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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sequence Bioinformatics **WS 2021/22**

Assignment 3 **Due: 10-Nov-2020, 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
CGTA
>S2
CGAT
>S3
GATAT
```

1 Counting (1 point)

How many edges are there between nucleotides that lie in different sequences in the alignment graph? How many simple mixed cycles are there?

Please write a Python program `ilp_aligment_YOUR_NAME.py` to solve the following tasks:

2 Simple mixed cycles (3 points)

In the following, use `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.

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

```
X11_21 + X12_21 < 1;
```

3 Objective function (1 point)

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

```
max 1*X11_21+4*X12_21+ ... ;
```

4 Run the ILP (4 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*X11_21+4*X12_21+ ...

X11_21 + X12_21 < 1;
 ...     (all simple mixed cycle constraints)

X11_21<1;
  ...    (all binary constraints)

int X11_21, X11_22, ... ;
        (specify all variables as integers)
```

Run this file using `lp_solve`.

5 Report the alignment (1 point)

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