12. Übung zur Vorlesung "Einführung in die Bioinformatik I, 1. Teil"

Wintersemester 2017/2018

Prof. Peter Dittrich, Emanuel Barth, Maximilian Collatz, Marcus Ludwig

Ausgabe: 31. Januar 2018,
Abgabe: 07. Februar 2018 zu Beginn der Übung

Aufgabe 1 (6 Punkte): Berechnen Sie mit Hilfe einer DP-Matrix alle optimalen globalen Alignments (mit Einheitskosten) von `AGTGCACACA` und `ATCACACTTA`. Wie viele sind es?

Aufgabe 2 (6 Punkte): Die vier organischen Basen Adenin, Cytosin, Guanin und Thymin der DNA lassen sich in sogenannte *Purine* (Adenin und Guanin) und *Pyrimidine* (Cytosin und Thymin) unterteilen. Bei Mutationen im genetischen Code ist es wahrscheinlicher, dass eine Purinbase duch eine andere oder eine Pyrimidinbase durch eine andere substituiert wird (*Transition*) als dass eine Purinbase duch eine Pyrimidinbase ersetzt wird oder umgekehrt (*Transversion*).

Diese Tatsache berücksichtigen wir in der folgenden Kostenfunktion δ:

$$\delta(a,b) = \begin{cases} 0 & a = b \\ 1 & a,b \in \{\texttt{A},\texttt{G}\} \text{ und } a \neq b \\ & a,b \in \{\texttt{C},\texttt{T}\} \text{ und } a \neq b \\ 2 & sonst \end{cases}$$

Berechnen Sie die DP-Matrix von `AGTGCACACA` und `ATCACACTTA` unter Verwendung dieser Kostenfunktion, und geben Sie wie oben alle optimalen Alignments an.

Hinweis: Wenn keine Einheitskosten verwendet werden, gilt für die DP-Matrix:

$$D[i,0] = \sum_{k=1}^{i} \delta(u_k, \texttt{-}) \qquad \text{und} \qquad D[0,j] = \sum_{k=1}^{j} \delta(\texttt{-}, v_k)$$

Aufgabe 3 (6 Punkte): Berechnen Sie mit der folgenden Rekurrenz die Anzahl aller globalen Alignments zweier Strings u, v für alle Längen von u und v bis einschließlich vier:

$$N_{\mathcal{A}}[i,0] = N_{\mathcal{A}}[0,j] = 1 \qquad N_{\mathcal{A}}[i,j] = N_{\mathcal{A}}[i-1,j] + N_{\mathcal{A}}[i,j-1] + N_{\mathcal{A}}[i-1,j-1]$$

Wieso funktioniert diese Rekurrenz?

Aufgabe 4 (2 Punkte): Im Verlaufe der Evolution haben sich Mensch und Maus aus einem gemeinsamen Vorfahren entwickelt. Daher wäre es praktischer, Gene von Mensch und Maus nicht miteinander, sondern mit denen des gemeinsamen Vorfahren zu alignieren. Warum vergleichen wir nicht gegen den Vorfahren? Warum ergibt es trotzdem Sinn, Gene von Mensch und Maus miteinander zu alignieren?

Bonusaufgabe (8 Punkte): Eine *längste gemeinsame Teilsequenz* zweier Strings u, v ist ein String größter Länge, der Teilsequenz [1] von u und von v ist.

Wie kann man mittels dynamischer Programmierung die Länge einer längsten gemeinsamen Teilsequenz zweier Strings bestimmen? Wie kann man aus der DP-Matrix die längsten Teilsequenzen bestimmen (Traceback)? Geben Sie eine Rekurrenzgleichung an und berechnen Sie alle längsten gemeinsamen Teilsequenzen von `AGGTCAT` und `ACGATA`.

[1] Teilsequenz eines Strings T ist jeder String, den man durch löschen beliebig vieler Buchstaben aus T erhält. Bsp. ACE ist Teilsequenz von ABCDEF.