

Übungen zur Vorlesung Sequenzanalyse

Universität Bielefeld, WS 2016/2017

Prof. Dr. Jens Stoye · M.Sc. Tizian Schulz

<https://gi.cebitec.uni-bielefeld.de/teaching/2016winter/sa>

Übungsblatt 1 vom 25.10.2016

Abgabe am 3.11.2016 vor Beginn der Vorlesung

Organisatorisches

Wichtige Hinweise: Bitte unbedingt im eKVV für die Vorlesung und die Übungen registrieren. Auf der Homepage der Veranstaltung steht eine elektronische Version des Skripts zum Download bereit. Die gedruckte Version wird in der ersten Übung verteilt.

Abgabe der Übungszettel: Die Bearbeitung der Aufgaben darf in Gruppen (maximal zwei Personen) erfolgen. Sie können entweder **vor** Beginn der Vorlesung am Dienstag im Hörsaal, im Briefkasten vor U10-151 oder in elektronischer Form als E-Mail an den jeweiligen Tutor abgegeben werden. Alternativ könnt ihr die Zettel auch euren Tutoren persönlich geben. Bitte den Tutorientermin bzw. den Tutor, sowie euren Namen und den eures Gruppenpartners deutlich auf der Abgabe vermerken.

Quellenangabe auf den Übungszetteln: Benutzt ihr bei der Bearbeitung der Übungsaufgaben andere Quellen als das Skript zur Vorlesung, so gebt diese bitte immer mit an. Ist nicht ersichtlich, woher ihr eure Informationen genommen habt, können wir euch leider keine Punkte geben.

Teilnahme an der Klausur: Die Klausur kann am Ende mitschreiben, wer bis dahin A&D bestanden, mindestens 50% der Übungspunkte erreicht und mindestens zweimal in den Übungen eine Aufgabe vorgerechnet hat.

Aufgaben

Aufgabe 1 (Diskrete Metrik)

(3 Punkte)

Auf jeder Menge \mathcal{X} lässt sich die diskrete Metrik $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ definieren:

$$d(x, y) := \begin{cases} 0, & \text{wenn } x = y, \\ 1, & \text{wenn } x \neq y. \end{cases}$$

Weise nach, dass es sich bei d um eine Metrik auf \mathcal{X} handelt, indem Du die Definitionen (3.1) - (3.3) aus dem Skript (S. 10) überprüfst.

Aufgabe 2 (Minimale und maximale Hammingdistanz)

(4 Punkte)

Bei der Translation werden Aminosäuren durch Basentriplets codiert.

Beispiel: Methionin wird nur durch AUG codiert, Arginin hingegen wird durch CGU, CGC, CGA, CGG, AGA und AGG codiert.¹

1. Gegeben seien zwei codierende DNA-Sequenzen der Länge $n = 3m$; ihre Hamming-Distanz betrage $h \in [0, n]$. Wie groß kann die Hamming-Distanz k der zugehörigen Aminosäuresequenzen der Länge m minimal und maximal sein? Gib einfache Sequenzbeispiele für die Extremfälle an.
2. Jetzt seien zwei Aminosäuresequenzen der Länge m mit Hammingdistanz k gegeben. Wie groß kann die Hammingdistanz h von zugehörigen codierenden DNA-Sequenzen der Länge $n = 3m$ minimal und maximal sein? Gib wiederum Beispiele an.

¹Mehr dazu unter: http://de.wikipedia.org/wiki/Genetischer_Code

Aufgabe 3 (Implementierung der Hamming-Distanz)

(4 Punkte)

In der Vorlesung wurde die Hamming-Distanz eingeführt. Implementiere eine Funktion, die als Parameter zwei Strings a und b übergeben bekommt und ihre Hamming-Distanz ausgibt. Wenn die beiden Strings eine unterschiedliche Länge haben, soll eine Fehlermeldung ausgegeben werden. Dein Tutor spricht mit dir ab, welche Programmiersprachen in Ordnung sind.

Aufgabe 4 (Edit-Operationen und Edit-Sequenz)

(6 Punkte)

Gegeben seien die beiden Strings $x = \text{LAUBHAUFEN}$ und $y = \text{RAUBFISCH}$.

1. Berechne den String $z = E(\text{RAUBFISCH}, S_W C D D I_F C I_E I_L I_E C C S_E S_N)$, gib alle Zwischenschritte an. Wie hoch sind die Kosten der Edit-Sequenz? Gib eine weitere Edit-Sequenz mit maximalen Kosten an, die den gleichen String S ergibt.
2. Gib mit Hilfe der Edit-Funktion eine möglichst kurze Edit-Sequenz an, die x in y überführt.