Übungen zur Vorlesung
Grundlagen der Sequenzanalyse

Universität Bielefeld, WS 2006/07
Dr. Sven Rahmann · Dipl.-Bioinf. Katharina Jahn

http://gi.cebitec.uni-bielefeld.de/teaching/2006winter/sequenzanalyse/

Blatt 2 vom 26.10.2006
Abgabe am 02.11.2006 vor der Vorlesung um 8:30 in H3

Aufgabe 1 Betrachte den genetischen Code (für Wirbeltiere).

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 Proteinsequenzen der Länge m minimal und maximal sein? Gib einfache Sequenzbeispiele für die Extremfälle an.

2. Jetzt seien zwei Proteinsequenzen 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.

Aufgabe 2 Zeige formal, dass folgende Transformationen, wenn sie gleichzeitig auf Sequenzen u und v angewendet werden, den Wert der Edit-Distanz $d(\cdot, \cdot)$ nicht ändern:

1. Umkehrung (d.h. Rückwärtslesen)

2. Permutation des Alphabets

Sind also u und v DNA-Sequenzen, so folgt, dass $d(rc(u), rc(v)) = d(u, v)$ ist, wobei $rc(x)$ das reverse Komplement von x ist.

Aufgabe 3 Berechne mit Hilfe der Edit-Matrix per Hand die LCS-Distanz zwischen $s =$ GTACA und $t =$ GATTAC. Gib alle optimalen Alignments an!

Aufgabe 4 Betrachte neben den Operationen, die die Edit-Distanz zwischen $x \in \Sigma^*$ und $y \in \Sigma^*$ definieren, die zusätzliche Operation

- Vertauschung der nächsten zwei Buchstaben (Kosten 1),

so dass z.B. die Distanz zwischen $x =$ ABCDE und $y =$ ACBDEF nur 2 (anstatt 3 für die Edit-Distanz) beträgt. Die Operationenfolge wäre hier KVKKE (Kopieren von A, Vertauschen von BC, Kopieren von D, Kopieren von E, Einfügen von F).

Beschreibe einen effizienten Algorithmus, der die so definierte Distanz berechnet und eine optimale Operationenfolge angibt. Zusatz: Beweise formal die Korrektheit deines Algorithmus.