Forschungstage Informatik 2015 – Bioinformatik Workshop

Prof. Dr. Volkhard Helms
Daria Gaidar, Kerstin Reuter
Chair of Computational Biology

Universität des Saarlandes
Zentrum für Bioinformatik

Aufgabenblatt 1

Wrap up your mind – leichter Einstieg

Aufgabe 1.1: Zählen von DNA Nukleotiden

Als Zeichenkette (*string*) bezeichnet man eine georderte Abfolge von Symbolen aus einem bestimmten Alphabet, die ein Wort bilden. Die Länge einer Zeichenkette ist die Anzahl der enthaltenen Symbole. Ein Beispiel für ein DNA Strang (welcher die Symbole 'A', 'C', 'G', and 'T' beinhaltet) mit einer Länge von 21 Nukleotiden (nt) ist "ATGCTTCAGAAAGGTCTTACG".

Input: Ein DNA Strang s.
Output: 4 Integer (ganze Zahlen) entsprechend der Anzahl von 'A', 'C', 'G' und 'T' in s.

Beispiel:

Input:
`AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC`

Output:
```
A: 20
C: 12
G: 17
T: 21
```

Aufgabe 1.2: Transkription: DNA ⇒ RNA

Die RNA kann als Zeichenkette, bestehend aus dem Alphabet 'A', 'C', 'G', und 'U', dargestellt werden. Mit der DNA als Vorlage (hier: kodierenden Strang) erhält man den RNA Strang durch Ersetzung aller Vorkommen von 'T' mit 'U'.

Input: Ein DNA Strang s.
Output: Die transkribierte RNA bezüglich s.

Beispiel:

Input:
`GATGGAACTTGACTACGTAAATT`

Output:
`GAUGGAACUUGACUACGUAAAUU`

Aufgabe 1.3: Translation: RNA ⇒ Protein

Die 20 natürlich vorkommende Aminosäuren werden mit 20 Buchstaben des Alphabetes abgekürzt (alle Buchstaben außer B, J, O, U, X und Z). Protein Zeichenketten werden mit Hilfe dieser Symbole generiert. Der genetische Code bezeichnet dabei die "Regeln" nach denen Codons (Nukleotid-Triplets) in Aminosäuren übersetzt werden.

Input: Ein RNA Strang s, welcher der mRNA entspricht.
Output: Das Protein (Zeichenkette), welches durch s kodiert wird.

Beispiel:

Input:
`AUGGCCAUGGCGCCCAGAACUGAGAUCAAUAGUACCCGUAUUAACGGGUGA`

Output:
`MAMAPRTEINSTRING`

Es wird ein wenig schwieriger...

Aufgabe 1.4: RNA Spleißen

Sind die Exons und Introns einer RNA bekannt, erhält man z.B. durch Entfernen der Introns (Spleißen) eine neue Zeichenkette für die Translation in ein Protein.

Input: Ein DNA Strang s und eine Menge an Substrings von s, bei denen es sich um die Introns handelt. Die Eingabe erfolgt im FASTA Format.
Output: Die Protein Zeichenkette, welches durch Transkription und Translation der gespleißten DNA s entsteht.

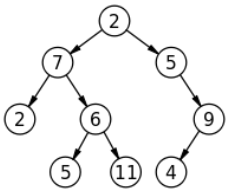
Beispiel:

Input:
```
>DNA
ATGGTCTACATAGCTGACAAACAGCACGTAGCAATCGGTCGAATCTCGAG
AGGCATATGGTCACATGATCGGTCGAGCGTGTTTCAAAGTTTGCGCCTAG
>INTRON_1
ATCGGTCGAA
>INTRON_2
ATCGGTCGAGCGTGT
```

Output:
`MVYIADKQHVASREAYGHMFKVCA`

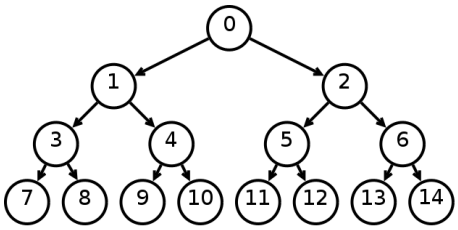
Aufgabe 1.5: Zählen phylogenetischer Vorfahren (Stift und Papier)

Ein binärer Baum wird definiert als Baum dessen Knoten einen Grad (Anzahl der Kanten mit denen der Knoten verbunden ist) von maximal 3 haben:



Die *Wurzel* ist der oberste Knoten. Als *Blätter* bezeichnet man Knoten, die einen Ausgangsgrad von Null aufweisen. Alle anderen Knoten sind *innere Knoten* (inkl. der Wurzel). Als *rooted binary tree* bezeichnet man einen binären Baum, in dem jeder innere Knoten einen Grad von 3

hat, während die Wurzel einen Grad von 2 hat. Die Blätter haben einen Grad von 1:



Input: Ein positiver Integer n (Anzahl der Blätter).
Output: Die Anzahl der inneren Knoten eines *rooted binary trees* mit n Blättern.

