bewertung eines optimalen Alignments für die Sequenzen $x_{[1,i]}$ und $y_{[1,j]}$ aus den Bewertungen der optimalen Alignments der Sequenzenpaare $(x_{[1,i-1]}, y_{[1,j-1]})$, $(x_{[1,i]}, y_{[1,j-1]})$ und $(x_{[1,i-1]}, y_{[1,j]})$ berechnet werden kann. Speichern Sie diese Werte in einem Array $T[i, j]$ der Größe $m \times n$.

Geben Sie nun einen Algorithmus basierend auf dem Prinzip der dynamischen Programmierung an, welcher ein bzw. alle optimalen Alignments bestimmt für gegebenes Sequenzenpaar (x, y) und Bewertungsfunktion s .

- (b) Es seien folgende Sequenzen und Bewertungsfunktion gegeben:

$$\begin{array}{rcl} x & = & ACGCTG \\ y & = & CATGT \\ s(a, b) & = & \begin{cases} 2 & \text{falls } a = b \text{ und } a, b \in \{A, C, G, T\} \\ -1 & \text{sonst} \end{cases} \end{array} .$$

Wenden Sie Ihren Algorithmus auf dieses Beispiel an. Stellen Sie die entsprechende Tabelle $T[i, j]$ auf und markieren Sie in dieser alle optimalen Alignments.

Wenn Sie wollen, implementieren Sie Ihren Algorithmus und lassen Sie sich die entsprechende Tabelle ausgeben.

- (c) Die obigen beiden Teilaufgaben haben den Fall eines *globalen* Alignments behandelt. Dies kann als der Fall interpretiert werden, in welchem beide gegebenen Sequenzen ein Gen codieren. Interessant ist daher auch der Fall, dass die genaue Position eines Gens auf einer DNA-Sequenz nicht bekannt ist. Dann ist ein optimales Alignment einer Teilsequenz $x_{[i,i+p]}$ von x mit einer Teilsequenz $y_{[j,j+q]}$ von y gesucht. Dabei sind i, j, p, q nicht vorgegeben (vgl. Aufgabe 2.2). Dies wird als *lokales* Alignment bezeichnet.

Wandeln Sie Ihren Algorithmus aus Teilaufgabe (a) entsprechend so ab, dass ein lokal optimales Alignment bestimmt wird (vgl. Aufgabe 2.2 !).

Berechnen Sie auch die Tabelle $T[i, j]$ für ein lokal optimales Alignment für das Beispiel aus der letzten Teilaufgabe.