Stockholm University
Department of Mathematics
Examiner: Marc Hellmuth

1. Homework "DA7065 Computational Biology"

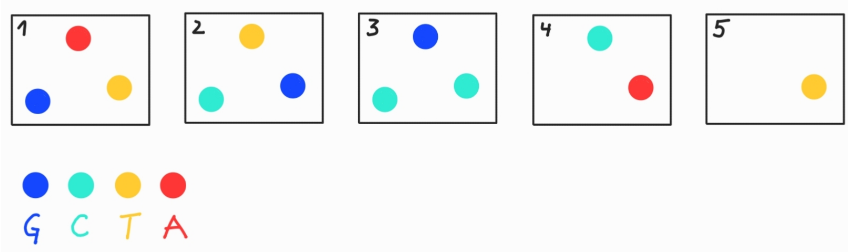
Exercises that are not *marked with ⋆ are for all participants. Exercises marked with ⋆ are intended as additional challenges for PhD students. However, all students are welcome to attempt solving ⋆-exercises to earn extra points.*

Exercise 1: 2.5+5+2.5 = 10p
Consider a gene as a subsequence of the DNA that encodes one protein and let S be a the protein (sequence of aminoacids) *CRICK* encoded by a "15-letter gene" g in a strand of DNA.

(a) Which aminoacids are encoded?
(b) How many different genes g can theoretically code for this sequence S?
(c) Write down one possible gene g encoding S.

Exercise 2: 7.5p
Recall the *Nirenberg-Matthaei-Experiment*: We introduced the technique of transcribing synthetic mRNA in order to solve some of the genetic code. The synthetic mRNA was periodic in nature: $XXXX\dots$, $XXYXXY\dots$, $XYYXYY\dots$, etc. Derive all the information you can about the genetic code using only two letters A and C. Clearly define the synthetic mRNA and their protein products.

Exercise 3: 2.5+2.5+2.5=7.5p
Given are the following "Illumina" photos in order $1, 2, \dots, 5$ showing the colored-glowing terminators.



(a) Determine the set ζ of reads you can determine based on the given photos.
(b) Draw the overlap graph for ζ (omit edges with weight 0).
(c) Apply the algorithm `Greedy_SCP` with input ζ and provide for each execution-step the resulting set ζ as well as the final superstring.

Exercise 4: 5p
Let $E = \{(S_1, S_2), (S_1, S_3), (S_1, S_4), (S_2, S_5), (S_3, S_5), (S_4, S_5)\}$ be the edge set of the overlap graph $G = (\{S_1, \dots, S_5\}, E, \text{ov}(,))$, where edges with weight 0 are omitted.
Find sequences $S_1, \dots, S_5$ that give rise to this graph - the particular weights you come up with are not important.

Exercise 5: 5+5=10p
Given is the sequence $S =$ `AATGATAGGCAGCCAC`.

(a) Draw the DeBruijn-graph G_k for $k = 3$.
(b) Determine all sequence reconstructions consistent with the Eulerian paths in G.

$\star$-*exercises*

Exercise 6*: 10p
Let X, Y, Z and Z' be distinct strings s.t. the set $\{X, Y, Z, Z'\}$ is substring-free.
Prove the following statement:

$$\text{If } \mathrm{ov}(X,Y) \geq \max\{\mathrm{ov}(X,Z), \mathrm{ov}(Z',Y)\}, \text{ then } \mathrm{ov}(X,Y) + \mathrm{ov}(Z',Z) \geq \mathrm{ov}(X,Z) + \mathrm{ov}(Z',Y).$$

Exercise 7*: 5p
Let us consider a protein simply as a sequence of aminoacids. Consider the set R of all DNA sequences of length $3n$ with $n \in \mathbb{N}$. Let $R' \subseteq R$ be the set of sequences $r \in R$ that can theoretically code for proteins. In particular, assume that each sequence $r = r_1 r_2 \ldots r_{3n} \in R'$ *begins* with the startcodon coding for `Met`, *ends* with one of the three stopcodons and none of the codons $r_i r_{i+1} r_{i+2}$ with $i \mod 3 = 1$ and $3 < i < 3n - 2$ corresponds to a start- or stopcodon.

Determine the cardinality $|R'|$ for $n = 1$, $n = 2$ and $n \neq 3$.

Deadline: Friday - Feb 2