Beispiel:

Input: `16`
Output: `?`

Hinweis: *Diese Aufgabe kann für das obige Beispiel zuerst mit Stift und Papier gelöst werden. Versucht danach eine allgemeingültige Formel für die Anzahl der inneren Knoten als Funktion der Blätter zu finden (gerne auch mit Hilfe einer Internetrecherche).*

Aufgabe 1.6: Zählen von Punktmutationen

Die Hamming Distanz $dH(s,t)$ zwischen zwei gleich langen Zeichenketten s und t ist definiert als die Anzahle der Symbole, die sich zwischen s und t unterscheiden (*mismatches*):

GAGCCTACTAACGGGAT
CATCGTAATGACGGCCT

Die Hamming Distanz zwischen diesen Zeichenketten ist 7. Nicht übereinstimmende Symbole (*mismatches*) sind rot dargestellt.

Beispiel:

Input:
`GAGCCTACTAACGGGAT`
`CATCGTAATGACGGCCT`

Output: `7`

Aufgabe 1.7: Genom–Assemblierung (Stift und Papier)

Eine zirkuläre Zeichenkette besitzt weder ein Anfangs- noch ein End-Element; man kann diese Zeichenkette als ein Art Halskette, bestehend aus Symbolen, vorstellen. Die zirkuläre DNA Zeichenkette (ACGTAC) kann daher auf identische Weise auch mit den folgenden Zeichenketten dargestellt werden: (CGTACA), (GTACAC), (TACACG), (ACACGT) und (CACGTA).

Input: Eine Menge an DNA k-mers (*reads*) aus dem selben Strang eines zirkulären Chromosoms.
Output: Ein zyklischer *superstring* minimaler Länge, welcher alle *reads* enthält (dies entspricht dann dem *candidate cyclic chromosome*).

Beispiel:

Input:
```
ATTAC
TACAG
GATTA
ACAGA
CAGAT
TTACA
AGATT
```

Output:
```
?
```

Aufgabe 1.8: Motif-Suche in der DNA

Gegeben sind zwei Zeichenketten $\boldsymbol{s}$ und $\boldsymbol{t}$, wobei $\boldsymbol{t}$ ein Substring von $\boldsymbol{s}$ ist. Voraussetzung: $\boldsymbol{t}$ ist in $\boldsymbol{s}$ enthalten (daher darf $\boldsymbol{t}$ nicht länger als $\boldsymbol{s}$ sein).
Die Position eines Symbol in einer Zeichenkette ist die Anzahl an Symbolen auf der linken Seite (das Symbol selbst mit eingeschlossen). Die Positionen des Auftretens von 'U' in "AUGCUUCAGAAAGGUCUUACG" sind beispielsweise 2, 5, 6, 15, 17 und 18. Das Symbol an Stelle $\boldsymbol{i}$ wird als $s[i]$ bezeichnet.
Ein Substring von $\boldsymbol{s}$ kann daher bezeichnet werden als $s[j:k]$ mit $\boldsymbol{j}$ und $\boldsymbol{k}$ definiert als sie Start- und Endpositionen des Substrings von $\boldsymbol{s}$. Wenn $\boldsymbol{s}$ = "AUGCUUCAGAAAGGUCUUACG", dann gilt s[2:5] = "UGCU". Dabei kann $\boldsymbol{t}$ auch an mehreren Positionen in $\boldsymbol{s}$ vorkommen.

Input: Zwei Zeichenketten $\boldsymbol{s}$ und $\boldsymbol{t}$.
Output: Alle Positionen von $\boldsymbol{t}$ in $\boldsymbol{s}$.

Beispiel:

Input:
```
GATATATGCATATACTT
ATAT
```

Output:
```
2 4 10
```

Quelle: rosalind.info