Abteilung für Genominformatik **Universität Hamburg**
Stefan Kurtz **Zentrum für Bioinformatik**

Genominformatik
Sommersemester 2014
Übungen zur Vorlesung: Ausgabe am 01.04.2014

Punkteverteilung: Aufgabe 1.1: 5 Punkte, Aufgabe 1.2: 5 Punkte

Abgabe bis zum 10.4.2014, 23:59 Uhr.

Aufgabe 1.1 Zeigen Sie, dass für die Einheitskostenfunktion δ die folgende Rekurrenz gilt:

$$E_\delta(i,j) = \begin{cases} E_\delta(i-1,j-1) & \text{if } u[i] = v[j] \\ 1 + E_\delta(i-1,j) & \text{else if } E_\delta(i-1,j) < E_\delta(i-1,j-1) \\ 1 + \min\{E_\delta(i,j-1), E_\delta(i-1,j-1)\} & \text{otherwise} \end{cases}$$

Tipp: Führen Sie in einer Fallunterscheidung jeden der drei Fälle auf die bekannte Rekurrenz für $E_\delta(i,j)$ zurück und beweisen Sie formal, dass die entsprechende Beziehung gilt. Die Ungleichungen in den Beobachtungen am Anfang des Kapitels über die schnelle Berechnung der Edit-Distanz sind dabei hilfreich.

Aufgabe 1.2 Schreiben Sie ein Programm, das die Edit-Distanz zweier Sequenzen u und v nach der Einheitskostenfunktion berechnet, wobei die zugrundeliegende DP-Matrix *lazy* ausgewertet werden soll. Benutzen Sie dabei die in Aufgabe 1.1 bewiesene Rekurrenz.

Für zwei Sequenzen u und v der jeweiligen Längen m und n beginnen Sie mit einer uninitialisierten $(m+1) \times (n+1)$-Matrix E_δ. Um den gewünschten Distanzwert an der Position $E_\delta(m,n)$ der Matrix zu erhalten, müssen Sie die entsprechenden Zellen der Matrix berechnen. Beachten Sie dabei, dass nicht in jedem Fall (entsprechend der obigen Rekurrenz) alle drei Vorgänger zu berechnen sind.

Zur Lösung der Aufgabe benötigen Sie einen Stack, der die noch zu berechnenden Einträge der Matrix E_δ enthält.

Beispiel: Für die zwei Sequenzen `aabaa` und `aaaba` würde die folgende Matrix berechnet werden. Dabei sind Einträge, die nicht berechnet werden, leer.

		a	a	a	b	a
	0	1				
a	1	0	1			
a		1	0	1		
b			1	1	1	
a				1	2	
a						2

Vergleichen Sie Ihre *lazy*-Methode mit dem Standard-DP-Verfahren. Wieviele Matrixeinträge werden jeweils bei den beiden Sequenzen aus der Datei `lazysequences.fas` (zu finden

in STiNE) berechnet? Im obigen Beispiel werden 14 von 36 Einträgen in E_δ berechnet. Testen Sie Ihr Programm, indem Sie die berechnete Distanz mit einer Distanz vergleichen, die von einem anderen Programm berechnet wurde.

Die Lösungen zu diesen Aufgaben werden am 14.04.2014 besprochen.