Prof. Dr. Daniel Huson
Algorithmen der Bioinformatik
Fachbereich Informatik
Mathematisch-Naturwissenschaftliche Fakultät

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sequence Bioinformatics **WS 2022/23**

Assignment 3 — Due: Nov-7, 10 am

In this assignment, we will investigate the idea of using Integer Linear Programming to compute a maximum scoring multiple sequence alignment for the following three sequences:

```
>S1
AGTC
>S2
AGCT
>S3
GCTCT
```

In all of the following tasks, you only need to develop a solution that works for *exactly three* input sequences. If you find it difficult to get this to work for three arbitrary input sequences, then please focus only on the three given input sequences.

Download the file `assignment03.zip` from Ilias to obtain the corresponding Java files to complete.

1 Counting (2 points)

Write a Java program `assignment03/CountEdgesSimpleMixedCycles_YOUR_NAME.java` that counts the following two quantities (for any three given sequence lengths);

- How many edges are there between nucleotides that lie in different sequences in the alignment graph? (For the three sequences above, the number is between 50 and 60)
- How many simple mixed cycles are there? (For the sequences above, the number is between 3100-3200.)

Please write a Java program `assignment03/AlignmentILP_YOUR_NAME.java` to solve tasks 2-4:

2 Simple mixed cycles (3 points)

In the following, use $\texttt{X}ij\_pq$ to denote the variable that represents the edge connecting the nucleotide $s_i(j)$ in sequence s_i, at position j, with the nucleotide $s_p(q)$ in sequence s_p, at position q. (This naming scheme assumes that there are never more than 10 sequences and no sequence contains more than 10 letters).

Generate the list of all simple mixed cycles for the three given sequences, using the following format (which can be parsed by `lp_solve`, note that `<` means "$\leq$"), e.g.:

```
X00_10 + X01_10 < 1;
```

First list all constraints based on two sequences, and then all constraints based on three sequences.

3 Objective function (1 point)

Using a match score of 4 and a mismatch score of 1, set up the objective function for the ILP, in the format:

```
max: 1*X00_10+4*X01_10+ ... ;
```

4 Run the ILP (3 points)

Download the program `lp_solve`, from `https://sourceforge.net/projects/lpsolve/` and install it. Setup the ILP in the format supported by the program, which looks like this:

```
max: 1*X00_10+4*X01_10+ ... ;

X00_10 + X00_11 < 1;

 ...     (all simple mixed cycle constraints involving two sequences)

 X00_10 + X10_20 + X00_20 < 2;

  ...     (all simple mixed cycle constraints involving three sequences)


X00_10<1;
  ...    (all binary constraints)

int X00_10, X00_11, ... ;
        (specify all variables as integers)
```

Run this file using `lp_solve`.

5 Report the alignment (1 point)

Indicate how to translate the output of `lp_solve` into an alignment and report the alignment.