

Algorithmen zur bioinformatischen Sequenzanalyse

Nr. 01 – 07/10/2024

Prof. Michael Sammeth

(1) Transkription

Implementieren Sie eine Python Funktion *transcribe(dna)*, die zu einer DNA-Sequenz *dna* die entsprechende RNA-Sequenz zurückgibt. Zur Vereinfachung soll *dna* lediglich die vier Symbole {A,C,G,T} beinhalten. Beachten Sie hier (sowie in allen weiteren Aufgaben), dass DNA-Sequenzen nicht immer (ausschließlich) in Großbuchstaben vorliegen.

Beispiel:

Eingabe= "ATGCCTTAG"

Ausgabe= "AUGCCUUAG"

(2) Reverses Komplement

Implementieren Sie eine Python Funktion *rc(dna)*, die zu einer DNA-Sequenz *dna* die revers-komplementäre Sequenz in 5'→3' Orientierung zurückgibt. Nehmen Sie zur Bearbeitung der Aufgabe an, dass *dna* ausschließlich Symbole aus {A,C,G,T,N} enthält.

Beispiel:

Eingabe= "AGTNCGG"

Ausgabe= "CCGNACT"

(3) IUPAC DNA-ambiguities

Im Seminar haben wir gesehen, dass in einer DNA-Sequenz neben den expliziten Basen {A,C,G,T,U} auch Symbole für verschiedene Basen an einer Position vorkommen können.

(a) Implementieren Sie eine Python Funktion *enumerateSequences(dna)*, die nach Eingabe einer ggf. mehrdeutigen IUPAC-Sequenz *dna* alle dadurch beschriebenen DNA-Sequenzen ausgibt:

Beispiel:

Eingabe= "WGGCCW"

Ausgabe= {"AGGCCA", "AGGCCT", "TGGCCA", "TGGCCT"}

(b) Implementieren Sie umgekehrt eine Python Funktion *encodeSequences(seq)*, die nach eingabe eines Tupels *seq* expliziter (d.h. ausschließlich aus Symbolen {A,C,G,T} bestehender) DNA-Sequenzen gleicher Länge eine IUPAC-Sequenz ausgibt, die an keiner Position mehr als die in *seq* beobachteten Basen beschreibt.

Beispiel:

Eingabe= ("AGGCCA", "AGGCCT")

Ausgabe= "AGGCCW" (und nicht: "AGGCCN")

(4) Translation

Implementieren Sie eine Python Funktion *translate(rna)*, die zu einer RNA-Sequenz *rna* die entsprechende Protein-Sequenz zurückgibt. Zur Vereinfachung soll die Eingabe *rna* lediglich die kanonischen Symbole {A,C,G,U} beinhalten, und Stopp-Codons in der Ausgabe mit dem Symbol “*” repräsentiert sein.

Eine maschinenlesbare Tabelle des genetischen Codes liegt dieser Aufgabe bei.

Beispiel:

Eingabe= “AUGCCUUAG”

Ausgabe= “MP*”